AnnoTree: visualization and exploration of a functionally annotated microbial tree of life

Kerrin Mendler[1],[†], Han Chen[1],[†], Donovan H. Parks[2], Briallen Lobb[1], Laura A. Hug[1] and Andrew C. Doxey[1],*

[1]Department of Biology, University of Waterloo, Waterloo, ON N2L 3G1, Canada and [2]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, University of Queensland, Queensland, Australia

ABSTRACT

Bacterial genomics has revolutionized our understanding of the microbial tree of life; however, mapping and visualizing the distribution of functional traits across bacteria remains a challenge. Here, we introduce AnnoTree—an interactive, functionally annotated bacterial tree of life that integrates taxonomic, phylogenetic and functional annotation data from over 27 000 bacterial and 1500 archaeal genomes. AnnoTree enables visualization of millions of precomputed genome annotations across the bacterial and archaeal phylogenies, thereby allowing users to explore gene distributions as well as patterns of gene gain and loss in prokaryotes. Using AnnoTree, we examined the phylogenomic distributions of 28 311 gene/protein families, and measured their phylogenetic conservation, patchiness, and lineage-specificity within bacteria. Our analyses revealed widespread phylogenetic patchiness among bacterial gene families, reflecting the dynamic evolution of prokaryotic genomes. Genes involved in phage infection/defense, mobile elements, and antibiotic resistance dominated the list of most patchy traits, as well as numerous intriguing metabolic enzymes that appear to have undergone frequent horizontal transfer. We anticipate that AnnoTree will be a valuable resource for exploring prokaryotic gene histories, and will act as a catalyst for biological and evolutionary hypothesis generation. AnnoTree is freely available at http://annotree.uwaterloo.ca

INTRODUCTION

Important biological and evolutionary insights can be generated by exploring the presence/absence of genes and functional annotations across species phylogenies. These include identifying unexpected taxonomic occurrences (1), uncovering the evolutionary origin of genes (2) and locating putative horizontal gene transfer (HGT) events (3,4). 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.

A number of strategies exist for merging taxonomic and functional information to create annotated phylogenies. For instance, homologs of a gene family retrieved using BLAST (5) or related methods can be manually mapped onto a custom species tree using tools such as iTOL (6) or GraPhlAn (7). Alternatively, several online bioinformatics databases offer precomputed summaries of taxonomic distributions for genes based on Linnean taxonomic classification or the NCBI taxonomy (8–11). However, there is a need for tools that allow users to explore gene/function distributions across a taxonomically curated and highly resolved tree of life.

Here, we present AnnoTree (annotree.uwaterloo.ca), a functionally annotated bacterial tree of life that enables interactive exploration of gene/function annotations across over 27 000 bacterial and 1500 archaeal genomes. The phylogeny and taxonomic nomenclature used within AnnoTree is derived from the recently developed Genome Taxonomy Database (GTDB; Release 03-RS86) (12). 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 prokaryotic genomes derived from metagenomic sources). This differentiates the GTDB taxonomy and AnnoTree from similar approaches that rely on the NCBI taxonomy (13), whose

hierarchy disagrees with several recent reconstructions of microbial phylogeny (14,15).

MATERIALS AND METHODS

Gene prediction, annotation and profile generation

Gene prediction was performed with Prodigal v2.6.3 (16). Prodigal was selected over other methods based on its top performance in a recent benchmarking study (17) and for consistency with GTDB's own annotation pipeline (https://github.com/Ecogenomics/GTDBTk). The predicted genes were annotated using the Pfam v27.0 (10), TIGRFAM v15.0 (18), and UniRef100 (19) (downloaded March 6, 2018) databases. Pfam and TIGRFAM protein families were identified using HMMER v3.1b1 (20) with model specific cutoff values for the Pfam (-cut_gc) and TIGRFAM (-cut_nc) 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 (21) 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-to-subject and subject-to-query 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.

