Large-scale phylogenomic visualization and analysis of functional traits in bacteria

by

Kerrin Mendler

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Biology

Waterloo, Ontario, Canada, 2019

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in this document. 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.

Statement of Contributions

The manuscript presented in this thesis is the work of Kerrin Mendler, in collaboration with her co-authors Han Chen, Donovan Parks, Laura Hug, and Andrew Doxey. The manuscript was submitted as a preprint to the journal bioRxiv and for publication to the journal Nucleic Acids Research. Text and figures from the manuscript were modified for use in this dissertation.

- Mendler, K., Chen, H., Parks, D. H., Hug, L. A. and Doxey, A. C. (2018). AnnoTree: visualization and exploration of a functionally annotated microbial tree of life. bioRxiv, *463455*

For the manuscript, H.C and K.M. built the front-end interface of AnnoTree and back-end database. K.M. performed data analysis. D.P. assisted with bioinformatic analysis and genome annotation. L.H. assisted with phylogenetics and tool design. A.D., K.M., and L.H. wrote the manuscript. A.D. conceived the project and tool design.

Additional contributors that were not included as co-authors on the paper are Briallen Lobb, who defined the 'saturation' and 'catchment' metrics for the classification of lineage-specific traits (unpublished) and Lee Bergstrand, for his assistance with the Docker-Compose implementation of the AnnoTree application.

Abstract

The growth of genomic information in public databases has dramatically improved our view of the tree of life and at the same time expanded our knowledge of protein diversity. Through the use of automated annotation pipelines, researchers can predict many of the functional capabilities of organisms directly from their genome sequence. Although there exist numerous phylogenetic and protein databases, there have been fewer attempts to combine these data, which is essential for the study of protein evolution. The web application AnnoTree (annotree.uwaterloo.ca) was developed as part of this thesis to facilitate the exploration and visualization of protein families (Pfams) and KEGG orthologs (KOs) on a phylogeny composed of nearly 24,000 bacterial genomes. The visualization includes an interactive tree of life, a summary of the taxonomic distribution of the query, basic taxonomic information, and annotation confidence scores. All protein sequences, visualizations, and summary information can be downloaded directly from the interface. The AnnoTree framework is open-source and can be modified to incorporate any custom tree, taxonomy, and proteome dataset. AnnoTree allows users to visualize the phylogenetic distribution of a Pfam of interest, which, in combination with obtained gain/loss data, promotes hypothesis-generation in the context of protein evolution. To identify functions that are more tightly associated with evolutionary mechanisms such as horizontal gene transfer and evolutionary conservation, the pre-computed annotation data were combined with the bacterial tree of life in a phylogenomics analysis. The phyletic patchiness of all Pfam and KO annotations

was measured using the normalized consistency index (CI), a measure of disagreement between the presence/absence states of traits across the tree and the tree topology. Pfams and KOs with the highest normalized CI represent functions known to be associated with mobile genetic elements and viral defence. These annotations were most commonly found within the genomes of symbiotic and pathogenic bacteria. The most highly conserved Pfams and KOs were functions related to core processes such as transcription, DNA replication, and protein synthesis as well as those required for oxygenic photosynthesis and sporulation. Lineage-specific Pfams and KOs were classified in many bacterial taxa, revealing many clade-defining functions in the *Bacillus_A* genus, the Oxyphotobacteria class, and the Actinobacteria class, among others. An additional phylogenomics analysis was performed to identify branches of a phylogeny encompassing representatives from all three domains of life undergoing the most Pfam gain and loss events. The branches dividing the three taxonomic domains had the highest density of gain events, all of which were associated with well-known clade-defining functions. Missing data influenced the frequency of Pfam losses in lower taxonomic levels, but some characterized genome streamlining events within Eukaryotes were uncovered. Ultimately, the development of AnnoTree and accompanying analyses provide new insights into large-scale bacterial phylogenomics and the evolution and distributions of bacterial protein domains and gene families.

Acknowledgements

I am most grateful for having Dr. Andrew Doxey as a supervisor, whose overwhelming kindness and enthusiasm made my experience in his lab a comfortable and enjoyable one. I would also like to thank the other members of my committee, Dr. Laura Hug and Dr. Brendan McConkey, whose further guidance allowed me to develop the critical thinking and technical skills required for the completion of this thesis.

My gratification is extended to the other members of the Doxey lab for the productive (and also unproductive) conversations that we had. It is rare to have such a knowledgeable and insightful group of people to bounce ideas off of. My final thanks go out to my friends and family for their unwavering support.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

API application programming interface

BM Brownian motion

CI consistency index

commamox complete ammonia oxidizer

CPR Candidate Phyla Radiation

CSV comma-separated values

DUF domain of unknown function

EC Enzyme Commission

GO Gene Ontology

GTDB genome taxonomy database

HGT horizontal gene transfer

HMM hidden Markov model

HSR homoplasy slope ratio

HTTP hypertext transfer protocol

ICE integrative and conjugative element

JSON JavaScript object notation

KEGG Kyoto Encyclopedia of Genes and Genomes

KO KEGG ortholog

MAG metagenome-assembled genome

MGE mobile genetic element

MSA multiple sequence alignment

NCBI National Center for Biotechnology Information

ncRNA non-coding RNA

ORF open reading frame

Pfam protein family

RDBMS relational database management system

REST representational state transfer

RI retention index

SVG scalable vector graphic

WSGI web server gateway interface

Chapter 1

Introduction

1.1 Overview of phylogenomics and trees of life

The growing number of genetic sequences submitted to public databases in the modern era and the simultaneous increase in contemporary computer hardware poses opportunities for studying the genomes of species at a large scale. Function prediction methods such as phylogenomic approaches use genomic or proteomics data to infer the function of proteins using evolutionary data derived from an accurate species tree (Eisen, 1998). Many large discoveries have been made through phylogenomics that support pre-existing and novel hypotheses that elucidate the mechanisms contributing to the evolution of genomes (Boucher et al., 2003; Szöllsi et al., 2015; Jiao et al., 2011; Spang et al., 2015). Models of evolutionary dependence have revealed that the

gain of functions occurs sequentially and that the gene content of extant species even has the potential to predict future gene acquisitions (Press et al., 2016), demonstrating the importance of studying the functional composition of extant genomes. Moreover, the taxonomic co-occurrence of uncharacterized genes and protein domains has the potential to deliver valuable insights into their function and provide a means to prioritize uncharacterized genetic features for experimental characterization (Barberán et al., 2017; Goodacre et al., 2014).

Phylogenomic analyses rely on the presence of functional data and a model of evolution in the form of a phylogenetic tree. Phylogenetic reconstruction is difficult in bacteria due to the prevalence of HGT between bacterial species (Ochman et al., 2000). The bias for HGT among species with similar genomes makes the reconstruction particularly difficult at lower taxonomic levels (Andam and Gogarten, 2011). Researchers have overcome this challenge through careful selection of universal single-copy orthologs from which to measure evolutionary distances (Patwardhan et al., 2014). The most recently-published tree of life containing members across all three domains is the first to include members of the Patescibacteria (previously classified as the CPR superphylum), a phylum of predominantly uncultivable bacteria identified through metagenomic sequencing (Hug et al., 2016). Currently, the phylogenetic tree with the highest resolution in the bacterial domain is made up of over 27,000 genomes and also includes members of the Patescibacteria (CPR) (GTDB Release 03-RS86; Parks et al., 2018).

1.2 Overview of protein functional annotation methods and databases

Functional annotations provide a standardized representation of the functional capabilities of a genome so that comparisons can be made between them. These comparisons are required in ecological contexts to describe the necessary functions in a stable microbial community (Louca et al., 2016; Coleman and Chisholm, 2010; Stapley et al., 2010), in medical contexts to identify proteins involved in human genetic disease (Wang et al., 2012), and in industrial biotechnology to identify novel enzymes and natural products that improve or replace conventional synthetic products and processes (Lorenz and Eck, 2005), among others.

The majority of functional events in a cell are the result of the actions of one or more proteins. The function of the protein is dictated by its structure, which is influenced substantially by the sequence of amino acids in its polypeptide chain. Similar patterns in polypeptide sequences make up the most basic functional unit, called a 'domain'. Protein domains are the building blocks of proteins and the fundamental units of protein structure, function, and evolution (Lees et al., 2016). As such, their characterization is regarded as one of the most essential steps in determining the functional capabilities of a newly sequenced genome.

In the current version of the popular protein family database, Pfam v32.0 (El-Gebali et al.,

2018), protein domains are classified into 17,929 families and 628 clans. Each Pfam entry includes a MSA of protein sequences determined semi-automatically by sequence similarity, expert knowledge, and other databases to exhibit a particular function (Sonnhammer et al., 1998). The MSA is used to produce a profile HMM, which is used to scan new protein sequences for the Pfam. Protein sequences that score beneath a curated *E*-value threshold for a Pfam annotation using a tool like pfamscan (Mistry et al., 2007) are assigned that Pfam annotation.

Some domains, termed domains of unknown function (DUFs), are conserved families of protein domains that have not yet been functionally characterized. Surprisingly, a considerable fraction (approximately 25%) of Pfam domains are labelled as DUFs, indicating the need for functional characterization. Despite a large number of uncharacterized entries in the database, Pfam remains one of the most widely used references for functional annotation of proteins. Alternative protein functional annotations include CATH (Dawson et al., 2017), COG (Galperin et al., 2015), ECOD (Cheng et al., 2014), SCOP (Chandonia et al., 2018), TIGRFAM (Haft et al., 2012), and CDD (Marchler-Bauer et al., 2017). The InterPro database (Finn et al., 2017) aims to combine these and other protein domain annotations in one place and is an excellent source for browsing protein domain data.

KEGG is a knowledge base for the biological interpretation of functional pathways, modules, and networks in organisms, with an emphasis on the genes and compounds therein (Kanehisa et al., 2018). Pathway mapping is performed using the KO system, whereby each unique function or

reaction, with or without a protein sequence, is assigned a KO identifier and is placed in the appropriate KEGG pathways, KEGG BRITE categories, KEGG modules, and KEGG networks based on experimental validation of the function or sufficient sequence similarity to an experimentally validated gene or protein. It is important to note that KEGG pathways include all characterized functions in a particular pathway and that a single species often does not contain every KO listed in a pathway. Currently, there are over 22,000 KOs in the KEGG database, in which 85% are linked to publications and 68% are further linked to sequence data. KOs are assigned to protein sequences through pairwise sequence similarity against a curated database of annotated proteins with programs such as BLASTP (Camacho et al., 2009), DIAMOND (Buchfink et al., 2015), or KEGG's own KOALA family of tools (Kanehisa et al., 2016).

The resources that KEGG provides are effective at providing a high-level interpretation of the functional capabilities of newly-sequenced organisms based solely on their repertoire of KOs. An alternative to KEGG pathways and modules are the MetaCyc superpathways and pathways (Caspi et al., 2018). MetaCyc contains a higher number of curated reactions and defined pathways that are closer to true biological pathways than those of KEGG. The analysis of pathway completion and enrichment is thus more suitable with the ontology described by MetaCyc than KEGG (Green and Karp, 2006); however annotation of new genomes with MetaCyc pathways is only possible for complete genomes pre-annotated with EC numbers and GO terms. These criteria make Metacyc unsuitable for the analysis of partially assembled novel genomes in a high-throughput manner

(Karp et al., 2011).

1.3 Overview of factors shaping functional diversity in bacterial genomes

The quickest way in which prokaryotes obtain new functions is by the procurement of exogenous genetic material from another organism or the environment, followed by its incorporation into the genome. Naturally competent bacteria are capable of taking up DNA directly from their environment. In most competent gram-negative and gram-positive bacteria, transport of DNA molecules across the outer and inner membranes requires the use of proteins associated with type IV pili and the type II secretion system (Chen and Dubnau, 2004; Sun, 2018).

Incorporation of imported DNA into the bacterial chromosome is performed through homologous recombination using both membrane and cytosolic proteins that are ubiquitous among bacteria (Claverys et al., 2009). Recombination of exogenous DNA is more likely to be successful when the donor and recipient molecules share regions of highly similar sequence, with the precise threshold of similarity differing between recipient species (Thomas and Nielsen, 2005).

Many bacterial genomes contain 'genomic islands', mobilizable clusters of genes that can be transmitted between bacterial cells by HGT (Rodriguez-Valera et al., 2016). They can harbour many kinds of MGEs, segments of DNA that encode enzymes and other proteins that mediate

the movement within and between cells (Frost et al., 2005). ICEs are one such element. It is also common for ICEs to be maintained extrachromosomally on a plasmid (Wozniak and Waldor, 2010). ICEs often encode a type IV secretion system-like machinery necessary for HGT via bacterial conjugation (Wozniak and Waldor, 2010; Sun, 2018) together with beneficial accessory proteins such as virulence factors and those conferring resistance to antibiotics and heavy metals (Wozniak and Waldor, 2010). An experiment involving adaptive evolutionary evolution of *Escherichia coli* in the presence of donor DNA showed that recombination success strongly depended on the donor-recipient strain combination and that the benefit of recombination on fitness depended on the environment and donor DNA (Chu et al., 2018).

Integrons are another class of MGE. They include an integrase of the tyrosine-recombinase family, a primary recombination site, and a promoter. The integron functions by capturing an ORF with the integrase and placing it downstream of the promoter for controlled expression (Mazel, 2006). Some of the biological functions associated with integrons are antibiotic resistance, secondary metabolism, plasmid maintenance, virulence, surface properties, and components of toxin-antitoxin systems (Gillings, 2014). They are often located within genomic islands and tend to be transferred within other mobile elements such as conjugative ICEs and bacteriophages.

Bacterial viruses, called bacteriophages, are also capable of transferring genes between bacterial hosts. In their integrated form, they are called prophages. Their DNA encodes for the bacteriophage tail and coat proteins, reverse transcriptase, and numerous accessory proteins.

These proteins have been associated with biofilm formation, antibiotic resistance, transcriptional regulation, and virulence factors, among others (Bobay et al., 2014). Bacteriophages have a broad host range, but the integration of the phage genome into its host is highly species-specific and very rarely occurs across large evolutionary distances (Popa et al., 2017).

There is less information about the mechanisms and causes of gene loss in bacteria, but some important discoveries have been made. Plasmid loss is one way in which organisms can lose genes quickly and is often seen among bacteria fostering plasmids with a high carriage cost (Ayala-Sanmartin and Gómez-Eichelmann, 1989). Plasmids can be lost due to the extra energy required to maintain them; however compensatory mutations can be made in the host to support relatively small plasmids that confer a fitness benefit (San Millan et al., 2014).

Genome streamlining, the reduction of a genome through gene loss, is thought to lessen the metabolic burden for fundamental cellular processes, resulting in a genotype that has an advantage over others that still bear those burdens (Giovannoni et al., 2005). Large-scale gene loss in an adaptive experimental evolution study did not give a fitness advantage to bacterial monocultures (Karcagi et al., 2016). However, bacteria do not often exist in nature as monocultures.

Null mutations of biosynthetic genes of an organism within a community have been shown to confer a fitness advantage that is shared amongst all members (D'Souza et al., 2014; Pande et al., 2014). The further loss of genes harbouring null mutations within a community is not unlikely if selective pressures are stable over a long enough period. Evidence of this theory is

exhibited by bacterial symbionts, whose co-operative adaptation with their host and co-symbionts allows for the distribution of essential functions between them (McCutcheon and Moran, 2012). The genomes of newly-emerging symbionts and pathogens are enriched in mobile elements, chromosome rearrangements, gene inactivation, and pseudogene accumulation (McCutcheon and Moran, 2012). These genomic characteristics, along with an increased rate of genetic drift in these populations, allow for the fixation of deleterious mutations (Kuo et al., 2009). Further community analyses have shown that functional diversity rather than taxonomic diversity in a community are better indicators of selective processes operating in an environment (Louca et al., 2016), supporting the thought that genes encoding a function that is abundant within a community are likely to be lost over long time periods in non-selective environments.

1.4 Measuring phylogenetic dispersion of binary traits

There are several ways in which the phylogenetic dispersion of a trait can be measured. Here, I will describe methods that can be used for features represented by a binary presence/absence profile (**Table 1.1**).

Parsimony-based metrics are commonly used to measure the quality of phylogenetic tree reconstructions by evaluating their agreement with trait profiles. Traits that do not agree with the phylogeny are said to be convergent or homoplasic. Parsimony-based metrics are the oldest of

the phylogenetic dispersion metrics presented here, but they are popular due to their availability in common taxonomic software packages and are still used in recent publications. They are all based on the principle of maximum parsimony, which favours the evolutionary solution with the least complexity. In the case of the phylogenetic distribution of a binary trait, the most parsimonious solution is the one that results in the minimum number of steps that explains the observed distribution on the phylogeny.

The CI, RI, and HSR are all measures of homoplasy that use the concept of maximum parsimony in their calculation. The CI is the ratio of state changes that can occur in the least homoplasic distribution of a trait (i.e. 1 for binary traits) against the observed parsimony score (Kluge and Farris, 1969). On the other hand, the RI represents the proportion of taxa that are not homoplasic (Farris, 1973, 1989). The homoplasy slope is a function that describes the relationship between the number of taxa in a phylogeny and the observed parsimony score (Meier et al., 1991). The HSR is obtained through comparison of the observed homoplasy slope to the average homoplasy slope of many presence/absence profiles obtained through random sampling. The metric relies on the presence of a linear relationship between the parsimony score of a random profile and the number of taxa. In the original paper, this was shown to occur between 4 and 40 taxa (Meier et al., 1991). It has yet to be demonstrated that such a relationship exists for modern phylogenies, which tend to be much larger.

The phylogenetic signal is another aspect of a phylogenetic distribution that can be measured.

It is the tendency for evolutionarily similar species to resemble each other more phenotypically than other species drawn randomly from a phylogeny (Münkemüller et al., 2012). This measure is often used in the fields of comparative analyses, community ecology, and macroecology to evaluate the correlation of traits within and between communities, among others (Münkemüller et al., 2012). Under the most straightforward evolutionary conditions involving only neutral genetic drift, the evolutionary rate does not correlate with the phylogenetic signal (Revell et al., 2008). Furthermore, the evolutionary signatures measured by phylogenetic signal may be the product of many different evolutionary processes, so phylogenetic signal alone is really only useful for suggesting the presence or absence of such a causative factor on the states of extant taxa unless the statistical experiment is modified in such a way that the elements can be tested independently (Revell et al., 2008).

The calculations for some of the phylogenetic signal metrics employ a comparison to characters simulated under the BM model (λ: Pagel, 1999; K: Blomberg et al., 2003; D: Fritz and Purvis, 2010). This model assumes that the value of a measured continuous trait changes randomly in direction and magnitude and that the trait values follow a normal distribution with mean 0 and variance proportional to the rate of evolution (Lande, 1976). Many evolutionary processes may give rise to such a distribution including pure genetic drift, randomly varying selection, varying stabilizing selection, and consistent directional selection (Hansen and Martins, 1996). A more appropriate model for the evolution of discrete traits is the Markov model, which models the

transition of one state to another based on its current state and a given transition rate (Pagel, 1994; Lewis, 2001; Paradis et al., 2004).

Pagel's λ is the scaling factor applied to the internal branches of the observed phylogeny that gives the best fit to a phylogeny fitted with the trait distribution under a BM model (Pagel, 1999). On the other hand, Blomberg's K is a scaled ratio of variance among observed trait values and the phylogeny for the observed data and the data expected under a BM model (Blomberg et al., 2003). Pagel's λ and Blomberg's K are the most common measures of phylogenetic signal, but the concept of binary traits are not always congruent with the underlying assumptions of those methods. They should be used cautiously or with modifications such that those assumptions are no longer violated for binary traits. Another phylogenetic signal metric, Fritz and Purvis' D, was designed for use with binary features (Fritz and Purvis, 2010). It compares the sum of the weighted values of internal nodes estimated from the observed trait profile and phylogeny against those of data generated through simulations under the BM model and by randomly shuffling the observed profile. The D statistic can effectively measure the phylogenetic signal of a binary trait, but it assumes that the trait is based on one or more evolved continuous traits such as body size and reproductive rates (Fritz and Purvis, 2010).

The next set of metrics described here are independent of evolutionary model and rely solely on the tip values and shape of the phylogeny to quantify phylogenetic signal. Moran's I is a measure of auto-correlation and can be used for a wide variety of distance metrics (Gittleman and

Kot, 1990). The interpretation of the measurement is dependent on the metric and the weighting function that quantifies the proximity of the taxa. Moran's I is dependent on trait prevalence, but this can be corrected through a rarefaction-like process (Lockwood et al., 2002). Abouheif's C_{mean} is a particular case of Moran's I using a topological distance-dependent weighting matrix with a non-zero diagonal (Pavoine et al., 2008).

Lastly, trait depth (τ_D) measures the average sequence similarity of trait-containing clades in the phylogeny and serves to directly link phylogenetic dispersion with the dispersion of traits (Martiny et al., 2013). This metric allows researchers to screen for lineages or features that are diverging at a particular evolutionary time (Martiny et al., 2013).

Table 1.1. Metrics for the quantification of phylogenetic dispersion of binary traits

Metric	Measurement	Variables	Interpretation	R implementation
Consistency Index (CI) (Kluge and Farris, 1969)	$CI = m/s$	m: minimum parsimony score in any tree; s: observed parsimony score	Ratio between the number of steps in a fully parsimonious tree with no homoplasy and the number of steps in the profile reconstruction. CI decreases as homoplasy increases	`phangorn::CI` (Schliep, 2011)
Retention Index (RI) (Farris, 1973, 1989)	$RI = \frac{g-s}{g-m}$	g: maximum parsimony score in any tree; s: observed parsimony score; m: minimum parsimony score in any tree	Proportion of taxa that are not homoplasic. RI decreases as homoplasy increases	`phangorn::RI` (Schliep, 2011)
Homoplasy Slope Ratio (HSR) (Meier et al., 1991)	$HSR = \frac{\frac{s_o-1}{t-3}}{\frac{1}{n}\sum_{i=1}^{n}\frac{s_i-1}{t-3}}$	s_o: observed parsimony score; t: number of taxa; n: number of random profiles to test; s_i: parsimony score of i^{th} random profile	The level of homoplasy observed in the profile relative to the average level of homoplasy in randomly-drawn profiles of the same size. HSR increases as homoplasy increases	None
Moran's I (Moran, 1950; Gittleman and Kot, 1990)	$I = \frac{t}{S_0} \times \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i-\bar{y})(y_j-\bar{y})}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$	t: number of taxa; S_0: sum of all pairwise weights (i.e. $\sum_{i=1}^{t}\sum_{j=1}^{t} w_{ij}$); w_{ij}: pairwise distance between taxa i and j as given in a matrix $\mathbf{W}$ with $w_{ii} = 0$; y_i and y_j: trait value of species i or j; $\bar{y}$: average trait value over all taxa	Measurement of the effect of proximity on the average value of trait-containing taxa in a phylogeny or taxonomic hierarchy	`ade4::gearymoran` (Bougeard and Dray, 2018), `adephylo::moran.idx` (Jombart et al., 2010), `ape::Moran.I` (Paradis et al., 2004), `phylosignal::phyloSignal` (Keck et al., 2016)
Abouheif's C_{mean} (Abouheif. 1999; Pavoine et al., 2008)	$C_{\text{mean}} = 1 - \frac{\frac{1}{T}\sum_{i=1}^{T} C_i}{2\sum_{j=1}^{t}(x_j-\bar{x})^2}$	t: number of taxa; T: number of topologies obtained by rotating all nodes in the phylogeny; $C_i = \sum_{j=1}^{n-1}(x_{K_i(j)} - x_{K_i(j+1)})^2$ where $x_{K_i(j)}$ and $x_{K_i(j+1)}$ are the trait values of adjacent taxa	Special case of Moran's I with proximities dependent on topological distance rather than branch length	`adephylo::abouheif.moran` (Jombart et al., 2010), `phylosignal::phyloSignal` (Keck et al., 2016)
Pagel's λ (Pagel, 1999)	Optimize λ such that $\lvert\lambda\times$ off-diagonal of $\mathrm{C}_o - \mathrm{C}_m\rvert$ is minimized	C_o: phylogenetic variance-covariance matrix of the observed traits on the phylogeny; C_m: phylogenetic variance-covariance matrix of a trait distribution obtained under the assumption of a given model on the phylogeny	The transformation of the phylogeny that gives the best fit of trait data to an evolutionary model. λ approaches 1 as it gets closer to the expectation of the given model	`geiger::fitDiscrete` (Harmon et al., 2008)
Blomberg's K (Blomberg et al., 2003)	$K = \frac{\text{observed}\frac{MSE_t}{MSE_m}}{\text{expected}\frac{MSE_t}{MSE_m}}$	MSE_t: mean squared error of the given trait values; MSE_m: mean squared error of the trait values based on the variance-covariance matrix of the phylogeny	A low MSE_m indicates good prediction of the trait value based on the phylogeny, thus giving a high 'signal'. $K = 1$ implies high agreement of the observed signal with the signal under the model	None
Fritz and Purvis' D (Fritz and Purvis, 2010)	$D = \frac{s_o - \frac{1}{n}\sum_{i=1}^{n}(s_b)_i}{\frac{1}{n}\sum_{i=1}^{n}(s_r)_j - \frac{1}{n}\sum_{i=1}^{n}(s_b)_i}$	s_o: sum of values at internal nodes, given as the weighted mean of the clade's tip values in the observed profile; n: number of profiles generated for each of the BM and random simulations; $(s_b)_i$: sum of values of internal nodes in the i^{th} profile permutation generated under the BM model; $(s_r)_j$: sum of values of internal nodes in the j^{th} profile permutation generated by shuffling the observed profile	$D = 1$: distribution of the trait is random; $D > 1$: trait is more dispersed than the random expectation; $D = 0$: trait is as expected under the Brownian motion model of evolution; $D < 0$: trait is more phylogenetically conserved than the Brownian expectation	`caper::phylo.d` (Orme et al., 2013)
Trait depth (τ_D) (Martiny et al., 2013)	$\tau_D = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{m}\sum_{i=1}^{m} d(S_{ij})$	n: number of clades where at least 90% of descendent taxa have the trait; m: number of descendent taxa in a clade where at least 90% descendents have the trait; $d(S_{ij})$: branch distance between the root of the i^{th} clade and the tip of the j^{th} descendent	The average sequence divergence of members containing the trait in the same units as the branches	`castor::get_trait_depth` (Louca et al., 2018)

1.5 Research aims and objectives

The impact phylogenomics methods have had on our understanding of bacterial evolution to date has been enormous (Boucher et al., 2003; Szöllsi et al., 2015; Jiao et al., 2011; Spang et al., 2015). This work aims to facilitate the exploration of functional traits within the context of bacterial evolution. This goal was accomplished through the development of an interactive functionally annotated bacterial tree of life, AnnoTree, available at annotree.uwaterloo.ca. Its concept and development is described in Chapter 2. AnnoTree's underlying functional annotations and phylogeny represent a wealth of data more substantial than any that has been explored in a phylogenomic context. The phylogenetic distribution of each Pfam and KEGG annotation was measured using a normalized homoplasy metric to evaluate the high-level functional trends between the most scattered and the most conserved traits in addition to the taxa that contain them. The methods and results of these analyses are described in Chapter 3. The webtool and complementing analyses promote hypothesis-generation in the context of protein evolution and have the potential to lead researchers to their next major discovery.

Chapter 2

The AnnoTree web application

2.1 Background

Important biological and evolutionary insights can be generated by exploring the presence of genes and functional annotations across species phylogenies. These include identifying unexpected taxonomic occurrences (Venter, 2004), uncovering the evolutionary origin of genes (Demuth and Hahn, 2009), and locating HGT events (Ravenhall et al., 2015). With the ongoing exponential increase in available genome sequences, including information from previously uncharacterized and uncultured lineages, online genomic repositories are becoming increasingly valuable collections of predicted genes and functional annotations. With this wealth of genomic data comes the opportunity for large-scale examinations of gene family distributions and evolutionary histories,

but databases are not easily accessed, updated, or visualized.

Many strategies exist for merging taxonomic and functional information to create annotated phylogenies. For instance, homologs of a gene family retrieved using BLAST (Camacho et al., 2009) or related methods can be manually mapped onto a custom species tree using tools such as iTOL (Letunic and Bork, 2016), GraPhlAn (Asnicar et al., 2015), Evolview (He et al., 2016), ETE (Huerta-Cepas et al., 2016), and PhyD3 (Kreft et al., 2017). Alternatively, several online bioinformatics databases offer pre-computed summaries of taxonomic distributions for genes based on Linnean taxonomic classification or the NCBI taxonomy (NCBI Resource Coordinators, 2016; Finn et al., 2016; Adebali and Zhulin, 2017; Chen et al., 2017; Wattam et al., 2017). However, there is a need for tools that allow users to explore the taxonomic distribution of functions across a curated and a highly resolved tree of life.

Here, I present AnnoTree (annotree.uwaterloo.ca), a web application that enables the interactive visualization of phylogenomic distributions of precomputed Pfams (Finn et al., 2016) and KO identifiers (Kanehisa et al., 2017) across a bacterial tree of life composed of nearly 24,000 genomes. The phylogeny and taxonomic nomenclature used within AnnoTree is derived from the recently developed GTDB (Parks et al., 2018). The GTDB overcomes several challenges with the construction of an annotated tree of life as it is *standardized* (its taxonomic nomenclature and phylogeny are made to be internally consistent) and *thorough* (it includes a large number of novel bacterial genomes derived from metagenomic sources). These important features differentiate

the GTDB taxonomy and AnnoTree from similar approaches that rely on the NCBI taxonomy (Federhen, 2012), whose hierarchy disagrees with several recent reconstructions of bacterial phylogeny (Bromberg et al., 2016; Hug et al., 2016).

2.2 Overview of AnnoTree: features and capabilities

AnnoTree divides the bacterial tree of life into distinct views by each major taxonomic level. A user can, therefore, explore the phylogenetic distribution of a trait anywhere from the phylum to genome level (**Figure 2.1**).

AnnoTree can be queried in several ways: by Pfam, KO ID, or NCBI taxon ID. Additionally, species that appear in a BLAST result can be visualized by uploading the BLAST XML2 output file directly. AnnoTree will then generate a highlighted phylogeny using root-to-tip colouring for all lineages containing matches to the query. Distribution summary statistics based on GTDB nomenclature complement visualizations by displaying the number of annotations within the ranks of each taxonomic level. Publication-quality SVG images, Newick formatted phylogenies, and taxonomic distribution tables of all queries can be downloaded for offline analysis or editing. When a highlighted node is selected on the tree, a node detail window appears. It displays basic taxonomic data and annotation confidence scores (i.e. *E*-values) that can be downloaded in CSV format. Annotated protein sequences can also be retrieved from this window.

The AnnoTree website currently hosts the visualization of functional annotations in the domain Bacteria, but its modular construction extends its use to any functionally annotated genomes given in a phylogenetic tree with a defined taxonomy.

2.3 Construction of AnnoTree

The AnnoTree application is a web application with front-end and back-end components designed by Han Chen, Andrew Doxey, and I. The development of the application was split between Han Chen and I, with Han Chen developing most of the front-end components and myself developing most of the back-end components. Required modifications to the front-end and back-end were performed as needed by both Han Chen and I. This section reports the details of each component.

2.3.1 Front-end

The front-end of the AnnoTree application is the portion of the codebase that generates the interface including the layout, images, and text (**Figure 2.1**). AnnoTree is a single page application using modern web frameworks such as D3.js, React.js, and Mobx.

The primary visualization in AnnoTree, the tree, is transformed to an SVG by D3.js after fetching the tree topology and taxonomy data from the back-end server and database. Search queries also result in the generation of an informative D3 donut chart. Elements from the React.js

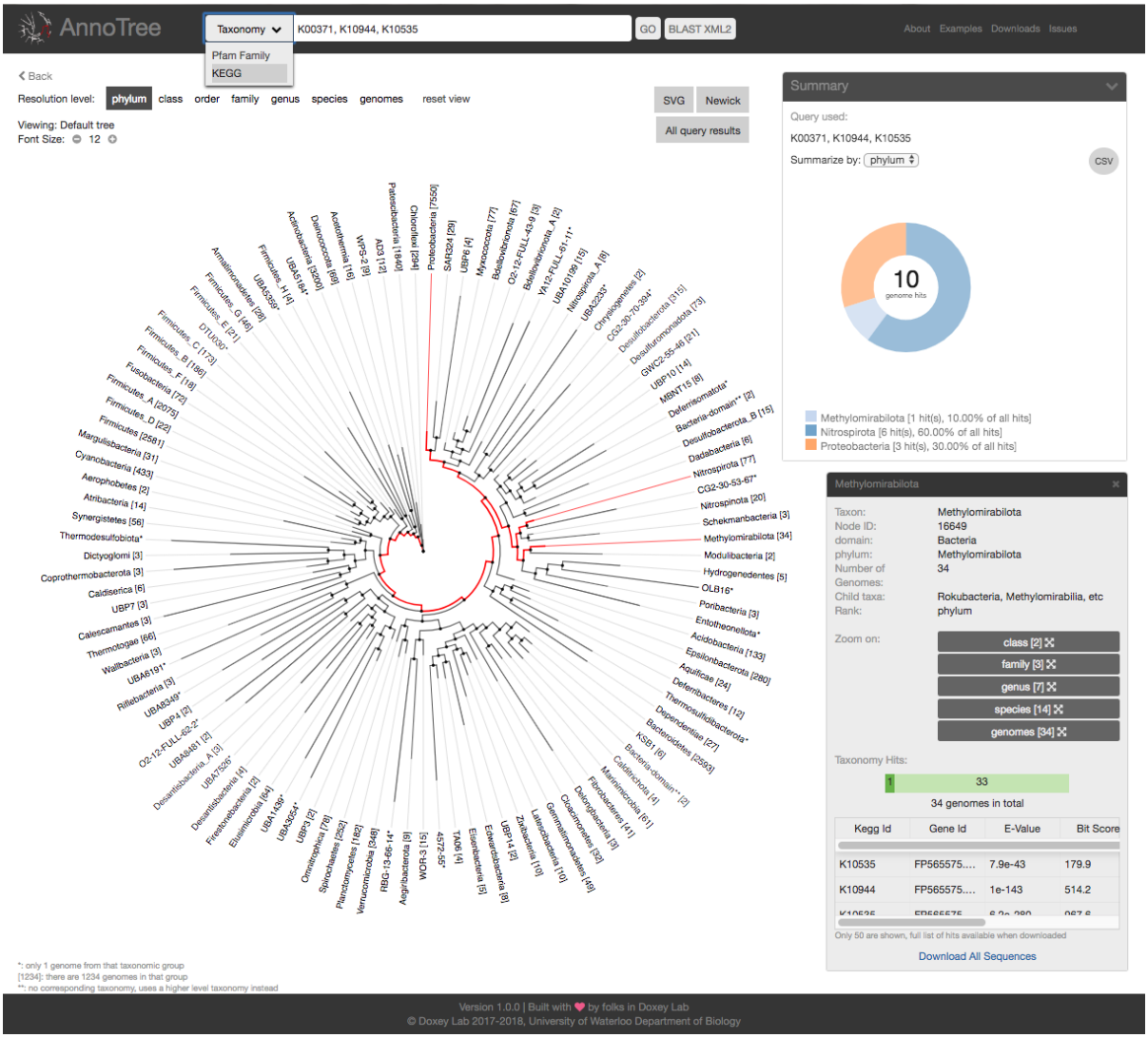

Figure 2.1. The AnnoTree v1.0.0 interface. AnnoTree can be queried with any number of KO IDs, Pfams, or NCBI Taxonomy IDs to display a mapping of those traits on the GTDB tree at any resolution level. Here, lineages containing at least one genome with all three KO genes required for commamox activity (K00371, K10944, and K10535) are highlighted in red. A donut chart displays a taxonomic distribution summary of the genomes containing hits to the query. When a highlighted node is clicked, a window appears displaying taxonomic information, zooming options, and downloadable annotation confidence scores and protein sequences.

library encode all other visual components.

User interactions including clicking action buttons, typing in the query bar, hovering over certain objects, and dragging the node detail window are all tracked by React.js and Mobx. These libraries immediately update the state of these objects and re-render the interface accordingly. In particular, the tree, query box, and summary box objects are marked for tracking by the user's browser with computed and observable properties. Users will perform actions that change the state of an object's observable properties, upon which Mobx updates dependent computed properties. For example, whenever the tree's list of hits (an observable property) is changed, the highlighting pattern (a dependent calculated property) is updated automatically. The combined use of Mobx and React.js is different from the sole use of React.js because the developer does not need to explicitly update dependent properties, leading to reduced code and more reliable performance.

2.3.2 Back-end

The back-end of the AnnoTree application is the portion of the codebase that allows the user's browser to communicate with the server and request data to be displayed. The back-end of AnnoTree is a Python Flask application that performs SQL queries to a MySQL database.

Data sources

All taxonomic, phylogenetic, molecular sequence, and annotation data used in AnnoTree v1.0.0 was obtained from the GTDB website and directly from Donovan Parks, a GTDB developer (Parks et al., 2018). The GTDB developers performed gene prediction on bacterial genomes with Prodigal v2.6.3 (Hyatt et al., 2010). Genes were annotated using the Pfam v27 (Finn et al., 2014) and UniRef100 (Suzek et al., 2015) databases (downloaded March 6, 2018). Pfam annotations were identified using HMMER v3.1b1 (Eddy, 2011) with model-specific cutoff values for the Pfam (`-cut_gc`) HMMs. Pfam annotations were assigned using the same methodology as the Sanger Institute, which accounts for homologous relationships between Pfam clans (see `pfam_scan.pl` on the Sanger Institute FTP site). UniRef100 was used to establish KO annotations by creating a DIAMOND v0.9.22 (Buchfink et al., 2015) database consisting of all UniRef100 clusters with one or more KO identifiers. KO identifiers were then assigned to predicted genes through homology with the following criteria: *E*-value cutoff $\leq$1e-5, percent identity $\geq$30%, and query and subject percent alignments $\geq$70%.

AnnoTree database

When choosing a RDBMS for the AnnoTree application, the nature of the data being stored and how the database would be used were carefully considered. The information is composed of text, integer, and decimal values in the form of tab- and comma-separated value files and FASTA

files. Regarding operation, writes to the database would only occur upon creation; otherwise, the database would only undergo read transactions. There is also a possibility of expanding the application to cover all Archaea and Eukaryotes. Ultimately, these specifications supported the selection of MySQL as the RDBMS. MySQL was chosen over the lightweight SQLite because the ability to scale the application to a larger dataset in the future will not be possible with SQLite without a significant decline in performance. PostgreSQL was not selected because although it is just as popular and well-supported as MySQL, it does not perform as well for applications undergoing read-heavy operations.

The data is arranged into a schema that allows for rapid retrieval of information required by the front-end application (**Figure 2.2**). Since the user explores most data through the tree visualization, all annotation tables are connected to a central `node` table, which also contains tree topology and taxonomy data. Query suggestions in the search bar are drawn from tables populated with data taken directly from the annotation sources. The annotation '`tree_count`' tables were used in combination with the `gtdb_node` table to produce the '`node_ids`' tables. These tables are critical for the rapid retrieval of the list of nodes to highlight in search queries. Annotation confidence scores and protein sequences are fetched from the corresponding '`top_hits`' and the `protein_sequences` tables when a user selects a highlighted node on the tree. The `db_config_data_files` table is used to store the metadata for the flat files that were used to create the database.

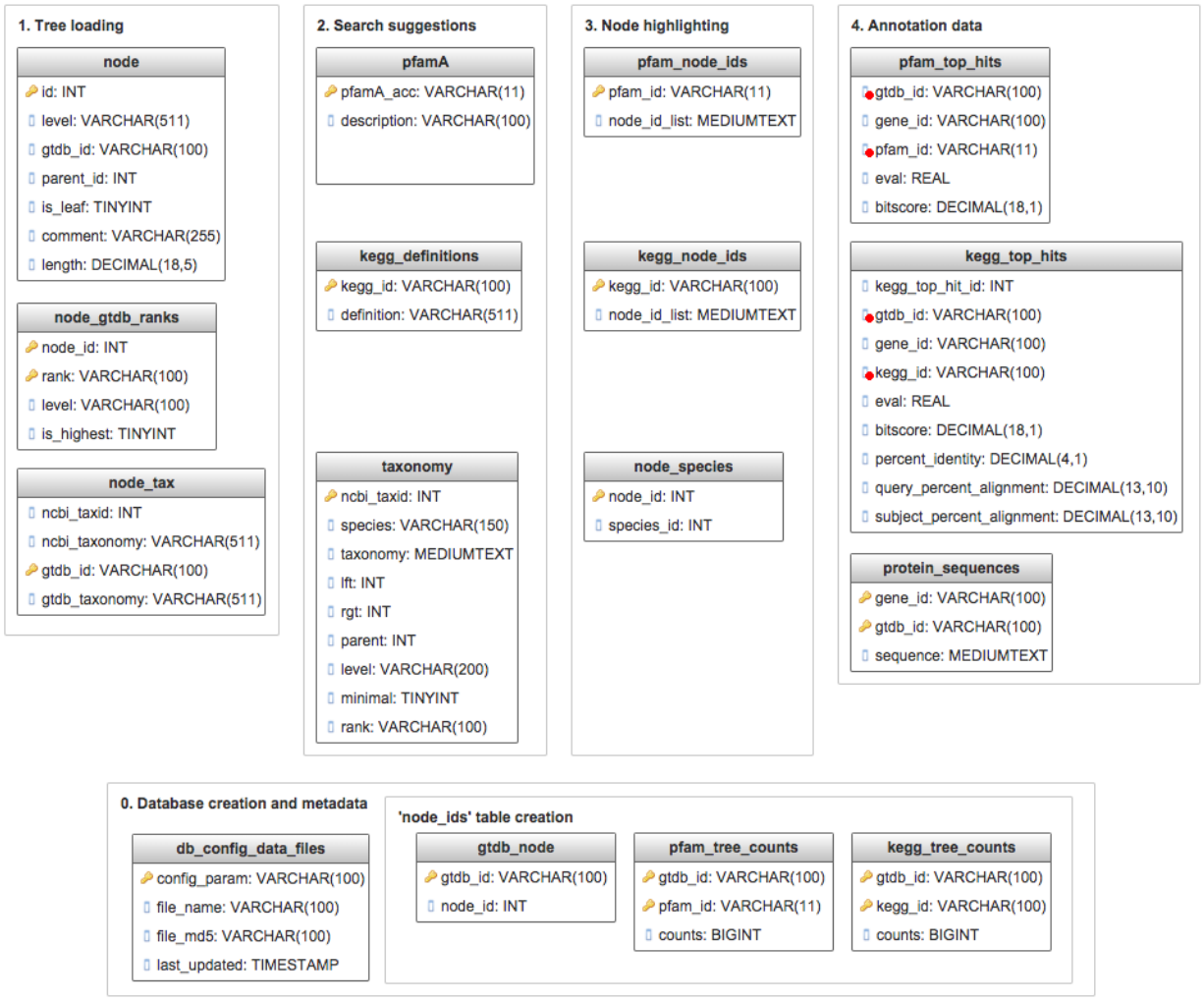

Figure 2.2. The AnnoTree v1.0.0 MySQL database schema. Tables are grouped according to the function in which their containing data serves the front-end application. The groups are numbered in the order in which they would be referenced in an annotation or taxonomy search. Yellow keys beside column names indicate the single or composite primary key within each table. Indexed columns that are not primary keys are indicated by a red dot.

Server-side application

The front-end application submits HTTP `GET` requests with JSON data to REST API endpoints specified by a Python Flask application that sits on top of an Apache web server (**Figure 2.3**). The Flask application generates the SQL statements that are required to obtain the requested data from the MySQL database. The data is returned to the user through an HTTP reply that is converted back to JSON format by JavaScript functions in the front-end. Since Python is not a native web language, the Apache WSGI module, mod_wsgi, was required to enable the communication between the Flask application and the Apache server.

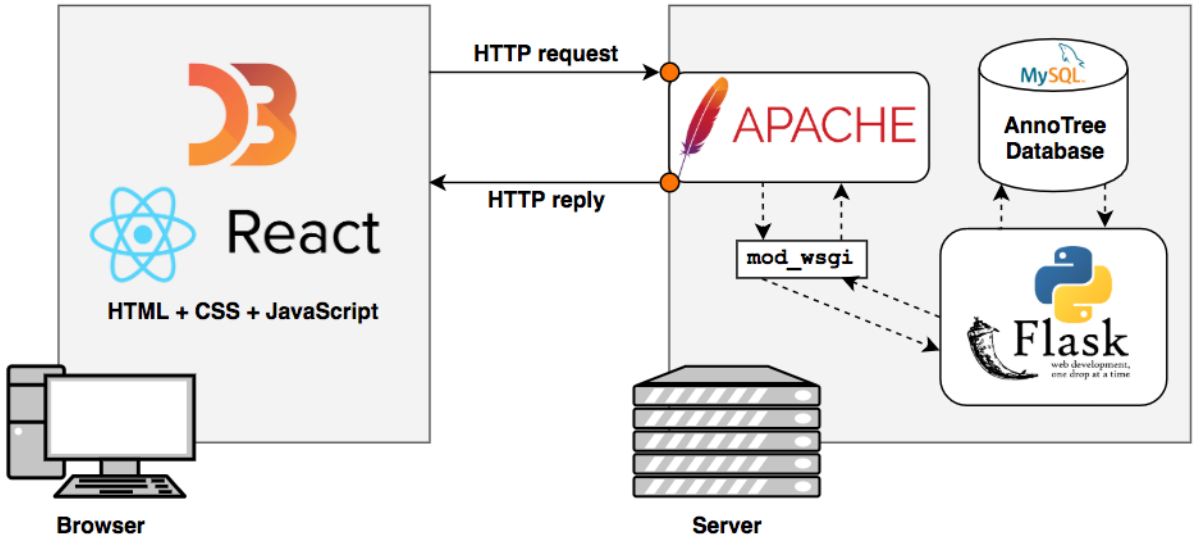

Figure 2.3. The AnnoTree v1.0.0 web application stack. The AnnoTree web application is serviced by a Python Flask application that communicates with an Apache web server through mod_wsgi. The user submits HTTP requests from a browser to obtain data from the MySQL database.

2.4 Installation and system requirements

All code required to set up a local database and run the latest version of AnnoTree is hosted in a Bitbucket repository for each application component (**Table 2.1**).

2.4.1 Setup with the latest AnnoTree data

Anyone running Docker and that has sufficient disk space for the newest database may download and host a local instance of AnnoTree using the scripts written by Han Chen and I in the Docker Compose repository (**Table 2.1**). This functionality is useful for those that wish to run the application on a computer that cannot connect to the Internet or for quick setup in case of a server outage. Briefly, the Docker Compose scripts perform the following steps: (1) Pull the `annotree-backend` and `annotree-frontend` repositories from BitBucket; (2) Configure a MySQL Docker image for hosting the data; (3) Set up a local network so that the front-end, back-end, and database containers may access one another; and (4) Download the latest AnnoTree database from the online data repository and store it as a Docker volume. Once finished, the

Table 2.1. AnnoTree repository locations

Content	URL
Scripts for database creation	bitbucket.org/doxeylab/annotree-scripts
Code for the interface	bitbucket.org/doxeylab/annotree-frontend
Code for the back-end server	bitbucket.org/doxeylab/annotree-backend
Docker Compose setup	bitbucket.org/doxeylab/annotree-docker-compose

application can be loaded in the browser through the port or URL specified in the configuration file.

2.4.2 Setup with custom or new data

When the GTDB releases a new version of their database, the AnnoTree database needs to be updated to reflect these changes. The modular nature of the AnnoTree front-end application allows it to work with any dataset that follows the formatting requirements, so custom data may also be used. I developed a Python3 script, `make_db.py`, that constructs the entire database when given a configuration file specifying the paths to properly-formatted taxonomy, metadata, tree, protein sequence, and annotation files. It is available in a BitBucket repository (**Table 2.1**) along with current system requirements, instructions, example input files, and helper scripts. When using new or custom data, the front-end and back-end components must be set up independently or by modifying the database component of the Docker Compose scripts.

2.5 AnnoTree use case examples

2.5.1 Visualizing the taxonomic distribution of annotations

The study of bacterial toxins, such as the botulinum and tetanus neurotoxins, is essential for the prevention of their toxicity to humans and their applications as protein therapeutics and cosmetic

enhancers. The discovery of homologs of important bacterial toxins outside of their respective bacterial lineages can be reproduced and visualized phylogenetically using simple AnnoTree queries. A query with Pfam PF01742 (botulinum neurotoxin protease) reveals a taxonomic distribution outside of *Clostridium* including the lineages *Weissella* and *Chryseobacterium*, consistent with earlier analyses (Mansfield et al., 2015, 2017) (http://annotree.uwaterloo.ca/app/#/?qtype=pfam&qstring=PF01742). AnnoTree also suggests the presence of the domain in a *Eubacterium* genome and a few unclassified genomes in the class Clostridia.

Similarly, a search with the diphtheria toxin domains (catalytic: PF02763 or translocation: PF02764) reveals homologs in related genera *Streptomyces* and *Austwickia*, again reproducing recent analyses (Mansfield et al., 2018) almost instantaneously (http://annotree.uwaterloo.ca/app/#/?qtype=pfam&qstring=PF02764). The search also reveals the presence of the diphtheria translocation domain (PF02764) in other genomes within the phylum Actinobacteria. The presence of even one domain may give rise to new hypotheses regarding the evolution of these multi-domain proteins.

Additionally, the recent metagenomics-driven discovery of commamox bacteria (van Kessel et al., 2015; Daims et al., 2015) can be reproduced through a simple AnnoTree query by searching for genomes possessing all three essential genes that act as a signature for commamox activity: KO terms K00371 (*nxrB*), K10944 (*amoA*), and K10535 (*hao*) (http://annotree.uwaterloo.ca/app/#/?qtype=kegg&qstring=K00371%2C%20K10944%2C%20K10535). Highlighted in the tree are

the known commamox species (i.e. organisms within the genus *Nitrospira*), along with several additional taxa implicated as having potential commamox-like activity (e.g., *Crenothrix*; **Figure 2.1**).

It is possible for a highlighted genome to contain an annotation due to the presence of a contaminating sequence in the assembly. Annotation confidence scores, protein sequences, and a link to the GTDB genome entry are made available in AnnoTree so that users can verify the annotation using other methods.

2.5.2 Observing the taxonomic distribution of an NCBI BLAST result

AnnoTree v1.0.0 contains many pre-calculated annotations, but it is limited to the annotations defined in the Pfam v27 and UniRef100 (March 6, 2018) databases. It is possible to work around this limitation by providing the result of an NCBI BLAST search to AnnoTree. The tree displays a blue root-to-tip mapping to genomes with an NCBI Taxonomy ID that matches a BLAST hit from the XML2 file. For example, the recently-characterized heliorhodopsin domain (Pushkarev et al., 2018; PF18761) only appears in the most recent Pfam release (v32) and is therefore not yet present in AnnoTree's internal database. A representative protein sequence from *Helicobacterium* sp. DL1 (GenBank accession AHG04535) was used as a query in a BLASTP search against the NCBI nr database with default parameters, restricted to the domain Bacteria (NCBI:txid2), and with an *E*-value cutoff of 1e-5. The search resulted in 394 significant hits. Results were

downloaded as a single-file XML2 to be uploaded to AnnoTree for visualization through the 'Taxonomy' option in the query bar. The taxonomic distribution of the BLAST query with the representative heliorhodopsin protein resulted in hits across 9% of all bacterial phyla in the GTDB taxonomy (Release 02-RS83), including all those reported by Pushkarev et al. (2018) (**Figure 2.4**).

2.6 Conclusion

The recent development of the standardized and complete GTDB taxonomy has made it possible to simultaneously explore bacterial taxonomy and phylogeny in a user-friendly way. In this work, I presented AnnoTree, the first application to offer this functionality in addition to the visualization of the phylogenetic distribution of millions of precomputed protein domain (Pfam) and functional ortholog (KO) annotations. Instructions for the setup and maintenance of AnnoTree with the GTDB data and a custom dataset were given for the purpose of upgrading the current database or applying the AnnoTree framework to a custom dataset. AnnoTree was used to replicate findings from three different publications within minutes, demonstrating its value as an exploratory tool in the study of bacterial and protein evolution.

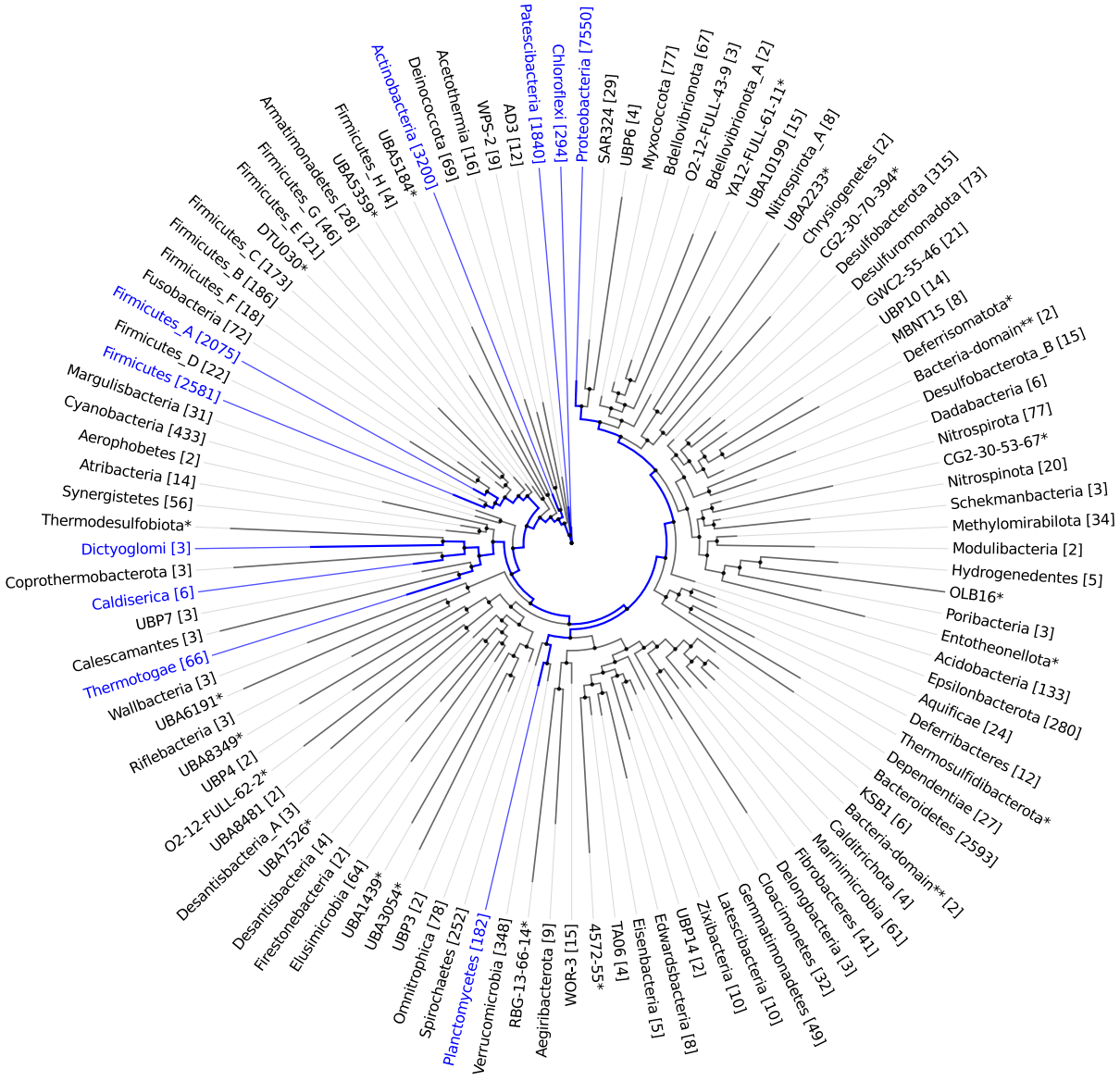

Figure 2.4. Phylogenetic distribution of heliorhodopsin BLAST hits. The AnnoTree web application can be used to instantly visualize the taxonomic diversity of functional annotations using the BLAST XML2 output file from NCBI. Taxa containing putative bacterial heliorhodopsins were obtained from a BLASTP search using the protein sequence identified in *Helicobacterium* sp. DL1 (GenBank accession AHG04535) and mapped to the GTDB tree in blue using AnnoTree's BLAST XML2 upload feature.

Chapter 3

The phylogenetic distribution of functional traits

3.1 Distribution of Pfam and KEGG annotations in bacteria

3.1.1 Background

Phylogenomic studies have been beneficial in providing supporting evidence for functions involved in the mechanisms of bacterial evolution (Chai et al., 2014). The incorporation of MAGs in the newest tree of life and the increase in annotation entries warrant a revisit to these analyses. The data generated for integration into the AnnoTree web application is the largest of its kind to

be analyzed in a phylogenomic context. These data include the GTDB tree (Release 02-RS83; Parks et al., 2018), which contains 23,936 fully annotated bacterial genomes from 109 different phyla. Results of these analyses aim to confirm and supplement previous efforts of characterizing high-level functional trends in bacterial evolution.

All Pfams and KOs were ranked by normalized CI, a measure of the patchiness of a trait on a phylogeny. The most homoplasic genes and protein families were screened for functions that may give insight into the mechanisms of HGT or convergent evolution. The least homoplasic traits and those that were highly conserved in a lineage were also noted, as they may indicate functional complexity or ancient events of genetic drift and natural selection.

3.1.2 Methods

Gene prediction, annotation, and profile generation

Gene prediction was performed on bacterial genomes obtained from the GTDB (Release 02-RS83) with Prodigal v2.6.3 (Hyatt et al., 2010). Genes were annotated using the Pfam v27 (Finn et al., 2014) and UniRef100 (Suzek et al., 2015) databases (downloaded March 6, 2018). Pfam annotations were identified using HMMER v3.1b1 (Eddy, 2011) with model specific cutoff values for the Pfam (`--cut_gc`) HMMs. Pfam annotations were assigned using the same methodology as the Sanger Institute, which accounts for homologous relationships between Pfam clans (see `pfam_scan.pl` on the Sanger Institute FTP site). UniRef100 was used to establish

KO annotations by creating a DIAMOND v0.9.22 (Buchfink et al., 2015) database consisting of all UniRef100 clusters with one or more KO identifiers. KO identifiers were then assigned to predicted genes through homology with the following criteria: *E*-value cutoff $\leq$1e-5, percent identity $\geq$30%, and query and subject percent alignments $\geq$70%.

A count matrix was computed for each trait and genome combination based on the annotation methods described above. The count matrices were converted to binary presence/absence profiles for all analyses, where a genome with at least one qualifying hit score for a trait was assigned '1' and '0' otherwise.

Benchmarking the measurement of homoplasy metrics in R

A sample of 120 Pfam and KO phylogenetic presence/absence profiles with evenly-distributed family sizes was taken from the full set of 28,311 Pfams and KOs for benchmarking on a Lenovo workstation (3.5 GHz Intel(R) Xeon(R) CPU E3-1275 V2 with 32 GB RAM) using a single processor for up to two days. Calculations were performed three times with default parameters unless otherwise noted. Some algorithms also had trouble with zero branch lengths in the phylogenetic tree, so branches meeting this criterion were assigned the height of the root node * 1e-6. The CI (Kluge and Farris, 1969) and RI (Farris, 1973, 1989) were calculated for each of the annotations and the GTDB bacterial tree using the `CI` and `RI` functions, respectively, in the phangorn R package (Schliep, 2011). The HSR was calculated similarly with a custom script that

utilizes the algorithm described in Meier et al. (1991) and functions in the phangorn R package (Schliep, 2011) (https://bitbucket.org/doxeylab/annotree-scripts/src/master/homoplasy/HSR.R). The random homoplasy slope was calculated using 100 randomly-drawn presence/absence profiles with equal probability of presence and absence. Abouheif's C_{mean} (Abouheif, 1999) was calculated using the `abouheif.moran` function in the adephylo R package (Jombart et al., 2010). Pagel's λ (Pagel, 1999) was calculated with the `fitDiscrete` function of the geiger R package through the use of the 'ARD' model (Harmon et al., 2008). Fritz and Purvis' D metric (Fritz and Purvis, 2010) for phylogenetic signal was calculated similarly using the `phylo.d` function of the caper R package (Orme et al., 2013). The trait depth (τ_D) (Martiny et al., 2013) was calculated using the `get_trait_depth` function of the castor R package (Louca et al., 2018).

Contamination sensitivity analysis of normalized CI

When screening for contaminating sequences in genome assemblies, It was assumed that, out of a set of taxa containing an annotation, an annotation on a contaminating sequence would most likely occur in the taxon that is the most evolutionarily distant from the others. First, the genomes containing the annotation of interest were extracted from the GTDB tree using functions in the APE package of R (Paradis et al., 2004). The tips with the longest edges were dropped one at a time until the desired number of tips, as determined by the indicated contamination level, had been removed. The phylogenetic profile was modified by changing the values in the

profile corresponding to the positions of the dropped tips from '1' to '0'. The normalized CI (ln(CI)/ln(family size)) for the annotation was re-calculated using the modified profile and the reduced family size using the methods specified in the benchmarking experiment. Re-calculations were performed for contamination levels of 1%, 5%, 10%, and a random value for each annotation drawn from a gamma distribution with α=0.7 and β=0.3.

Kendall's W, a measure of concordance, was calculated for the ranked annotations for each contamination level using the `kendall.global` and `kendall.post` functions in the vegan R package (Kendall and Smith, 1939; Dixon, 2003).

Calculating the significance of phylogenetic conservation

The trait depth (τ_D) for each annotation profile on the GTDB tree was calculated similarly to how they were done in the benchmarking experiments above. A trait was classified as phylogenetically conserved if the probability of encountering a profile with such a τ_D or higher is less than 5% (ie. $P<0.05$) based on 1000 different independently- and randomly-drawn binary presence/absence profiles where the probability of a tip exhibiting the trait is equal to the proportion of positive states in the trait's profile.

Classification of lineage-specific traits

Traits were classified as lineage-specific if there was at least one clade in the tree where at least 95% of presence states occurred in at least 95% of the taxa in that clade and that no more than half of the genomes in the tree contained the trait. The node furthest from the root of the GTDB tree passing these criteria was assigned the root of the lineage-specific clade for that trait. The trait's taxonomic rank was selected as the lowest taxonomic rank shared between all genomes of the lineage-specific clade.

Taxonomic rank homoplasy enrichment analysis

Annotations contained within fewer than 50 genomes were removed before verifying taxonomic enrichment of homoplasic traits for each annotation type. Taxonomic rank presence/absence profiles for each trait were generated for each taxonomic rank by combining the profiles of all genomes in the rank; '1' was assigned if at least one genome possessed the trait and '0' otherwise. Next, traits were ordered by increasing ln(CI)/ln(family size) with CIs calculated as in the benchmarking experiments. Each taxonomic rank at each taxonomic level was tested for over-enrichment within the 5% most homoplasic traits in Bacteria (KO: 618; Pfam: 552) using the hypergeometric test. The tests were conducted similarly to those done in Nasir et al. (2012). *P* values were obtained using the `fisher.test` function of R with the 'alternative' option set to 'greater' (R Core Team, 2018). The contingency table is given in **Table 3.1**. P values

Table 3.1. Taxon enrichment contingency table

	Category 1 ($\in$ rank)	**Category 2 ($\notin$ rank)**
Class 1 ($\in$ homoplasic trait)	k	$n - k$
Class 2 ($\notin$ homoplasic trait)	$M - k$	$N - M - n + k$

The number of different homoplasic traits within the rank is k, n is the number of ranks that contain at least one of the homoplasic traits, M is the total number of different traits within the rank, and N is the total number of different traits.

were corrected for multiple tests at each taxonomic level using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995).

3.1.3 Results and Discussion

Homoplasy and phylogenetically scattered traits

Phylogenetic patchiness metric selection

Due to the possibility of incorporating the phylogenetic patchiness data in the AnnoTree webtool in future versions, it was beneficial to select metrics that are used widely in recent studies, and that can be easily recalculated for database updates. Candidate homoplasy metrics were selected based on their use in recent publications and mention in recent reviews (Rodríguez-Torres et al., 2017; Comte et al., 2014; Speed and Arbuckle, 2017). Computational runtime and algorithm availability were used to assess whether candidate metrics would be easy to apply to current and future AnnoTree data sets. Calculations of each candidate metric were performed on 120 Pfam and KO binary presence/absence profiles and the GTDB tree to evaluate the computational

runtime (**Appendix A**). Preliminary analyses showed that some calculations had a high variance in runtime that was correlated with the presence/absence ratio, so annotations were selected such that profiles were representing the full range of presence/absence state proportions. Some algorithms also had trouble with zero branch lengths in the phylogenetic tree, so branches meeting this criterion were assigned the height of the root node * 1e-6. Only algorithms that could be performed without a user interface were considered so that the AnnoTree database update pipeline could remain automated and hands-free. All calculations were run independently on the same Lenovo workstation (3.5 GHz Intel(R) Xeon(R) CPU E3-1275 V2 with 32 GB RAM) using a single processor for a maximum of two days.

The calculations of the CI, RI, HSR, and trait depth (τ_D) metrics all finished in reasonable times with RI being the fastest and HSR being the slowest of these (**Table 3.2**). The calculation times of Abouheif's C_{mean}, Fritz and Purvis' D, and Pagel's λ exceeded the allotted time of two days for the set of 120 annotations. Due to the size of the tree, these functions also had substantial memory requirements and could not be run on a workstation with less than 16 GB of RAM. Based on their performance in the benchmarking experiment and their popularity in the literature, CI and Martiny's τ_D were selected for incorporation into a future version of AnnoTree and were used in further analyses.

A trait's CI shows a strong negative correlation with the number of genomes containing the trait, here termed family size, so it could not be used directly for ranking (**Figure 3.1a**). By

Table 3.2. Results of the homoplasy metric benchmarking experiment

Metric	Average Elapsed Time (seconds/trait)	Standard Deviation
RI	0.486	2.763E-03
CI	0.501	4.132E-03
Trait depth (τ_D)	1.337	6.014E-03
HSR	48.695	8.040E-01
Abouheif's C_{mean}	>1440	-
Fritz and Purvis' D	>1440	-
Pagel's λ	>1440	-

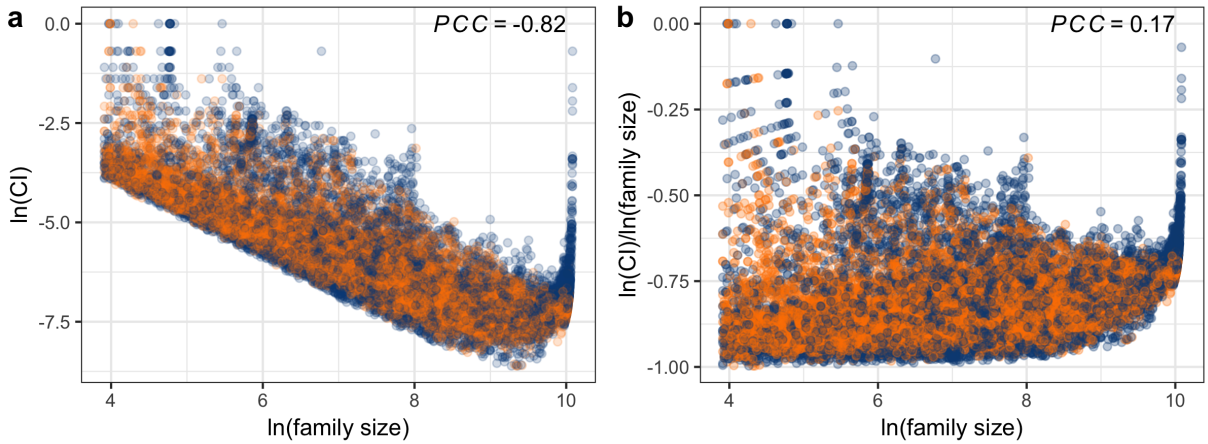

Figure 3.1. The effect of family size on CI. Visual representation of the correlation of family size with the CI metric for all Pfam (blue) and KEGG (orange) annotations before (**a**) and after (**b**) normalization. *PCC* = Pearson's correlation coefficient.

applying a natural logarithmic transformation and dividing the CI by family size, the correlation between the two variables is notably reduced (**Figure 3.1b**). This normalized metric was used in further analyses.

Sensitivity analysis of phylogenetic patchiness rankings to contaminating sequences

Contaminating sequences are a problem in all forms of genome assembly, but especially in MAGs

(Parks et al., 2015). Since a large proportion of genomes in the GTDB tree are derived from metagenomes, it is necessary to test the sensitivity of contaminating sequences on the ranking of annotations by the selected phylogenetic patchiness score: normalized CI. All Pfam and KO annotations were re-analyzed assuming the presence of varying amounts of contaminating sequences.

In most cases, contaminants are identified by their comparatively long taxonomic distance from all other annotation-containing genomes. Here, the most distant taxa were labelled as contaminants and were changed from '1' to '0' in the phylogenetic profile of the annotation. The modified profile was used to re-calculate the normalized CI. Since a functional characterization of contaminating sequences in MAGs has not been reported, simulated contamination levels were selected with the assumption that all functions have an equal chance of being characterized on a contaminating sequence. Thus, annotation contamination levels of 1%, 5%, 10%, and a random value drawn from a gamma distribution modelled after true sequence contamination levels of MAGs reported in Parks et al. (2017) were selected for simulation (**Figure 3.2**).

A concordance test using Kendall's W (Kendall and Smith, 1939) showed that values from all simulated functional contamination levels were in concordance in all cases (W per annotation$>$0.98). There is a markedly high agreement in rank at the highest (i.e. more conserved) and lowest (i.e. more homoplasic) ranks, meaning that conclusions based on these annotations are valid for traits with contamination levels$\leq$10%. Due to the high agreement of the ranking of

contamination-filtered traits by homoplasy score and the ranking of unfiltered traits by the same metric, all further analyses were performed without filtering of traits by contamination level.

Analysis of the phylogenetic patchiness of functional traits

As an initial exploration of the data within AnnoTree, the distributions of all 77,004,395 Pfam and KO annotations were examined when mapped onto the GTDB bacterial tree of life. Based on the phylogenetic conservation score, trait depth (τ_D) (Martiny et al., 2013), 68.1% of KOs and 60.0% of Pfams had significantly non-random phylogenomic distributions ($P<0.05$), revealing a greater phylogenetic congruency for KO annotations than Pfam annotations obtained using their standard confidence score thresholds.

Although 60-68% of functional annotations show a significant phylogenetic signal when mapped onto the tree, more surprising are the remaining 30-40% that show more random phylogenetic distributions, potentially reflecting the widespread horizontal transfer and/or frequent gene gain/loss that is known to occur in bacterial genomes (Ochman et al., 2000). To investigate this further, all Pfam and KEGG annotations were ranked according to their phylogenetic patchiness as determined by the normalized CI metric (ln(CI)/ln(family size)). The normalized CI ranges from -1 to 0, where homoplasic traits tend toward -1.

For a visual comparison of high-level trends, KOs present in at least 50 genomes were grouped into their high-level functional categories (**Figure 3.3, Appendix Table A1**). The functional

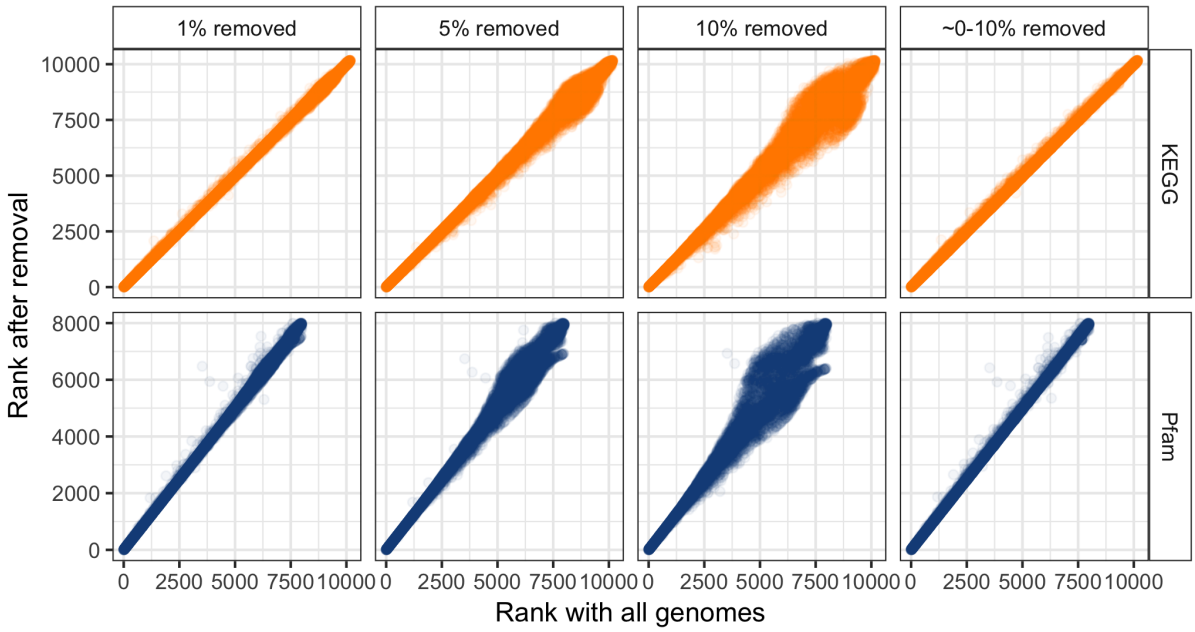

Figure 3.2. Homoplasy rank sensitivity to contamination. Normalized CIs were calculated for each annotation after removal of a proportion of the annotations that are most likely to be identified within contaminating sequences. These annotations were identified as the those that were present in the genome that is the most taxonomically distant from all other genomes containing the annotation. Homoplasy ranks from each simulated contamination level are plotted against the homoplasy ranks from unmodified presence/absence profiles. In the '~0-10%' contamination level simulation, the contamination proportion for each annotation was randomly selected from a gamma distribution modelled after typical contamination levels reported in MAGs (Parks et al., 2017).

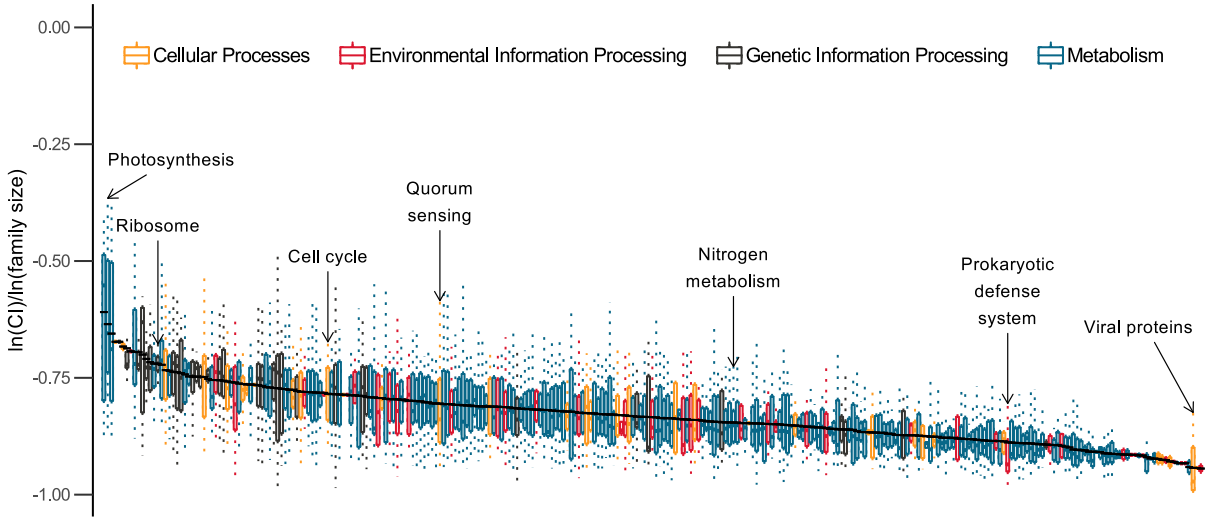

Figure 3.3. Homoplasy of KEGG categories. Phylogenetic patchiness was computed for each KO using the CI, a common homoplasy metric representing the inverse of the minimum possible number of state changes (trait gain or loss) given the tree topology. The final phylogenetic patchiness score is equal to ln(CI)/ln(family size) where family size is the total number of genomes containing the trait. Mean-sorted box plots comprised of normalized CI scores from KOs in their respective KEGG pathways and KEGG BRITE categories. The black horizontal lines through each of the box plots represent the mean normalized CI of a set of KOs in a KEGG pathway or KEGG BRITE category.

categories within each of the highest functional classes were evenly distributed by homoplasy rank except for 'Genetic Information Processing', whose encompassing categories tended to be less patchy (i.e. normalized CI closer to 0).

Not surprisingly, viral functions for gene mobility and bacteriophage defence were among the most homoplasic KEGG functional categories. These functions are also present in the list of the most homoplasic KO terms (**Appendix Table A2**). This list is dominated by transposases and CRISPR- and bacteriophage-associated gene families. The extreme phylogenetic patchiness of

bacteriophage and CRISPR genes further supports the existence of an ongoing arms race between these two opposing biological forces (Levin et al., 2013). Other biologically relevant members of the most highly scattered KOs include K19057-K19059 (*merC*, *merD*, and *merR*) for mercury resistance; K19155 and K19156, components of a toxin-antitoxin system characterized in *E. coli*; K15943, K15945, and K16411 for polyketide antibiotic biosynthesis; and K19173-K19175 for DNA backbone S-modification (phosphorothioation).

Since Pfams are characterized from a more taxonomically diverse sequence database than KOs, there was an initial expectation that the functional trends in the least and most homoplasic annotations would differ slightly between them. The most homoplasic Pfams were similar to the most homoplasic KOs in that most were associated with mobile genetic elements and bacteriophages (**Appendix Table A4**). There were several homoplasic DUFs, but further inspection of a large number of them suggests that they are also associated with bacteriophages. There were no CRISPR-associated Pfams among the most homoplasic domains, contrary to the KEGG term results. The absence of these domains may be due to the inclusion of more uncultivated species in the Pfam database relative to the KEGG database and the observed lack of CRISPR-Cas systems in uncultivable symbionts (Burstein et al., 2016). Prophage-related Pfams encoded by these species appear to be pushing the CRISPR-related domains down the list.

The top 5% of highly scattered KOs showed significant over-representation among the genera *Pseudomonas_E*, *Streptomyces*, *Mycobacterium*, and *Nocardia*, suggesting that these taxa may be

hotspots of HGT (**Appendix Table A6**). There are many more genera significantly enriched with the top 5% of homoplasic Pfams than the top 5% of homoplasic KOs. Among the most enriched taxa in homoplasic Pfams are *Pseudomonas_E*, *Acinetobacter*, *Enterobacter*, *Burkholderia*, and *Xenorhabdus* (**Appendix Table A7**). All of the genera most significantly enriched in homoplasic KOs and Pfams include species that are symbionts or pathogens. Such relationships require evolutionary flexibility that is provided efficiently through horizontal transfer of genes encoded by bacteriophages (Bondy-Denomy and Davidson, 2014).

KEGG functional categories exhibiting the least phylogenetic patchiness include photosynthesis and core processes such as transcription, DNA replication, and protein synthesis. KOs involved in sporulation were not seen in the high-level analysis, but they are among the least homoplasic functions based on inspection of individually ranked terms (**Appendix Table A3**). The bias for research in biomedically relevant species is apparent in the least homoplasic KOs since this list contains a large number of species-specific genes. For example, more than half of the top 20 terms ranked by normalized CI are adhesion and virulence genes identified in *Helicobacter pylori* (K11028, K15843-K15848) and are seemingly conserved due to the strict taxonomic specificity of the annotation.

The least homoplasic Pfams were associated with biofilm formation as well as the sporulation-related functions and core genetic processes identified in the least homoplasic KO terms (**Appendix Table A5**). A higher proportion of the least homoplasic Pfams than the most homoplasic Pfams

are uncharacterized. This observation is likely the result of prioritized research efforts, which promote the characterization of functions that are nearly ubiquitous across all life or that appear in high-throughput analyses in heavily-studied biological and ecological systems such as the human body and biological wastewater treatment (Galperin and Koonin, 2004; Chang et al., 2016).

Lineage-specific traits

The distributions of Pfam and KEGG annotations were analyzed to identify those with strong lineage-specificity that may have contributed to the lineage's evolutionary divergence from its ancestors. Classification of lineage-specific annotations within a clade has not been formally defined in the literature. Here, a trait, t, is classified as lineage-specific within a clade, C, if the catchment and saturation of t in C are below a given threshold. The catchment is defined here as the proportion of genomes containing t that are present within C. It represents the exclusiveness of the trait to the lineage. Saturation is defined here as the proportion of genomes in C that contain t. This parameter is necessary to filter out annotations that are not conserved, leaving only traits that are more likely to have had an active role in the evolutionary separation of that lineage from its ancestors. An example demonstrating how catchment and saturation are determined for a tree and annotation profile is given in **Figure 3.4**.