Web application development

AnnoTree has three components: a front-end, back-end, and a MySQL database. The latest AnnoTree database stores annotation data in the form of Pfam, TIGRFAM and KEGG confidence scores, protein sequence files, and the GTDB (Release 03-RS86) taxonomy and phylogenetic tree. The back-end is a Python Flask application to serve REST API endpoints. It converts JSON query to SQL statements. The front-end is a single page application using modern web frameworks such as D3, React, and Mobx. The tree and summary chart is drawn using D3.js, while other UI components are encapsulated by React. Mobx is a state management engine that triggers UI update whenever state variables change.

Calculation of phylogenetic conservation

The trait depth (τ_D) for each Pfam and KEGG annotation profile on the GTDB tree (Release 02-RS83) was calculated using the consenTRAIT algorithm (22) implemented in the castor R package (23). A trait was classified as phylogenetically conserved if the probability of encountering a profile with such a τ_D or higher is <5% (i.e., $P < 0.05$) 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

Lineage-specificity of a trait within a clade was measured using methods employed in statistical analysis of binary classification results. The precision of a lineage-specific classifier indicates the degree to which the trait is conserved within a lineage whereas the sensitivity indicates the exclusivity of the trait to a lineage. The precision and sensitivity of a trait T within clade C of GTDB tree P are calculated as follows:

- precision = [number of T-containing genomes in clade C] ÷ [number of genomes in clade C]
- sensitivity = [number of T-containing genomes in clade C] ÷ [number of T-containing genomes in tree P]

The F1 score combines the measures of precision and sensitivity to evaluate the ability of the clade to predict the occurrence of a trait within a phylogenetic tree. It is calculated as follows:

$$F1 = 2 \times [precision \times sensitivity] / [precision + sensitivity]$$

Here, Pfam domains and KEGG genes were classified as lineage-specific if there was at least one node whose precision and sensitivity were both $\geq$95%. The node with the greatest F1 score was assigned the root of the lineage-specific clade for that trait. The trait's taxonomic rank was selected as the lowest identical taxonomic rank between all genomes of the lineage-specific clade.

Calculation of homoplasy metrics

Parsimony-based homoplasy metrics were used to quantify phylogenetic scatter of traits. The consistency index (CI) and retention index (RI) were calculated for each Pfam and KEGG annotation profile with the GTDB tree (Release 02-RS83) using the phangorn R package (24). The homoplasy slope ratio (HSR) was calculated similarly with a custom script ('HSR.R' in https://bitbucket.org/doxeylabcrew/annotree-scripts) that utilizes the algorithm described in Meier *et al*. (25). The random homoplasy slope was calculated using 100 randomly-drawn presence/absence profiles with equal probability of presence and absence.

Taxonomic rank homoplasy enrichment analysis

Annotations contained within <50 genomes were removed before verifying taxonomic enrichment of homoplasic Pfam domains and KEGG genes. Taxonomic rank presence/absence profiles for each trait were generated for each taxonomic rank by combining the profiles of all encompassing genomes; '1' was assigned if at least one genome possessed the trait and '0' otherwise. Next, traits were ranked by increasing -ln(CI)/ln(family size). 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 by Nasir *et al*. (26). P values were obtained using the fisher.test function of R with the 'alternative' option set to 'greater'. The contingency table was given as follows:

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

where k is the number of different homoplasic traits within the rank, n is the number of different 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. P values were corrected for multiple tests at each taxonomic level using the Benjamini-Hochberg method (27).