To determine the appropriate threshold values for saturation and catchment, combinations of each variable were tested on the GTDB bacterial tree with all KOs and Pfams present in no

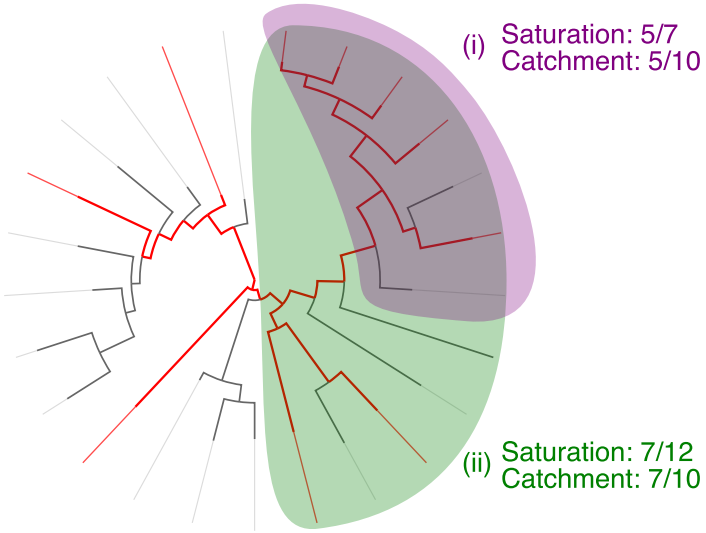

Figure 3.4. Saturation and catchment of a trait on a phylogeny. Tree exported from AnnoTree visualizing the KO term K00371 (*nxrB*: nitrate reductase/nitrite reductase) within the phylum Nitrospirota at the family level. The tree is painted red from root to tip to indicate the genomes containing the trait. Catchment and saturation values for K00371 in clades (i) and (ii) are indicated to the right of clade labels, where catchment is the proportion of K00371-containing genomes present within the clade and saturation is the proportion of genomes within the clade that possess K00371.

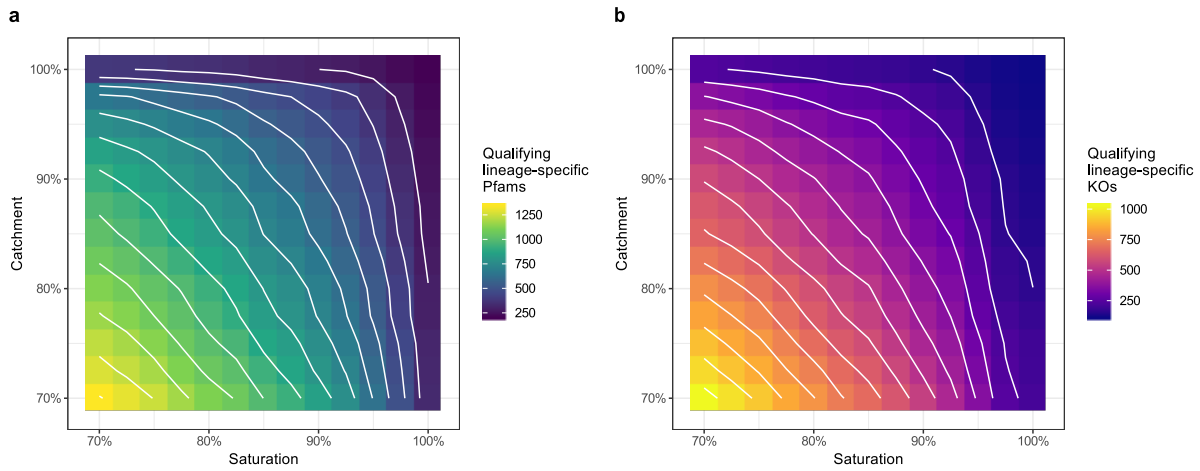

Figure 3.5. Sensitivity of lineage-specific annotations to catchment and saturation. Saturation and catchment cutoffs were applied to the GTDB tree and annotation profiles (Pfam: a; KEGG: b). The heat map displays the number of annotations passing the catchment and saturation filters. The white contour lines show an estimation of the catchment and saturation values for a single amount of qualifying annotations. The change in the number of qualifying annotations is consistent between consecutive contour lines.

more than half of the total genomes in the GTDB tree. This initial filtration eliminates protein families and genes that are necessary for essential household functions (i.e. protein synthesis, DNA replication) as well as those involved in Bacteria-defining processes (i.e. peptidoglycan synthesis), which cannot be differentiated with this data set alone. The number and identities of annotations passing the cutoff thresholds were recorded for each experiment to evaluate the effects of saturation and catchment on these values (**Figure 3.5**).

The contour plots show that saturation is slightly more strict than catchment for the trait distributions tested, but the difference between the numbers of qualifying annotations between

experiments with the same catchment and saturation thresholds is not significant by paired t-test (data not shown). This indicates that function exclusivity within a lineage (i.e. catchment) is more often observed within bacterial evolution than function retention (i.e. saturation).

The cutoff thresholds were set to 95% for catchment and 95% for saturation to obtain a list of the most conserved lineage-specific annotations. Based on these criteria, 358 (3.2%) lineage-specific Pfams and 152 (0.9%) lineage-specific KOs were identified in Bacteria (**Appendix Tables B1, B2**). Lineage-specific KOs and Pfams increased in frequency from higher (e.g., phylum) to lower (e.g., species) taxonomic levels (**Figure 3.6**), consistent with the idea that gene family taxonomic distributions tend to diversify over time and that HGT impacts evolution over short evolutionary timescales (McDonald and Currie, 2017). Although lineage-specific annotations are relatively rare at high taxonomic levels, these cases may represent ancient, clade-defining bacterial innovations. Examples include PF06781, a domain within the CrgA protein required by actinomycetes for sporulation septation in aerial hyphae (Del Sol et al., 2003), and numerous photosynthesis-related genes within the Cyanobacteria (class Oxyphotobacteria).

The classification of lineage-specific traits within the bacterial domain uncovered several instances of apparent cross-domain gene transfer from the archaeal and eukaryotic domains. The most evident example of a taxon in which this is occurring is the *Endozoicomonas* subtree, a clade of endosymbiotic bacteria that inhabit numerous marine eukaryotic hosts (Neave et al., 2016). The lineage-specific protein families and KEGG genes detected within this clade appear to be of

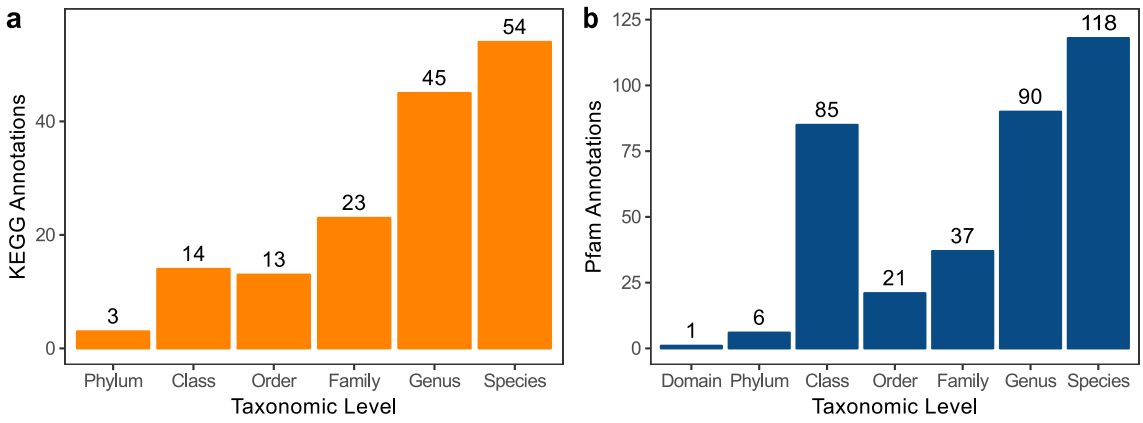

Figure 3.6. Frequency of lineage-specific annotations at each taxonomic level. Internal nodes of the GTDB tree were classified as lineage-specific for a trait if at least 95% of the genomes containing the trait were contained within the clade (i.e. catchment$\geq$0.95), at least 95% of the genomes in the clade had the trait (i.e. saturation$\geq$0.95), and the trait was present in no more than half of the genomes in the GTDB tree. The taxonomic level of each root of a lineage-specific node was classified as the lowest common taxonomic rank between all encompassing genomes. The number of lineage-specific KO annotations (a) and Pfams (b) are displayed in decreasing taxonomic order. The abundance of lineage-specific Pfam annotations at the class level is due to numerous photosynthesis-related Pfams unique to the Oxyphotobacteria.

eukaryotic origin and include those involved in cytoskeletal organization (PF01302), eukaryotic cell-cell signalling (PF00812), apoptosis inhibition (K010343, K010344, K04725, PF07525), and eukaryotic proteolysis (K01378). The presence of these traits within many independently sequenced *Endozoicomonas* genomes suggests that these sequences are not likely contaminants. The uptake and retention of these traits also indicate that they may be contributing positively to the fitness of the bacteria or their host.

Another example of a putative cross-domain gene transfer uncovered in this analysis is the presence of a the eukaryotic-like histone domain within a transcription factor (K11275) encoded by two species of the *Myxococcus* genus. *Myxococcus* spp. are prokaryotes that undergo multicellular development through tightly controlled mechanisms of gene expression, cell movement, and differentiation (Kroos, 2005). The eukaryotic-like histone domain-containing protein in *Myxococcus xanthus*, CarD, has been shown to regulate the expression of many genes required for multicellular development and carotenogenesis (García-Heras et al., 2013).

A third example of a putative cross-domain gene transfer is the presence of a thymosin β_4 homolog (K05764) within the genome of *Roseofilum reptotaenium*, a suspected causative agent of the Caribbean black band disease in corals (Casamatta et al., 2012). The disease can be identified by a band or ring of bacteria with black pigmentation that actively lyse coral tissue (Casamatta et al., 2012). The thymosin's ability to sequester free actin monomers prevents the polymerization of the actin cytoskeleton (Goldstein and Badamchian, 2004), therefore abnormal

levels of thymosin may interrupt cytoskeletal organization in corals, leading to cell death.

Lineage-specific gene families can provide insight into the unique biology of their respective organisms. Some notable examples of lineage-specific annotations that characterize the biology of their respective organisms are the virulence factors and toxins of the well-studied pathogenic genera *Bordetella*, *Helicobacter*, *Legionella*, *Mycoplasma*, and *Vibrio*.

There are also several lineages that are abundant with lineage-specific DUFs. The taxonomic ranks with the highest number of lineage-specific DUFs are *Bacillus_A* (genus; 37), Oxyphotobacteria (class; 30), Actinobacteria (class; 14), and Enterobacteriaceae (family; 5). A DUF is considered to be essential if it is present in a gene that is necessary for the survival of an organism under favourable growth conditions (Goodacre et al., 2014). Some lineage-specific DUFs overlap with a recent list of essential DUFs (**Table 3.3**). Knowing that these traits are lineage-specific can lead to the generation of more accurate hypotheses regarding their function.

Surprisingly, many MAGs also contained lineage-specific annotations (**Appendix Figures B1, B3**). Through a simple database search of these annotations, it was discovered that they are primarily found within eukaryotic and archaeal sequences (data not shown). Due to the low number of genomes representing these taxonomic ranks (<3 for 81% of occurrences) and the degree of contamination permitted in most of these MAGs (Parks et al., 2017), it is likely that these sequences are derived from contaminating sequences.

Table 3.3. Lineage-specific Pfams also listed as essential DUFs

Pfam ID	DUF #	Lowest Common Rank
PF04217	DUF412	Enterobacterales
PF06242	DUF1013	Alphaproteobacteria
PF07288	DUF1447	Bacilli
PF07372	DUF1491	Alphaproteobacteria
PF10398	DUF2443	Helicobacteraceae
PF10954	DUF2755	Enterobacteriaceae
PF10969	DUF2771	Corynebacteriales
PF11268	DUF3071	Actinobacteria
PF11297	DUF3098	Bacteroidetes
PF11826	DUF3346	Vibrionaceae
PF12506	DUF3713	*Mycoplasma_C*
PF13179	DUF4006	Campylobacterales
PF13829	DUF4191	Actinobacteria
PF14123	DUF4290	Bacteroidia

Essential DUFs are regarded as prominent candidates for characterization due to their importance for the survival of an organism under favourable growth conditions. The lineage-specific Pfams listed here are also listed as essential DUFs in Goodacre et al. (2014).

3.1.4 Conclusion

In this work, the evolutionary dynamics of 28,311 Pfams and KOs were analyzed through quantification of their phylogenetic distribution across 23,936 bacterial genomes. High-level functional trends of the most and least homoplasic traits were determined through the ranking of the traits in terms of a normalized CI (ln(CI)/ln(family size)). The set of taxa that are enriched with the most homoplasic terms, related to bacteriophages and MGEs, were determined. Two new terms, saturation and catchment, were defined to systematically classify lineage-specific traits based on their evolutionary preservation and their exclusivity within a lineage, respectively. High cutoff thresholds of these terms led to the identification of previously-characterized clade-defining innovations within bacteria as well as putative instances of cross-domain HGT.

3.2 Distribution of protein families across all life

3.2.1 Background

Detecting, visualizing, and analyzing the distributions of protein domain families across species is imperative for understanding their evolutionary history and functional importance in different lineages (Yang and Bourne, 2009). By examining the presence/absence and abundance of protein domains across species phylogenies, it is possible to reconstruct their histories and identify key

evolutionary events such as gains, losses, and horizontal transfer events (Yang and Bourne, 2009).

The tree of life recently published by Hug et al. (2016) is the most comprehensive model of species evolution that is currently available. The tree was generated using concatenated ribosomal protein sequences from high-quality genomes of 2684 bacterial, 169 eukaryotic, and 230 archaeal species from different genera. Here, the evolutionary histories of all pre-computed protein families from the Pfam database (v31.0; Finn et al., 2016) are mapped onto the tree of life using modern statistical methods to identify large-scale genome modifications in ancestral taxa.

3.2.2 Methods

Data retrieval and profile generation

The phylogenetic tree from Hug et al. (2016) was obtained directly from the Nature Microbiology website, and Pfam annotations were obtained from the Pfam database (v31.0; Finn et al., 2016). Presence/absence profiles were constructed for each Pfam annotation and genus combination, where '1' indicates the presence of the Pfam in at least one genome within the genus and '0' otherwise. If a genus in the phylogenetic tree did not have a genomic representative in the Pfam database, it was removed from the tree using the `drop.tip` function of the APE R package (Paradis et al., 2004). The resulting pruned tree was used for all further analyses. The final tree includes species representatives from 812 bacterial, 123 eukaryotic, and 83 archaeal genera.

Evolutionary model estimation and stochastic mapping

The rate of gain and loss was estimated for each Pfam phylogenetic profile on the pruned tree using the gainLoss v1.266 command line program (Cohen et al., 2010). Gain and loss rates were estimated by maximum likelihood assuming a variable gain/loss ratio with rates drawn from independent continuous gamma distributions approximated with four discrete rate categories. Pfam gain and loss events were mapped onto the phylogenetic tree through a continuous time Markov process using the estimated gain/loss rates and phylogenetic profiles with gainLoss v1.266 (Cohen et al., 2010). Gain and loss events along a branch were counted if the posterior probability of the state change along the branch was greater than 0.7.

3.2.3 Results and Discussion

The phylogenetic distribution of Pfam gains and losses

The ancestral states of extant genera were reconstructed to determine the lineages experiencing the highest level of gain and loss of genetically-encoded functions. By mapping the gain and loss events for all 15,282 Pfams present in at least one genus back onto the tree of life (**Figures 3.7, 3.8**), inferences can be made about ancient evolutionary events and the general trends in functional evolution.

Most Pfam gain and loss events occur at the tips of the tree, at the division between genera.

Since this is a genus-resolved phylogenetic tree, state changes that occur at tips in this phylogeny might be happening at the species or strain level. Thus, these branches encompass a more substantial evolutionary time than other branches of the same length. The observation that functional gains appear to be occurring more often in recent evolutionary time contradicts results from the inspection of SCOP annotations mapped to a less resolved phylogeny than the one used here that also included members from all three domains of life (Yang and Bourne, 2009). The authors suggested that the HGT events resulting from the endosymbiosis of mitochondria and chloroplasts pulled the gain events of many protein domains closer to the root of the tree. The opposite result may be seen here because of the imprecise nature of Pfam annotations relative to SCOP annotations (Xu et al., 2012) and the large proportion of missing proteomic data missing from many genera in the data set (see below). Missing data would produce an artificially high number of loss events and push gain events further from the root of the phylogeny.

Of the genera represented here, the ones listed in **Table 3.4** underwent the most gain events at their terminal branch. The bacterial genera *Clostridium*, *Streptomyces*, and *Bacillus* are at the top of the list. Sampling bias may be playing a role since these genera are all very large and have many independently sequenced strains. However, these genera are also under high environmental stress relative to their neighbouring taxa, and readily obtain new functions through HGT or DNA uptake to overcome them (He et al., 2010; Doroghazi and Buckley, 2010; Brito et al., 2018). The rest of the most functionally innovative taxa relative to their neighbours at the genus level are

Table 3.4. Genera exhibiting the most Pfam gains

NCBI Taxonomy ID	Genus Lineage String	# Gain Events
1485	d__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae; g__Clostridium	1269
1883	d__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Streptomycetales; f__Streptomycetaceae; g__Streptomyces	1240
1386	d__Bacteria; p__Firmicutes; c__Bacilli; o__Bacillales; f__Bacillaceae; g__Bacillus	1225
44249	d__Bacteria; p__Firmicutes; c__Bacilli; o__Bacillales; f__Paenibacillaceae; g__Paenibacillus	702
1578	d__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Lactobacillaceae; g__Lactobacillus	693
157	d__Bacteria; p__Spirochaetes; c__Spirochaetia; o__Spirochaetales; f__Spirochaetaceae; g__Treponema	644
33882	d__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Micrococcales; f__Microbacteriaceae; g__Microbacterium	567
55528	d__Eukaryota; p__Cryptophyta; c__Cryptophyceae; o__Pyrenomonadales; f__Geminigeraceae; g__Guillardia	558
1678	d__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium	538
3987	d__Eukaryota; k__Viridiplantae; p__Streptophyta; u__Magnoliophyta; o__Malpighiales; f__Euphorbiaceae; t__Acalypheae; g__Ricinus	523
2093	d__Bacteria; p__Tenericutes; c__Mollicutes; o__Mycoplasmatales; f__Mycoplasmataceae; g__Mycoplasma	482
374	d__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rhizobiales; f__Bradyrhizobiaceae; g__Bradyrhizobium	449
286	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Pseudomonadaceae; g__Pseudomonas	439
68287	d__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rhizobiales; f__Phyllobacteriaceae; g__Mesorhizobium	424
1298	d__Bacteria; p__Deinococcus-Thermus; c__Deinococci; o__Deinococcales; f__Deinococcaceae; g__Deinococcus	424
170636	d__Bacteria; p__Gemmatimonadetes; c__Gemmatimonadetes; o__Gemmatimonadales; f__Gemmatimonadaceae; g__Gemmatirosa	423
1562	d__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Peptococcaceae; g__Desulfotomaculum	421
2745	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Oceanospirillales; f__Halomonadaceae; g__Halomonas	409
22	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Alteromonadales; f__Shewanellaceae; g__Shewanella	393
6237	d__Eukaryota; u__Opisthokonta; k__Metazoa; p__Nematoda; c__Chromadorea; o__Rhabditida; f__Rhabditidae; g__Caenorhabditis	388

Lineage strings are derived from the NCBI taxonomy accessed November 29, 2018. c=class, d=domain, f=family, g=genus, k=kingdom, o=order, p=phylum, t=tribe, u=unranked.

predominantly bacteria in the phyla Proteobacteria and Firmicutes.

The internal branches with the highest number of Pfam gains are those separating the three domains: the branches splitting Bacteria from Eukaryota and Archaea and Eukaryota from Archaea (**Table 3.5**). The protein domains that were gained in the bacterial lineage are predominantly DUFs and domains associated with the outer membrane. The set of protein domains that associate with the evolutionary separation of eukaryotes from archaea includes many cytoskeletal elements and domains involved with nucleus formation and intracellular trafficking. Furthermore, the number of protein domains required for the production and maintenance of bones places the first Euteleost ancestor high on the list of functionally innovative taxa. The first member of the bacterial phylum Cyanobacteria also contains a large number of new Pfams, many of which contribute to their unique ability, as bacteria, to produce energy through photosynthesis.

On the other hand, genera exhibiting the highest number of Pfam losses (**Table 3.6**) are mostly

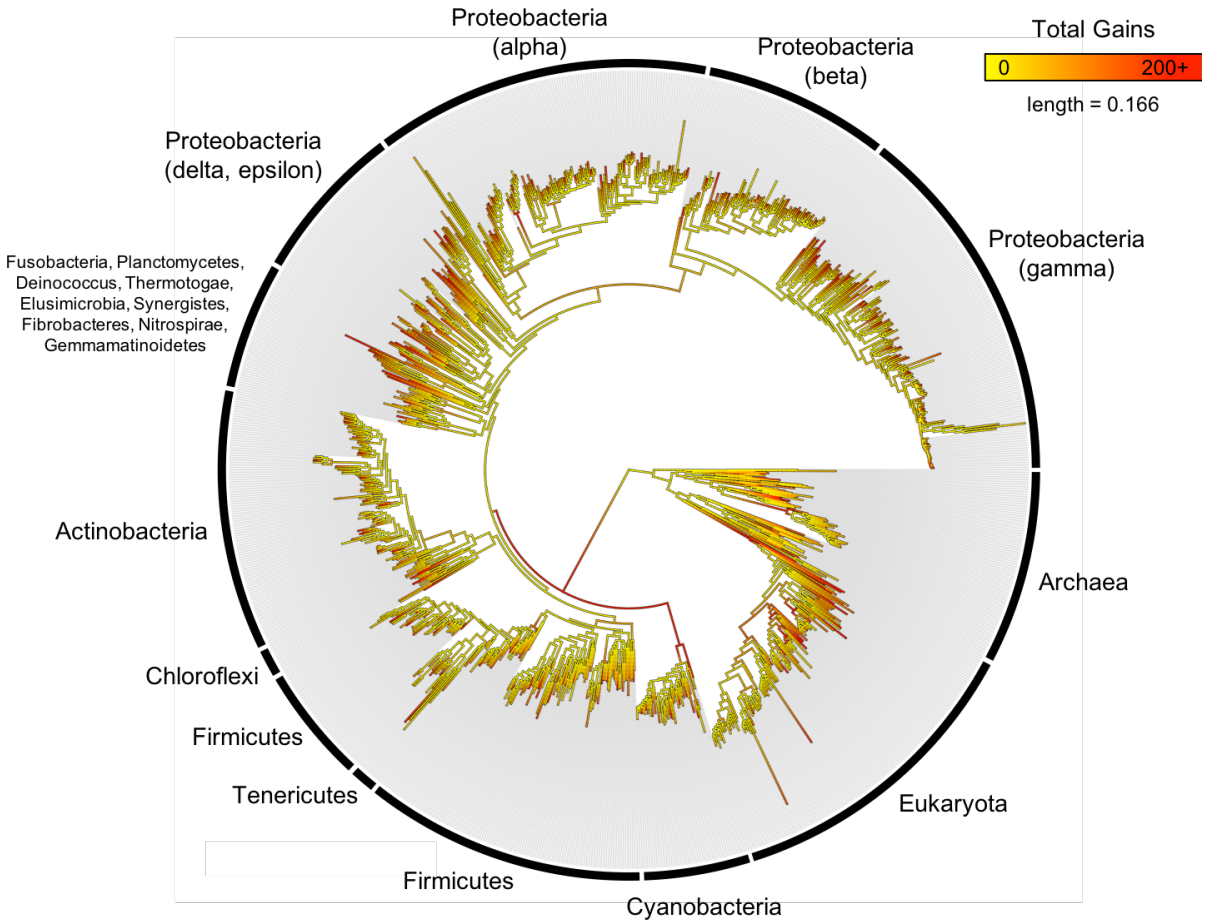

Figure 3.7. Density of Pfam gain events on the tree of life. The gainLoss program (Cohen et al., 2010) was used to infer the location of Pfam gain events based on a binary presence/absence profile for the Pfam and the tree of life. The few branches with a sum of gain events exceeding 200 were truncated down to 200 to increase the resolution of a greater proportion of branches with less than 200 total gain events.

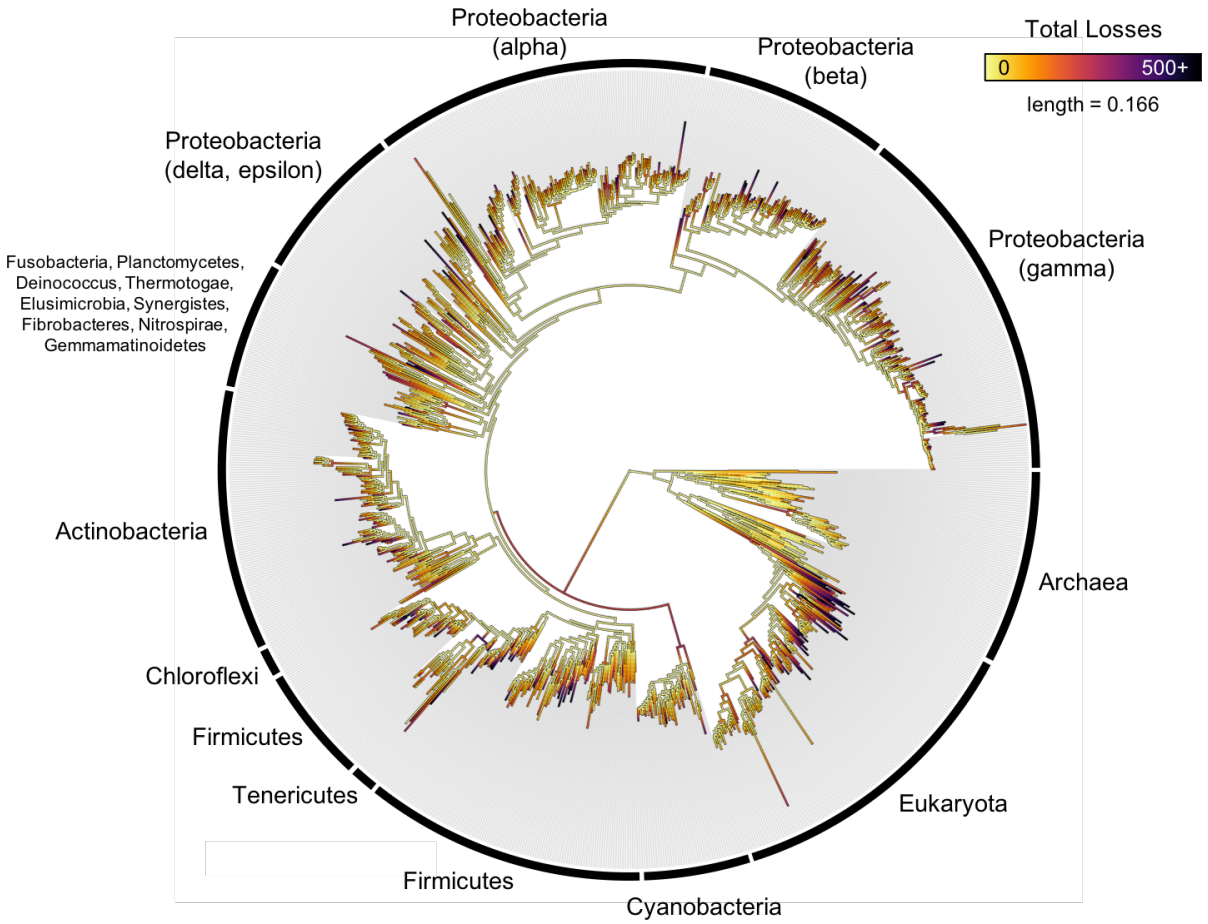

Figure 3.8. Density of Pfam loss events on the tree of life. The gainLoss program (Cohen et al., 2010) was used to infer the location of Pfam loss events based on a binary presence/absence profile for the Pfam and the tree of life. The few branches with a sum of loss events exceeding 500 were truncated down to 500 to increase the resolution of a greater proportion of branches with less than 500 total loss events.

Table 3.5. Internal branches exhibiting the most Pfam gains

NCBI Taxonomy ID of the LCA	Lineage String of the LCA	# Gain Events
2	d__Bacteria	1036
2759	d__Eukaryota	718
117571	d__Eukaryota; k__Metazoa; p__Chordata; u__Craniata; u__Vertebrata; u__Euteleostomi	513
40674	d__Eukaryota; k__Metazoa; p__Chordata; u__Craniata; u__Vertebrata; u__Euteleostomi; c__Mammalia	341
204457	d__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Sphingomonadales	339
33208	d__Eukaryota; k__Metazoa	283
1117	d__Bacteria; p__Cyanobacteria	278
58023	d__Eukaryota; k__Viridiplantae; p__Streptophyta; u__Embryophyta; u__Trachiophyta	260
2236	d__Archaea; p__Euryarchaeota; c__Halobacteria; o__Halobacteriales; f__Halobacteriaceae	217
543	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae	212
91061	d__Bacteria; p__Firmicutes; c__Bacilli	144
2259	d__Archaea; p__Euryarchaeota; c__Thermococci; o__Thermococcales; f__Thermococcaceae	141
4751	d__Eukaryota; u__Opisthokonta; k__Fungi	135
33154	d__Eukaryota; u__Opisthokonta	131
118883	d__Archaea; p__Crenarchaeota; c__Thermoprotei; o__Sulfolobales; f__Sulfolobaceae	127
2062	d__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Streptomycetales; f__Streptomycetaceae	126
80811	d__Bacteria; p__Proteobacteria; c__Deltaproteobacteria; o__Myxococcales; subo__Cystobacterineae	125
7711	d__Eukaryota; k__Metazoa; p__Chordata	124
543	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae	115
194924	d__Bacteria; p__Proteobacteria; c__Deltaproteobacteria; o__Desulfovibrionales; f__Desulfovibrionaceae	115

Lineage strings are derived from the NCBI taxonomy accessed November 29, 2018. c=class, d=domain, f=family, g=genus, k=kingdom, LCA=lowest common ancestor, o=order, p=phylum, u=unranked.

genera that were not well represented in the Pfam database, thus will not be discussed here in the context of function loss. Fortunately, internal branches relying on the input of multiple genera are influenced less by this lack of data. Many of the inner branches exhibiting the highest number of Pfam loss events reflect known ancient large-scale genomic modification events (**Table 3.7**). For example, the genus representatives of the eukaryotic family *Trypanosomatidae* are obligate parasites that have lost many genes necessary for free-living that are maintained within the genomes of their taxonomic neighbours within Kinetoplastida (Jackson, 2015).

3.2.4 Conclusion

Here, the evolutionary history of 15,282 Pfams was reconstructed directly onto the most recently-published three-domain tree of life (Hug et al., 2016) to identify ancestral taxa undergoing high

Table 3.6. Genera exhibiting the most Pfam losses

NCBI Taxonomy ID	Genus Lineage String	# Loss Events
456492	d__Bacteria; p__Firmicutes; c__Bacilli; o__Bacillales; f__Paenibacillaceae; g__Saccharibacillus	1811
1381133	d__Bacteria; p__Proteobacteria; c__Betaproteobacteria; g__Profftella	1799
5873	d__Eukaryota; p__Apicomplexa; c__Aconoidasida; o__Piroplasmida; f__Theileriidae; g__Theileria	1787
1076727	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Vibrionales; f__Vibrionaceae; g__Photodesmus	1665
104341	d__Eukaryota; u__Opisthokonta; k__Fungi; p__Basidiomycota; c__Agaricomycetes; o__Polyporales; f__Dacryobolaceae; g__Postia	1654
414715	d__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Catenulisporineae; f__Actinospicaceae; g__Actinospica	1632
224135	d__Bacteria; p__Proteobacteria; c__Betaproteobacteria; o__Burkholderiales; f__Burkholderiaceae; g__Glomeribacter	1455
1048757	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae; g__Moranella	1417
5758	d__Eukaryota; p__Amoebozoa; c__Archamoebae; f__Entamoebidae; g__Entamoeba	1403
12967	d__Eukaryota; u__Stramenopiles; c__Blastocystae; o__Blastocystida; f__Blastocystidae; g__Blastocystis	1326
691882	d__Eukaryota; u__Opisthokonta; c__Cristidiscoidea; o__Fonticulida; f__Fonticulaceae; g__Fonticula	1287
637	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae; g__Arsenophonus	1237
1763546	d__Bacteria; g__Vermiphilus	1234
93827	d__Eukaryota; u__Opisthokonta; k__Fungi; p__Basidiomycota; c__Agaricomycetes; o__Agaricales; f__Tricholomataceae; g__Gymnopus	1211
45156	d__Eukaryota; p__Rhodophyta; c__Bangiophyceae; o__Cyanidiales; f__Cyanidiaceae; g__Cyanidioschyzon	1148
541	d__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Sphingomonadales; f__Sphingomonadaceae; g__Zymomonas	1122
34763	d__Eukaryota; u__Opisthokonta; k__Metazoa; p__Chordata; c__Appendicularia; o__Copelata; f__Oikopleuridae; g__Oikopleura	1106
81525	d__Eukaryota; u__Opisthokonta; c__Choanoflagellatea; o__Choanoflagellida; f__Salpingoecidae; g__Monosiga	1103
160674	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae; g__Raoultella	1100
33055	d__Bacteria; p__Proteobacteria; c__Betaproteobacteria; g__Kinetoplastibacterium	1072

Lineage strings are derived from the NCBI taxonomy accessed November 29, 2018. c=class, d=domain, f=family, g=genus, k=kingdom, o=order, p=phylum, u=unranked.

Table 3.7. Internal branches exhibiting the most Pfam losses

NCBI Taxonomy ID of the LCA	Lineage String of the LCA	# Loss Events
5654	d__Eukaryota; p__Euglenozoa; o__Kinetoplastida; f__Trypanosomatidae	1140
2836	d__Eukaryota; u__Heterokonta; p__Bacillariophyta	618
33630	d__Eukaryota; u__Alveolata	521
868	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Cardiobacteriales; f__Cardiobacteriaceae	518
186828	d__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Carnobacteriaceae	488
544448/1239	d__Bacteria; p__Tenericutes/Firmicutes	409
33634/33154	d__Eukaryota; u__Heterokonta/u__Opisthokonta	409
712	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pasteurellales; f__Pasteurellaceae	357
85030	d__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Geodermatophilales; f__Geodermatophilaceae	353
543	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae	346
31979	d__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiacea	310
1706372	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Cellvibrionales; f__Halieaceae	306
91347/135625	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales/Pasteurellales	303
995019	d__Bacteria; p__Proteobacteria; c__Betaproteobacteria; o__Burkholderiales; f__Sutterellaceae	295
468	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Moraxellaceae	293
4751	d__Eukaryota; u__Opisthokonta; k__Fungi	290
32011	d__Bacteria; p__Proteobacteria; c__Betaproteobacteria; o__Nitrosomonadales; f__Methylophilaceae	283
213468	d__Bacteria; p__Proteobacteria; c__Deltaproteobacteria; o__Syntrophobacterales; f__Syntrophaceae	279
301297	d__Bacteria; p__Chloroflexi; c__Dehalococcoidia	277
1117	d__Bacteria; p__Cyanobacteria	271

Lineage strings are derived from the NCBI taxonomy accessed November 29, 2018. c=class, d=domain, f=family, g=genus, k=kingdom, LCA=lowest common ancestor, o=order, p=phylum, u=unranked.

rates of functional gain and loss. Unfortunately, there was a lack of annotation data for many of the genera in the phylogenetic tree that required the removal of many of the most novel genera from the phylogeny. The missing data within genera that were kept in the analysis inflated the number of Pfam gains and losses at the terminal branches of the tree and led to many false positive hits when screening for putative instances of HGT and taxa undergoing genomic streamlining. Despite the missing data in many genera, functional trends in gained and lost functions at the domain level were representative of functions that are known to be unique to those lineages.

Chapter 4

Conclusion and Future Directions

4.1 Contributions

The impact phylogenomics methods have had on our understanding of bacterial evolution to date has been enormous (Boucher et al., 2003; Szöllsi et al., 2015; Jiao et al., 2011; Spang et al., 2015). The work presented in this dissertation was performed to facilitate the exploration of functional traits within the context of bacterial evolution. This goal was accomplished through the development of the AnnoTree webtool, which serves as a much-needed access point and exploration tool for functional annotation data associated with genomes in the GTDB. AnnoTree's underlying functional annotations and phylogeny represent a wealth of data more substantial than any that has been explored in a phylogenomic context. I performed high-level analyses of

these data to determine the high-level functional trends between the most scattered and the most conserved traits. The webtool and complementing analyses promote hypothesis-generation in the context of protein evolution and have the potential to lead researchers to their next major discovery.

4.1.1 AnnoTree web application

Prior to the development of the AnnoTree web application, all other popular bacterial taxonomy exploration tools employed the NCBI taxonomy, whose hierarchy is inconsistent with many phylogenetic reconstructions of the bacterial domain (Bromberg et al., 2016; Hug et al., 2016). AnnoTree is the first phylogenetic exploration tool to integrate the new GTDB taxonomy (Parks et al., 2018) as one of its options for navigation, allowing for the simultaneous exploration of bacterial evolution, as determined by the phylogenetic tree, and the taxonomic ranks that have been assigned to the bacterial genomes. More functionality was added to the web application through the incorporation of search features that allow for the visualization of phylogenetic distribution patterns of Pfam protein domain families and KEGG Ortholog functional orthologs within all genomes in the GTDB's non-redundant phylogenetic tree. Downloadable visualizations, data summaries, and data tables were made available to facilitate follow-up analyses in other tools, which may be better suited to answer many of the different kinds of biological hypotheses that can stem from the user's observations within AnnoTree. In this work, it was demonstrated how

AnnoTree could reproduce the findings of three different research articles in just seconds. These examples show that AnnoTree has the potential to serve well for researchers that are curious about the evolutionary dynamics of bacteria and their functions.

4.1.2 Phylogenetic distribution of functional traits

Phylogenomic studies have been beneficial in providing supporting evidence for functions involved in the mechanisms of bacterial evolution (Chai et al., 2014). The incorporation of MAGs in the newest bacterial tree of life (Parks et al., 2017) and the increase in entries in functional annotation databases warrant a revisit to these analyses. The data generated for integration into the AnnoTree web application is the largest of its kind to be analyzed in a phylogenomic context. These data include the GTDB tree (Release 02-RS83; Parks et al., 2018), which contains 23,936 fully annotated bacterial genomes from 109 different phyla.

The evolutionary dynamics of 28,311 Pfams and KOs were analyzed through quantification of their phylogenetic distribution across bacterial genomes in the GTDB tree. The most and least homoplasic traits were determined through the ranking of the traits in terms of a normalized CI (ln(CI)/ln(family size)), which was shown to be resistant to contaminating sequences at levels previously observed in MAGs. The high-level functional trends seen within the ranked traits are representative of those which have been identified in previous studies (Chai et al., 2014; Cohen et al., 2010). The set of taxa that are enriched with the most homoplasic terms, often related to

bacteriophages and MGEs, were determined. These taxa tend to be associated with environments with highly dynamic selective pressures, such as pathogens, or highly conserved environments for which genomic reduction offers a competitive advantage, such as early endosymbiosis (McCutcheon and Moran, 2012).

Two new terms, saturation and catchment, were defined to systematically classify lineage-specific traits based on their evolutionary preservation and their exclusivity within a lineage, respectively. High cutoff thresholds of these terms led to the identification of previously-characterized clade-defining innovations within bacteria as well as putative instances of cross-domain HGT. Further inspection of some of the putative instances of cross-domain HGT produced hypotheses regarding the interactions of particular bacterial species with their eukaryotic hosts.

The results of these analyses confirm and supplement previous efforts that characterized high-level functional trends in bacterial evolution (Chai et al., 2014; Cohen et al., 2010; Martiny et al., 2013; Yang and Bourne, 2009; Barberán et al., 2017; Press et al., 2016). The normalized homoplasy metric, novel lineage-specificity terms, list of traits ranked by their phylogenetic distribution, and list of the most lineage-specific traits, in particular, are valuable contributions to the fields of functional and bacterial evolution.

4.2 Future Directions

4.2.1 AnnoTree web application

There are many ways in which AnnoTree can be improved and expanded. For instance, the modular functionality of the front-end visualization can be applied to other phylogenetic trees, taxonomies, and annotations. The GTDB has a novel archaeal taxonomy that is as standardized and complete as their bacterial taxonomy that can be visualized using the AnnoTree framework. It will be beneficial to include a eukaryotic version of the tool, too. The functional annotations that would offer the greatest value to AnnoTree are those of ncRNAs, which are not described within the Pfam or KEGG Orthology databases. The Rfam database is a good resource for browsing defined ncRNA families and for obtaining the tools that can be used to identify them in genomic sequences (Griffiths-Jones et al., 2003).

Improvements may also be made to AnnoTree in terms of the analysis features that it provides. I believe that users will benefit greatly from being able to perform a homology search by BLAST or protein domain search through hmmscan directly against the amino acid sequences in the AnnoTree database. It would also be beneficial to allow users to add their own genomes to the visualization for the purpose of taxonomic classification. The GTDB team is working towards a tool that can offer this functionality (https://github.com/Ecogenomics/GtdbTk), which can be

integrated into AnnoTree's framework in a future release.

Together, these future additions to AnnoTree will contribute towards the high-resolution mapping of the evolution of genomic traits in microbes and the continued exploration of the tree of life.

4.2.2 Phylogenetic distribution of functional traits

The results of the high-level biological analyses that were described in Chapter 3 largely confirmed those of previous analyses. The same analyses can be applied to individual clades of the GTDB tree to identify essential functions and drivers of evolution within those lineages. The recently-sequenced genomes from the Patescibacteria (CPR) are enticing candidates of such analyses due to their recent addition to the tree of life.

The large dataset of Pfam and KO annotations generated as part of this work can be applied in other large-scale analyses. An application that comes to mind is that of a co-occurrence analysis, which would combine the phylogenetic tree with the presence/absence profiles to group traits into groups of functional modules. This analysis would be particularly useful for the characterization of DUFs, since their function can be inferred by the characterized functions of other Pfams that they are associated with. The combination of this annotation data with geographical and habitat metadata from the Joint Genome Institute's Integrated Microbial Genomes and Microbiomes database (Chen et al., 2017) can offer insights into selective pressures in those environments.

Lastly, the information gained from the ancestral reconstructions of Pfams across all of life a the domain and phylum levels in Chapter 3.2 were quite insightful. By repeating the analysis with Pfam annotations from all genome representatives of the tree of life (Hug et al., 2016), the biological functions gained and lost at lower taxonomic levels can be properly detected.

References

Abouheif, E. (1999). A method for testing the assumption of phylogenetic independence in comparative data. Evolutionary Ecology Research *1*, 895–909.

Adebali, O. and Zhulin, I. B. (2017). Aquerium: a web application for comparative exploration of domain-based protein occurrences on the taxonomically clustered genome tree. Proteins *85*, 72–77.

Andam, C. P. and Gogarten, J. P. (2011). Biased gene transfer in microbial evolution. Nature Reviews Microbiology *9*, 543–555.

Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C. and Segata, N. (2015). Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ *3*, e1029.

Ayala-Sanmartin, J. and Gómez-Eichelmann, M. C. (1989). Stability of ColE1-like and pBR322-like plasmids in Escherichia coli. Molecular Microbiology *3*, 1745–1752.

Barberán, A., Caceres Velazquez, H., Jones, S. and Fierer, N. (2017). Hiding in plain sight: mining bacterial species records for phenotypic trait information. mSphere *2*, e00237–17.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Statistical Methodology) *57*, 289–300.

Blomberg, S. P., Garland, T. and Ives, A. R. (2003). Testing for phylogenetic signal in comparative data: behavioural traits are more labile. Evolution *57*, 717–745.

Bobay, L.-M., Touchon, M. and Rocha, E. P. C. (2014). Pervasive domestication of defective prophages by bacteria. Proceedings of the National Academy of Sciences *111*, 12127–12132.

Bondy-Denomy, J. and Davidson, A. R. (2014). When a virus is not a parasite: the beneficial effects of prophages on bacterial fitness.

Boucher, Y., Douady, C. J., Papke, R. T., Walsh, D. A., Boudreau, M. E. R., Nesbø, C. L., Case, R. J. and Doolittle, W. F. (2003). Lateral gene transfer and the origins of prokaryotic groups. Annual Review of Genetics *37*, 283–328.

Bougeard, S. and Dray, S. (2018). Supervised Multiblock Analysis in R with the ade4 Package. Journal of Statistical Software *86*, 1–17.

Brito, P. H., Chevreux, B., Serra, C. R., Schyns, G., Henriques, A. O. and Pereira-Leal, J. B.

(2018). Genetic competence drives genome diversity in *Bacillus subtilis*. Genome Biology and Evolution *10*, 108–124.

Bromberg, R., Grishin, N. V. and Otwinowski, Z. (2016). Phylogeny reconstruction with alignment-free method that corrects for horizontal gene transfer. PLoS Computational Biology *12*, e1004985.

Buchfink, B., Xie, C. and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. Nature Methods *12*, 59–60.

Burstein, D., Sun, C. L., Brown, C. T., Sharon, I., Anantharaman, K., Probst, A. J., Thomas, B. C. and Banfield, J. F. (2016). Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. Nature Communications *7*, 10613.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T. L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

Casamatta, D., Stanić, D., Gantar, M. and Richardson, L. L. (2012). Characterization of *Roseofilum reptotaenium* (Oscillatoriales, Cyanobacteria) gen. et sp. nov. isolated from Caribbean black band disease. Phycologia *51*, 489–499.

Caspi, R., Billington, R., Fulcher, C. A., Keseler, I. M., Kothari, A., Krummenacker, M., Latendresse, M., Midford, P. E., Ong, Q., Ong, W. K., Paley, S., Subhraveti, P. and Karp, P. D.

(2018). The MetaCyc database of metabolic pathways and enzymes. Nucleic Acids Research *46*, D633–D639.

Chai, J., Kora, G., Ahn, T.-H., Hyatt, D. and Pan, C. (2014). BMC Evolutionary Biology *14*, 207.

Chandonia, J.-M., Fox, N. K. and Brenner, S. E. (2018). SCOPe: classification of large macromolecular structures in the structural classification of proteins-extended database. Nucleic Acids Research *gky1134*, doi:10.1093/nar/gky1134.

Chang, Y.-C., Hu, Z., Rachlin, J., Anton, B. P., Kasif, S., Roberts, R. J. and Steffen, M. (2016). COMBREX-DB: an experiment centered database of protein function: knowledge, predictions and knowledge gaps. Nucleic Acids Research *44*, D330–D335.

Chen, I. and Dubnau, D. (2004). DNA uptake during bacterial transformation. Nature Reviews Microbiology *2*, 241–249.

Chen, I.-M. A., Markowitz, V. M., Chu, K., Palaniappan, K., Szeto, E., Pillay, M., Ratner, A., Huang, J., Andersen, E., Huntemann, M., Varghese, N., Hadjithomas, M., Tennessen, K., Nielsen, T., Ivanova, N. N. and Kyrpides, N. C. (2017). IMG/M: integrated genome and metagenome comparative data analysis system. Nucleic Acids Research *45*, D507–D516.

Cheng, H., Schaeffer, R. D., Liao, Y., Kinch, L. N., Pei, J., Shi, S., Kim, B.-H. and Grishin, N. V.

(2014). ECOD: an evolutionary classification of protein domains. PLoS Computational Biology *10*, e1003926.

Chu, H. Y., Sprouffske, K. and Wagner, A. (2018). Assessing the benefits of horizontal gene transfer by laboratory evolution and genome sequencing. BMC Evolutionary Biology *18*, 54.

Claverys, J.-P., Martin, B. and Polard, P. (2009). The genetic transformation machinery: composition, localization, and mechanism. FEMS Microbiology Reviews *33*, 643–656.

Cohen, O., Ashkenazy, H., Belinky, F., Huchon, D. and Pupko, T. (2010). GLOOME: gain loss mapping engine. Bioinformatics *26*, 2914–2915.

Coleman, M. L. and Chisholm, S. W. (2010). Ecosystem-specific selection pressures revealed through comparative population genomics. Proceedings of the National Academy of Sciences *107*, 18634–18639.

Comte, L., Murienne, J. and Grenouillet, G. (2014). Species traits and phylogenetic conservatism of climate-induced range shifts in stream fishes. Nature Communications *5*, 5023.

Daims, H., Lebedeva, E. V., Pjevac, P., Han, P., Herbold, C., Albertsen, M., Jehmlich, N., Palatinszky, M., Vierheilig, J., Bulaev, A., Kirkegaard, R. H., von Bergen, M., Rattei, T., Bendinger, B., Nielsen, P. H. and Wagner, M. (2015). Complete nitrification by *Nitrospira* bacteria. Nature *528*, 504–509.

Dawson, N. L., Lewis, T. E., Das, S., Lees, J. G., Lee, D., Ashford, P., Orengo, C. A. and Sillitoe, I. (2017). CATH: an expanded resource to predict protein function through structure and sequence. Nucleic Acids Research *45*, D289–D295.

Del Sol, R., Pitman, A., Herron, P. and Dyson, P. (2003). The product of a developmental gene, *crgA*, that coordinates reproductive growth in *Streptomyces* belongs to a novel family of small actinomycete-specific proteins. Journal of Bacteriology *185*, 6678–6685.

Demuth, J. P. and Hahn, M. W. (2009). The life and death of gene families. BioEssays *31*, 29–39.

Dixon, P. (2003). VEGAN, a package of R functions for community ecology. Journal of Vegetation Science *14*, 927–930.

Doroghazi, J. R. and Buckley, D. H. (2010). Widespread homologous recombination within and between *Streptomyces* species. The ISME Journal *4*, 1136–1143.

D'Souza, G., Waschina, S., Pande, S., Bohl, K., Kaleta, C. and Kost, C. (2014). Less is more: selective advantages can explain the prevalent loss of biosynthetic genes in bacteria. Evolution *68*, 2559–2570.

Eddy, S. R. (2011). Accelerated profile HMM searches. PLoS Computational Biology *7*, e1002195.

Eisen, J. A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Research *8*, 163–167.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E. and Finn, R. D. (2018). The Pfam protein families database in 2019. Nucleic Acids Research *gky995*, doi:10.1093/nar/gky995.

Farris, J. S. (1973). On comparing the shapes of taxonomic trees. Systematic Zoology *22*, 50–54.

Farris, J. S. (1989). The retention index and homoplasy excess. Systematic Zoology *38*, 406–407.

Federhen, S. (2012). The NCBI Taxonomy database. Nucleic Acids Research *40*, D136–D143.

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., Chang, H.-Y., Dosztányi, Z., El-Gebali, S., Fraser, M., Gough, J., Haft, D., Holliday, G. L., Huang, H., Huang, X., Letunic, I., Lopez, R., Lu, S., Marchler-Bauer, A., Mi, H., Mistry, J., Natale, D. A., Necci, M., Nuka, G., Orengo, C. A., Park, Y., Pesseat, S., Piovesan, D., Potter, S. C., Rawlings, N. D., Redaschi, N., Richardson, L., Rivoire, C., Sangrador-Vegas, A., Sigrist, C., Sillitoe, I., Smithers, B., Squizzato, S., Sutton, G., Thanki, N., Thomas, P. D., Tosatto, S., Wu, C. H., Xenarios, I., Yeh, L.-S., Young, S.-Y. and Mitchell, A. L. (2017). InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Research *45*, D190–D199.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J. and Punta, M. (2014). Pfam: the protein families database. Nucleic Acids Research *42*, D222–D230.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J. and Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research *44*, D279–D285.

Fritz, S. A. and Purvis, A. (2010). Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. Conservation Biology *24*, 1042–1051.

Frost, L. S., Leplae, R., Summers, A. O. and Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. Nature Reviews Microbiology *3*, 722–732.

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

Galperin, M. Y., Makarova, K. S., Wolf, Y. I. and Koonin, E. V. (2015). Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Research *43*, D261–D269.

García-Heras, F., Abellón-Ruiz, J., Murillo, F. J., Padmanabhan, S. and Elías-Arnanz, M. (2013).

High-mobility-group A-like CarD binds to a DNA site optimized for affinity and position and to RNA polymerase to regulate a light-inducible promoter in *Myxococcus xanthus*. Journal of Bacteriology *195*, 378–388.

Gillings, M. R. (2014). Integrons: past, present, and future. Microbiology and Molecular Biology Reviews *78*, 257–277.

Giovannoni, S. J., Tripp, H. J., Givan, S., Podar, M., Vergin, K. L., Baptista, D., Bibbs, L., Eads, J., Richardson, T. H., Noordewier, M., Rappé, M. S., Short, J. M., Carrington, J. C. and Mathur, E. J. (2005). Genome streamlining in a cosmopolitan oceanic bacterium. Science *309*, 1242–1245.

Gittleman, J. L. and Kot, M. (1990). Adaptation: statistics and a null model for estimating phylogenetic effects. Systematic Zoology *39*, 227–241.

Goldstein, A. L. and Badamchian, M. (2004). Thymosins: chemistry and biological properties in health and disease. Expert Opinion on Biological Therapy *4*, 559–573.

Goodacre, N. F., Gerloff, D. L. and Uetz, P. (2014). Protein domains of unknown function are essential in bacteria. mBio *5*, e00744.

Green, M. L. and Karp, P. D. (2006). The outcomes of pathway database computations depend on pathway ontology. Nucleic Acids Research *34*, 3687–3697.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. and Eddy, S. R. (2003). Rfam: an RNA family database. Nucleic Acids Research *31*, 439–441.

Haft, D. H., Selengut, J. D., Richter, R. A., Harkins, D., Basu, M. K. and Beck, E. (2012). TIGRFAMs and Genome Properties in 2013. Nucleic Acids Research *41*, D387–D395.

Hansen, T. F. and Martins, E. P. (1996). Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. Evolution *50*, 1404–1417.

Harmon, L. J., Weir, J. T., Brock, C. D., Glor, R. E. and Challenger, W. (2008). GEIGER: investigating evolutionary radiations. Bioinformatics *24*, 129–131.

He, M., Sebaihia, M., Lawley, T. D., Stabler, R. A., Dawson, L. F., Martin, M. J., Holt, K. E., Seth-Smith, H. M. B., Quail, M. A., Rance, R., Brooks, K., Churcher, C., Harris, D., Bentley, S. D., Burrows, C., Clark, L., Corton, C., Murray, V., Rose, G., Thurston, S., van Tonder, A., Walker, D., Wren, B. W., Dougan, G. and Parkhill, J. (2010). Evolutionary dynamics of *Clostridium difficile* over short and long time scales. Proceedings of the National Academy of Sciences *107*, 7527–7532.

He, Z., Zhang, H., Gao, S., Lercher, M. J., Chen, W.-H. and Hu, S. (2016). Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. Nucleic Acids Research *44*, W236–W241.

Huerta-Cepas, J., Serra, F. and Bork, P. (2016). ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Molecular Biology and Evolution *33*, 1635–1638.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C. and Banfield, J. F. (2016). A new view of the tree of life. Nature Microbiology *1*, 16048.

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W. and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119.

Jackson, A. P. (2015). Genome evolution in trypanosomatid parasites. Parasitology *142*, S40–S56.

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., Tomsho, L. P., Hu, Y., Liang, H., Soltis, P. S., Soltis, D. E., Clifton, S. W., Schlarbaum, S. E., Schuster, S. C., Ma, H., Leebens-Mack, J. and DePamphilis, C. W. (2011). Ancestral polyploidy in seed plants and angiosperms. Nature *473*, 97–100.

Jombart, T., Balloux, F. and Dray, S. (2010). adephylo: new tools for investigating the phylogenetic signal in biological traits. Bioinformatics *26*, 1907–1909.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. and Morishima, K. (2017). KEGG: new

perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research *45*, D353–D361.

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. and Tanabe, M. (2018). New approach for understanding genome variations in KEGG. Nucleic Acids Research *gky962*, doi:10.1093/nar/gky962.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. Nucleic Acids Research *44*, D457–D462.

Karcagi, I., Draskovits, G., Umenhoffer, K., Fekete, G., Kovács, K., Méhi, O., Balikó, G., Szappanos, B., Györfy, Z., Fehér, T., Bogos, B., Blattner, F. R., Pál, C., Pósfai, G. and Papp, B. (2016). Indispensability of horizontally transferred genes and its impact on bacterial genome streamlining. Molecular Biology and Evolution *33*, 1257–1269.

Karp, P. D., Latendresse, M. and Caspi, R. (2011). The pathway tools pathway prediction algorithm. Standards in Genomic Sciences *5*, 424–429.

Keck, F., Rimet, F., Bouchez, A. and Franc, A. (2016). phylosignal: an R package to measure, test, and explore the phylogenetic signal. Ecology and Evolution *6*, 2774–2780.

Kendall, M. G. and Smith, B. B. (1939). The problem of *m* rankings. The Annals of Mathematical Statistics *10*, 275–287.

Kluge, A. G. and Farris, J. S. (1969). Quantitative phyletics and the evolution of anurans. Systematic Zoology *18*, 1–32.

Kreft, Ł., Botzki, A., Coppens, F., Vandepoele, K. and Van Bel, M. (2017). PhyD3: a phylogenetic tree viewer with extended phyloXML support for functional genomics data visualization. Bioinformatics *33*, 2946–2947.

Kroos, L. (2005). Eukaryotic-like signaling and gene regulation in a prokaryote that undergoes multicellular development. Proceedings of the National Academy of Sciences *102*, 2681–2682.

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

Lande, R. (1976). Natural selection and random genetic drift in phenotypic evolution. Evolution *30*, 314–334.

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

Letunic, I. and Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Research *44*, W242–W245.

Levin, B. R., Moineau, S., Bushman, M. and Barrangou, R. (2013). The population and evolutionary dynamics of phage and bacteria with CRISPR-mediated immunity. PLoS Genetics *9*, e1003312.

Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. Systematic Biology *50*, 913–925.

Lockwood, J. L., Russell, G. J., Gittleman, J. L., Daehler, C. C., McKinney, M. L. and Purvis, A. (2002). A metric for analyzing taxonomic patterns of extinction risk. Conservation Biology *16*, 1137–1142.

Lorenz, P. and Eck, J. (2005). Metagenomics and industrial applications. Nature Reviews Microbiology *3*, 510–516.

Louca, S., Doebeli, M. and Valencia, A. (2018). Efficient comparative phylogenetics on large trees. Bioinformatics *34*, 1053–1055.

Louca, S., Jacques, S. M. S., Pires, A. P. F., Leal, J. S., Srivastava, D. S., Parfrey, L. W., Farjalla, V. F. and Doebeli, M. (2016). High taxonomic variability despite stable functional structure across microbial communities. Nature Ecology & Evolution *1*, 0015.

Mansfield, M. J., Adams, J. B. and Doxey, A. C. (2015). Botulinum neurotoxin homologs in non-*Clostridium* species. FEBS Letters *589*, 342–348.

Mansfield, M. J., Sugiman-Marangos, S. N., Melnyk, R. A. and Doxey, A. C. (2018). Identification of a diphtheria toxin-like gene family beyond the *Corynebacterium* genus. FEBS Letters *592*, 2693–2705.

Mansfield, M. J., Wentz, T. G., Zhang, S., Lee, E. J., Dong, M., Sharma, S. K. and Doxey, A. C. (2017). Newly identified relatives of botulinum neurotoxins shed light on their molecular evolution. bioRxiv, *220806*.

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Wang, Z., Yamashita, R. A., Zhang, D., Zheng, C., Geer, L. Y. and Bryant, S. H. (2017). CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Research *45*, D200–D203.

Martiny, A. C., Treseder, K. and Pusch, G. (2013). Phylogenetic conservatism of functional traits in microorganisms. The ISME Journal *7*, 830–838.

Mazel, D. (2006). Integrons: agents of bacterial evolution. Nature Reviews Microbiology *4*, 608–620.

McCutcheon, J. P. and Moran, N. A. (2012). Extreme genome reduction in symbiotic bacteria. Nature Reviews Microbiology *10*, 13–26.

McDonald, B. R. and Currie, C. R. (2017). Lateral gene transfer dynamics in the ancient bacterial genus *Streptomyces*. mBio *8*, e00644.

Meier, R., Kores, P. and Darwin, S. (1991). Homoplasy slope ratio: a better measurement of observed homoplasy in cladistic analyses. Systematic Biology *40*, 74–88.

Mendler, K., Chen, H., Parks, D. H., Hug, L. A. and Doxey, A. C. (2018). AnnoTree: visualization and exploration of a functionally annotated microbial tree of life. bioRxiv, *463455*.

Mistry, J., Bateman, A. and Finn, R. D. (2007). Predicting active site residue annotations in the Pfam database. BMC Bioinformatics *8*, 298.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. Biometrika *37*, 17–23.

Münkemüller, T., Lavergne, S., Bzeznik, B., Dray, S., Jombart, T., Schiffers, K. and Thuiller, W. (2012). How to measure and test phylogenetic signal. Methods in Ecology and Evolution *3*, 743–756.

Nasir, A., Kim, K. and Caetano-Anolles, G. (2012). Giant viruses coexisted with the cellular ancestors and represent a distinct supergroup along with superkingdoms Archaea, Bacteria and Eukarya. BMC Evolutionary Biology *12*, 156.

NCBI Resource Coordinators (2016). Database resources of the National Center for Biotechnology Information. Nucleic Acids Research *44*, D7–D19.

Neave, M. J., Apprill, A., Ferrier-Pagès, C. and Voolstra, C. R. (2016). Diversity and function of prevalent symbiotic marine bacteria in the genus *Endozoicomonas*. Applied Microbiology and Biotechnology *100*, 8315–8324.

Ochman, H., Lawrence, J. G. and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. Nature *405*, 299–304.

Orme, C., Freckleton, R., Thomas, G., Petzoldt, T., Fritz, S. and Isaac, N. (2013). CAPER: comparative analyses of phylogenetics and evolution in R. Methods in Ecology and Evolution *3*, 145–151.

Pagel, M. (1994). Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proceedings of the Royal Society B: Biological Sciences *255*, 37–45.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. Nature *401*, 877–884.

Pande, S., Merker, H., Bohl, K., Reichelt, M., Schuster, S., de Figueiredo, L. F., Kaleta, C. and Kost, C. (2014). Fitness and stability of obligate cross-feeding interactions that emerge upon gene loss in bacteria. The ISME Journal *8*, 953–962.

Paradis, E., Claude, J. and Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. Bioinformatics *20*, 289–290.

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A. and Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature Biotechnology *36*, 996–1004.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research *25*, 1043–1055.

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B. J., Evans, P. N., Hugenholtz, P. and Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature Microbiology *2*, 1533–1542.

Patwardhan, A., Ray, S. and Roy, A. (2014). Molecular markers in phylogenetic studies - a review. Journal of Phylogenetics & Evolutionary Biology *2*, 131.

Pavoine, S., Ollier, S., Pontier, D. and Chessel, D. (2008). Testing for phylogenetic signal in phenotypic traits: new matrices of phylogenetic proximities. Theoretical Population Biology *73*, 79–91.

Popa, O., Landan, G. and Dagan, T. (2017). Phylogenomic networks reveal limited phylogenetic range of lateral gene transfer by transduction. The ISME Journal *11*, 543–554.

Press, M. O., Queitsch, C. and Borenstein, E. (2016). Evolutionary assembly patterns of prokaryotic genomes. Genome Research *26*, 826–833.

Pushkarev, A., Inoue, K., Larom, S., Flores-Uribe, J., Singh, M., Konno, M., Tomida, S., Ito, S., Nakamura, R., Tsunoda, S. P., Philosof, A., Sharon, I., Yutin, N., Koonin, E. V., Kandori, H. and Béjà, O. (2018). A distinct abundant group of microbial rhodopsins discovered using functional metagenomics. Nature *558*, 595–599.

R Core Team (2018). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria.

Ravenhall, M., Škunca, N., Lassalle, F. and Dessimoz, C. (2015). Inferring horizontal gene transfer. PLoS Computational Biology *11*, e1004095.

Revell, L. J., Harmon, L. J. and Collar, D. C. (2008). Phylogenetic signal, evolutionary process, and rate. Systematic Biology *57*, 591–601.

Rodríguez-Torres, D. M., Islas-Robles, Á., Gómez-Lunar, Z., Delaye, L., Hernández-González, I., Souza, V., Travisano, M. and Olmedo-Álvarez, G. (2017). Phenotypic microdiversity and phylogenetic signal analysis of traits related to social interaction in *Bacillus* spp. from sediment communities. Frontiers in Microbiology *8*, 29.

Rodriguez-Valera, F., Martin-Cuadrado, A.-B. and López-Pérez, M. (2016). Flexible genomic islands as drivers of genome evolution. Current Opinion in Microbiology *31*, 154–160.

San Millan, A., Peña-Miller, R., Toll-Riera, M., Halbert, Z. V., McLean, A. R., Cooper, B. S. and MacLean, R. C. (2014). Positive selection and compensatory adaptation interact to stabilize non-transmissible plasmids. Nature Communications *5*, 5208.

Schliep, K. P. (2011). phangorn: phylogenetic analysis in R. Bioinformatics *27*, 592–593.

Sonnhammer, E., Eddy, S. R., Birney, E., Bateman, A. and Durbin, R. (1998). Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Research *26*, 320–322.

Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., van Eijk, R., Schleper, C., Guy, L. and Ettema, T. J. G. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature *521*, 173–179.

Speed, M. P. and Arbuckle, K. (2017). Quantification provides a conceptual basis for convergent evolution. Biological Reviews *92*, 815–829.

Stapley, J., Reger, J., Feulner, P. G., Smadja, C., Galindo, J., Ekblom, R., Bennison, C., Ball, A. D., Beckerman, A. P. and Slate, J. (2010). Adaptation genomics: the next generation. Trends in Ecology & Evolution *25*, 705–712.

Sun, D. (2018). Pull in and push out: mechanisms of horizontal gene transfer in bacteria. Frontiers in Microbiology *9*, 2154.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B. and Wu, C. H. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics *31*, 926–932.

Szöllsi, G. J., Davín, A. A., Tannier, E., Daubin, V. and Boussau, B. (2015). Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. Philosophical Transactions of the Royal Society B: Biological Sciences *370*, 20140335.

Thomas, C. M. and Nielsen, K. M. (2005). Mechanisms of and barriers to, horizontal gene transfer between bacteria. Nature Reviews Microbiology *3*, 711–721.

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

Venter, J. C. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. Science *304*, 66–74.

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S. M. and Yu, H. (2012). Three-dimensional

reconstruction of protein networks provides insight into human genetic disease. Nature Biotechnology *30*, 159–164.

Wattam, A. R., Davis, J. J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., Conrad, N., Dietrich, E. M., Disz, T., Gabbard, J. L., Gerdes, S., Henry, C. S., Kenyon, R. W., Machi, D., Mao, C., Nordberg, E. K., Olsen, G. J., Murphy-Olson, D. E., Olson, R., Overbeek, R., Parrello, B., Pusch, G. D., Shukla, M., Vonstein, V., Warren, A., Xia, F., Yoo, H. and Stevens, R. L. (2017). Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic Acids Research *45*, D535–D542.

Wozniak, R. A. F. and Waldor, M. K. (2010). Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. Nature Reviews Microbiology *8*, 552–563.

Xu, Q., Dunbrack, R. L. and Jr (2012). Assignment of protein sequences to existing domain and family classification systems: Pfam and the PDB. Bioinformatics *28*, 2763–2772.

Yang, S. and Bourne, P. E. (2009). The evolutionary history of protein domains viewed by species phylogeny. PLoS ONE *4*, e8378.

APPENDICES

Appendix A

Measured phylogenetic dispersion of functional annotations

Table A1. Homoplasy ranking of KEGG categories. Phylogenetic patchiness was computed for each KO using the CI, a common homoplasy metric representing the inverse of the minimum possible number of state changes (trait gain or loss) given the tree topology. The final phylogenetic patchiness score is equal to ln(CI)/ln(family size) where family size is the total number of genomes containing the trait. Each KO was grouped into its KEGG pathway and KEGG BRITE category for comparison of higher-level functional trends.

Category ID	Category Name	Pathway Class	Mean ln(CI)/ln(family size)
ko00196	Photosynthesis - antenna proteins	Metabolism	-0.60928
ko00195	Photosynthesis	Metabolism	-0.63508
BR:ko00194	Photosynthesis proteins	Metabolism	-0.65534
ko04111	Cell cycle - yeast	Cellular Processes	-0.67286
ko04120	Ubiquitin mediated proteolysis	Genetic Information Processing	-0.67286
ko04114	Oocyte meiosis	Cellular Processes	-0.68302
ko03440	Homologous recombination	Genetic Information Processing	-0.68684
ko04015	Rap1 signaling pathway	Environmental Information Processing	-0.69318
ko00572	Arabinogalactan biosynthesis - Mycobacterium	Metabolism	-0.69448
ko03030	DNA replication	Genetic Information Processing	-0.69493
ko03050	Proteasome	Genetic Information Processing	-0.70029
ko03430	Mismatch repair	Genetic Information Processing	-0.70945
ko03020	RNA polymerase	Genetic Information Processing	-0.71647
ko00550	Peptidoglycan biosynthesis	Metabolism	-0.71965
ko03010	Ribosome	Genetic Information Processing	-0.71973
ko00571	Lipoarabinomannan (LAM) biosynthesis	Metabolism	-0.72210
ko04112	Cell cycle - Caulobacter	Cellular Processes	-0.73354
BR:ko01011	Peptidoglycan biosynthesis and degradation proteins	Metabolism	-0.73487
BR:ko03009	Ribosome biogenesis	Genetic Information Processing	-0.73737
BR:ko03032	DNA replication proteins	Genetic Information Processing	-0.73799
ko03008	Ribosome biogenesis in eukaryotes	Genetic Information Processing	-0.73988
ko00061	Fatty acid biosynthesis	Metabolism	-0.74265
ko00785	Lipoic acid metabolism	Metabolism	-0.74705
ko03060	Protein export	Genetic Information Processing	-0.74716
ko00970	Aminoacyl-tRNA biosynthesis	Genetic Information Processing	-0.74761
ko03022	Basal transcription factors	Genetic Information Processing	-0.74786
BR:ko04812	Cytoskeleton proteins	Cellular Processes	-0.74826
ko03018	RNA degradation	Genetic Information Processing	-0.75249
ko00780	Biotin metabolism	Metabolism	-0.75466
ko04070	Phosphatidylinositol signaling system	Environmental Information Processing	-0.75472
BR:ko03016	Transfer RNA biogenesis	Genetic Information Processing	-0.75511
BR:ko03021	Transcription machinery	Genetic Information Processing	-0.75630
ko02026	Biofilm formation - Escherichia coli	Cellular Processes	-0.75779
ko00290	Valine, leucine and isoleucine biosynthesis	Metabolism	-0.75936
BR:ko02042	Bacterial toxins	Environmental Information Processing	-0.76002
ko00710	Carbon fixation in photosynthetic organisms	Metabolism	-0.76225
BR:ko02035	Bacterial motility proteins	Cellular Processes	-0.76323
ko02040	Flagellar assembly	Cellular Processes	-0.76420
BR:ko01006	Prenyltransferases	Metabolism	-0.76425
ko04215	Apoptosis - multiple species	Cellular Processes	-0.76498
BR:ko03400	DNA repair and recombination proteins	Genetic Information Processing	-0.76569
BR:ko03019	Messenger RNA Biogenesis	Genetic Information Processing	-0.76736
ko00473	D-Alanine metabolism	Metabolism	-0.76998
ko00770	Pantothenate and CoA biosynthesis	Metabolism	-0.77049
ko03420	Nucleotide excision repair	Genetic Information Processing	-0.77177
BR:ko03036	Chromosome and associated proteins	Genetic Information Processing	-0.77201
BR:ko03012	Translation factors	Genetic Information Processing	-0.77363
ko00900	Terpenoid backbone biosynthesis	Metabolism	-0.77402
ko00240	Pyrimidine metabolism	Metabolism	-0.77556
ko04113	Meiosis - yeast	Cellular Processes	-0.77560
ko04122	Sulfur relay system	Genetic Information Processing	-0.77829
ko05111	Biofilm formation - Vibrio cholerae	Cellular Processes	-0.77974
ko04020	Calcium signaling pathway	Environmental Information Processing	-0.78026
ko00130	Ubiquinone and other terpenoid-quinone biosynthesis	Metabolism	-0.78037
ko00230	Purine metabolism	Metabolism	-0.78102
ko00190	Oxidative phosphorylation	Metabolism	-0.78155
ko00591	Linoleic acid metabolism	Metabolism	-0.78197
ko04151	PI3K-Akt signaling pathway	Environmental Information Processing	-0.78225
ko04110	Cell cycle	Cellular Processes	-0.78424
ko00400	Phenylalanine, tyrosine and tryptophan biosynthesis	Metabolism	-0.78483
BR:ko03029	Mitochondrial biogenesis	Genetic Information Processing	-0.78504
ko00540	Lipopolysaccharide biosynthesis	Metabolism	-0.78569
ko04510	Focal adhesion	Cellular Processes	-0.78638
ko04512	ECM-receptor interaction	Environmental Information Processing	-0.78638

Table A1 continued from previous page

Category ID	Category Name	Pathway Class	Mean ln(CI)/ln(family size)
ko00480	Glutathione metabolism	Metabolism	-0.78739
ko04066	HIF-1 signaling pathway	Environmental Information Processing	-0.78739
ko01040	Biosynthesis of unsaturated fatty acids	Metabolism	-0.78846
BR:ko03110	Chaperones and folding catalysts	Genetic Information Processing	-0.78876
ko03410	Base excision repair	Genetic Information Processing	-0.79091
ko00471	D-Glutamine and D-glutamate metabolism	Metabolism	-0.79181
ko00909	Sesquiterpenoid and triterpenoid biosynthesis	Metabolism	-0.79182
BR:ko04090	CD Molecules	Environmental Information Processing	-0.79321
ko02030	Bacterial chemotaxis	Cellular Processes	-0.79331
ko00300	Lysine biosynthesis	Metabolism	-0.79489
ko04016	MAPK signaling pathway - plant	Environmental Information Processing	-0.79561
BR:ko01007	Amino acid related enzymes	Metabolism	-0.79576
BR:ko02044	Secretion system	Environmental Information Processing	-0.79587
ko00020	Citrate cycle (TCA cycle)	Metabolism	-0.79679
ko03450	Non-homologous end-joining	Genetic Information Processing	-0.79788
ko04152	AMPK signaling pathway	Environmental Information Processing	-0.79908
ko00220	Arginine biosynthesis	Metabolism	-0.79926
ko00250	Alanine, aspartate and glutamate metabolism	Metabolism	-0.79976
ko00670	One carbon pool by folate	Metabolism	-0.80020
ko00340	Histidine metabolism	Metabolism	-0.80233
ko00730	Thiamine metabolism	Metabolism	-0.80254
ko00260	Glycine, serine and threonine metabolism	Metabolism	-0.80469
ko00280	Valine, leucine and isoleucine degradation	Metabolism	-0.80489
ko02024	Quorum sensing	Cellular Processes	-0.80505
ko00261	Monobactam biosynthesis	Metabolism	-0.80625
ko00564	Glycerophospholipid metabolism	Metabolism	-0.80630
ko03070	Bacterial secretion system	Environmental Information Processing	-0.80643
BR:ko01001	Protein kinases	Metabolism	-0.80802
ko00281	Geraniol degradation	Metabolism	-0.80804
BR:ko01004	Lipid biosynthesis proteins	Metabolism	-0.80877
ko00010	Glycolysis / Gluconeogenesis	Metabolism	-0.80908
ko00515	Mannose type O-glycan biosynthesis	Metabolism	-0.80945
ko00521	Streptomycin biosynthesis	Metabolism	-0.81051
ko00860	Porphyrin and chlorophyll metabolism	Metabolism	-0.81090
ko00270	Cysteine and methionine metabolism	Metabolism	-0.81114
ko00450	Selenocompound metabolism	Metabolism	-0.81126
ko04218	Cellular senescence	Cellular Processes	-0.81145
ko00903	Limonene and pinene degradation	Metabolism	-0.81161
ko04022	cGMP-PKG signaling pathway	Environmental Information Processing	-0.81170
ko00983	Drug metabolism - other enzymes	Metabolism	-0.81253
ko02020	Two-component system	Environmental Information Processing	-0.81253
ko00620	Pyruvate metabolism	Metabolism	-0.81337
BR:ko01002	Peptidases	Metabolism	-0.81499
BR:ko03000	Transcription factors	Genetic Information Processing	-0.81534
ko00072	Synthesis and degradation of ketone bodies	Metabolism	-0.81606
ko00500	Starch and sucrose metabolism	Metabolism	-0.81661
ko00630	Glyoxylate and dicarboxylate metabolism	Metabolism	-0.81685
ko00790	Folate biosynthesis	Metabolism	-0.81772
ko04068	FoxO signaling pathway	Environmental Information Processing	-0.81804
ko00966	Glucosinolate biosynthesis	Metabolism	-0.81829
ko00740	Riboflavin metabolism	Metabolism	-0.81869
ko00908	Zeatin biosynthesis	Metabolism	-0.81895
ko00561	Glycerolipid metabolism	Metabolism	-0.82013
ko00071	Fatty acid degradation	Metabolism	-0.82064
ko00720	Carbon fixation pathways in prokaryotes	Metabolism	-0.82123
ko00514	Other types of O-glycan biosynthesis	Metabolism	-0.82186
ko04146	Peroxisome	Cellular Processes	-0.82196
ko00750	Vitamin B6 metabolism	Metabolism	-0.82241
ko00650	Butanoate metabolism	Metabolism	-0.82306
BR:ko02000	Transporters	Environmental Information Processing	-0.82488
ko00960	Tropane, piperidine and pyridine alkaloid biosynthesis	Metabolism	-0.82509
ko04217	Necroptosis	Cellular Processes	-0.82552
ko04216	Ferroptosis	Cellular Processes	-0.82600
ko00030	Pentose phosphate pathway	Metabolism	-0.82600
ko00330	Arginine and proline metabolism	Metabolism	-0.82760
ko00310	Lysine degradation	Metabolism	-0.82767
ko00640	Propanoate metabolism	Metabolism	-0.82831
ko00660	C5-Branched dibasic acid metabolism	Metabolism	-0.82848
ko00999	Biosynthesis of secondary metabolites - unclassified	Metabolism	-0.82850
ko02025	Biofilm formation - Pseudomonas aeruginosa	Cellular Processes	-0.82974
ko04014	Ras signaling pathway	Environmental Information Processing	-0.82974
ko02010	ABC transporters	Environmental Information Processing	-0.83044
BR:ko04147	Exosome	Cellular Processes	-0.83159
ko00410	beta-Alanine metabolism	Metabolism	-0.83188
ko04141	Protein processing in endoplasmic reticulum	Genetic Information Processing	-0.83321
ko00603	Glycosphingolipid biosynthesis - globo and isoglobo series	Metabolism	-0.83334

Table A1 continued from previous page

Category ID	Category Name	Pathway Class	Mean ln(CI)/ln(family size)
ko00051	Fructose and mannose metabolism	Metabolism	-0.83360
BR:ko04121	Ubiquitin system	Genetic Information Processing	-0.83383
ko02060	Phosphotransferase system (PTS)	Environmental Information Processing	-0.83402
BR:ko01003	Glycosyltransferases	Metabolism	-0.83457
ko00592	alpha-Linolenic acid metabolism	Metabolism	-0.83608
ko00680	Methane metabolism	Metabolism	-0.83630
ko00380	Tryptophan metabolism	Metabolism	-0.83639
ko00906	Carotenoid biosynthesis	Metabolism	-0.83644
ko04115	p53 signaling pathway	Cellular Processes	-0.83737
ko00472	D-Arginine and D-ornithine metabolism	Metabolism	-0.83832
ko04371	Apelin signaling pathway	Environmental Information Processing	-0.83892
ko00100	Steroid biosynthesis	Metabolism	-0.83893
ko04024	cAMP signaling pathway	Environmental Information Processing	-0.84009
ko04214	Apoptosis - fly	Cellular Processes	-0.84019
ko04013	MAPK signaling pathway - fly	Environmental Information Processing	-0.84028
ko00562	Inositol phosphate metabolism	Metabolism	-0.84162
ko00984	Steroid degradation	Metabolism	-0.84302
ko00930	Caprolactam degradation	Metabolism	-0.84317
BR:ko01005	Lipopolysaccharide biosynthesis proteins	Metabolism	-0.84403
ko00350	Tyrosine metabolism	Metabolism	-0.84431
ko03013	RNA transport	Genetic Information Processing	-0.84507
ko00430	Taurine and hypotaurine metabolism	Metabolism	-0.84532
ko00360	Phenylalanine metabolism	Metabolism	-0.84546
ko00910	Nitrogen metabolism	Metabolism	-0.84555
ko00920	Sulfur metabolism	Metabolism	-0.84580
ko04080	Neuroactive ligand-receptor interaction	Environmental Information Processing	-0.84690
ko00253	Tetracycline biosynthesis	Metabolism	-0.84690
BR:ko04091	Lectins	Cellular Processes	-0.84708
ko00982	Drug metabolism - cytochrome P450	Metabolism	-0.84744
ko00520	Amino sugar and nucleotide sugar metabolism	Metabolism	-0.84756
ko00565	Ether lipid metabolism	Metabolism	-0.84762
ko01053	Biosynthesis of siderophore group nonribosomal peptides	Metabolism	-0.84870
ko04072	Phospholipase D signaling pathway	Environmental Information Processing	-0.84874
ko00460	Cyanoamino acid metabolism	Metabolism	-0.84913
ko04071	Sphingolipid signaling pathway	Environmental Information Processing	-0.84915
ko00052	Galactose metabolism	Metabolism	-0.84964
ko00760	Nicotinate and nicotinamide metabolism	Metabolism	-0.85061
ko00401	Novobiocin biosynthesis	Metabolism	-0.85106
ko00643	Styrene degradation	Metabolism	-0.85175
ko04138	Autophagy - yeast	Cellular Processes	-0.85211
ko01056	Biosynthesis of type II polyketide backbone	Metabolism	-0.85276
ko00511	Other glycan degradation	Metabolism	-0.85400
BR:ko00536	Glycosaminoglycan binding proteins	Environmental Information Processing	-0.85454
ko00510	N-Glycan biosynthesis	Metabolism	-0.85457
ko00363	Bisphenol degradation	Metabolism	-0.85574
ko00333	Prodigiosin biosynthesis	Metabolism	-0.85595
ko00513	Various types of N-glycan biosynthesis	Metabolism	-0.85645
BR:ko01504	Antimicrobial resistance genes	Environmental Information Processing	-0.85781
ko00965	Betalain biosynthesis	Metabolism	-0.85912
ko00600	Sphingolipid metabolism	Metabolism	-0.85917
BR:ko01009	Protein phosphatase and associated proteins	Metabolism	-0.86025
ko03015	mRNA surveillance pathway	Genetic Information Processing	-0.86037
BR:ko04131	Membrane trafficking	Genetic Information Processing	-0.86061
ko00073	Cutin, suberine and wax biosynthesis	Metabolism	-0.86071
ko00332	Carbapenem biosynthesis	Metabolism	-0.86135
ko00405	Phenazine biosynthesis	Metabolism	-0.86410
ko00633	Nitrotoluene degradation	Metabolism	-0.86604
ko00121	Secondary bile acid biosynthesis	Metabolism	-0.86608
BR:ko04040	Ion channels	Cellular Processes	-0.86656
ko04210	Apoptosis	Cellular Processes	-0.86696
ko00791	Atrazine degradation	Metabolism	-0.86704
ko00981	Insect hormone biosynthesis	Metabolism	-0.86778
ko00053	Ascorbate and aldarate metabolism	Metabolism	-0.86778
ko00040	Pentose and glucuronate interconversions	Metabolism	-0.86966
ko00604	Glycosphingolipid biosynthesis - ganglio series	Metabolism	-0.86999
ko00531	Glycosaminoglycan degradation	Metabolism	-0.87223
ko00980	Metabolism of xenobiotics by cytochrome P450	Metabolism	-0.87247
ko03040	Spliceosome	Genetic Information Processing	-0.87249
BR:ko00199	Cytochrome P450	Metabolism	-0.87284
ko04150	mTOR signaling pathway	Environmental Information Processing	-0.87286
BR:ko00537	Glycosylphosphatidylinositol (GPI)-anchored proteins	Cellular Processes	-0.87301
ko00231	Puromycin biosynthesis	Metabolism	-0.87378
ko04145	Phagosome	Cellular Processes	-0.87439
ko00945	Stilbenoid, diarylheptanoid and gingerol biosynthesis	Metabolism	-0.87494
ko00590	Arachidonic acid metabolism	Metabolism	-0.87590
ko04142	Lysosome	Cellular Processes	-0.87608

Table A1 continued from previous page

Category ID	Category Name	Pathway Class	Mean ln(CI)/ln(family size)
ko00362	Benzoate degradation	Metabolism	-0.87634
ko00062	Fatty acid elongation	Metabolism	-0.87818
ko00627	Aminobenzoate degradation	Metabolism	-0.87825
ko00905	Brassinosteroid biosynthesis	Metabolism	-0.87842
ko00364	Fluorobenzoate degradation	Metabolism	-0.87846
ko04011	MAPK signaling pathway - yeast	Environmental Information Processing	-0.87986
ko00830	Retinol metabolism	Metabolism	-0.88015
ko00940	Phenylpropanoid biosynthesis	Metabolism	-0.88092
ko01055	Biosynthesis of vancomycin group antibiotics	Metabolism	-0.88125
ko00522	Biosynthesis of 12-, 14- and 16-membered macrolides	Metabolism	-0.88136
BR:ko01008	Polyketide biosynthesis proteins	Metabolism	-0.88255
ko00404	Staurosporine biosynthesis	Metabolism	-0.88322
ko01057	Biosynthesis of type II polyketide products	Metabolism	-0.88396
ko00311	Penicillin and cephalosporin biosynthesis	Metabolism	-0.88486
ko00625	Chloroalkane and chloroalkene degradation	Metabolism	-0.88503
ko04010	MAPK signaling pathway	Environmental Information Processing	-0.88513
ko00950	Isoquinoline alkaloid biosynthesis	Metabolism	-0.88532
ko04144	Endocytosis	Cellular Processes	-0.88604
BR:ko02048	Prokaryotic Defense System	Environmental Information Processing	-0.88813
ko00943	Isoflavonoid biosynthesis	Metabolism	-0.88853
ko00120	Primary bile acid biosynthesis	Metabolism	-0.88929
ko00440	Phosphonate and phosphinate metabolism	Metabolism	-0.88941
ko01059	Biosynthesis of enediyne antibiotics	Metabolism	-0.88993
ko00901	Indole alkaloid biosynthesis	Metabolism	-0.89010
ko00626	Naphthalene degradation	Metabolism	-0.89027
ko00941	Flavonoid biosynthesis	Metabolism	-0.89115
ko01051	Biosynthesis of ansamycins	Metabolism	-0.89211
ko00525	Acarbose and validamycin biosynthesis	Metabolism	-0.89217
ko03460	Fanconi anemia pathway	Genetic Information Processing	-0.89286
BR:ko04031	GTP-binding proteins	Environmental Information Processing	-0.89352
ko00232	Caffeine metabolism	Metabolism	-0.89372
ko00902	Monoterpenoid biosynthesis	Metabolism	-0.89421
ko04075	Plant hormone signal transduction	Environmental Information Processing	-0.89538
ko01052	Type I polyketide structures	Metabolism	-0.89726
ko00361	Chlorocyclohexane and chlorobenzene degradation	Metabolism	-0.89903
ko00140	Steroid hormone biosynthesis	Metabolism	-0.90338
ko00523	Polyketide sugar unit biosynthesis	Metabolism	-0.90395
ko00622	Xylene degradation	Metabolism	-0.90480
ko00254	Aflatoxin biosynthesis	Metabolism	-0.90528
ko00621	Dioxin degradation	Metabolism	-0.90798
ko00524	Neomycin, kanamycin and gentamicin biosynthesis	Metabolism	-0.90810
ko00532	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	Metabolism	-0.90844
ko00365	Furfural degradation	Metabolism	-0.90882
ko00624	Polycyclic aromatic hydrocarbon degradation	Metabolism	-0.91158
ko00623	Toluene degradation	Metabolism	-0.91175
ko01054	Nonribosomal peptide structures	Metabolism	-0.91263
ko00904	Diterpenoid biosynthesis	Metabolism	-0.91297
ko00331	Clavulanic acid biosynthesis	Metabolism	-0.91359
ko04330	Notch signaling pathway	Environmental Information Processing	-0.91419
ko00944	Flavone and flavonol biosynthesis	Metabolism	-0.91428
ko04064	NF-kappa B signaling pathway	Environmental Information Processing	-0.91484
ko04370	VEGF signaling pathway	Environmental Information Processing	-0.91484
ko00402	Benzoxazinoid biosynthesis	Metabolism	-0.91607
ko04668	TNF signaling pathway	Environmental Information Processing	-0.91627
ko00642	Ethylbenzene degradation	Metabolism	-0.91988
ko00601	Glycosphingolipid biosynthesis - lacto and neolacto series	Metabolism	-0.92249
ko00534	Glycosaminoglycan biosynthesis - heparan sulfate / heparin	Metabolism	-0.92289
ko04137	Mitophagy - animal	Cellular Processes	-0.92358
ko04530	Tight junction	Cellular Processes	-0.92542
BR:ko04030	G protein-coupled receptors	Environmental Information Processing	-0.92605
ko04140	Autophagy - animal	Cellular Processes	-0.92891
ko04390	Hippo signaling pathway	Environmental Information Processing	-0.92973
ko01058	Acridone alkaloid biosynthesis	Metabolism	-0.93253
ko04310	Wnt signaling pathway	Environmental Information Processing	-0.93275
ko04341	Hedgehog signaling pathway - fly	Environmental Information Processing	-0.93283
ko00563	Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	Metabolism	-0.94100
BR:ko03200	Viral proteins	Cellular Processes	-0.94270
ko04139	Mitophagy - yeast	Cellular Processes	-0.94291
ko04340	Hedgehog signaling pathway	Environmental Information Processing	-0.94415

Table A2. Most homoplasic KO annotations. Phylogenetic patchiness was computed for each KO present in at least 50 genomes using the CI, a common homoplasy metric representing the inverse of the minimum possible number of state changes (trait gain or loss) given the tree topology. The final phylogenetic patchiness score is equal to ln(CI)/ln(family size) where family size is the total number of genomes containing the trait. This table contains the 200 most homoplasic KO terms.

KEGG Orthology ID	Definition	Family Size	ln(CI)/ln(family size)
K21502	Gp4; DNA primase/helicase [EC:2.7.7.- 3.6.4.12]	77	-0.99699
K21238	Gp5; T7virus DNA-directed DNA polymerase [EC:2.7.7.7]	75	-0.99689
K21525	Gp3; endonuclease I [EC:3.1.21.2]	58	-0.99572
K17398	DNMT3A; DNA (cytosine-5)-methyltransferase 3A [EC:2.1.1.37]	82	-0.99440
K10841	ERCC6, CSB, RAD26; DNA excision repair protein ERCC-6	77	-0.98772
K10908	POLRMT, RPO41; DNA-directed RNA polymerase, mitochondrial [EC:2.7.7.6]	69	-0.98590
K18959	uvsW; ATP-dependent DNA helicase UvsW [EC:3.6.4.12]	53	-0.98532
K21313	1; T7 RNA polymerase [EC:2.7.7.6]	72	-0.98317
K07505	repA; regulatory protein RepA	482	-0.98265
K19175	dptH; DNA phosphorothioation-dependent restriction protein DptH	290	-0.98142
K19173	dptF; DNA phosphorothioation-dependent restriction protein DptF	244	-0.98116
K07453	K07453; putative restriction endonuclease	163	-0.98105
K19174	dptG; DNA phosphorothioation-dependent restriction protein DptG	243	-0.97940
K19001	HELLS, DDM1; ATP-dependent DNA helicase	116	-0.97904
K15858	ascF; CDP-3, 6-dideoxy-D-glycero-L-glycero-4-hexulose-4-reductase [EC:1.1.1.-]	165	-0.97870
K19169	dndB; DNA sulfur modification protein DndB	789	-0.97837
K11665	INO80, INOC1; DNA helicase INO80 [EC:3.6.4.12]	182	-0.97763
K15943	snoaL, dnrD, dauD, aknH; nogalonic acid methyl ester cyclase / aklanonic acid methyl ester cyclase [EC:5.5.1.26 5.5.1.23]	75	-0.97731
K15045	RSAD2; radical S-adenosyl methionine domain-containing protein 2	186	-0.97707
K15192	BTAF1, MOT1; TATA-binding protein-associated factor [EC:3.6.4.-]	210	-0.97630
K06922	K06922; uncharacterized protein	644	-0.97589
K12914	phpK; P-methyltransferase [EC:2.1.1.326]	51	-0.97376
K22363	etnE; 2-hydroxypropyl-CoM lyase [EC:4.4.1.23]	58	-0.97311
K15359	hspA; 6-hydroxy-3-succinoylpyridine 3-monooxygenase [EC:1.14.13.163]	170	-0.97302
K18960	gp49; recombination endonuclease VII	126	-0.97192
K07445	K07445; putative DNA methylase	864	-0.97180
K21511	GPC; capsid assembly protease [EC:3.4.21.-]	361	-0.97138
K02315&K11144	dnaC; DNA replication protein DnaC & dnaI; primosomal protein DnaI	62	-0.97097
K14580	nahAd, ndoC, nbzAd, dntAd; naphthalene 1,2-dioxygenase subunit beta [EC:1.14.12.12 1.14.12.23 1.14.12.24]	62	-0.97097
K02334	dpo; DNA polymerase bacteriophage-type [EC:2.7.7.7]	850	-0.97037
K21527	mom; adenine modification enzyme [EC:2.3.1.-]	68	-0.97034
K07132	MuB; ATP-dependent target DNA activator [EC:3.6.1.3]	2083	-0.97014
K19057	merD; MerR family transcriptional regulator, mercuric resistance operon regulatory protein	259	-0.96981
K19630	tsaC1; 4-formylbenzenesulfonate dehydrogenase [EC:1.2.1.62]	67	-0.96976
K21328	calS13, atmS13; dTDP-4-amino-4,6-dideoxy-D-glucose/dTDP-4-amino-2,4-dideoxy-beta-L-xylose transaminase [EC:2.6.1.33 2.6.1.-]	196	-0.96966
K09843	CYP707A; (+)-abscisic acid 8'-hydroxylase [EC:1.14.13.93]	73	-0.96933
K15945	snoaL2; C-1 hydroxylase	182	-0.96914
K13003	wbtG; glycosyltransferase [EC:2.4.1.-]	59	-0.96903
K21512	gpA; terminase, large subunit [EC:3.1.21.4]	884	-0.96893
K17677	IRC3; ATP-dependent helicase IRC3 [EC:3.6.4.-]	186	-0.96878
K07495	K07495; putative transposase	1535	-0.96836
K07504	K07504; predicted type IV restriction endonuclease	1580	-0.96819
K21183	sgcC2, mdpC2, kedY2; peptidyl carrier protein	51	-0.96817
K08280	wbbJ; lipopolysaccharide O-acetyltransferase [EC:2.3.1.-]	586	-0.96771
K20156	sgcG; 2-amino-4-deoxychorismate dehydrogenase [EC:1.3.99.24]	164	-0.96757
K22302	dicC; transcriptional repressor of cell division inhibition gene dicB	301	-0.96747
K09124	K09124; uncharacterized protein	690	-0.96642
K09144	K09144; uncharacterized protein	2314	-0.96599
K18916	ptxD; phosphonate dehydrogenase [EC:1.20.1.1]	317	-0.96558
K16112	blmIV; nonribosomal peptide synthetase protein BlmIV	233	-0.96545
K19059	merE; mercuric ion transport protein	244	-0.96473
K01160	rusA; crossover junction endodeoxyribonuclease RusA [EC:3.1.22.4]	983	-0.96456
K01156&K07316	res; type III restriction enzyme [EC:3.1.21.5] & mod; adenine-specific DNA-methyltransferase [EC:2.1.1.72]	300	-0.96449
K14692	SLC30A5_7, ZNT5_7, MTP, MSC2; solute carrier family 30 (zinc transporter), member 5/7	53	-0.96432
K01127	E3.1.4.50; glycosylphosphatidylinositol phospholipase D [EC:3.1.4.50]	59	-0.96426
K21490	yokJ; antitoxin YokJ	131	-0.96416
K20170	nicA1; nicotine oxidoreductase [EC:1.5.3.-]	621	-0.96379

Table A2 continued from previous page

KEGG Orthology ID	Definition	Family Size	ln(CI)/ln(family size)
K19147	mcrC; 5-methylcytosine-specific restriction enzyme subunit McrC	2545	-0.96366
K19058	merC; mercuric ion transport protein	711	-0.96364
K01992&K19310	ABC-2.P; ABC-2 type transport system permease protein & bcrB; bacitracin transport system permease protein	103	-0.96357
K19167	abiQ; protein AbiQ	428	-0.96346
K12376	ARSK; arylsulfatase K [EC:3.1.6.-]	58	-0.96345
K07339	hicA; mRNA interferase HicA [EC:3.1.-.-]	854	-0.96339
K07475	cas3; CRISPR-associated endonuclease Cas3-HD [EC:3.1.-.-]	1217	-0.96305
K01143	E3.1.11.3; exodeoxyribonuclease (lambda-induced) [EC:3.1.11.3]	465	-0.96279
K07474	xtmA; phage terminase small subunit	1794	-0.96264
K19156	prlF, sohA; antitoxin PrlF	939	-0.96262
K10700	ebdA; ethylbenzene hydroxylase subunit alpha [EC:1.17.99.2]	57	-0.96259
K08356	aoxB; arsenite oxidase large subunit [EC:1.20.2.1 1.20.9.1]	279	-0.96259
K07452	mcrB; 5-methylcytosine-specific restriction enzyme B [EC:3.1.21.-]	2845	-0.96254
K19171	dndD; DNA sulfur modification protein DndD	1172	-0.96247
K10954	zot; zona occludens toxin	564	-0.96224
K07454	K07454; putative restriction endonuclease	2523	-0.96197
K14747	bal; benzoylacetate-CoA ligase [EC:6.2.1.-]	164	-0.96184
K12228	trbB; TrbB protein	56	-0.96171
K22358	amoE; alkene monooxygenase beta subunit [EC:1.14.13.69]	67	-0.96156
K19145	csx16; CRISPR-associated protein Csx16	134	-0.96155
K07392	PRS2; AAA family ATPase	129	-0.96152
K22360	amoC; alkene monooxygenase ferredoxin subunit	114	-0.96150
K12062	trbI; conjugal transfer pilin signal peptidase TrbI	695	-0.96144
K11946	phdG; hydratase-aldolase [EC:4.1.2.-]	153	-0.96140
K18275	phdK; 2-formylbenzoate dehydrogenase [EC:1.2.1.78]	316	-0.96137
K07741	antB; anti-repressor protein	2259	-0.96135
K19155	yhaV; toxin YhaV [EC:3.1.-.-]	684	-0.96121
K15241	pcpC; tetrachloro-p-hydroquinone reductive dehalogenase [EC:2.5.1.-]	61	-0.96117
K00152	nahF; salicylaldehyde dehydrogenase [EC:1.2.1.65]	156	-0.96083
K21732	SLD; acyl-lipid (11-3)-desaturase [EC:1.14.19.4]	66	-0.96078
K16411	stiG; stigmatellin polyketide synthase StiG	211	-0.96072
K09960	K09960; uncharacterized protein	556	-0.96044
K12453	rfbS; CDP-paratose synthetase [EC:1.1.1.342]	271	-0.96033
K20765	camK; 6-oxocamphor hydrolase [EC:3.7.1.18]	441	-0.96016
K19138	csm2; CRISPR-associated protein Csm2	613	-0.95961
K17825	FTMF; verruculogen synthase [EC:1.14.11.38]	138	-0.95944
K19139	csm4; CRISPR-associated protein Csm4	626	-0.95940
K21674	cdhB; caffeine dehydrogenase subunit beta [EC:1.17.5.2]	210	-0.95938
K21376	aziB1; 5-methyl-1-naphthoate 3-hydroxylase [EC:1.14.13.189]	266	-0.95936
K19170	dndC; DNA sulfur modification protein DndC	1003	-0.95928
K07464&K15342	cas4; CRISPR-associated exonuclease Cas4 [EC:3.1.12.1] & cas1; CRISP-associated protein Cas1	165	-0.95924
K19068	wbjC; UDP-2-acetamido-2,6-beta-L-arabino-hexul-4-ose reductase [EC:1.1.1.367]	1593	-0.95918
K19136	csx17; CRISPR-associated protein Csx17	113	-0.95880
K17058	EOMT1, CVOMT1; eugenol/chavicol O-methyltransferase [EC:2.1.1.146]	53	-0.95879
K01156	res; type III restriction enzyme [EC:3.1.21.5]	4789	-0.95878
K22014	Nu1; terminase small subunit	258	-0.95859
K19821	SERPINB2, PAI2; plasminogen activator inhibitor 2	103	-0.95856
K09132	K09132; uncharacterized protein	521	-0.95855
K19299	aph3-III; aminoglycoside 3'-phosphotransferase III [EC:2.7.1.95]	236	-0.95838
K21725	pnpA; 4-nitrocatechol/4-nitrophenol 4-monooxygenase [EC:1.14.13.166 1.14.13.167]	374	-0.95822
K16395	epoB; epothilone synthetase B	144	-0.95822
K00251	AKR1D1; 3-oxo-5-beta-steroid 4-dehydrogenase [EC:1.3.1.3]	83	-0.95819
K00089	AKR1C2; 3alpha-hydroxysteroid 3-dehydrogenase [EC:1.1.1.213 1.1.1.357]	68	-0.95818
K12070	traI; conjugal transfer pilus assembly protein TraI	243	-0.95806
K07451	mcrA; 5-methylcytosine-specific restriction enzyme A [EC:3.1.21.-]	4173	-0.95778
K15173	TTF2; transcription termination factor 2 [EC:3.6.4.-]	57	-0.95749
K18227	cmtAa; p-cumate 2,3-dioxygenase ferredoxin reductase component [EC:1.18.1.3]	164	-0.95744
K14585	nahE; trans-o-hydroxybenzylidenepyruvate hydratase-aldolase [EC:4.1.2.45]	177	-0.95744
K09002	csm3; CRISPR-associated protein Csm3	774	-0.95740
K17068	fdm; formaldehyde dismutase [EC:1.2.98.1]	119	-0.95723
K12630	pur4; puromycin biosynthesis protein Pur4 [EC:2.-.-.-]	159	-0.95721
K11949	phdJ; 4-(2-carboxyphenyl)-2-oxobut-3-enoate aldolase [EC:4.1.2.34]	172	-0.95721
K19141	cmr5; CRISPR-associated protein Cmr5	467	-0.95721
K21018	fumD; fumonisin B1 esterase [EC:3.1.1.87]	206	-0.95720
K21224	kedN5; radical SAM C-methyltransferase	91	-0.95713
K18842	chpS, chpBI; antitoxin ChpS	967	-0.95709
K19132	csb2; CRISPR-associated protein Csb2	303	-0.95700
K13306	fcd; dTDP-4-dehydro-6-deoxyglucose reductase [EC:1.1.1.266]	118	-0.95675
K20566	kanI; paromamine 6'-oxidase / 2'-deamino-2'-hydroxyparomamine 6'-oxidase [EC:1.1.3.43 1.1.3.-]	149	-0.95674
K09961	K09961; uncharacterized protein	689	-0.95664
K19143	csx1; CRISPR-associated protein Csx1	249	-0.95662
K16003	pikAIV; narbonolide synthase [EC:2.3.1.240]	131	-0.95656
K19160	yafO; mRNA interferase YafO [EC:3.1.-.-]	170	-0.95655

Table A2 continued from previous page

KEGG Orthology ID	Definition	Family Size	ln(CI)/ln(family size)
K07487	K07487; transposase	3497	-0.95629
K07497&K07483	K07497; putative transposase & K07483; transposase	2527	-0.95627
K14440	SMARCAL1, HARP; SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A-like protein 1 [EC:3.6.4.12]	117	-0.95626
K07016	csm1, cas10; CRISPR-associated protein Csm1	639	-0.95603
K18841	chpB, chpBK; mRNA interferase ChpB [EC:3.1.-.-]	1460	-0.95593
K00670	NAA30, MAK3; N-alpha-acetyltransferase 30 [EC:2.3.1.256]	75	-0.95591
K16419	mlsB; mycolactone side chain polyketide synthase	89	-0.95585
K12069	traA; conjugal transfer pilus assembly protein TraA	303	-0.95552
K06231	HHIP; hedgehog interacting protein	60	-0.95547
K12720	cloN5, couN5; peptidyl carrier protein	50	-0.95543
K19117	csd1, cas8c; CRISPR-associated protein Csd1	1838	-0.95535
K19118	csd2, cas7; CRISPR-associated protein Csd2	1937	-0.95530
K19119	cas5d; CRISPR-associated protein Cas5d	1948	-0.95516
K00590	E2.1.1.113; site-specific DNA-methyltransferase (cytosine-N4-specific) [EC:2.1.1.113]	2204	-0.95502
K16111	blmVII; nonribosomal peptide synthetase protein BlmVII	131	-0.95462
K12055	K12055, parA; chromosome partitioning related protein ParA	651	-0.95456
K11395	kdpgA; 2-dehydro-3-deoxy-phosphogluconate/2-dehydro-3-deoxy-6-phosphogalactonate aldolase [EC:4.1.2.55]	96	-0.95451
K13670	pimF; putative glycosyltransferase [EC:2.4.-.-]	228	-0.95440
K21325	calS11; dTDP-rhamnose C3-O-methyltransferase [EC:2.1.1.-]	405	-0.95440
K12743	PCBAB; N-(5-amino-5-carboxypentanoyl)-L-cysteinyl-D-valine synthase [EC:6.3.2.26]	100	-0.95424
K12064	traV; conjugal transfer pilus assembly protein TraV	608	-0.95409
K06909	xtmB; phage terminase large subunit	1705	-0.95391
K17831	lodA; L-lysine 6-oxidase [EC:1.4.3.20]	108	-0.95382
K19856	aveBVII, avrH; 3-O-methyltransferase [EC:2.1.1.-]	419	-0.95380
K15764	tmoE, tbuA2, touE; toluene monooxygenase system protein E [EC:1.14.13.236 1.14.13.-]	186	-0.95366
K01155	E3.1.21.4; type II restriction enzyme [EC:3.1.21.4]	3996	-0.95357
K03395	aac3-I; aminoglycoside 3-N-acetyltransferase I [EC:2.3.1.60]	435	-0.95353
K07493	K07493; putative transposase	4893	-0.95346
K18159	NDUFAF1, CIA30; NADH dehydrogenase [ubiquinone] 1 alpha subcomplex assembly factor 1	58	-0.95339
K12071	traD; conjugal transfer pilus assembly protein TraD	921	-0.95338
K09952	csn1, cas9; CRISPR-associated endonuclease Csn1 [EC:3.1.-.-]	1943	-0.95337
K16106	blmVI; nonribosomal peptide synthetase protein BlmVI	273	-0.95323
K13965	SERPINB8; serpin B8	204	-0.95314
K12983	waaV; UDP-glucose:(glucosyl)LPS beta-1,3-glucosyltransferase [EC:2.4.1.-]	1909	-0.95305
K18827	wbdD, wbbD; O-antigen chain-terminating methyltransferase [EC:2.1.1.- 2.1.1.294 2.7.1.181]	477	-0.95302
K19140	csm5; CRISPR-associated protein Csm5	472	-0.95282
K12059	trbC; conjugal transfer pilus assembly protein TrbC	506	-0.95274
K12065	traB; conjugal transfer pilus assembly protein TraB	718	-0.95268
K16382	amphA, nysA; polyene macrolide polyketide synthase, loading module	159	-0.95254
K18610	pdla; 4-pyridoxolactonase [EC:3.1.1.27]	190	-0.95239
K00221	E4.99.1.2; alkylmercury lyase [EC:4.99.1.2]	309	-0.95225
K20681	ftdC; dTDP-3-amino-3,6-dideoxy-alpha-D-galactopyranose 3-N-acetyltransferase [EC:2.3.1.197]	193	-0.95211
K12823	DDX5, DBP2; ATP-dependent RNA helicase DDX5/DBP2 [EC:3.6.4.13]	66	-0.95210
K08687	E3.5.1.59; N-carbamoylsarcosine amidase [EC:3.5.1.59]	253	-0.95201
K15930	lndM2; bifunctional oxygenase/reductase	79	-0.95181
K07534	badK; cyclohex-1-ene-1-carboxyl-CoA hydratase [EC:4.2.1.-]	388	-0.95174
K14746	ped; (S)-1-phenylethanol dehydrogenase [EC:1.1.1.311]	813	-0.95165
K15913	pglD; UDP-N-acetylbacillosamine N-acetyltransferase [EC:2.3.1.203]	251	-0.95150
K19543	aph3-VII; aminoglycoside 3'-phosphotransferase [EC:2.7.1.95]	168	-0.95145
K20590	genD2, gtmC, gntC; gentamicin A2 dehydrogenase [EC:1.1.1.-]	213	-0.95124
K19128	csy2; CRISPR-associated protein Csy2	714	-0.95116
K21675	cdhC; caffeine dehydrogenase subunit gamma [EC:1.17.5.2]	353	-0.95112
K19545	lnuA_C_D_E, lin; lincosamide nucleotidyltransferase A/C/D/E	401	-0.95103
K07482	K07482; transposase, IS30 family	4034	-0.95097
K12068	traL; conjugal transfer pilus assembly protein TraL	600	-0.95080
K21364	wfeD; UDP-Gal:alpha-D-GlcNAc-diphosphoundecaprenol beta-1,4-galactosyltransferase [EC:2.4.1.304]	139	-0.95073
K20577	aprQ; paromamine/lividamine 6'-oxidase [EC:1.1.3.43 1.1.3.-]	135	-0.95070
K07316	mod; adenine-specific DNA-methyltransferase [EC:2.1.1.72]	5998	-0.95070
K14623	dinD; DNA-damage-inducible protein D	1974	-0.95061
K17680	PEO1; twinkle protein [EC:3.6.4.12]	459	-0.95056
K15857	ascE; CDP-3, 6-dideoxy-D-glycero-D-glycero-4-hexulose-5-epimerase [EC:5.1.-.-]	229	-0.95051
K16109	blmIX; nonribosomal peptide synthetase protein BlmIX	192	-0.95049
K19165	phd; antitoxin Phd	754	-0.95045
K13308	desI, eryCIV; dTDP-4-amino-4,6-dideoxy-D-glucose transaminase [EC:2.6.1.33]	587	-0.95045
K11944	nidB; PAH dioxygenase small subunit [EC:1.13.11.-]	397	-0.95037
K19590	araDH; D-arabinose 1-dehydrogenase [EC:1.1.1.117]	73	-0.95037
K19115	csh2; CRISPR-associated protein Csh2	1295	-0.95031
K07061	cmr1; CRISPR-associated protein Cmr1	526	-0.95019

Table A2 continued from previous page

KEGG Orthology ID	Definition	Family Size	ln(CI)/ln(family size)
K10550	alsC; D-allose transport system permease protein	288	-0.95001
K14602	flnD1; 2'-carboxy-2,3-dihydroxybiphenyl 1,2-dioxygenase large subunit	209	-0.94999

Table A3. Least homoplasic KO annotations. Phylogenetic patchiness was computed for each KO present in at least 50 genomes using the CI, a common homoplasy metric representing the inverse of the minimum possible number of state changes (trait gain or loss) given the tree topology. The final phylogenetic patchiness score is equal to ln(CI)/ln(family size) where family size is the total number of genomes containing the trait. This table contains the 200 least homoplasic KO terms.

KEGG Orthology ID	Definition	Family Size	ln(CI)/ln(family size)
K11028	vacA; vacuolating cytotoxin	54	0.00000
K15843	hopC, alpA; outer membrane protein HopC/AlpA	54	0.00000
K15844	hopB, alpB; outer membrane protein HopB/AlpB	54	0.00000
K15848	sabA; outer membrane protein SabA	54	0.00000
K06437	yknT; sigma-E controlled sporulation protein	73	0.00000
K12210	icmF; intracellular multiplication protein IcmF	81	-0.15773
K16919	ytrC_D; acetoin utilization transport system permease protein	80	-0.15818
K12215	icmM, dotJ; intracellular multiplication protein IcmM	76	-0.16005
K15845	hopZ; outer membrane protein HopZ	53	-0.17458
K15847	babA; outer membrane protein BabA	53	-0.17458
K10922	toxS; transmembrane regulatory protein ToxS	236	-0.25372
K06427	sspJ; small acid-soluble spore protein J (minor)	67	-0.26128
K02972	sra; stationary-phase-induced ribosome-associated protein	180	-0.26696
K06389	spoIISB; stage II sporulation protein SB	81	-0.31546
K19432	slrA; anti-repressor of SlrR	76	-0.32011
K01760&K01739	metC; cystathionine beta-lyase [EC:4.4.1.8] & metB; cystathionine gamma-synthase [EC:2.5.1.48]	108	-0.34374
K07268	oapA; opacity associated protein	106	-0.34512
K16654	K16654; spore-specific protein	105	-0.34582
K18132	K18132, porA; major outer membrane protein P.IA	104	-0.34653
K12051	comB7; ComB7 competence protein	52	-0.35085
K13630	marB; multiple antibiotic resistance protein MarB	214	-0.36264
K07676	rcsD; two-component system, NarL family, sensor histidine kinase RcsD [EC:2.7.13.3]	552	-0.36471
K02720	psbV; photosystem II cytochrome c550	295	-0.36565
K12221	icmS; intracellular multiplication protein IcmS	130	-0.36810
K06364	rapF; response regulator aspartate phosphatase F [EC:3.1.-.-]	78	-0.36942
K01947	birA-coaX; biotin—[acetyl-CoA-carboxylase] ligase / type III pantothenate kinase [EC:6.3.4.15 2.7.1.33]	71	-0.37756
K12551	sgtA; monofunctional glycosyltransferase [EC:2.4.1.129]	69	-0.38011
K15846	hpaA; neuraminyllactose-binding hemagglutinin	68	-0.38143
K20337	psmB; phenol-soluble modulin beta	64	-0.38699
K03922	desA2; acyl-[acyl-carrier-protein] desaturase [EC:1.14.19.2]	383	-0.38712
K18955	whiB1_2_3_4; WhiB family transcriptional regulator, redox-sensing transcriptional regulator	3041	-0.39096
K13301	secM; secretion monitor	555	-0.39325
K06321	cgeC; spore maturation protein CgeC	54	-0.40347
K20707	E5.1.1.5; lysine racemase [EC:5.1.1.5]	54	-0.40347
K02093	apcB; allophycocyanin beta subunit	284	-0.40761
K02097	apcF; phycobilisome core component	284	-0.40761
K19162	tomB; hha toxicity modulator TomB	540	-0.40768
K02255	ftnB; ferritin-like protein 2	216	-0.40876
K15474	enhC; enhanced entry protein EnhC	80	-0.40889
K12225	icmX; intracellular multiplication protein IcmX	78	-0.41126
K06340	cotV; spore coat protein V	75	-0.41500
K16711	wcaM; colanic acid biosynthesis protein WcaM	199	-0.41510
K02692	psaD; photosystem I subunit II	353	-0.42358
K02634	petA; apocytochrome f	352	-0.42378
K11609&K09458	kas; beta-ketoacyl ACP synthase [EC:2.3.1.-] & fabF; 3-oxoacyl-[acyl-carrier-protein] synthase II [EC:2.3.1.179]	286	-0.42396
K02344	holD; DNA polymerase III subunit psi [EC:2.7.7.7]	1031	-0.42438
K02095	apcD; allophycocyanin-B	282	-0.42502
K06314	rsfA; prespore-specific regulator	1017	-0.42521
K09146	K09146; uncharacterized protein	887	-0.42582
K09911	K09911; uncharacterized protein	943	-0.42990
K12216	icmN, lphA, dotK; intracellular multiplication protein IcmN	90	-0.43244
K02704	psbB; photosystem II CP47 chlorophyll apoprotein	351	-0.43765
K08903	psb28; photosystem II 13kDa protein	349	-0.43807
K20338	rot; MarR family transcriptional regulator, global regulator for virulence	83	-0.44037
K06144	uspB; universal stress protein B	799	-0.44056
K02092	apcA; allophycocyanin alpha subunit	281	-0.44071
K12147	msyB; acidic protein MsyB	385	-0.44330
K06363	rapE; response regulator aspartate phosphatase E [EC:3.1.-.-]	80	-0.44407
K19433	tapA; TasA anchoring/assembly protein	80	-0.44407

Table A3 continued from previous page

KEGG Orthology ID	Definition	Family Size	ln(CI)/ln(family size)
K07027&K14205	K07027; glycosyltransferase 2 family protein & mprF, fmtC; phosphatidylglycerol lysyltransferase [EC:2.3.2.3]	108	-0.44412
K18956	whiB5; WhiB family transcriptional regulator, redox-sensing transcriptional regulator	177	-0.44485
K05839	hha; haemolysin expression modulating protein	574	-0.44599
K02621&K02469	parC; topoisomerase IV subunit A [EC:5.99.1.-] & gyrA; DNA gyrase subunit A [EC:5.99.1.3]	172	-0.44732
K19688	bssR; biofilm regulator BssR	211	-0.44805
K13620	wcaD; putative colanic acid polymerase	205	-0.45048
K02698	psaK; photosystem I subunit X	350	-0.45051
K02716	psbO; photosystem II oxygen-evolving enhancer protein 1	347	-0.45117
K02719	psbU; photosystem II PsbU protein	286	-0.45349
K03764	metJ; MetJ family transcriptional regulator, methionine regulon repressor	1300	-0.45440
K02290	cpcG; phycobilisome rod-core linker protein	281	-0.45491
K12148	bssS; biofilm regulator BssS	565	-0.45612
K14159	rnhA-dnaQ; ribonuclease HI / DNA polymerase III subunit epsilon [EC:3.1.26.4 2.7.7.7]	228	-0.45768
K11386	embB; arabinosyltransferase B [EC:2.4.2.-]	834	-0.45955
K12219	icmQ; intracellular multiplication protein IcmQ	92	-0.45987
K11033	nheA; non-hemolytic enterotoxin A	118	-0.46057
K04478	sgtB; monofunctional glycosyltransferase [EC:2.4.1.129]	91	-0.46099
K08902	psb27; photosystem II Psb27 protein	344	-0.46366
K02693	psaE; photosystem I subunit IV	342	-0.46412
K06430	sspM; small acid-soluble spore protein M (minor)	65	-0.46615
K12288	hofM; pilus assembly protein HofM	546	-0.46718
K09904	K09904; uncharacterized protein	1229	-0.46840
K02699	psaL; photosystem I subunit XI	356	-0.47194
K02425	fliZ; regulator of sigma S factor FliZ	502	-0.47349
K09899	K09899; uncharacterized protein	1299	-0.47441
K05382	cpeS; phycoerythrin-associated linker protein	345	-0.47447
K09901	K09901; uncharacterized protein	1065	-0.47803
K12151	bhsA; multiple stress resistance protein BhsA	525	-0.47829
K21474	ripB; peptidoglycan DL-endopeptidase RipB [EC:3.4.-.-]	328	-0.47861
K08480	kaiA; circadian clock protein KaiA	281	-0.48029
K06365	rapG; response regulator aspartate phosphatase G [EC:3.1.-.-]	208	-0.48055
K18657	zapC; cell division protein ZapC	1098	-0.48096
K15723	syd; SecY interacting protein Syd	1226	-0.48288
K03971	pspD; phage shock protein D	441	-0.48356
K02859	ribT; riboflavin biosynthesis RibT protein	1279	-0.48446
K06360	rapB; response regulator aspartate phosphatase B [EC:3.1.-.-]	93	-0.48476
K02285	cpcB; phycocyanin beta chain	345	-0.48485
K03468	aaeB; p-hydroxybenzoic acid efflux pump subunit AaeB	481	-0.48507
K02694	psaF; photosystem I subunit III	343	-0.48533
K02707	psbE; photosystem II cytochrome b559 subunit alpha	342	-0.48557
K02713&K02707	psbL; photosystem II PsbL protein & psbE; photosystem II cytochrome b559 subunit alpha	342	-0.48557
K19032	PSRP3; 30S ribosomal protein 3	342	-0.48557
K02691	psaC; photosystem I subunit VII	341	-0.48581
K02284	cpcA; phycocyanin alpha chain	296	-0.48724
K05378	cpeC, mpeC; phycoerythrin-associated linker protein	296	-0.48724
K19229	sapD; cationic peptide transport system ATP-binding protein	1293	-0.48802
K19693	tfoS; AraC family transcriptional regulator, chitin signaling transcriptional activator	162	-0.48842
K02286	cpcC; phycocyanin-associated rod linker protein	289	-0.48930
K11926	crl; sigma factor-binding protein Crl	890	-0.49066
K09682	hpr; MarR family transcriptional regulator, protease production regulatory protein HPr	759	-0.49127
K02629	pecB; phycoerythrocyanin beta chain	276	-0.49331
K06604	flaI; flagellar rod protein FlaI	178	-0.49499
K07269	ytfB; uncharacterized protein	817	-0.49693
K03645	seqA; negative modulator of initiation of replication	1277	-0.49709
K10911	luxU; two-component system, phosphorelay protein LuxU	264	-0.49724
K06369	rapK; response regulator aspartate phosphatase K [EC:3.1.-.-]	82	-0.49861
K11387	embC; arabinosyltransferase C [EC:2.4.2.-]	848	-0.49939
K19227	sapB; cationic peptide transport system permease protein	1304	-0.49957
K11920	envY; AraC family transcriptional regulator	81	-0.50000
K02345	holE; DNA polymerase III subunit theta [EC:2.7.7.7]	575	-0.50014
K20073	mapZ, locZ; mid-cell-anchored protein Z	317	-0.50190
K06361	rapC; response regulator aspartate phosphatase C [EC:3.1.-.-]	216	-0.50380
K02094	apcC; phycobilisome core linker protein	276	-0.50409
K03633	mukF; chromosome partition protein MukF	1021	-0.50465
K14518	pipX; PII interaction protein X	339	-0.50540
K06368	rapJ; response regulator aspartate phosphatase J [EC:3.1.-.-]	92	-0.50922
K07751	pepB; PepB aminopeptidase [EC:3.4.11.23]	1312	-0.51029
K11037	hblD; hemolysin BL lytic component L1	130	-0.51051
K15744	Z-ISO; zeta-carotene isomerase [EC:5.2.1.12]	345	-0.51266
K00687	pbp2B, penA; penicillin-binding protein 2B	860	-0.51292
K01510	ENTPD1_3_8, CD39; apyrase [EC:3.6.1.5]	89	-0.51298
K12202	dotA; defect in organelle trafficking protein DotA	89	-0.51298
K02709	psbH; photosystem II PsbH protein	341	-0.51368

Table A3 continued from previous page

KEGG Orthology ID	Definition	Family Size	ln(CI)/ln(family size)
K12224	icmW; intracellular multiplication protein IcmW	126	-0.51381
K02637	petD; cytochrome b6-f complex subunit 4	373	-0.51414
K02705	psbC; photosystem II CP43 chlorophyll apoprotein	337	-0.51472
K06349	kbaA; KinB signaling pathway activation protein	1039	-0.51591
K07781	rcsA; LuxR family transcriptional regulator, capsular biosynthesis positive transcription factor	363	-0.51651
K06347	kapB; kinase-associated protein B	755	-0.51820
K18805	undefined	320	-0.51934
K07175&K00057	phoH2; PhoH-like ATPase & gpsA; glycerol-3-phosphate dehydrogenase (NAD(P)+) [EC:1.1.1.94]	260	-0.51979
K00410	fbcH; ubiquinol-cytochrome c reductase cytochrome b/c1 subunit	158	-0.52129
K13273	PMVK; phosphomevalonate kinase [EC:2.7.4.2]	54	-0.52130
K07490	feoC; ferrous iron transport protein C	444	-0.52135
K05584	ndhM; NAD(P)H-quinone oxidoreductase subunit M [EC:1.6.5.3]	342	-0.52179
K03755	adiY; AraC family transcriptional regulator, transcriptional activator of adiA	82	-0.52252
K21905	gadX; AraC family transcriptional regulator, glutamate-dependent acid resistance regulator	82	-0.52252
K00211	TYR1; prephenate dehydrogenase (NADP+) [EC:1.3.1.13]	67	-0.52256
K04017&K04018	nrfF; formate-dependent nitrite reductase complex subunit NrfF & nrfG; formate-dependent nitrite reductase complex subunit NrfG	113	-0.52564
K21463	tcaA; membrane-associated protein TcaA	93	-0.52903
K02628	pecA; phycoerythrocyanin alpha chain	286	-0.52966
K09698&K01885	gltX; nondiscriminating glutamyl-tRNA synthetase [EC:6.1.1.24] & EARS, gltX; glutamyl-tRNA synthetase [EC:6.1.1.17]	145	-0.53028
K05585	ndhN; NAD(P)H-quinone oxidoreductase subunit N [EC:1.6.5.3]	339	-0.53056
K06318	bofC; forespore regulator of the sigma-K checkpoint	764	-0.53119
K12098	cag13; cag pathogenicity island protein 13	50	-0.53155
K06047	TTL; tubulin—tyrosine ligase [EC:6.3.2.25]	62	-0.53239
K08985	K08985; putative lipoprotein	789	-0.53297
K03822	K03822; putative long chain acyl-CoA synthase [EC:6.2.1.-]	589	-0.53323
K06866	grcA; autonomous glycyl radical cofactor	1010	-0.53325
K16709	amsF; amylovoran biosynthesis protein AmsF	75	-0.53332
K08884&K12132	K08884; serine/threonine protein kinase, bacterial [EC:2.7.11.1] & prkC, stkP; eukaryotic-like serine/threonine-protein kinase [EC:2.7.11.1]	159	-0.53425
K12208	icmD, dotP; intracellular multiplication protein IcmD	120	-0.53576
K06286	ezrA; septation ring formation regulator	1876	-0.53643
K01100	E3.1.3.37; sedoheptulose-bisphosphatase [EC:3.1.3.37]	60	-0.53665
K12582	wecF, rffT; dTDP-N-acetylfucosamine:lipid II N-acetylfucosaminyltransferase [EC:2.4.1.325]	599	-0.53696
K09896	K09896; uncharacterized protein	998	-0.53775
K06178&K06183	rluB; 23S rRNA pseudouridine2605 synthase [EC:5.4.99.22] & rsuA; 16S rRNA pseudouridine516 synthase [EC:5.4.99.19]	194	-0.53783
K06080	rcsF; RcsF protein	782	-0.53792
K11611	inhA; enoyl ACP reductase [EC:1.3.1.9]	782	-0.53792
K09916	K09916; uncharacterized protein	2407	-0.53809
K03891	qcrB; ubiquinol-cytochrome c reductase cytochrome b subunit	2674	-0.53838
K21906	gadW; AraC family transcriptional regulator, glutamate-dependent acid resistance regulator	72	-0.53841
K06401	spoIVFA; stage IV sporulation protein FA	893	-0.53919
K18111	gcoA; (+)-beta-caryophyllene/(+)-caryolan-1-ol synthase [EC:4.2.3.89 4.2.1.138]	85	-0.53974
K02724	psbZ; photosystem II PsbZ protein	331	-0.54041
K07476	yusF; toprim domain protein	1097	-0.54057
K14606	cruP; lycopene cyclase CruP [EC:5.5.1.19]	277	-0.54134
K02086	dnaD; DNA replication protein	2489	-0.54147
K21489	yokI; toxin YokI [EC:3.1.-.-]	70	-0.54198
K02717	psbP; photosystem II oxygen-evolving enhancer protein 2	323	-0.54269
K19828	MTG1; mitochondrial GTPase 1	374	-0.54334
K18765	csrD; RNase E specificity factor CsrD	878	-0.54427
K06431	sspN; small acid-soluble spore protein N (minor)	482	-0.54505
K06426	sspI; small acid-soluble spore protein I (minor)	1074	-0.54543
K02630	pecC; phycoerythrocyanin-associated rod linker protein	264	-0.54601
K18958	whiB7; WhiB family transcriptional regulator, redox-sensing transcriptional regulator	1941	-0.54724
K02383	flbB; flagellar protein FlbB	196	-0.54761
K06328	cotE; spore coat protein E	1124	-0.54809
K13591	popA; two-component system, cell cycle response regulator PopA	123	-0.54841
K06375	spo0B; stage 0 sporulation protein B (sporulation initiation phosphotransferase) [EC:2.7.-.-]	827	-0.54912
K10156	eizS; epi-isozizaene synthase [EC:4.2.3.37]	277	-0.54961
K02250	comK; competence protein ComK	974	-0.54991
K05803	nlpI; lipoprotein NlpI	1402	-0.55053
K00051	E1.1.1.82; malate dehydrogenase (NADP+) [EC:1.1.1.82]	54	-0.55082
K21465	pbpA; penicillin-binding protein A	1937	-0.55152
K11522	pixG; two-component system, chemotaxis family, response regulator PixG	208	-0.55165
K16565	exoX; exopolysaccharide production repressor protein	228	-0.55177
K02096	apcE; phycobilisome core-membrane linker protein	270	-0.55213
K03591	ftsN; cell division protein FtsN	867	-0.55250

Table A3 continued from previous page

KEGG Orthology ID	Definition	Family Size	ln(CI)/ln(family size)
K06948	yqeH; 30S ribosome assembly GTPase	2572	-0.55318
K01802&K03769	E5.2.1.8; peptidylprolyl isomerase [EC:5.2.1.8] & ppiC; peptidyl-prolyl cis-trans isomerase C [EC:5.2.1.8]	64	-0.55365

Table A4. Most homoplasic Pfam annotations. Phylogenetic patchiness was computed for each Pfam present in at least 50 genomes using the CI, a common homoplasy metric representing the inverse of the minimum possible number of state changes (trait gain or loss) given the tree topology. The final phylogenetic patchiness score is equal to ln(CI)/ln(family size) where family size is the total number of genomes containing the trait. This table contains the 200 most homoplasic Pfam annotations.

Pfam ID	Description	Family Size	ln(CI)/ln(family size)
PF03175.8	DNA polymerase type B, organellar and viral	71	-0.99667
PF03420.8	Prohead core protein serine protease	50	-0.99484
PF07673.9	Domain of unknown function (DUF1602)	235	-0.99204
PF15140.1	Domain of unknown function (DUF4573)	50	-0.98956
PF14700.1	DNA-directed RNA polymerase N-terminal	105	-0.98952
PF00940.14	DNA-dependent RNA polymerase	135	-0.98915
PF12532.3	Protein of unknown function (DUF3732)	653	-0.98797
PF04693.7	Archaeal putative transposase ISC1217	89	-0.98712
PF14390.1	Putative PD-(D/E)XK family member, (DUF4420)	1218	-0.98630
PF14130.1	Domain of unknown function (DUF4297)	548	-0.98610
PF10712.4	NAD-specific glutamate dehydrogenase	195	-0.98587
PF14529.1	Endonuclease-reverse transcriptase	169	-0.98564
PF13910.1	Domain of unknown function (DUF4209)	307	-0.98517
PF09566.5	SacI restriction endonuclease	91	-0.98488
PF15532.1	Bacterial toxin 30	78	-0.98479
PF10053.4	Uncharacterized conserved protein (DUF2290)	194	-0.98472
PF14462.1	Prokaryotic E2 family E	343	-0.98432
PF15650.1	Restriction endonuclease fold toxin 9	74	-0.98375
PF08878.6	Domain of unknown function (DUF1837)	1118	-0.98354
PF04555.8	Restriction endonuclease XhoI	288	-0.98329
PF10546.4	P63C domain	393	-0.98298
PF01498.13	Transposase	105	-0.98297
PF14461.1	Prokaryotic E2 family B	569	-0.98274
PF10463.4	Peptidase U49	194	-0.98259
PF15640.1	Metallopeptidase toxin 4	70	-0.98256
PF15636.1	GHH signature containing HNH/Endo VII superfamily nuclease toxin	144	-0.98249
PF10800.3	Protein of unknown function (DUF2528)	81	-0.98249
PF11429.3	Colicin D	163	-0.98237
PF11753.3	Protein of unknwon function (DUF3310)	638	-0.98201
PF05887.6	Procyclic acidic repetitive protein (PARP)	79	-0.98192
PF09570.5	SinI restriction endonuclease	68	-0.98190
PF05367.6	Phage endonuclease I	99	-0.98166
PF14457.1	Prokaryotic E2 family A	373	-0.98135
PF09217.5	Restriction endonuclease EcoRII, N-terminal	252	-0.98110
PF14082.1	Domain of unknown function (DUF4263)	1154	-0.98110
PF11523.3	Protein of unknown function (DUF3223)	208	-0.98106
PF10593.4	Z1 domain	1652	-0.98043
PF13009.1	Putative phage integrase	244	-0.98034
PF10899.3	Putative abortive phage resistance protein AbiGi, antitoxin	332	-0.98023
PF12703.2	Toxin of toxin-antitoxin type 1 system	63	-0.98004
PF14022.1	Protein of unknown function (DUF4238)	1765	-0.97975
PF08747.6	Domain of unknown function (DUF1788)	1194	-0.97969
PF06504.6	Replication protein C (RepC)	145	-0.97960
PF09545.5	AccI restriction endonuclease	71	-0.97929
PF10592.4	AIPR protein	2348	-0.97927
PF14355.1	Abortive infection C-terminus	1366	-0.97890
PF13337.1	Putative ATP-dependent Lon protease	1400	-0.97884
PF08849.6	Putative inner membrane protein (DUF1819)	1122	-0.97883
PF08357.6	SEFIR domain	598	-0.97874
PF12083.3	Domain of unknown function (DUF3560)	458	-0.97871
PF13182.1	Protein of unknown function (DUF4007)	958	-0.97858
PF02305.12	Capsid protein (F protein)	171	-0.97837
PF09509.5	Protein of unknown function (Hypoth_ymh)	1095	-0.97819
PF12358.3	Protein of unknown function (DUF3644)	805	-0.97782
PF11985.3	Protein of unknown function (DUF3486)	830	-0.97781
PF10088.4	Uncharacterised protein conserved in bacteria (DUF2326)	874	-0.97761
PF10065.4	Uncharacterized conserved protein (DUF2303)	578	-0.97752
PF10549.4	ORF11CD3 domain	181	-0.97748
PF15611.1	EH_Signature domain	550	-0.97743
PF13166.1	AAA domain	2468	-0.97699
PF15524.1	Novel toxin 17	56	-0.97677
PF04218.8	CENP-B N-terminal DNA-binding domain	161	-0.97667
PF14437.1	MafB19-like deaminase	153	-0.97659
PF09039.6	Mu DNA binding, I gamma subdomain	424	-0.97659

Table A4 continued from previous page

Pfam ID	Description	Family Size	ln(CI)/ln(family size)
PF02923.10	Restriction endonuclease BamHI	90	-0.97659
PF13856.1	ATP-binding sugar transporter from pro-phage	296	-0.97653
PF01815.11	Rop protein	81	-0.97634
PF05144.9	Phage replication protein CRI	286	-0.97621
PF00910.17	RNA helicase	256	-0.97592
PF10888.3	Protein of unknown function (DUF2742)	127	-0.97589
PF11133.3	Head fiber protein	200	-0.97587
PF05183.7	RNA dependent RNA polymerase	185	-0.97575
PF09019.6	EcoRII C terminal	543	-0.97572
PF13752.1	Domain of unknown function (DUF4165)	54	-0.97564
PF14427.1	Pput_2613-like deaminase	54	-0.97564
PF14452.1	Multiubiquitin	777	-0.97548
PF14367.1	Domain of unknown function (DUF4411)	1050	-0.97542
PF15653.1	URI fold toxin 2	125	-0.97540
PF02407.11	Putative viral replication protein	102	-0.97533
PF10123.4	Mu-like prophage I protein	931	-0.97523
PF10547.4	P22_AR N-terminal domain	480	-0.97486
PF00165.18	Bacterial regulatory helix-turn-helix proteins, AraC family	975	-0.97484
PF09517.5	Eco29kI restriction endonuclease	349	-0.97473
PF09355.5	Phage protein Gp19/Gp15/Gp42	330	-0.97472
PF11134.3	Phage stabilisation protein	115	-0.97472
PF07471.7	Phage DNA packaging protein Nu1	423	-0.97470
PF15639.1	Metallopeptidase toxin 3	100	-0.97470
PF09077.6	Mu B transposition protein, C terminal	232	-0.97458
PF04687.7	Microvirus H protein (pilot protein)	136	-0.97452
PF02914.10	Bacteriophage Mu transposase	333	-0.97440
PF15545.1	Bacterial toxin 8	60	-0.97427
PF05371.7	Phage major coat protein, Gp8	68	-0.97425
PF11363.3	Protein of unknown function (DUF3164)	909	-0.97411
PF12477.3	Sex factor F TraW protein N terminal	215	-0.97402
PF05155.10	Phage X family	260	-0.97400
PF09956.4	Uncharacterized conserved protein (DUF2190)	774	-0.97398
PF04726.8	Microvirus J protein	51	-0.97376
PF08483.6	IstB-like ATP binding N-terminal	1094	-0.97368
PF05621.6	Bacterial TniB protein	1028	-0.97343
PF07460.6	NUMOD3 motif (2 copies)	298	-0.97333
PF15605.1	Bacterial toxin 28	117	-0.97324
PF15526.1	Novel toxin 21	171	-0.97322
PF13250.1	Domain of unknown function (DUF4041)	877	-0.97293
PF14311.1	Probable Zinc-ribbon domain	628	-0.97278
PF12183.3	Restriction endonuclease NotI	340	-0.97271
PF09571.5	XcyI restriction endonuclease	65	-0.97270
PF13876.1	Phage protein (N4 Gp49/phage Sf6 gene 66) family	328	-0.97269
PF12476.3	Protein of unknown function (DUF3696)	1091	-0.97264
PF13565.1	Homeodomain-like domain	1096	-0.97248
PF14253.1	Bacteriophage abortive infection AbiH	1043	-0.97241
PF06613.6	KorB C-terminal beta-barrel domain	107	-0.97228
PF08822.6	Protein of unknown function (DUF1804)	283	-0.97228
PF13020.1	Domain of unknown function (DUF3883)	2501	-0.97219
PF10124.4	Mu-like prophage major head subunit gpT	1047	-0.97204
PF14335.1	Domain of unknown function (DUF4391)	1015	-0.97197
PF11651.3	P22 coat protein - gene protein 5	786	-0.97196
PF13512.1	Tetratricopeptide repeat	99	-0.97188
PF14487.1	Domain of unknown function (DUF4433)	1341	-0.97186
PF05063.9	MT-A70	1226	-0.97177
PF15649.1	Restriction endonuclease fold toxin 7	189	-0.97177
PF09563.5	LlaJI restriction endonuclease	464	-0.97170
PF10926.3	Protein of unknown function (DUF2800)	886	-0.97166
PF09436.5	Domain of unknown function (DUF2016)	176	-0.97165
PF01870.13	Archaeal holliday junction resolvase (hjc)	63	-0.97157
PF00261.15	Tropomyosin	70	-0.97143
PF08721.6	TnsA endonuclease C terminal	829	-0.97139
PF07057.6	DNA helicase TraI	124	-0.97134
PF06252.7	Protein of unknown function (DUF1018)	1172	-0.97132
PF06763.6	Prophage minor tail protein Z (GPZ)	403	-0.97117
PF14281.1	PD-(D/E)XK nuclease superfamily	1796	-0.97098
PF14594.1	Siphovirus ReqiPepy6 Gp37-like protein	554	-0.97091
PF06152.6	Phage minor capsid protein 2	817	-0.97086
PF05772.7	NinB protein	526	-0.97078
PF02924.9	Bacteriophage lambda head decoration protein D	1148	-0.97072
PF03090.12	Replicase family	387	-0.97070
PF06528.7	Phage P2 GpE	315	-0.97061
PF11679.3	Protein of unknown function (DUF3275)	315	-0.97061
PF02831.10	gpW	336	-0.97049
PF03864.10	Phage major capsid protein E	2045	-0.97049
PF02171.12	Piwi domain	362	-0.97036

Table A4 continued from previous page

Pfam ID	Description	Family Size	ln(CI)/ln(family size)
PF09195.6	Restriction endonuclease BglII	648	-0.97011
PF14338.1	Mrr N-terminal domain	2534	-0.97006
PF10991.3	Protein of unknown function (DUF2815)	874	-0.96994
PF14236.1	Domain of unknown function (DUF4338)	452	-0.96989
PF12684.2	PDDEXK-like domain of unknown function (DUF3799)	713	-0.96962
PF10137.4	Predicted nucleotide-binding protein containing TIR-like domain	1406	-0.96961
PF07030.7	Protein of unknown function (DUF1320)	1320	-0.96960
PF05261.6	TraM protein, DNA-binding	93	-0.96952
PF04465.7	Protein of unknown function (DUF499)	1073	-0.96945
PF10711.4	Hypothetical protein (DUF2513)	689	-0.96937
PF06634.7	Protein of unknown function (DUF1156)	1125	-0.96934
PF04404.7	ERF superfamily	1244	-0.96922
PF06074.7	Protein of unknown function (DUF935)	1670	-0.96922
PF11114.3	Minor capsid protein	293	-0.96922
PF01446.12	Replication protein	459	-0.96918
PF06854.6	Bacteriophage Gp15 protein	499	-0.96916
PF03837.9	RecT family	2355	-0.96905
PF05701.6	Weak chloroplast movement under blue light	52	-0.96897
PF08707.6	Primase C terminal 2 (PriCT-2)	1200	-0.96897
PF11867.3	Domain of unknown function (DUF3387)	3840	-0.96887
PF02303.12	Helix-destabilising protein	134	-0.96878
PF07278.6	Protein of unknown function (DUF1441)	263	-0.96877
PF05373.6	L-proline 3-hydroxylase, C-terminal	78	-0.96851
PF13479.1	AAA domain	1375	-0.96849
PF12721.2	RIP homotypic interaction motif	58	-0.96832
PF08937.6	MTH538 TIR-like domain (DUF1863)	2040	-0.96820
PF10805.3	Protein of unknown function (DUF2730)	408	-0.96820
PF09373.5	Pseudomurein-binding repeat	51	-0.96817
PF13643.1	Domain of unknown function (DUF4145)	3586	-0.96810
PF03071.10	GNT-I family	102	-0.96808
PF11426.3	Tn7 transposition regulator TnsC	166	-0.96807
PF07102.7	Protein of unknown function (DUF1364)	441	-0.96803
PF03288.11	Poxvirus D5 protein-like	1495	-0.96786
PF11459.3	Transcriptional regulator, AbiEi antitoxin, Type IV TA system	1125	-0.96776
PF12635.2	Protein of unknown function (DUF3780)	203	-0.96771
PF09194.5	Restriction endonuclease BsobI	107	-0.96768
PF14301.1	Domain of unknown function (DUF4376)	609	-0.96768
PF14267.1	Domain of unknown function (DUF4357)	1809	-0.96763
PF15590.1	Immunity protein 27	70	-0.96761
PF09565.5	NgoFVII restriction endonuclease	392	-0.96758
PF12324.3	Helix-turn-helix domain of alkylmercury lyase	255	-0.96752
PF09556.5	HaeIII restriction endonuclease	202	-0.96751
PF12738.2	twin BRCT domain	76	-0.96742
PF15535.1	Bacterial toxin 37	76	-0.96742
PF15570.1	Immunity protein 43	82	-0.96731
PF04927.7	Seed maturation protein	63	-0.96722
PF15633.1	HYD1 signature containing ADP-ribosyltransferase	157	-0.96719
PF07693.9	KAP family P-loop domain	3679	-0.96714
PF13175.1	AAA ATPase domain	3832	-0.96706
PF05565.6	Siphovirus Gp157	825	-0.96700
PF06616.6	BsuBI/PstI restriction endonuclease C-terminus	770	-0.96691
PF07352.7	Bacteriophage Mu Gam like protein	737	-0.96682
PF08706.6	D5 N terminal like	2874	-0.96678
PF11284.3	Protein of unknown function (DUF3085)	423	-0.96677
PF14294.1	Domain of unknown function (DUF4372)	1262	-0.96657
PF12959.2	Protein of unknown function (DUF3848)	149	-0.96649
PF14411.1	A nuclease of the HNH/ENDO VII superfamily with conserved LHH	488	-0.96625
PF06356.6	Protein of unknown function (DUF1064)	712	-0.96587
PF05869.6	DNA N-6-adenine-methyltransferase (Dam)	803	-0.96584
PF10544.4	T5orf172 domain	3712	-0.96583

Table A5. Least homoplasic Pfam annotations. Phylogenetic patchiness was computed for each Pfam present in at least 50 genomes using the CI, a common homoplasy metric representing the inverse of the minimum possible number of state changes (trait gain or loss) given the tree topology. The final phylogenetic patchiness score is equal to ln(CI)/ln(family size) where family size is the total number of genomes containing the trait. This table contains the 200 least homoplasic Pfam annotations.

Pfam ID	Description	Family Size	ln(CI)/ln(family size)
PF10811.3	Protein of unknown function (DUF2532)	53	0.00000
PF12334.3	Rickettsia outer membrane protein B	53	0.00000
PF10878.3	Protein of unknown function (DUF2672)	58	0.00000
PF10859.3	Protein of unknown function (DUF2660)	60	0.00000
PF13221.1	Protein of unknown function (DUF4029)	102	0.00000
PF13043.1	Domain of unknown function (DUF3903)	118	0.00000
PF13065.1	Protein of unknown function (DUF3928)	118	0.00000
PF13074.1	Protein of unknown function (DUF3938)	118	0.00000
PF13050.1	Protein of unknown function (DUF3911)	119	0.00000
PF13063.1	Protein of unknown function (DUF3925)	119	0.00000
PF13110.1	Protein of unknown function (DUF3966)	119	0.00000
PF13142.1	Domain of unknown function (DUF3960)	119	0.00000
PF13294.1	Domain of unknown function (DUF4075)	119	0.00000
PF13141.1	Protein of unknown function (DUF3979)	126	0.00000
PF10954.3	Protein of unknown function (DUF2755)	236	0.00000
PF00005.22	ABC transporter	23934	-0.06874
PF08866.5	Putative amino acid metabolism	874	-0.10234
PF08930.5	Domain of unknown function (DUF1912)	286	-0.12255
PF13983.1	YsaB-like lipoprotein	232	-0.12726
PF13069.1	Protein of unknown function (DUF3933)	121	-0.14453
PF13049.1	Protein of unknown function (DUF3910)	120	-0.14478
PF13134.1	Protein of unknown function (DUF3948)	120	-0.14478
PF13052.1	Protein of unknown function (DUF3913)	118	-0.14529
PF13054.1	Protein of unknown function (DUF3915)	118	-0.14529
PF13066.1	Protein of unknown function (DUF3929)	118	-0.14529
PF13077.1	Protein of unknown function (DUF3909)	118	-0.14529
PF13105.1	Protein of unknown function (DUF3959)	118	-0.14529
PF13033.1	Protein of unknown function (DUF3894)	117	-0.14555
PF13051.1	Protein of unknown function (DUF3912)	117	-0.14555
PF13210.1	Domain of unknown function (DUF4018)	116	-0.14582
PF13153.1	Protein of unknown function (DUF3985)	105	-0.14894
PF00009.22	Elongation factor Tu GTP binding domain	23931	-0.15962
PF09281.5	Taq polymerase, exonuclease	70	-0.16315
PF11609.3	Protein of unknown function (DUF3248)	70	-0.16315
PF11482.3	Protein of unknown function (DUF3208)	68	-0.16427
PF12723.2	Protein of unknown function (DUF3809)	68	-0.16427
PF10875.3	Protein of unknown function (DUF2670)	60	-0.16929
PF12574.3	120 KDa Rickettsia surface antigen	59	-0.16999
PF15437.1	Plasminogen-binding protein pgbA C-terminal	55	-0.17297
PF00271.26	Helicase conserved C-terminal domain	23925	-0.19300
PF15513.1	Domain of unknown function (DUF4651)	259	-0.19770
PF10953.3	Protein of unknown function (DUF2754)	234	-0.20138
PF11448.3	Protein of unknown function (DUF3005)	223	-0.20318
PF03144.20	Elongation factor Tu domain 2	23927	-0.21792
PF13055.1	Protein of unknown function (DUF3917)	119	-0.22988
PF13082.1	Protein of unknown function (DUF3931)	119	-0.22988
PF13067.1	Protein of unknown function (DUF3930)	117	-0.23070
PF13068.1	Protein of unknown function (DUF3932)	117	-0.23070
PF13120.1	Domain of unknown function (DUF3974)	117	-0.23070
PF13140.1	Domain of unknown function (DUF3980)	107	-0.23511
PF10827.3	Protein of unknown function (DUF2552)	267	-0.24812
PF11388.3	Phagosome trafficking protein DotA	73	-0.25606
PF11497.3	NADH-quinone oxidoreductase chain 15	67	-0.26128
PF11065.3	Protein of unknown function (DUF2866)	199	-0.26190
PF09390.5	Protein of unknown function (DUF1999)	64	-0.26416
PF14507.1	CppA C-terminal	381	-0.27082
PF02691.10	Vacuolating cyotoxin	56	-0.27292
PF03677.8	Uncharacterised protein family (UPF0137)	50	-0.28083
PF11674.3	Protein of unknown function (DUF3270)	284	-0.28491
PF13268.1	Protein of unknown function (DUF4059)	283	-0.28509
PF15436.1	Plasminogen-binding protein pgbA N-terminal	269	-0.28767
PF13112.1	Protein of unknown function (DUF3965)	122	-0.28857
PF13071.1	Protein of unknown function (DUF3935)	121	-0.28906
PF13219.1	Protein of unknown function (DUF4027)	118	-0.29059

Table A5 continued from previous page

Pfam ID	Description	Family Size	ln(CI)/ln(family size)
PF11826.3	Protein of unknown function (DUF3346)	249	-0.29170
PF13135.1	Protein of unknown function (DUF3947)	115	-0.29216
PF13080.1	Protein of unknown function (DUF3926)	114	-0.29270
PF13059.1	Protein of unknown function (DUF3992)	111	-0.29436
PF10818.3	Protein of unknown function (DUF2547)	104	-0.29849
PF10398.4	Protein of unknown function (DUF2443)	96	-0.30372
PF08181.6	DegQ (SacQ) family	87	-0.31042
PF10216.4	CO2 hydration protein (ChpXY)	298	-0.31450
PF11364.3	Protein of unknown function (DUF3165)	273	-0.31942
PF12163.3	DNA replication regulator	270	-0.32005
PF11486.3	Protein of unknown function (DUF3212)	70	-0.32630
PF13058.1	Protein of unknown function (DUF3920)	134	-0.32860
PF08264.8	Anticodon-binding domain of tRNA	23908	-0.33051
PF10877.3	Protein of unknown function (DUF2671)	66	-0.33089
PF07176.6	Alpha/beta hydrolase of unknown function (DUF1400)	358	-0.33091
PF12046.3	Cofactor assembly of complex C subunit B	357	-0.33106
PF13397.1	RNA polymerase-binding protein	2871	-0.33144
PF13060.1	Protein of unknown function (DUF3921)	125	-0.33333
PF10808.3	Protein of unknown function (DUF2542)	63	-0.33460
PF13062.1	Protein of unknown function (DUF3924)	121	-0.33559
PF11076.3	Putative inner membrane protein YbhQ	327	-0.33608
PF13121.1	Domain of unknown function (DUF3976)	119	-0.33676
PF00004.24	ATPase family associated with various cellular activities (AAA)	23905	-0.33736
PF01479.20	S4 domain	23905	-0.33736
PF00133.17	tRNA synthetases class I (I, L, M and V)	23904	-0.33736
PF00575.18	S1 RNA binding domain	23887	-0.33738
PF01926.18	50S ribosome-binding GTPase	23901	-0.34062
PF10913.3	Protein of unknown function (DUF2706)	58	-0.34141
PF10879.3	Protein of unknown function (DUF2674)	54	-0.34753
PF06558.7	Secretion monitor precursor protein (SecM)	553	-0.34792
PF09366.5	Protein of unknown function (DUF1997)	394	-0.34795
PF09456.5	RcsC Alpha-Beta-Loop (ABL)	548	-0.34842
PF13053.1	Protein of unknown function (DUF3914)	100	-0.34949
PF02518.21	Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase	23902	-0.34978
PF14250.1	AbrB-like transcriptional regulator	367	-0.35213
PF07614.6	Protein of unknown function (DUF1577)	51	-0.35258
PF03129.15	Anticodon binding domain	23900	-0.35266
PF14217.1	Domain of unknown function (DUF4327)	236	-0.35614
PF14147.1	Sporulation protein YhaL	609	-0.35912
PF12071.3	Protein of unknown function (DUF3551)	146	-0.35953
PF10762.4	Protein of unknown function (DUF2583)	559	-0.36398
PF07305.7	Protein of unknown function (DUF1454)	558	-0.36408
PF10766.4	Multidrug efflux pump-associated protein AcrZ	552	-0.36471
PF11269.3	Protein of unknown function (DUF3069)	297	-0.36522
PF14495.1	Cytochrome c-550 domain	297	-0.36522
PF01383.16	CpcD/allophycocyanin linker domain	288	-0.36720
PF14185.1	Antitoxin SpoIISB, type II toxin-antitoxin system	80	-0.36728
PF14182.1	YgaB-like protein	395	-0.36750
PF12087.3	Protein of unknown function (DUF3564)	192	-0.37012
PF09475.5	Dot/Icm secretion system protein (dot_icm_IcmQ)	126	-0.37048
PF03077.9	Putative vacuolating cytotoxin	77	-0.37051
PF07288.6	Protein of unknown function (DUF1447)	1774	-0.37062
PF00270.24	DEAD/DEAH box helicase	23820	-0.37320
PF14159.1	CAAD domains of cyanobacterial aminoacyl-tRNA synthetase	352	-0.37472
PF13980.1	Uncharacterised protein family (UPF0370)	592	-0.37564
PF13123.1	Protein of unknown function (DUF3978)	113	-0.37902
PF13987.1	YedD-like protein	328	-0.37929
PF09575.5	Small spore protein J (Spore_SspJ)	69	-0.38011
PF07423.6	Protein of unknown function (DUF1510)	847	-0.38046
PF14148.1	YhdB-like protein	544	-0.38068
PF06619.6	Protein of unknown function (DUF1149)	543	-0.38079
PF13078.1	Protein of unknown function (DUF3942)	110	-0.38119
PF13139.1	Domain of unknown function (DUF3981)	110	-0.38119
PF05211.7	Neuraminyllactose-binding hemagglutinin precursor (NLBH)	68	-0.38143
PF13126.1	Protein of unknown function (DUF3975)	109	-0.38193
PF14162.1	YozD-like protein	414	-0.38212
PF11184.3	Protein of unknown function (DUF2969)	806	-0.38328
PF12227.3	Protein of unknown function (DUF3603)	647	-0.38393
PF10730.4	Protein of unknown function (DUF2521)	513	-0.38426
PF13984.1	MsyB protein	390	-0.38594
PF07338.8	Protein of unknown function (DUF1471)	601	-0.38835
PF11226.3	Protein of unknown function (DUF3022)	209	-0.38924
PF00421.14	Photosystem II protein	365	-0.39028
PF14011.1	EspG family	709	-0.39077
PF11240.3	Protein of unknown function (DUF3042)	703	-0.39128
PF11375.3	Protein of unknown function (DUF3177)	356	-0.39193

Table A5 continued from previous page

Pfam ID	Description	Family Size	ln(CI)/ln(family size)
PF10939.3	Protein of unknown function (DUF2631)	999	-0.39209
PF13834.1	Domain of unknown function (DUF4193)	2891	-0.39344
PF12452.3	Protein of unknown function (DUF3685)	348	-0.39346
PF10757.4	Biofilm formation regulator YbaJ	542	-0.39473
PF04220.7	Der GTPase activator (YihI)	1287	-0.39570
PF06643.6	Protein of unknown function (DUF1158)	328	-0.39748
PF00587.20	tRNA synthetase class II core domain (G, H, P, S and T)	23879	-0.39931
PF13220.1	Protein of unknown function (DUF4028)	86	-0.40225
PF02341.10	RbcX protein	234	-0.40277
PF00905.17	Penicillin binding protein transpeptidase domain	23433	-0.40355
PF13829.1	Domain of unknown function (DUF4191)	2865	-0.40436
PF07639.6	YTV	83	-0.40548
PF13999.1	MarB protein	225	-0.40568
PF00679.19	Elongation factor G C-terminus	23873	-0.40616
PF01336.20	OB-fold nucleic acid binding domain	23873	-0.40616
PF02531.11	PsaD	354	-0.40855
PF02605.10	Photosystem I reaction centre subunit XI	354	-0.40855
PF11210.3	Protein of unknown function (DUF2996)	352	-0.40894
PF10799.3	Biofilm formation protein (YliH/bssR)	215	-0.40912
PF03912.9	Psb28 protein	351	-0.40914
PF08848.6	Domain of unknown function (DUF1818)	351	-0.40914
PF06799.6	Protein of unknown function (DUF1230)	349	-0.40954
PF11152.3	Cofactor assembly of complex C subunit B, CCB2/CCB4	348	-0.40974
PF14233.1	Domain of unknown function (DUF4335)	347	-0.40994
PF12502.3	Protein of unknown function (DUF3710)	2802	-0.41519
PF10949.3	Protein of unknown function (DUF2777)	481	-0.41532
PF06569.6	Protein of unknown function (DUF1128)	1175	-0.41653
PF07307.6	Heptaprenyl diphosphate synthase (HEPPP synthase) subunit 1	1169	-0.41683
PF08796.5	Protein of unknown function (DUF1797)	1155	-0.41754
PF04686.7	Streptomyces sporulation and cell division protein, SsgA	1125	-0.41911
PF02467.11	Transcription factor WhiB	3059	-0.41956
PF12021.3	Protein of unknown function (DUF3509)	539	-0.41958
PF08838.5	Protein of unknown function (DUF1811)	981	-0.41959
PF14183.1	YwpF-like protein	724	-0.42106
PF09654.5	Protein of unknown function (DUF2396)	234	-0.42208
PF01716.13	Manganese-stabilising protein / photosystem II polypeptide	352	-0.42378
PF11332.3	Protein of unknown function (DUF3134)	351	-0.42399
PF11016.3	Protein of unknown function (DUF2854)	350	-0.42420
PF12065.3	Protein of unknown function (DUF3545)	590	-0.42445
PF10969.3	Protein of unknown function (DUF2771)	905	-0.42456
PF08180.6	B melanoma antigen family	226	-0.42479
PF11012.3	Protein of unknown function (DUF2850)	226	-0.42479
PF12484.3	Polymorphic PE/PPE proteins C terminal	176	-0.42496
PF11341.3	Protein of unknown function (DUF3143)	346	-0.42503
PF13942.1	YfhG lipoprotein	565	-0.42735
PF11241.3	Protein of unknown function (DUF3043)	2858	-0.42740
PF14165.1	YtzH-like protein	564	-0.42747
PF12084.3	Protein of unknown function (DUF3561)	478	-0.42775
PF01856.12	Helicobacter outer membrane protein	94	-0.42830
PF14017.1	Protein of unknown function (DUF4233)	2586	-0.42853
PF13317.1	Protein of unknown function (DUF4088)	269	-0.42860
PF13106.1	Domain of unknown function (DUF3961)	127	-0.42927
PF08741.5	YwhD family	1200	-0.42941
PF10896.3	Protein of unknown function (DUF2714)	92	-0.43034
PF14506.1	CppA N-terminal	320	-0.43079
PF07865.6	Protein of unknown function (DUF1652)	456	-0.43104
PF10625.4	Universal stress protein B (UspB)	800	-0.43239
PF10788.4	Protein of unknown function (DUF2603)	203	-0.43337
PF10801.3	Protein of unknown function (DUF2537)	678	-0.43460
PF03040.9	CemA family	298	-0.43617

Table A6. Taxa significantly enriched with homoplasic KO annotations. A hypergeometric test was performed to identify taxa that are significantly enriched with the 5% most homoplasic KOs. The total KO set is the set of KOs present in the top 5% of homoplasic terms. The total taxon counts is the number of different KOs contained in any genome within the rank. The total counts are the number of KOs that are present in at least one genome in any rank. The number of successes in the sample are the number of different KO terms within the homoplasic set that is contained in at least one of the genomes within the rank. The fold-enrichment is calculated as the successes in the sample as a fraction of all the possible successes versus the total number of KOs represented by the rank out of all KOs represented by any rank. Raw *P* values and Benjamini-Hochberg/FDR-adjusted *P* values are also given. Only taxa with Benjamini-Hochberg/FDR-adjusted $P<0.05$ are listed here.

Taxonomic Level	Taxonomic Rank	Total KO Set Counts	Total Taxon Counts	Total Counts	Success in Sample	Fold-enrichment	*P*	BH *P*
genus	*Pseudomonas_E*	618	6289	12351	409	1.2997	2.70E-15	1.54E-11
genus	*Streptomyces*	618	6720	12351	426	1.2669	2.97E-14	1.69E-10
genus	*Mycobacterium*	618	5389	12351	335	1.2424	3.91E-08	2.23E-04
genus	*Nocardia*	618	5238	12351	317	1.2095	3.13E-06	1.78E-02
family	Burkholderiaceae	618	8079	12351	526	1.3012	1.44E-29	2.39E-26
family	Pseudomonadaceae	618	6698	12351	435	1.2979	2.73E-17	4.53E-14
family	Corynebacteriaceae	618	6613	12351	422	1.2753	1.46E-14	2.42E-11
family	Streptomycetaceae	618	6825	12351	427	1.2504	3.84E-13	6.36E-10
family	Pseudonocardiaceae	618	6551	12351	408	1.2447	1.37E-11	2.26E-08
family	Sphingomonadaceae	618	6154	12351	387	1.2568	3.58E-11	5.94E-08
family	Streptosporangiaceae	618	6023	12351	354	1.1746	8.23E-06	1.37E-02
order	Betaproteobacteriales	618	8545	12351	545	1.2747	4.13E-30	3.04E-27
order	Rhizobiales	618	7789	12351	495	1.2701	3.90E-21	2.87E-18
order	Pseudomonadales	618	8382	12351	518	1.2351	2.18E-20	1.60E-17
order	Corynebacteriales	618	7574	12351	476	1.2560	1.29E-17	9.50E-15
order	Streptomycetales	618	6865	12351	428	1.2460	6.71E-13	4.93E-10
order	Sphingomonadales	618	6305	12351	389	1.2330	6.61E-10	4.86E-07
order	Methylococcales	618	4939	12351	309	1.2504	1.51E-07	1.11E-04
order	Streptosporangiales	618	6023	12351	354	1.1746	8.23E-06	6.05E-03
class	Alphaproteobacteria	618	8892	12351	559	1.2564	7.08E-31	1.91E-28
class	Gammaproteobacteria	618	10069	12351	589	1.1691	2.52E-25	6.81E-23
class	Actinobacteria	618	8480	12351	521	1.2279	5.60E-20	1.51E-17
class	Bacilli_A	618	6480	12351	374	1.1535	2.17E-05	5.85E-03
phylum	Proteobacteria	618	10437	12351	597	1.1432	2.52E-23	2.79E-21
phylum	Actinobacteria	618	8967	12351	541	1.2058	3.06E-20	3.40E-18
phylum	Firmicutes	618	7707	12351	450	1.1669	1.36E-08	1.51E-06
phylum	Chloroflexi	618	7092	12351	407	1.1469	6.69E-06	7.43E-04
phylum	Cyanobacteria	618	6623	12351	380	1.1467	3.13E-05	3.48E-03
phylum	Planctomycetes	618	6587	12351	376	1.1408	6.79E-05	7.54E-03
phylum	Bacteroidetes	618	7898	12351	437	1.1058	1.59E-04	1.77E-02

Table A7. Taxa significantly enriched with homoplasic Pfam annotations. A hypergeometric test was performed to identify taxa that are significantly enriched with the 5% most homoplasic Pfams. The total Pfam set is the set of Pfams present in the top 5% of homoplasic domains. The total taxon counts are the number of different Pfams contained in any genome within the rank. The total counts are the number of Pfams that are present in at least one genome in any rank. The number of successes in the sample is the number of different Pfam domains within the homoplasic set that are contained in at least one of the genomes within the rank. The fold-enrichment is calculated as the successes in the sample as a fraction of all the possible successes versus the total number of Pfams represented by the rank out of all Pfams represented by any rank. Raw *P* values and Benjamini-Hochberg/FDR-adjusted *P* values are also given. Only taxa with Benjamini-Hochberg/FDR-adjusted $P<0.05$ are listed here.

Taxonomic Level	Taxonomic Rank	Total Pfam Set Counts	Total Taxon Counts	Total Counts	Success in Sample	Fold-enrichment	*P*	BH *P*
species	*Escherichia coli*	552	3259	11047	212	1.3018	2.60E-06	2.58E-02
genus	*Pseudomonas_E*	552	4895	11047	401	1.6394	1.60E-43	9.14E-40
genus	*Acinetobacter*	552	3481	11047	303	1.7420	1.79E-31	1.02E-27
genus	*Enterobacter*	552	3559	11047	299	1.6813	7.59E-28	4.32E-24
genus	*Burkholderia*	552	4024	11047	319	1.5865	1.05E-25	6.00E-22
genus	*Xenorhabdus*	552	3001	11047	258	1.7205	5.03E-24	2.87E-20
genus	*Vibrio*	552	4684	11047	345	1.4740	1.57E-22	8.93E-19
genus	*Yersinia*	552	3563	11047	284	1.5952	7.90E-22	4.50E-18
genus	*Stenotrophomonas*	552	3545	11047	274	1.5468	1.26E-18	7.15E-15
genus	*Prevotella*	552	3236	11047	255	1.5770	5.33E-18	3.04E-14
genus	*Nitrosomonas*	552	3103	11047	247	1.5930	7.87E-18	4.48E-14
genus	*Pantoea*	552	3642	11047	273	1.5001	1.50E-16	8.52E-13
genus	*Cupriavidus*	552	3847	11047	282	1.4670	6.64E-16	3.78E-12
genus	*Halomonas*	552	3793	11047	279	1.4721	7.15E-16	4.08E-12
genus	*Gilliamella*	552	2473	11047	203	1.6428	2.50E-15	1.42E-11
genus	*Providencia*	552	3172	11047	243	1.5331	2.87E-15	1.63E-11
genus	*Serratia*	552	3508	11047	260	1.4833	8.75E-15	4.99E-11
genus	*Sphingobium*	552	3494	11047	257	1.4720	4.05E-14	2.31E-10
genus	*Streptococcus*	552	2920	11047	225	1.5421	4.18E-14	2.38E-10
genus	*Caballeronia*	552	3852	11047	274	1.4235	1.90E-13	1.08E-09
genus	*Paraburkholderia*	552	3984	11047	276	1.3864	4.52E-12	2.57E-08
genus	*Citrobacter*	552	3408	11047	245	1.4387	4.70E-12	2.68E-08
genus	*Erwinia*	552	3577	11047	254	1.4211	5.26E-12	3.00E-08
genus	*Escherichia*	552	3430	11047	245	1.4295	9.88E-12	5.63E-08
genus	*Photorhabdus*	552	2751	11047	207	1.5059	1.14E-11	6.48E-08
genus	*Bacteroides*	552	3106	11047	226	1.4562	2.05E-11	1.17E-07
genus	*Rhizobium*	552	3888	11047	265	1.3640	1.50E-10	8.55E-07
genus	*Pseudomonas_A*	552	3466	11047	242	1.3973	1.89E-10	1.07E-06
genus	*Variovorax*	552	3879	11047	264	1.3620	2.02E-10	1.15E-06
genus	*Thauera*	552	3170	11047	225	1.4205	3.07E-10	1.75E-06
genus	*Achromobacter*	552	3244	11047	229	1.4127	3.18E-10	1.81E-06
genus	*Marinobacter*	552	3644	11047	250	1.3730	4.64E-10	2.65E-06
genus	*Moraxella*	552	2424	11047	182	1.5026	5.67E-10	3.23E-06
genus	*Rodentibacter*	552	2305	11047	175	1.5194	6.10E-10	3.47E-06
genus	*Sphingomonas*	552	3895	11047	261	1.3410	1.69E-09	9.61E-06
genus	*Acidovorax*	552	3323	11047	230	1.3852	2.04E-09	1.16E-05
genus	*Xanthomonas*	552	3057	11047	214	1.4010	4.49E-09	2.56E-05
genus	*Aeromonas*	552	3569	11047	241	1.3514	6.89E-09	3.93E-05
genus	*Nitrosospira*	552	2620	11047	188	1.4360	1.12E-08	6.36E-05
genus	*Psychrobacter*	552	2785	11047	197	1.4156	1.28E-08	7.27E-05
genus	*Pseudoalteromonas*	552	4015	11047	262	1.3059	2.63E-08	1.50E-04
genus	*Novosphingobium*	552	3715	11047	245	1.3198	4.70E-08	2.68E-04
genus	*Ruminococcus_B*	552	3036	11047	208	1.3711	5.74E-08	3.27E-04
genus	*Dorea*	552	2856	11047	197	1.3804	1.00E-07	5.70E-04
genus	*Blautia_A*	552	2845	11047	193	1.3576	5.31E-07	3.02E-03
genus	*Agrobacterium*	552	3314	11047	218	1.3165	6.66E-07	3.80E-03
genus	*Photobacterium*	552	3996	11047	254	1.2721	7.06E-07	4.02E-03
genus	*Paenibacillus*	552	4233	11047	266	1.2576	8.15E-07	4.64E-03
genus	*Neisseria*	552	2064	11047	147	1.4253	1.58E-06	9.01E-03
genus	*Chryseobacterium*	552	3239	11047	211	1.3037	2.52E-06	1.44E-02
genus	*Clostridium_M*	552	2864	11047	190	1.3277	3.42E-06	1.95E-02
genus	*Acetobacter*	552	2866	11047	190	1.3267	3.59E-06	2.04E-02
genus	*Hungatella*	552	3072	11047	201	1.3094	3.90E-06	2.22E-02

Table A7 continued from previous page

Taxonomic Level	Taxonomic Rank	Total Pfam Set Counts	Total Taxon Counts	Total Counts	Success in Sample	Fold-enrichment	*P*	BH *P*
genus	*Klebsiella_A*	552	3240	11047	210	1.2971	3.98E-06	2.27E-02
genus	*Faecalibacterium*	552	2137	11047	149	1.3954	4.53E-06	2.58E-02
genus	*Methylomonas*	552	3048	11047	199	1.3066	5.28E-06	3.01E-02
genus	*Flavobacterium*	552	3815	11047	239	1.2537	7.39E-06	4.21E-02
family	Enterobacteriaceae	552	5346	11047	448	1.6771	8.83E-60	1.46E-56
family	Burkholderiaceae	552	6084	11047	470	1.5460	9.25E-54	1.53E-50
family	Pseudomonadaceae	552	5191	11047	423	1.6308	3.61E-48	6.00E-45
family	Xanthomonadaceae	552	4575	11047	369	1.6141	3.80E-35	6.31E-32
family	Moraxellaceae	552	4103	11047	336	1.6389	3.96E-31	6.56E-28
family	Pasteurellaceae	552	3114	11047	278	1.7866	1.20E-29	1.99E-26
family	Vibrionaceae	552	5008	11047	375	1.4986	4.01E-28	6.66E-25
family	Sphingomonadaceae	552	4613	11047	351	1.5228	2.71E-26	4.49E-23
family	Lachnospiraceae	552	4978	11047	367	1.4754	2.13E-25	3.54E-22
family	Rhodobacteraceae	552	4884	11047	359	1.4710	4.64E-24	7.69E-21
family	Alteromonadaceae	552	4987	11047	362	1.4527	3.03E-23	5.02E-20
family	Methylomonaceae	552	4071	11047	315	1.5485	3.61E-23	5.99E-20
family	Nitrosomonadaceae	552	3402	11047	277	1.6295	1.50E-22	2.49E-19
family	Halomonadaceae	552	4372	11047	328	1.5014	3.87E-22	6.42E-19
family	Rhodocyclaceae	552	4509	11047	335	1.4869	4.04E-22	6.70E-19
family	Neisseriaceae	552	3212	11047	264	1.6449	1.11E-21	1.84E-18
family	Bacteroidaceae	552	3793	11047	296	1.5618	1.25E-21	2.08E-18
family	Ruminococcaceae	552	3909	11047	298	1.5257	4.18E-20	6.94E-17
family	Desulfovibrionaceae	552	3934	11047	294	1.4956	2.84E-18	4.70E-15
family	Clostridiaceae	552	4399	11047	317	1.4422	8.37E-18	1.39E-14
family	Acetobacteraceae	552	4146	11047	300	1.4481	1.42E-16	2.36E-13
family	DTU089	552	3760	11047	279	1.4850	1.98E-16	3.29E-13
family	Rhizobiaceae	552	4820	11047	329	1.3660	7.86E-15	1.30E-11
family	Paenibacillaceae	552	4848	11047	329	1.3581	2.05E-14	3.41E-11
family	Streptococcaceae	552	3161	11047	239	1.5131	3.10E-14	5.15E-11
family	Oscillospiraceae	552	3610	11047	263	1.4580	5.00E-14	8.30E-11
family	Aeromonadaceae	552	3920	11047	276	1.4091	5.40E-13	8.96E-10
family	Hahellaceae	552	4106	11047	279	1.3598	3.52E-11	5.83E-08
family	Flavobacteriaceae	552	4984	11047	322	1.2930	1.13E-10	1.88E-07
family	Weeksellaceae	552	3747	11047	257	1.3726	2.00E-10	3.32E-07
family	Corynebacteriaceae	552	4665	11047	304	1.3041	3.23E-10	5.36E-07
family	Selenomonadaceae	552	3070	11047	219	1.4276	4.06E-10	6.73E-07
family	Rhodanobacteraceae	552	4091	11047	273	1.3355	6.75E-10	1.12E-06
family	Planococcaceae	552	4003	11047	266	1.3298	2.44E-09	4.05E-06
family	Gallionellaceae	552	3074	11047	216	1.4062	2.50E-09	4.16E-06
family	Caulobacteraceae	552	3991	11047	265	1.3288	2.97E-09	4.92E-06
family	Beijerinckiaceae	552	4418	11047	284	1.2865	1.55E-08	2.57E-05
family	Xanthobacteraceae	552	4426	11047	284	1.2841	1.90E-08	3.15E-05
family	Homeothermaceae	552	2745	11047	194	1.4144	1.93E-08	3.20E-05
family	Stappiaceae	552	3740	11047	247	1.3217	3.37E-08	5.59E-05
family	Cellvibrionaceae	552	4351	11047	278	1.2787	5.40E-08	8.95E-05
family	Erysipelotrichaceae	552	3248	11047	219	1.3494	7.56E-08	1.25E-04
family	Nitrincolaceae	552	3984	11047	253	1.2709	8.33E-07	1.38E-03
family	Sphingobacteriaceae	552	4046	11047	255	1.2613	1.43E-06	2.37E-03
family	Gastranaerophilaceae	552	2014	11047	144	1.4309	1.69E-06	2.81E-03
family	Lactobacillaceae	552	3357	11047	216	1.2877	4.45E-06	7.38E-03
family	Desulfitobacteriaceae	552	3439	11047	219	1.2744	7.88E-06	1.31E-02
family	Chitinophagaceae	552	4205	11047	258	1.2279	1.24E-05	2.05E-02
family	Thiomicrospiraceae	552	2704	11047	178	1.3174	1.41E-05	2.33E-02
family	Chromobacteriaceae	552	3700	11047	230	1.2440	2.39E-05	3.96E-02
order	Enterobacterales	552	6718	11047	498	1.4835	6.34E-57	4.66E-54
order	Betaproteobacteriales	552	6501	11047	490	1.5084	9.67E-57	7.11E-54
order	Pseudomonadales	552	6510	11047	490	1.5063	1.73E-56	1.27E-53
order	Xanthomonadales	552	4939	11047	387	1.5681	3.67E-35	2.70E-32
order	Methylococcales	552	4435	11047	357	1.6109	7.75E-33	5.70E-30
order	Oscillospirales	552	4571	11047	355	1.5543	7.28E-29	5.35E-26
order	Sphingomonadales	552	4702	11047	356	1.5152	1.67E-26	1.22E-23
order	Lachnospirales	552	5134	11047	377	1.4696	2.24E-26	1.64E-23
order	Rhizobiales	552	5588	11047	397	1.4218	1.04E-25	7.67E-23
order	Rhodobacterales	552	4895	11047	359	1.4677	7.58E-24	5.57E-21
order	Desulfovibrionales	552	4167	11047	311	1.4936	7.23E-20	5.32E-17
order	Bacteroidales	552	5267	11047	364	1.3831	5.87E-19	4.32E-16
order	Clostridiales	552	4497	11047	324	1.4419	1.88E-18	1.38E-15
order	Acetobacterales	552	4146	11047	300	1.4481	1.42E-16	1.04E-13
order	Peptostreptococcales	552	4181	11047	296	1.4168	8.75E-15	6.43E-12
order	Tissierellales	552	3999	11047	286	1.4313	1.16E-14	8.53E-12
order	Bacillales	552	5637	11047	368	1.3065	1.84E-14	1.35E-11
order	Paenibacillales	552	4848	11047	329	1.3581	2.05E-14	1.51E-11
order	Lactobacillales	552	4450	11047	307	1.3807	6.47E-14	4.76E-11
order	Caulobacterales	552	4478	11047	308	1.3765	8.66E-14	6.37E-11
order	Campylobacterales	552	3907	11047	274	1.4035	1.24E-12	9.12E-10
order	Flavobacteriales	552	5492	11047	351	1.2790	1.19E-11	8.73E-09

Table A7 continued from previous page

Taxonomic Level	Taxonomic Rank	Total Pfam Set Counts	Total Taxon Counts	Total Counts	Success in Sample	Fold-enrichment	*P*	BH *P*
order	Actinomycetales	552	4910	11047	318	1.2961	1.33E-10	9.76E-08
order	Desulfotomaculales	552	3460	11047	240	1.3882	4.89E-10	3.60E-07
order	Chitinophagales	552	4599	11047	299	1.3011	7.89E-10	5.80E-07
order	Ectothiorhodospirales	552	3445	11047	238	1.3826	9.51E-10	6.99E-07
order	Corynebacteriales	552	5265	11047	332	1.2620	1.06E-09	7.76E-07
order	Cytophagales	552	4852	11047	309	1.2745	3.61E-09	2.66E-06
order	Erysipelotrichales	552	3497	11047	238	1.3620	4.36E-09	3.21E-06
order	Selenomonadales	552	3282	11047	224	1.3659	1.52E-08	1.12E-05
order	Chromatiales	552	4136	11047	269	1.3016	1.88E-08	1.38E-05
order	Desulfitobacteriales	552	3534	11047	237	1.3421	2.06E-08	1.52E-05
order	Nostocales	552	4561	11047	290	1.2725	3.01E-08	2.21E-05
order	Veillonellales	552	2725	11047	190	1.3954	8.76E-08	6.44E-05
order	Geobacterales	552	3492	11047	228	1.3067	5.38E-07	3.95E-04
order	Thiomicrospirales	552	3136	11047	208	1.3274	8.00E-07	5.88E-04
order	Acetivibrionales	552	3478	11047	226	1.3004	9.49E-07	6.97E-04
order	Sphingobacteriales	552	4046	11047	255	1.2613	1.43E-06	1.05E-03
order	Treponematales	552	3648	11047	230	1.2618	8.08E-06	5.94E-03
order	Desulfobulbales	552	3806	11047	236	1.2409	2.01E-05	1.48E-02
order	Rhodospirillales	552	4076	11047	249	1.2226	3.01E-05	2.21E-02
order	Nitrosococcales	552	3216	11047	203	1.2632	4.08E-05	3.00E-02
order	Coriobacteriales	552	3147	11047	199	1.2655	4.60E-05	3.38E-02
order	Desulfobacterales	552	4161	11047	252	1.2120	5.05E-05	3.71E-02
order	Gastranaerophilales	552	2350	11047	155	1.3200	6.16E-05	4.53E-02
class	Alphaproteobacteria	552	6623	11047	481	1.4534	6.16E-47	1.66E-44
class	Gammaproteobacteria	552	8168	11047	530	1.2986	1.65E-45	4.46E-43
class	Clostridia	552	6215	11047	430	1.3846	1.01E-27	2.72E-25
class	Desulfovibrionia	552	4167	11047	311	1.4936	7.23E-20	1.95E-17
class	Bacteroidia	552	6780	11047	433	1.2781	1.21E-18	3.27E-16
class	Bacilli	552	6578	11047	424	1.2900	1.34E-18	3.62E-16
class	Bacilli_A	552	5190	11047	359	1.3843	1.57E-18	4.23E-16
class	Negativicutes	552	3971	11047	294	1.4817	1.27E-17	3.44E-15
class	Actinobacteria	552	6180	11047	396	1.2824	3.47E-15	9.38E-13
class	Campylobacteria	552	3914	11047	275	1.4061	8.34E-13	2.25E-10
class	Desulfotomaculia	552	3539	11047	246	1.3911	1.83E-10	4.93E-08
class	Oxyphotobacteria	552	5025	11047	320	1.2744	1.07E-09	2.90E-07
class	Desulfuromonadia	552	4085	11047	266	1.3032	2.22E-08	6.00E-06
class	Desulfitobacteriia	552	3586	11047	239	1.3338	3.06E-08	8.26E-06
class	Spirochaetia	552	4200	11047	271	1.2913	3.62E-08	9.78E-06
class	Desulfobacteria	552	4257	11047	264	1.2411	3.25E-06	8.78E-04
class	Coriobacteriia	552	3368	11047	215	1.2775	8.63E-06	2.33E-03
class	Desulfobulbia	552	3806	11047	236	1.2409	2.01E-05	5.43E-03
class	Planctomycetia	552	4400	11047	261	1.1871	1.60E-04	4.32E-02
class	Verrucomicrobiae	552	4877	11047	285	1.1695	1.74E-04	4.71E-02
phylum	Proteobacteria	552	8550	11047	530	1.2406	5.40E-36	5.99E-34
phylum	Firmicutes_A	552	6255	11047	430	1.3758	6.35E-27	7.04E-25
phylum	Desulfobacterota	552	5432	11047	386	1.4221	3.28E-24	3.64E-22
phylum	Firmicutes_B	552	4410	11047	326	1.4794	1.07E-20	1.19E-18
phylum	Firmicutes	552	6939	11047	441	1.2719	4.49E-19	4.99E-17
phylum	Bacteroidetes	552	6989	11047	440	1.2599	7.24E-18	8.03E-16
phylum	Firmicutes_C	552	3971	11047	294	1.4817	1.27E-17	1.41E-15
phylum	Actinobacteria	552	6553	11047	415	1.2674	8.09E-16	8.98E-14
phylum	Cyanobacteria	552	5528	11047	360	1.3033	1.18E-13	1.31E-11
phylum	Verrucomicrobia	552	5438	11047	347	1.2770	2.68E-11	2.98E-09
phylum	Epsilonbacterota	552	4022	11047	275	1.3683	2.76E-11	3.06E-09
phylum	Planctomycetes	552	5273	11047	334	1.2676	4.39E-10	4.87E-08
phylum	Spirochaetes	552	4896	11047	310	1.2671	6.79E-09	7.54E-07
phylum	Desulfuromonadota	552	4085	11047	266	1.3032	2.22E-08	2.46E-06
phylum	Patescibacteria	552	4249	11047	271	1.2764	1.23E-07	1.36E-05
phylum	Chloroflexi	552	5090	11047	307	1.2071	2.54E-06	2.82E-04
phylum	Nitrospirota	552	4078	11047	242	1.1876	3.60E-04	3.99E-02

Appendix B

Lineage-specific annotations

Table B1. Lineage-specific KO annotations. KO annotations with at least one clade in the bacterial GTDB tree with a catchment and saturation of at least 95% and that were present in no more than half of the genomes in the bacterial GTDB tree were classified as lineage-specific and are listed in this table.

KEGG Orthology ID	Definition	# Genomes	Lowest Common Level	Lowest Common Rank
K00630	glycerol-3-phosphate O-acyltransferase [EC:2.3.1.15]	47	order	Chlamydiales
K00687	penicillin-binding protein 2B	860	order	Lactobacillales
K00753	glycoprotein 3-alpha-L-fucosyltransferase [EC:2.4.1.214]	2	species	*Mesorhizobium* sp1
K00796&K01633&K00950	dihydropteroate synthase [EC:2.5.1.15]; 7,8-dihydroneopterin aldolase/epimerase/oxygenase [EC:4.1.2.25 5.1.99.8 1.13.11.81]; 2-amino-4-hydroxy-6-hydroxymethyldihydropteridine diphosphokinase [EC:2.7.6.3]	2	species	*Tropheryma whipplei*
K01057&K00851	6-phosphogluconolactonase [EC:3.1.1.31] ; gluconokinase [EC:2.7.1.12]	6	species	*Bradyrhizobium* spp.
K01112	phosphohistidine phosphatase [EC:3.9.1.3]	2	species	2-01-FULL-43-22 sp1
K01378	chymosin [EC:3.4.23.4]	5	genus	*Endozoicomonas*
K01760&K01739	cystathionine beta-lyase [EC:4.4.1.8]; cystathionine gamma-synthase [EC:2.5.1.48]	108	order	Betaproteobacteriales
K01885&K09698	glutamyl-tRNA synthetase [EC:6.1.1.17] ; nondiscriminating glutamyl-tRNA synthetase [EC:6.1.1.24]	42	order	Leptospirales
K01947	biotin—[acetyl-CoA-carboxylase] ligase / type III pantothenate kinase [EC:6.3.4.15 2.7.1.33]	71	family	Neisseriaceae
K02244	competence protein ComGB	1775	class	Bacilli
K02344	DNA polymerase III subunit psi [EC:2.7.7.7]	1031	order	Enterobacterales
K02382	flagellar protein FlbA	34	family	Borreliaceae
K02384	flbC; flagellar protein FlbC	35	family	Borreliaceae
K02489	two-component system, glycerol uptake and utilization sensor kinase [EC:2.7.13.3]	35	family	Borreliaceae
K02527&K03439	3-deoxy-D-manno-octulosonic-acid transferase [EC:2.4.99.12 2.4.99.13 2.4.99.14 2.4.99.15]; tRNA (guanine-N7-)-methyltransferase [EC:2.1.1.33]	20	genus	*Fusobacterium*
K02621&K02469	topoisomerase IV subunit A [EC:5.99.1.-]; DNA gyrase subunit A [EC:5.99.1.3]	172	family	Clostridiaceae
K02634	apocytochrome f	352	class	Oxyphotobacteria
K02637	cytochrome b6-f complex subunit 4	373	class	Oxyphotobacteria
K02692	photosystem I subunit II	353	class	Oxyphotobacteria
K02698	photosystem I subunit X	350	class	Oxyphotobacteria
K02699	photosystem I subunit XI	356	class	Oxyphotobacteria
K02704	photosystem II CP47 chlorophyll apoprotein	351	class	Oxyphotobacteria
K02716	photosystem II oxygen-evolving enhancer protein 1	347	class	Oxyphotobacteria
K03071	preprotein translocase subunit SecB	7540	phylum	Proteobacteria
K03600	stringent starvation protein B	4742	class	Gammaproteobacteria
K03764	MetJ family transcriptional regulator, methionine regulon repressor	1300	order	Enterobacterales
K04420	mitogen-activated protein kinase kinase kinase 2 [EC:2.7.11.25]	2	species	*Tatlockia jamestowniensis*
K04725	E3 ubiquitin-protein ligase XIAP [EC:2.3.2.27]	2	species	*Endozoicomonas ascidiicola*
K04952	cyclic nucleotide gated channel beta 1	2	species	*Leptospira mayottensis*
K04954	hyperpolarization activated cyclic nucleotide-gated potassium channel 1	2	species	*Leptospira_A biflexa*
K04956	hyperpolarization activated cyclic nucleotide-gated potassium channel 3	2	genus	UBA3465
K05382	phycoerythrin-associated linker protein	345	class	Oxyphotobacteria
K05747	Wiskott-Aldrich syndrome protein	2	species	*Burkholderia gladioli*
K05764	thymosin, beta 4	3	species	*Roseofilum reptotaenium*
K06352	phosphatase RapA inhibitor	35	genus	*Bacillus*
K06353	phosphatase RapC regulator	35	genus	*Bacillus*
K06437	sigma-E controlled sporulation protein	73	genus	*Bacillus*
K06865	ATPase	7	family	X112
K07158	uncharacterized protein	4	genus	*Mycoplasma_C*
K07268	opacity associated protein	106	family	Pasteurellaceae
K07497&K07498	putative transposase; putative transposase	2	species	*Corynebacterium glutamicum*
K07503	endonuclease [EC:3.1.-.-]	3035	phylum	Actinobacteria
K07580	Zn-ribbon RNA-binding protein	2	species	*Paenibacillus_G napthalenovorans*
K07805	putative membrane protein PagD	10	genus	*Salmonella*
K08265	heterodisulfide reductase subunit E [EC:1.8.98.1]	2	species	UBA2210 sp1
K08274	late transcription unit A protein	21	family	Chlamydiaceae
K08275	ltuB; late transcription unit B protein	22	family	Chlamydiaceae

Table B1 continued from previous page

KEGG Orthology ID	Definition	# Genomes	Lowest Common Level	Lowest Common Rank
K08605	coccolysin [EC:3.4.24.30]	2	species	*Enterococcus faecalis*
K08719	outer membrane protein B	22	family	Chlamydiaceae
K08754	fatty acid-binding protein 5, epidermal	3	species	*Planktothrix* sp1
K08833	mitogen-activated protein kinase kinase kinase kinase 5 [EC:2.7.11.1]	2	species	*Verrucomicrobium spinosum*
K08902	photosystem II Psb27 protein	344	class	Oxyphotobacteria
K08903	photosystem II 13kDa protein	349	class	Oxyphotobacteria
K08942	photosystem P840 reaction center cytochrome c551	22	order	Chlorobiales
K08943	photosystem P840 reaction center protein PscD	22	order	Chlorobiales
K08945	chlorosome envelope protein A	22	order	Chlorobiales
K08946	chlorosome envelope protein B	20	family	Chlorobiaceae
K08947	chlorosome envelope protein C	19	family	Chlorobiaceae
K08949	chlorosome envelope protein E	20	family	Chlorobiaceae
K08951	chlorosome envelope protein H	19	family	Chlorobiaceae
K08952	csmI; chlorosome envelope protein I	22	order	Chlorobiales
K08954	chlorosome envelope protein X	23	order	Chlorobiales
K09146	uncharacterized protein	887	order	Corynebacteriales
K09377	cysteine and glycine-rich protein	2	species	*Planktothrix* sp1
K09487	heat shock protein 90kDa beta	2	species	UBA4811 sp1
K09552	spastic paraplegia 7 [EC:3.4.24.-]	2	species	UBA5776 sp1
K09987	uncharacterized protein	2466	class	Alphaproteobacteria
K10343	SPRY domain-containing SOCS box protein 1/4	2	species	*Endozoicomonas elysicola*
K10344	SPRY domain-containing SOCS box protein 2	2	species	*Endozoicomonas elysicola*
K10919	toxin coregulated pilus biosynthesis protein H	2	species	*Vibrio cholerae*
K10922	toxS; transmembrane regulatory protein ToxS	236	family	Vibrionaceae
K10923	tcpN, toxT; AraC family transcriptional regulator, TCP pilus virulence regulatory protein	2	species	*Vibrio cholerae*
K10935	toxin coregulated pilus biosynthesis protein F	2	species	*Vibrio cholerae*
K10963	toxin coregulated pilus biosynthesis protein R	2	species	*Vibrio cholerae*
K11010	superantigen YpmA	3	genus	*Yersinia*
K11025	ptxC; pertussis toxin subunit 3	11	genus	*Bordetella*
K11026	pertussis toxin subunit 4	11	genus	*Bordetella*
K11027	ptxE; pertussis toxin subunit 5	11	genus	*Bordetella*
K11028	vacuolating cytotoxin	54	genus	*Helicobacter*
K11033	nheA; non-hemolytic enterotoxin A	118	genus	*Bacillus_A*
K11046	streptolysin S associated protein	2	species	*Streptococcus pyogenes*
K11061	probable enterotoxin C	2	species	*Clostridium_P perfringens*
K11207	glutathione peroxidase-type tryparedoxin peroxidase [EC:1.11.1.-]	5	genus	*Ureaplasma*
K11275	histone H1/5	2	genus	*Myxococcus*
K11609&K09458	beta-ketoacyl ACP synthase [EC:2.3.1.-]; 3-oxoacyl-[acyl-carrier-protein] synthase II [EC:2.3.1.179]	286	genus	*Mycobacterium*
K12051	ComB7 competence protein	52	genus	*Helicobacter*
K12082	type IV secretion system protein PtlG	11	genus	*Bordetella*
K12084	type IV secretion system protein PtlD	11	genus	*Bordetella*
K12210	intracellular multiplication protein IcmF	81	family	Legionellaceae
K12215	intracellular multiplication protein IcmM	76	family	Legionellaceae
K12220	intracellular multiplication protein IcmR	5	genus	*Legionell*a
K12225	intracellular multiplication protein IcmX	78	family	Legionellaceae
K12314	actin, alpha cardiac muscle	2	genus	UBA3465
K12547	polysaccharidase protein	41	genus	*Rhizobium*
K12548	rap; autoaggregation protein RapA/B/C	41	genus	*Rhizobium*
K12551	monofunctional glycosyltransferase [EC:2.4.1.129]	69	genus	*Staphylococcus*
K12681	prn; pertactin	11	genus	*Bordetella*
K12695	raucaffricine beta-D-glucosidase / vomilenine glucosyltransferase [EC:3.2.1.125 2.4.1.219]	2	genus	*Luminiphilus*
K12809	espG2; T3SS secreted effector EspG-like protein	2	species	*Escherichia albertii*
K13032	prunasin beta-glucosidase [EC:3.2.1.118]	2	species	*Pseudomonas_L hussainii*
K13338	peroxin-1	2	species	*Flavobacterium psychrophilum_A*
K13563	ribostamycin:4-(gamma-L-glutamylamino)-(S)-2-hydroxybutanoyl-[BtrI acyl-carrier protein] 4-(gamma-L-glutamylamino)-(S)-2-hydroxybutanoate transferase [EC:2.3.2.19]	2	species	*Paenibacillus_F chitinolyticus*
K13630	multiple antibiotic resistance protein MarB	214	family	Enterobacteriaceae
K13739	secreted effector protein SopD	10	genus	*Salmonella*
K13749	solute carrier family 24 (sodium/potassium/calcium exchanger), member 1	2	species	UBA1450 sp1
K14159	ribonuclease HI / DNA polymerase III subunit epsilon [EC:3.1.26.4 2.7.7.7]	228	family	Moraxellaceae

Table B1 continued from previous page

KEGG Orthology ID	Definition	# Genomes	Lowest Common Level	Lowest Common Rank
K14203	FPRL1 inhibitory protein	7	genus	*Staphylococcus*
K14254	aminotransferase	2832	class	Actinobacteria
K14428	solute carrier family 12 (potassium/chloride transporters), member 8	6	species	UBA12451 sp1
K14494	DELLA protein	6	genus	*Sorangium*
K15474	enhanced entry protein EnhC	80	family	Legionellaceae
K15480	effector protein DrrA/SidM	2	species	*Legionella pneumophila*
K15566	tRNA (adenine9-N1/guanine9-N1)-methyltransferase [EC:2.1.1.218 2.1.1.221]	4	family	Desulfurobacteriaceae
K15630	GDP-D-glucose phosphorylase [EC:2.7.7.78]	2	species	*Sulfuricella denitrificans*
K15843	outer membrane protein HopC/AlpA	54	genus	*Helicobacter*
K15844	outer membrane protein HopB/AlpB	54	genus	*Helicobacter*
K15845	outer membrane protein HopZ	53	genus	*Helicobacter*
K15847	outer membrane protein BabA	53	genus	*Helicobacter*
K15848	outer membrane protein SabA	54	genus	*Helicobacter*
K16060	baculoviral IAP repeat-containing protein 2/3	2	species	*Endozoicomonas ascidiicola*
K16463	centrosomal protein CEP170	5	species	*Treponema_D* sp2
K16654	spore-specific protein	105	genus	*Bacillus_A*
K16712	EPS I polysaccharide export inner membrane protein EpsE	6	genus	*Ralstonia*
K16713	EPS I polysaccharide export inner membrane protein EpsF	6	genus	*Ralstonia*
K16883	heat-stable enterotoxin STa/STI	2	species	*Escherichia coli_A*
K16919	acetoin utilization transport system permease protein	80	family	Bacillaceae
K17211	curcumin/demethoxycurcumin synthase [EC:2.3.1.217 2.3.1.219]	2	species	*Sphingopyxis* sp3
K17307	fibulin 1/2	2	species	UBA5124 sp2
K17533	mitogen-activated protein kinase kinase kinase 19 [EC:2.7.11.25]	2	species	*Tatlockia jamestowniensis*
K17605	PPP2R4, PTPA; serine/threonine-protein phosphatase 2A activator	2	species	*Streptomyces hydroscopicus*
K17815	exonuclease V [EC:3.1.-.-]	2	species	UBA1671 sp1
K18132	major outer membrane protein P.IA	104	order	Betaproteobacteriales
K18150	antirepressor for MexR	2	species	*Pseudomonas aeruginosa_A*
K18268	podocin	2	species	UBA6899 sp1
K18561	FAD-dependent fumarate reductase [EC:1.3.8.-]	5	species	UBA12409 sp1
K18861	4-hydroxybutyrate—CoA ligase (AMP-forming) [EC:6.2.1.40]	2	species	UBA8473 sp1
K18886	gibberellin A4 carboxyl methyltransferase [EC:2.1.1.276]	2	species	*Labrenzia* sp1
K18955	WhiB family transcriptional regulator, redox-sensing transcriptional regulator	3041	phylum	Actinobacteria
K19004	processive 1,2-diacylglycerol beta-glycosyltransferase [EC:2.4.1.315 2.4.1.-]	4	genus	*Mycoplasma_C*
K19007	lysophosphatidiate acyltransferase [EC:2.3.1.51]	2	species	XYA12-FULL-58-9 sp1
K19094	antitoxin ParD2	2	species	*Mycobacterium tuberculosis*
K19237	gingipain K [EC:3.4.22.47]	2	species	*Porphyromonas gingivalis*
K20117&K20116	PTS system, glucose-specific IIB component [EC:2.7.1.199] ; PTS system, glucose-specific IIA component [EC:2.7.1.199]	4	genus	*Coprococcus*
K20312	TRAPP-associated protein TCA17	5	species	UBA12409 sp1
K20389	phosphatase RapH regulator	22	genus	*Bacillus*
K20592	genN; SAM-dependent 3"-N-methyltransferase [EC:2.1.1.-]	3	genus	*Micromonospora*
K20593	cobalamin-dependent radical SAM methyltransferase [EC:2.1.1.-]	3	genus	*Micromonospora*
K20717	mitogen-activated protein kinase kinase kinase YODA [EC:2.7.11.25]	2	species	*Tatlockia jamestowniensis*
K21121	N-alpha-acetyltransferase 60 [EC:2.3.1.259]	2	species	*Bacillus_A weihenstephanensis_A*
K21274	hexaprenyl-diphosphate synthase small subunit [EC:2.5.1.83]	5	genus	*Macrococcus*
K21474	peptidoglycan DL-endopeptidase RipB [EC:3.4.-.-]	328	genus	*Mycobacterium*

Table B2. Lineage-specific Pfam annotations. Pfam annotations with at least one clade in the bacterial GTDB tree with a catchment and saturation of at least 95% and that were present in no more than half of the genomes in the bacterial GTDB tree were classified as lineage-specific and are listed in this table.

Pfam ID	Definition	# Genomes	Lowest Common Level	Lowest Common Rank
PF00283.14	Cytochrome b559, alpha (gene psbE) and beta (gene psbF)subunits	344	class	Oxyphotobacteria
PF00341.12	PDGF/VEGF domain	2	species	*Leptospira interrogans*
PF00379.18	Insect cuticle protein	3	species	*Megasphaera massiliensis*
PF00421.14	Photosystem II protein	365	class	Oxyphotobacteria
PF00504.16	Chlorophyll A-B binding protein	374	class	Oxyphotobacteria
PF00715.12	Interleukin 2	11	species	UBA12075 sp1
PF00737.15	Photosystem II 10 kDa phosphoprotein	345	class	Oxyphotobacteria
PF00778.12	DIX domain	2	genus	*Eubacterium_S*
PF00812.12	Ephrin	2	species	*Endozoicomonas montiporae*
PF00864.14	ATP P2X receptor	2	species	*Clostridium_L* sp1
PF00868.15	Transglutaminase family	2	genus	*Thiohalocapsa*
PF00938.12	Lipoprotein	4	genus	*Mycoplasma_C*
PF00971.13	EIAV coat protein, gp90	2	species	*Pseudoalteromonas piscicida*
PF01283.14	Ribosomal protein S26e	2	species	*Caulobacter* sp3
PF01302.20	CAP-Gly domain	2	species	*Endozoicomonas ascidiicola*
PF01333.14	Apocytochrome F, C-terminal	358	class	Oxyphotobacteria
PF01340.15	Met Apo-repressor, MetJ	1338	order	Enterobacterales
PF01352.22	KRAB box	2	species	*Roseofilum reptotaenium*
PF01466.14	Skp1 family, dimerisation domain	2	species	*Lactobacillus_H suebicus*
PF01621.12	Cell fusion glycoprotein K	2	species	*Mycoplasma_C genitalium*
PF01716.13	Manganese-stabilising protein / photosystem II polypeptide	352	class	Oxyphotobacteria
PF01846.14	FF domain	2	species	GWB1-40-14 sp2
PF02177.11	Amyloid A4 N-terminal heparin-binding	2	species	UBA8515 sp3
PF02330.11	Mitochondrial glycoprotein	2	species	*Gemmata massiliana*
PF02467.11	Transcription factor WhiB	3059	phylum	Actinobacteria
PF02507.10	Photosystem I reaction centre subunit III	344	class	Oxyphotobacteria
PF02531.11	PsaD	354	class	Oxyphotobacteria
PF02605.10	Photosystem I reaction centre subunit XI	354	class	Oxyphotobacteria
PF02672.10	CP12 domain	357	class	Oxyphotobacteria
PF02691.10	Vacuolating cytotoxin	56	genus	*Helicobacter*
PF02722.10	Major Outer Sheath Protein C-terminal domain	12	family	Treponemataceae
PF02966.11	Mitosis protein DIM1	3	species	*Calescibacterium nevades*
PF03000.9	NPH3 family	2	species	*Lactobacillus gigeriorum*
PF03072.9	MG032/MG096/MG288 family 1	4	genus	*Mycoplasma_C*
PF03086.9	MG032/MG096/MG288 family 2	4	genus	*Mycoplasma_C*
PF03108.10	MuDR family transposase	2	species	*Obscuribacter phosphatis*
PF03257.8	Mycoplasma adhesin P1	4	genus	*Mycoplasma_C*
PF03268.9	Caenorhabditis protein of unknown function, DUF267	2	species	*Mycoplasma_E ovipneumoniae*
PF03373.9	Octapeptide repeat	7	genus	*Staphylococcus*
PF03384.9	Drosophila protein of unknown function, DUF287	2	genus	*Pseudomonas_E*
PF03429.8	Major surface protein 1B	3	genus	*Anaplasma*
PF03503.8	Chlamydia cysteine-rich outer membrane protein 3	22	family	Chlamydiaceae
PF03579.8	Small hydrophobic protein	2	species	NA
PF03622.8	IBV 3B protein	2	species	*Pseudomonas_B luteola*
PF03635.12	Vacuolar protein sorting-associated protein 35	2	genus	NK4A144
PF03661.8	Uncharacterised protein family (UPF0121)	2	species	UBA2993 sp1
PF03672.8	Uncharacterised protein family (UPF0154)	2450	phylum	Firmicutes
PF03912.9	Psb28 protein	351	class	Oxyphotobacteria
PF03931.10	Skp1 family, tetramerisation domain	4	species	UBA12411 sp1
PF03943.8	TAP C-terminal domain	2	species	*Pseudoalteromonas piscicida*
PF04022.7	Staphylocoagulase repeat	7	genus	*Staphylococcus*
PF04036.7	NA	4	genus	*Nocardia*

Table B2 continued from previous page

Pfam ID	Description	# Genomes	Lowest Common Level	Lowest Common Rank
PF04088.8	Peroxin 13, N-terminal region	3	species	*Nonlabens ulvanivorans*
PF04139.8	Rad9	2	species	2-01-FULL-40-13 sp1
PF04217.8	Protein of unknown function, DUF412	1318	order	Enterobacterales
PF04297.9	Putative helix-turn-helix protein, YlxM / p13 like	5068	domain	Bacteria
PF04382.8	SAB domain	2	species	*Gilliamella apicola_D*
PF04386.8	Stringent starvation protein B	7353	phylum	Proteobacteria
PF04420.9	CHD5-like protein	2	species	*Lactobacillus_G rapi*
PF04483.7	Protein of unknown function (DUF565)	341	class	Oxyphotobacteria
PF04501.7	Baculovirus major capsid protein VP39	2	species	*Marinobacter psychrophilus*
PF04521.8	ssRNA positive strand viral 18kD cysteine rich protein	2	species	*Bacillus xiamenensis*
PF04546.8	Sigma-70, non-essential region	7247	phylum	Proteobacteria
PF04559.7	Herpesvirus UL17 protein	2	genus	*Actinomyces*
PF04637.7	Herpesvirus phosphoprotein 85 (HHV6-7 U14/HCMV UL25)	3	species	UBA1364 sp1
PF04648.7	Yeast mating factor alpha hormone	2	species	*Arthrospira maxima*
PF04716.9	ETC complex I subunit conserved region	2	species	*Dysgonomonas macrotermitis*
PF04746.7	Protein of unknown function (DUF575)	4	species	*Chlamydophila pneumoniae*
PF04763.7	Protein of unknown function (DUF562)	4	species	*Chlamydophila pneumoniae*
PF04795.7	PAPA-1-like conserved region	2	species	UBA8366 sp1
PF04801.8	Sin-like protein conserved region	6	species	Zag1 sp1
PF04839.8	Plastid and cyanobacterial ribosomal protein (PSRP-3 / Ycf65)	344	class	Oxyphotobacteria
PF05018.8	Protein of unknown function (DUF667)	2	species	2-01-FULL-40-13 sp3
PF05020.10	NPL4 family, putative zinc binding region	2	species	MS4 sp1
PF05024.10	N-acetylglucosaminyl transferase component (Gpi1)	2	genus	UBA5862
PF05071.11	NADH ubiquinone oxidoreductase subunit NDUFA12	2418	class	Alphaproteobacteria
PF05109.8	Herpes virus major outer envelope glycoprotein (BLLF1)	2	genus	UBA5734
PF05174.7	Cysteine-rich D. radiodurans N terminus	2	species	*Deinococcus radiodurans*
PF05251.7	Oligosaccharyltransferase subunit 5	2	species	*Persicobacter* sp1
PF05271.6	Tobravirus 2B protein	4	genus	*Paenibacillus*
PF05302.6	Protein of unknown function (DUF720)	22	family	Chlamydiaceae
PF05310.7	Tenuivirus movement protein	2	species	*Prevotella* sp31
PF05340.7	Protein of unknown function (DUF740)	2	species	UBA12087 sp1
PF05540.6	Serpulina hyodysenteriae variable surface protein	12	genus	*Brachyspira*
PF05542.6	Protein of unknown function (DUF760)	351	class	Oxyphotobacteria
PF05555.6	Coxiella burnetii protein of unknown function (DUF762)	2	species	*Coxiella burnetii*
PF05627.6	Cleavage site for pathogenic type III effector avirulence factor Avr	2	species	*Fuchsiella alkaliacetigena*
PF05660.6	Coxiella burnetii protein of unknown function (DUF807)	2	species	*Coxiella burnetii*
PF05672.6	MAP7 (E-MAP-115) family	2	species	UBA11549 sp2
PF05795.6	Plasmodium vivax Vir protein	2	species	*Robinsoniella peoriensis*
PF05821.6	NADH-ubiquinone oxidoreductase ASHI subunit (CI-ASHI or NDUFB8)	2	species	*Bifidobacterium saguini*
PF05910.7	Plant protein of unknown function (DUF868)	2	species	*Nonlabens sediminis*
PF06011.7	Transient receptor potential (TRP) ion channel	4	genus	*Chryseobacterium*
PF06242.6	Protein of unknown function (DUF1013)	2507	class	Alphaproteobacteria
PF06261.6	Actinobacillus actinomycetemcomitans leukotoxin activator LktC	2	species	*Aggregatibacter actinomycetemcomitans*
PF06340.6	Vibrio cholerae toxin co-regulated pilus biosynthesis protein F	2	species	*Vibrio cholerae*
PF06394.8	Pepsin inhibitor-3-like repeated domain	7	genus	*Vibrio*
PF06399.8	GTP cyclohydrolase I feedback regulatory protein (GFRP)	2	species	*Caviibacter abscessus*
PF06427.6	UDP-glucose:Glycoprotein Glucosyltransferase	2	species	*Enterococcus_D gallinarum*
PF06485.6	Protein of unknown function (DUF1092)	348	class	Oxyphotobacteria
PF06519.6	TolA C-terminal	1335	order	Enterobacterales
PF06548.6	NA	2	species	*Lactobacillus_H reuteri*
PF06563.6	Protein of unknown function (DUF1125)	3	genus	*Lactococcus*
PF06582.7	Repeat of unknown function (DUF1136)	2	species	*Brochothrix thermosphacta*
PF06587.6	Protein of unknown function (DUF1137)	22	family	Chlamydiaceae
PF06588.6	Muskelin N-terminus	2	species	UBA6024 sp1
PF06643.6	Protein of unknown function (DUF1158)	328	family	Enterobacteriaceae
PF06663.8	Protein of unknown function (DUF1170)	2	species	*Corynebacterium camporealensis*
PF06697.7	Protein of unknown function (DUF1191)	2	species	*Enterococcus silesiacus*
PF06781.7	Cell division protein CrgA	2964	class	Actinobacteria
PF06799.6	Protein of unknown function (DUF1230)	349	class	Oxyphotobacteria
PF07028.6	Protein of unknown function (DUF1319)	2	species	*Parageobacillus thermoglucosidasius*

Table B2 continued from previous page

Pfam ID	Description	# Genomes	Lowest Common Level	Lowest Common Rank
PF07037.6	Putative transcription regulator (DUF1323)	340	family	Enterobacteriaceae
PF07078.6	Forty-two-three protein	2	species	*Vibrio hyugaensis*
PF07082.6	Protein of unknown function (DUF1350)	347	class	Oxyphotobacteria
PF07122.6	Variable length PCR target protein (VLPT)	2	species	*Ehrlichia chaffeensis*
PF07123.7	Photosystem II reaction centre W protein (PsbW)	2	species	*Photorhabdus temperata*
PF07176.6	Alpha/beta hydrolase of unknown function (DUF1400)	358	class	Oxyphotobacteria
PF07197.7	Protein of unknown function (DUF1409)	2	species	CAG-65 sp2
PF07199.6	Protein of unknown function (DUF1411)	2	species	*Bacillus_A thuringiensis_A*
PF07200.8	Modifier of rudimentary (Mod(r)) protein	2	species	UBA10025 sp1
PF07279.6	Protein of unknown function (DUF1442)	2	species	*Frankia inefficax*
PF07288.6	Protein of unknown function (DUF1447)	1774	class	Bacilli
PF07306.6	Protein of unknown function (DUF1455)	2	species	*Anaplasma marginale*
PF07327.6	Neuroparsin	2	species	*Corynebacterium* sp2
PF07372.7	Protein of unknown function (DUF1491)	2194	class	Alphaproteobacteria
PF07404.6	Telomere-binding protein beta subunit (TEBP beta)	2	species	*Bacteriovorax* sp1
PF07444.6	Ycf66 protein N-terminus	357	class	Oxyphotobacteria
PF07525.11	SOCS box	3	genus	*Endozoicomonas*
PF07542.6	ATP12 chaperone protein	2289	class	Alphaproteobacteria
PF07571.8	TAF6 C-terminal HEAT repeat domain	3	genus	*Planktothrix*
PF07577.6	Domain of Unknown Function (DUF1547)	22	family	Chlamydiaceae
PF07621.6	Protein of unknown function (DUF1582)	5	genus	*Rhodopirellula*
PF07623.6	Protein of unknown function (DUF1584)	5	genus	*Rhodopirellula*
PF07765.7	KIP1-like protein	2	species	UBA10001 sp1
PF07860.6	WisP family C-Terminal Region	2	species	*Tropheryma whipplei*
PF07861.6	WisP family N-Terminal Region	2	species	*Tropheryma whipplei*
PF07877.6	Protein of unknown function (DUF1661)	4	genus	*Porphyromonas*
PF07937.6	Protein of unknown function (DUF1686)	2	species	UBA5169 sp1
PF08043.7	Xin repeat	2	species	*Smaragdicoccus niigatensis*
PF08129.6	Alpha/beta enterocin family	2	species	*Enterococcus_A pallens*
PF08149.6	BING4CT (NUC141) domain	2	species	*Lactobacillus hominis*
PF08181.6	DegQ (SacQ) family	87	family	Bacillaceae
PF08203.6	Yeast RNA polymerase I subunit RPA14	2	species	2-12-FULL-69-37 sp1
PF08272.6	Topoisomerase I zinc-ribbon-like	2726	class	Gammaproteobacteria
PF08300.8	Hepatitis C virus non-structural 5a zinc finger domain	2	family	Thermovenabulaceae
PF08320.7	PIG-X / PBN1	2	species	*Intestinibacter bartlettii*
PF08391.5	Ly49-like protein, N-terminal region	2	species	*Wohlfahrtiimonas chitiniclastica*
PF08430.7	Forkhead N-terminal region	5	genus	*Paenibacillus*
PF08434.6	Calcium-activated chloride channel N terminal	2	species	*Pseudobacteroides cellulosolvens*
PF08499.7	3'5'-cyclic nucleotide phosphodiesterase N-terminal	2	species	UBA1557 sp1
PF08624.5	Chromatin remodelling complex Rsc7/Swp82 subunit	2	genus	*Pedobacter*
PF08683.6	Microtubule-binding calmodulin-regulated spectrin-associated	2	species	*Staphylococcus schleiferi*
PF08848.6	Domain of unknown function (DUF1818)	351	class	Oxyphotobacteria
PF08855.5	Domain of unknown function (DUF1825)	352	class	Oxyphotobacteria
PF08866.5	Putative amino acid metabolism	874	order	Lactobacillales
PF08930.5	Domain of unknown function (DUF1912)	286	family	Streptococcaceae
PF08954.6	Trimerisation motif	2	species	*Marinomonas gallaica*
PF09042.6	Titin Z	2	species	*Rhodococcus triatomae*
PF09090.6	MIF4G like	2	species	*Sphaerochaeta* sp2
PF09105.5	Elongation factor SelB, winged helix	6	genus	*Moorella*
PF09144.5	Yersinia pseudo-tuberculosis mitogen	3	genus	*Yersinia*
PF09229.6	Activator of Hsp90 ATPase, N-terminal	2	species	*Croceibacter atlanticus*
PF09237.6	GAGA factor	2	species	*Caldithrix abyssi*
PF09255.5	Caf1 Capsule antigen	2	species	*Yersinia pestis*
PF09281.5	Taq polymerase, exonuclease	70	order	Deinococcales
PF09288.5	Fungal ubiquitin-associated domain	6	species	UBA10105 sp1
PF09321.5	Domain of unknown function (DUF1978)	3	species	*Chlamydophila pneumoniae*
PF09324.5	Domain of unknown function (DUF1981)	2	species	UBA1557 sp1
PF09326.6	NADH-ubiquinone oxidoreductase subunit G, C-terminal	2394	class	Alphaproteobacteria
PF09341.5	Transcription factor Pcc1	2	species	*Staphylococcus_A vitulinus*
PF09353.5	Domain of unknown function (DUF1995)	345	class	Oxyphotobacteria
PF09367.5	CpeS-like protein	357	class	Oxyphotobacteria
PF09437.5	Pombe specific 5TM protein	2	species	GWA2-41-24 sp1
PF09510.5	Rtt102p-like transcription regulator protein	2	species	CAG-288 sp1
PF09644.5	Mg296 protein	13	genus	*Mycoplasma_C*
PF09736.4	Pre-mRNA-splicing factor of RES complex	2	species	*Lachnoclostridium phytofermentans_A*

Table B2 continued from previous page

Pfam ID	Description	# Genomes	Lowest Common Level	Lowest Common Rank
PF09793.4	Anticodon-binding domain	2	species	UBA2883 sp2
PF09840.4	Uncharacterized protein conserved in archaea (DUF2067)	2	species	*Fabibacter* sp1
PF10044.4	Retinal tissue protein	3	species	GWC1-27-15 sp1
PF10103.4	Zincin-like metallopeptidase	3063	phylum	Actinobacteria
PF10148.4	Schwannomin-interacting protein 1	2	species	*Tatlockia massiliensis*
PF10183.4	ESSS subunit of NADH:ubiquinone oxidoreductase (complex I)	2	species	RF16 sp12
PF10231.4	Uncharacterised conserved protein (DUF2315)	2	species	*Peptoclostridium litorale*
PF10350.4	Putative death-receptor fusion protein (DUF2428)	2	species	*Protochlamydia amoebophila*
PF10398.4	Protein of unknown function (DUF2443)	96	family	Helicobacteraceae
PF10453.4	Nuclear fragile X mental retardation-interacting protein 1 (NUFIP1)	2	genus	*Streptomyces*
PF10508.4	Proteasome non-ATPase 26S subunit	2	species	*Lactobacillus psittaci*
PF10611.4	Protein of unknown function (DUF2469)	2848	class	Actinobacteria
PF10642.4	Mitochondrial import receptor subunit or translocase	2	species	*Taylorella equigenitalis*
PF10664.4	Cyanobacterial and plastid NDH-1 subunit M	346	class	Oxyphotobacteria
PF10742.4	Protein of unknown function (DUF2555)	344	class	Oxyphotobacteria
PF10808.3	Protein of unknown function (DUF2542)	63	family	Enterobacteriaceae
PF10811.3	Protein of unknown function (DUF2532)	53	genus	*Rickettsia*
PF10814.3	Cell wall synthesis protein CwsA	314	genus	*Mycobacterium*
PF10818.3	Protein of unknown function (DUF2547)	104	family	Pasteurellaceae
PF10859.3	Protein of unknown function (DUF2660)	60	family	Rickettsiaceae
PF10875.3	Protein of unknown function (DUF2670)	60	family	Rickettsiaceae
PF10878.3	Protein of unknown function (DUF2672)	58	family	Rickettsiaceae
PF10879.3	Protein of unknown function (DUF2674)	54	genus	Rickettsia
PF10915.3	Protein of unknown function (DUF2709)	48	class	Chlamydiia
PF10939.3	Protein of unknown function (DUF2631)	999	order	Corynebacteriales
PF10953.3	Protein of unknown function (DUF2754)	234	family	Enterobacteriaceae
PF10954.3	Protein of unknown function (DUF2755)	236	family	Enterobacteriaceae
PF10969.3	Protein of unknown function (DUF2771)	905	order	Corynebacteriales
PF10976.3	Protein of unknown function (DUF2790)	532	family	Pseudomonadaceae
PF10999.3	Protein of unknown function (DUF2839)	365	class	Oxyphotobacteria
PF11016.3	Protein of unknown function (DUF2854)	350	class	Oxyphotobacteria
PF11061.3	Protein of unknown function (DUF2862)	346	class	Oxyphotobacteria
PF11065.3	Protein of unknown function (DUF2866)	199	family	Burkholderiaceae
PF11076.3	Putative inner membrane protein YbhQ	327	family	Enterobacteriaceae
PF11082.3	Protein of unknown function (DUF2880)	24	genus	*Cupriavidus*
PF11098.3	Chlorosome envelope protein C	22	order	Chlorobiales
PF11116.3	Protein of unknown function (DUF2624)	665	order	Bacillales
PF11131.3	Rap-phr extracellular signalling	35	genus	*Bacillus*
PF11152.3	Cofactor assembly of complex C subunit B, CCB2/CCB4	348	class	Oxyphotobacteria
PF11165.3	Protein of unknown function (DUF2949)	351	class	Oxyphotobacteria
PF11184.3	Protein of unknown function (DUF2969)	806	order	Lactobacillales
PF11189.3	Protein of unknown function (DUF2973)	357	class	Oxyphotobacteria
PF11210.3	Protein of unknown function (DUF2996)	352	class	Oxyphotobacteria
PF11224.3	Protein of unknown function (DUF3023)	10	genus	*Ehrlichia*
PF11226.3	Protein of unknown function (DUF3022)	209	family	Burkholderiaceae
PF11228.3	Protein of unknown function (DUF3027)	2833	class	Actinobacteria
PF11237.3	Protein of unknown function (DUF3038)	345	class	Oxyphotobacteria
PF11238.3	Protein of unknown function (DUF3039)	2927	class	Actinobacteria
PF11241.3	Protein of unknown function (DUF3043)	2858	class	Actinobacteria
PF11252.3	Protein of unknown function (DUF3051)	4	species	UBA12393 sp1
PF11263.3	Borrelia burgdorferi attachment protein P66	35	family	Borreliaceae
PF11264.3	Thylakoid formation protein	356	class	Oxyphotobacteria
PF11267.3	Domain of unknown function (DUF3067)	345	class	Oxyphotobacteria
PF11268.3	Protein of unknown function (DUF3071)	2830	class	Actinobacteria
PF11273.3	Protein of unknown function (DUF3073)	2832	class	Actinobacteria
PF11285.3	Protein of unknown function (DUF3086)	358	class	Oxyphotobacteria
PF11297.3	Protein of unknown function (DUF3098)	2599	phylum	Bacteroidetes
PF11305.3	Protein of unknown function (DUF3107)	2917	class	Actinobacteria
PF11332.3	Protein of unknown function (DUF3134)	351	class	Oxyphotobacteria
PF11334.3	Protein of unknown function (DUF3136)	105	family	Cyanobiaceae
PF11341.3	Protein of unknown function (DUF3143)	346	class	Oxyphotobacteria
PF11344.3	Protein of unknown function (DUF3146)	337	class	Oxyphotobacteria
PF11364.3	Protein of unknown function (DUF3165)	273	genus	*Streptococcus*
PF11375.3	Protein of unknown function (DUF3177)	356	class	Oxyphotobacteria
PF11377.3	Protein of unknown function (DUF3180)	2818	class	Actinobacteria
PF11388.3	Phagosome trafficking protein DotA	73	family	Legionellaceae
PF11452.3	Protein of unknown function (DUF3000)	2763	class	Actinobacteria
PF11482.3	Protein of unknown function (DUF3208)	68	order	Deinococcales
PF11497.3	NADH-quinone oxidoreductase chain 15	67	order	Deinococcales
PF11507.3	Ebola virus-specific transcription factor VP30	3	genus	*Bdellovibrio*
PF11516.3	Protein of unknown function (DUF3120)	11	genus	*Bordetella*
PF11586.3	Protein of unknown function (DUF3242)	36	order	Thermotogales

Table B2 continued from previous page

Pfam ID	Description	# Genomes	Lowest Common Level	Lowest Common Rank
PF11609.3	Protein of unknown function (DUF3248)	70	order	Deinococcales
PF11631.3	Protein of unknown function (DUF3255)	34	genus	*Bacillus*
PF11674.3	Protein of unknown function (DUF3270)	284	family	Streptococcaceae
PF11690.3	Protein of unknown function (DUF3287)	2	species	*Pediococcus acidilactici*
PF11691.3	Protein of unknown function (DUF3288)	344	class	Oxyphotobacteria
PF11801.3	Tom37 C-terminal domain	3	genus	*Tatlockia*
PF11826.3	Protein of unknown function (DUF3346)	249	family	Vibrionaceae
PF11833.3	Protein CHAPERONE-LIKE PROTEIN OF POR1-like	345	class	Oxyphotobacteria
PF11909.3	NADH-quinone oxidoreductase cyanobacterial subunit N	344	class	Oxyphotobacteria
PF11910.3	Cyanobacterial and plant NDH-1 subunit O	348	class	Oxyphotobacteria
PF11947.3	Protein of unknown function (DUF3464)	349	class	Oxyphotobacteria
PF12021.3	Protein of unknown function (DUF3509)	539	family	Pseudomonadaceae
PF12027.3	Protein of unknown function (DUF3514)	10	genus	*Ehrlichia*
PF12046.3	Cofactor assembly of complex C subunit B	357	class	Oxyphotobacteria
PF12049.3	Protein of unknown function (DUF3531)	345	class	Oxyphotobacteria
PF12071.3	Protein of unknown function (DUF3551)	146	family	Xanthobacteraceae
PF12089.3	Transmembrane domain of unknown function (DUF3566)	2918	class	Actinobacteria
PF12095.3	Protein CHLORORESPIRATORY REDUCTION 7	345	class	Oxyphotobacteria
PF12113.3	SVM protein signal sequence	14	genus	*Phytoplasma*
PF12135.3	Sialidase enzyme penultimate C terminal domain	2	species	*Clostridium_P perfringens*
PF12145.3	Eukaryotic Mediator 12 subunit domain	2	species	*Listeria grayi*
PF12163.3	DNA replication regulator	270	order	Campylobacterales
PF12178.3	Chromosome passenger complex (CPC) protein INCENP N terminal	2	species	*Pseudoalteromonas luteoviolacea*
PF12211.3	Low molecular weight S layer protein N terminal	7	genus	*Clostridioides*
PF12227.3	Protein of unknown function (DUF3603)	647	order	Bacillales
PF12240.3	Angiomotin C terminal	2	species	NA
PF12334.3	Rickettsia outer membrane protein B	53	genus	*Rickettsia*
PF12378.3	Trypsin-sensitive surface-exposed protein	13	genus	*Mycoplasma_C*
PF12422.3	Condensin II non structural maintenance of chromosomes subunit	2	species	*Lactobacillus_G collinoides*
PF12452.3	Protein of unknown function (DUF3685)	348	class	Oxyphotobacteria
PF12502.3	Protein of unknown function (DUF3710)	2802	class	Actinobacteria
PF12506.3	Protein of unknown function (DUF3713)	13	genus	*Mycoplasma_C*
PF12527.3	Protein of unknown function (DUF3727)	354	class	Oxyphotobacteria
PF12574.3	120 KDa Rickettsia surface antigen	59	family	Rickettsiaceae
PF12600.3	Protein of unknown function (DUF3769)	357	class	Oxyphotobacteria
PF12626.2	Polymerase A arginine-rich C-terminus	4866	class	Gammaproteobacteria
PF12723.2	Protein of unknown function (DUF3809)	68	order	Deinococcales
PF13033.1	Protein of unknown function (DUF3894)	117	genus	*Bacillus_A*
PF13043.1	Domain of unknown function (DUF3903)	118	genus	*Bacillus_A*
PF13049.1	Protein of unknown function (DUF3910)	120	genus	*Bacillus_A*
PF13050.1	Protein of unknown function (DUF3911)	119	genus	*Bacillus_A*
PF13051.1	Protein of unknown function (DUF3912)	117	genus	*Bacillus_A*
PF13052.1	Protein of unknown function (DUF3913)	118	genus	*Bacillus_A*
PF13053.1	Protein of unknown function (DUF3914)	100	genus	*Bacillus_A*
PF13054.1	Protein of unknown function (DUF3915)	118	genus	*Bacillus_A*
PF13055.1	Protein of unknown function (DUF3917)	119	genus	*Bacillus_A*
PF13059.1	Protein of unknown function (DUF3992)	111	genus	*Bacillus_A*
PF13060.1	Protein of unknown function (DUF3921)	125	family	Bacillaceae_G
PF13062.1	Protein of unknown function (DUF3924)	121	genus	*Bacillus_A*
PF13063.1	Protein of unknown function (DUF3925)	119	genus	*Bacillus_A*
PF13065.1	Protein of unknown function (DUF3928)	118	genus	*Bacillus_A*
PF13066.1	Protein of unknown function (DUF3929)	118	genus	*Bacillus_A*
PF13067.1	Protein of unknown function (DUF3930)	117	genus	*Bacillus_A*
PF13068.1	Protein of unknown function (DUF3932)	117	genus	*Bacillus_A*
PF13069.1	Protein of unknown function (DUF3933)	121	genus	*Bacillus_A*
PF13071.1	Protein of unknown function (DUF3935)	121	genus	*Bacillus_A*
PF13074.1	Protein of unknown function (DUF3938)	118	genus	*Bacillus_A*
PF13077.1	Protein of unknown function (DUF3909)	118	genus	*Bacillus_A*
PF13080.1	Protein of unknown function (DUF3926)	114	genus	*Bacillus_A*
PF13082.1	Protein of unknown function (DUF3931)	119	genus	*Bacillus_A*
PF13105.1	Protein of unknown function (DUF3959)	118	genus	*Bacillus_A*
PF13110.1	Protein of unknown function (DUF3966)	119	genus	*Bacillus_A*
PF13112.1	Protein of unknown function (DUF3965)	122	genus	*Bacillus_A*
PF13120.1	Domain of unknown function (DUF3974)	117	genus	*Bacillus_A*
PF13121.1	Domain of unknown function (DUF3976)	119	genus	*Bacillus_A*
PF13126.1	Protein of unknown function (DUF3975)	109	genus	*Bacillus_A*
PF13134.1	Protein of unknown function (DUF3948)	120	genus	*Bacillus_A*
PF13135.1	Protein of unknown function (DUF3947)	115	genus	*Bacillus_A*
PF13140.1	Domain of unknown function (DUF3980)	107	genus	*Bacillus_A*
PF13141.1	Protein of unknown function (DUF3979)	126	family	Bacillaceae_G
PF13142.1	Domain of unknown function (DUF3960)	119	genus	*Bacillus_A*

Table B2 continued from previous page

Pfam ID	Description	# Genomes	Lowest Common Level	Lowest Common Rank
PF13153.1	Protein of unknown function (DUF3985)	105	genus	*Bacillus_A*
PF13169.1	Poxvirus B22R protein N-terminal	2	species	*Listeria fleischmannii*
PF13179.1	Family of unknown function (DUF4006)	262	order	Campylobacterales
PF13210.1	Domain of unknown function (DUF4018)	116	genus	*Bacillus_A*
PF13219.1	Protein of unknown function (DUF4027)	118	genus	*Bacillus_A*
PF13221.1	Protein of unknown function (DUF4029)	102	genus	*Bacillus_A*
PF13268.1	Protein of unknown function (DUF4059)	283	family	Streptococcaceae
PF13294.1	Domain of unknown function (DUF4075)	119	genus	*Bacillus_A*
PF13326.1	Photosystem II Pbs27	347	class	Oxyphotobacteria
PF13355.1	Protein of unknown function (DUF4101)	366	class	Oxyphotobacteria
PF13397.1	RNA polymerase-binding protein	2871	class	Actinobacteria
PF13652.1	Putative quorum-sensing-regulated virulence factor	577	family	Pseudomonadaceae
PF13721.1	SecD export protein N-terminal TM region	4780	class	Gammaproteobacteria
PF13763.1	Domain of unknown function (DUF4167)	2302	class	Alphaproteobacteria
PF13829.1	Domain of unknown function (DUF4191)	2865	class	Actinobacteria
PF13834.1	Domain of unknown function (DUF4193)	2891	class	Actinobacteria
PF13983.1	YsaB-like lipoprotein	232	family	Enterobacteriaceae
PF13987.1	YedD-like protein	328	family	Enterobacteriaceae
PF13999.1	MarB protein	225	family	Enterobacteriaceae
PF14029.1	Protein of unknown function (DUF4244)	2770	class	Actinobacteria
PF14123.1	Domain of unknown function (DUF4290)	2381	class	Bacteroidia
PF14159.1	CAAD domains of cyanobacterial aminoacyl-tRNA synthetase	352	class	Oxyphotobacteria
PF14233.1	Domain of unknown function (DUF4335)	347	class	Oxyphotobacteria
PF14250.1	AbrB-like transcriptional regulator	367	class	Oxyphotobacteria
PF14409.1	Ribosomally synthesized peptide in Herpetosiphon	3	genus	*Herpetosiphon*
PF14507.1	CppA C-terminal	381	order	Lactobacillales
PF14632.1	Acidic N-terminal SPT6	4	genus	*Actinomyces_A*
PF14702.1	Central domain of human glycogen debranching enzyme	2	species	*Aneurinibacillus migulanus*
PF14738.1	Solute carrier (proton/amino acid symporter), TRAMD3 or PAT1	2	species	UBA7694 sp1
PF14795.1	Leucine-tRNA synthetase-specific domain	34	order	Deinococcales
PF14801.1	tRNA methyltransferase complex GCD14 subunit N-term	2858	class	Actinobacteria
PF14886.1	FAM183A and FAM183B related	2	species	UBA6595 sp1
PF15111.1	TMEM101 protein family	2	species	UBA1668 sp1
PF15196.1	Activator of apoptosis harakiri	2	species	*Nocardiopsis prasina*
PF15271.1	Spindle pole body component BBP1, Mps2-binding protein	2	species	*Brevibacillus parabrevis*
PF15436.1	Plasminogen-binding protein pgbA N-terminal	269	order	Campylobacterales
PF15437.1	Plasminogen-binding protein pgbA C-terminal	55	genus	*Helicobacter*
PF15513.1	Domain of unknown function (DUF4651)	259	genus	*Streptococcus*

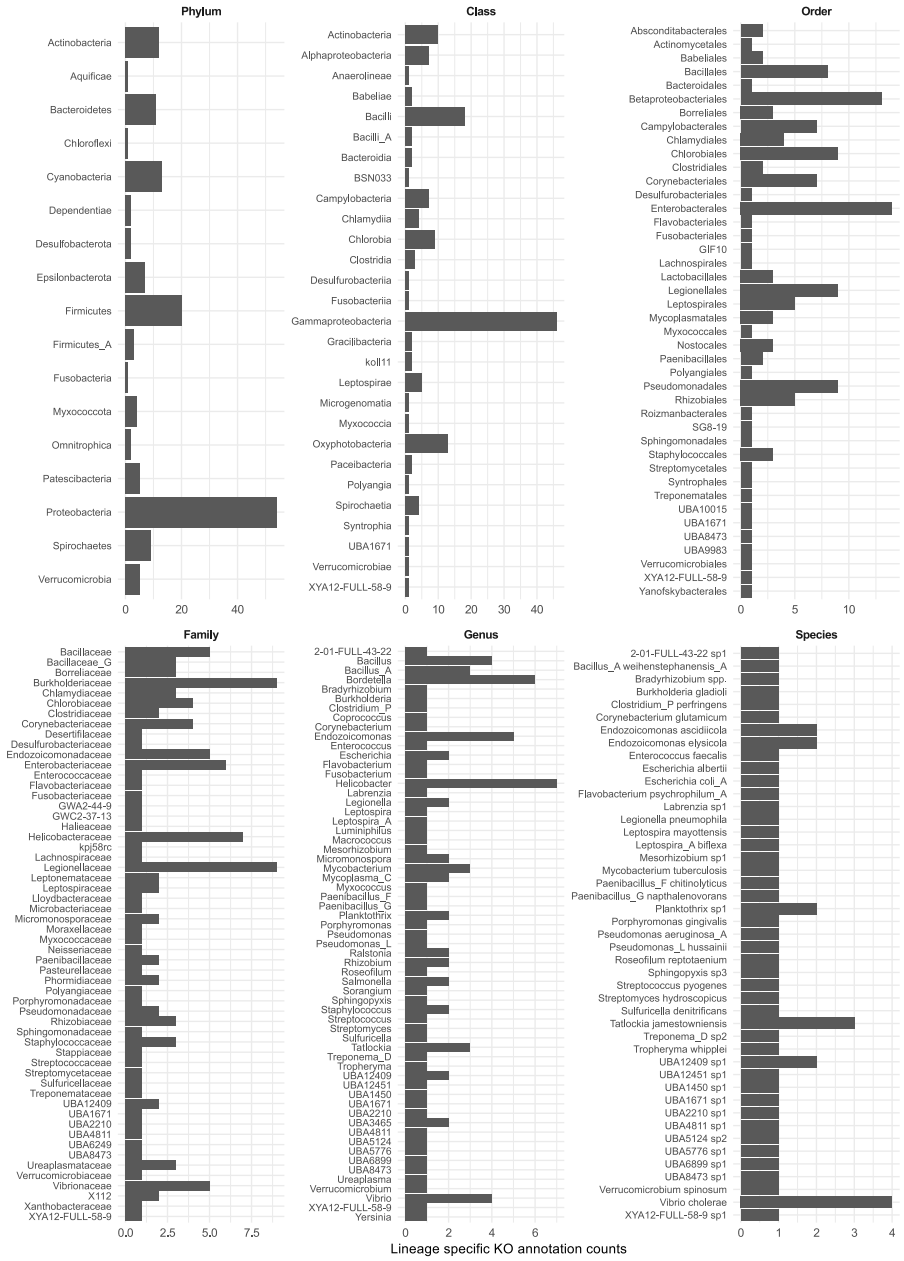

Figure B1. The taxonomic distribution of lineage-specific KO annotations. The taxonomic identity of each lineage-specific KO annotation was determined and counted at each level. Counts at higher levels include all lineage-specific traits at that level and all child levels.

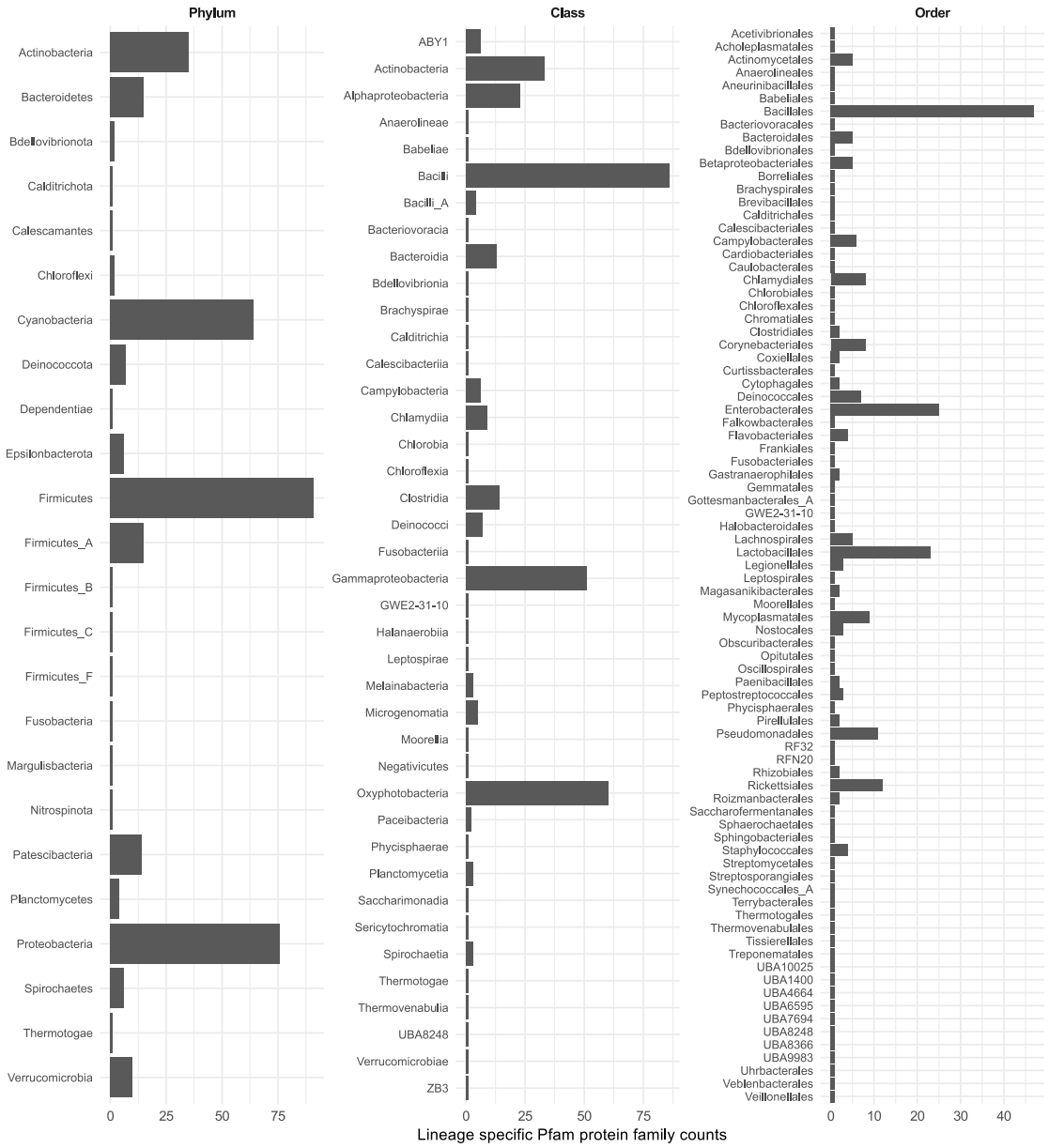

Figure B2. The taxonomic distribution of lineage-specific Pfam annotations in higher taxa. The taxonomic identity of each lineage-specific Pfam was determined and counted at each level. Counts at higher levels include all lineage-specific traits at that level and all child levels. See **Figure B3** for Pfam counts at the family, genus, and species levels.

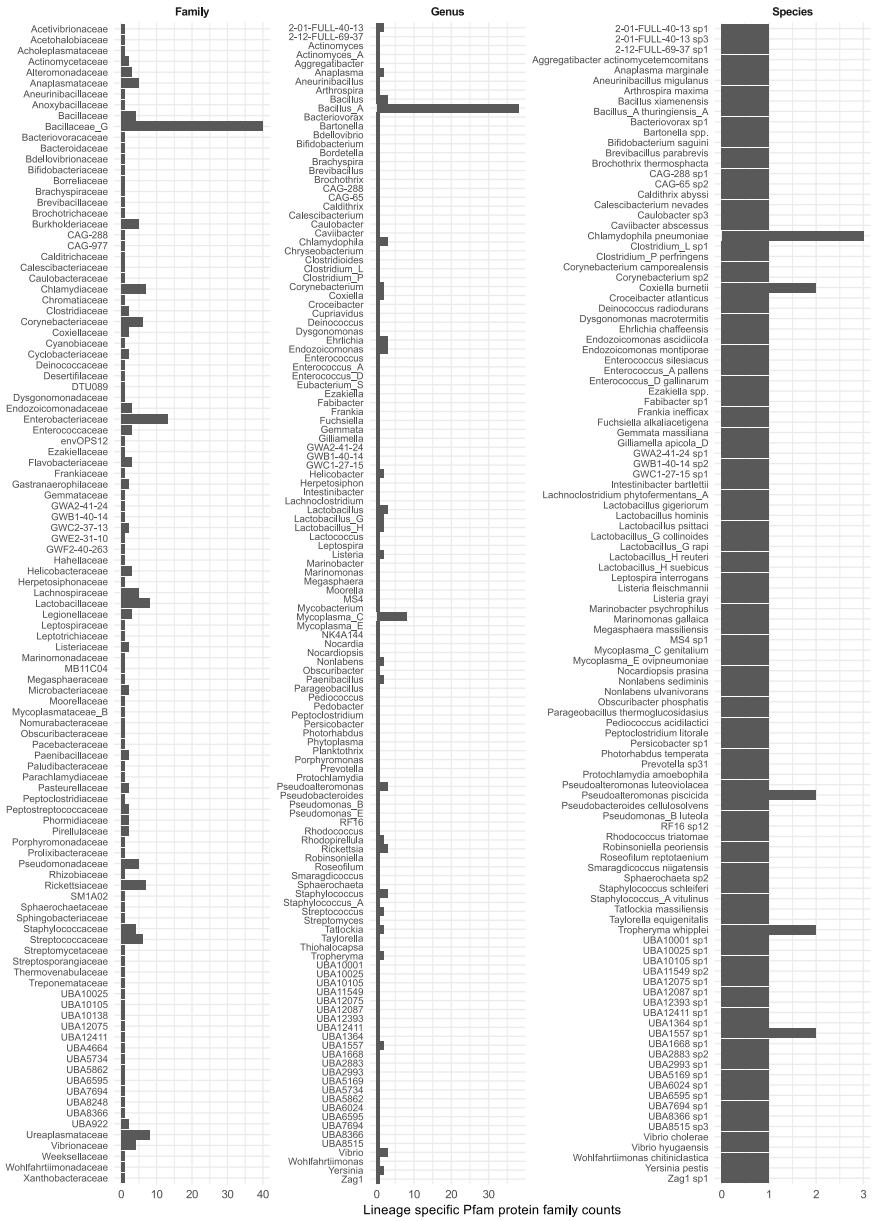

Figure B3. The taxonomic distribution of lineage-specific Pfam annotations in lower taxa. The taxonomic identity of each lineage-specific Pfam was determined and counted at each level. Counts at higher levels include all lineage-specific traits at that level and all child levels. See **Figure B2** for Pfam counts at the phylum, class, and order levels.