RESULTS

To construct the AnnoTree database, we re-annotated all 28 941 prokaryotic genomes in the GTDB (Release 03-RS86) using a consistent annotation pipeline. Following gene prediction, we assigned functional annotations [Pfam protein families (10), TIGRFAM protein families (18) and KEGG Orthology (KO) identifiers (28)] to protein sequences using standard confidence score thresholds, resulting in 106 856 093 Pfam, 27 624 080 TIGRFAM, and 67 878 984 KEGG annotations. All taxonomic information, protein sequences, and functional annotations are stored in a back-end MySQL database for rapid retrieval by the front-end AnnoTree application (Figure 1). To enable phylogenetic visualization of all 28 941 prokaryotic genomes, AnnoTree divides the bacterial and archaeal trees of life into distinct views by each major taxonomic level. A user can explore the phylogenetic distribution of a trait anywhere from the phylum to genome level in either taxonomic domain. Additionally, AnnoTree can be used to explore custom trees and datasets (see Data Availability).

AnnoTree can be queried in several ways: by Pfam protein family, TIGRFAM protein family, KO term, or taxonomic name/id. Annotation queries can be filtered by their corresponding confidence scores such as E-value and percent alignment. Additionally, species that appear in a BLAST result can be visualized by uploading the BLAST XML2 output file directly. AnnoTree will then generate a 'painted' phylogeny using root-to-tip coloring for all lineages containing matches to the query (Figure 2). Visualizations are also accompanied by basic taxonomic information and distribution summary statistics based on GTDB nomenclature (Figure 2). Publication-quality SVG images, Newick formatted phylogenies for any selected subset of the tree, and taxonomic distribution tables of all queries can be downloaded for offline analysis or editing. Confidence scores (E-values) and options for downloading protein sequences for each annotation in a genome or lineage are displayed within a pop-up window when a colored node is selected on the tree.

Since all data is precomputed, users can explore the phylogenomic distribution of any combination of gene families within seconds. As an example, the recent metagenomics-driven discovery of commamox bacteria (29,30) can be reproduced through a simple AnnoTree query by searching for genomes possessing all three key genes that act as a signature for commamox activity: KO terms K00371 (*nxrB*), K10944 (*amoA*) and K10535 (*hao*). 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*) (Supplementary Figure S1).

As a second example, the recent discoveries 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 (31,32) (Supplementary Figure S2). Similarly, a search with the diphtheria toxin domains (PF02763 or PF02764) reveals homologs in related genera *Streptomyces* and *Austwickia*, again reproducing recent analyses (33) almost instantaneously (Supplementary Figure S3). These examples illustrate the use of AnnoTree as a hypothesis-generating tool by revealing distributions of gene families that may be new or unexpected to users.

Lineage-specific gene families

As an initial exploration of the data within AnnoTree, we examined the distributions of all 77 004 395 bacterial Pfam and KO annotations when mapped onto the bacterial GTDB tree of life (Release 02-RS83). Based on the phylogenetic conservation score (τ_D) (22), 68.1% of KO identifiers and 60.0% of Pfam protein families had significantly non-random phylogenomic distributions ($P < 0.05$), revealing a greater phylogenetic congruency for KO predictions than Pfam predictions. Next, we analyzed the distributions of Pfam and KO annotations, and used standard binary classification metrics to identify those with strong lineage-specificity (see Methods) (Supplementary Data File S1). Extremely lineage-specific families were identified as those with both very high ($\geq 95\%$) *precision* (percentage of genomes in the clade containing a trait) and very high ($\geq 95\%$) *sensitivity* (percentage of a trait-containing genomes occurring in the clade). Based on these criteria, we identified 358 (3.2%) Pfam protein families and 152 (0.9%) KO identifiers with lineage-specific distributions in Bacteria. We observed a trend in which lineage-specific KO identifiers and Pfam protein families increase in frequency from higher (e.g. phylum) to lower (e.g. species) taxonomic levels (Supplementary Figure S4), consistent with the idea that gene family taxonomic distributions tend to diversify over time and that HGT impacts evolution over short evolutionary timescales (34). Although lineage-specific families are relatively rare at high taxonomic levels, these cases often represent ancient, clade-defining bacterial innovations. Examples include K18955 (WhiB family transcriptional regulator) in the Actinobacteria, PF07542 (ATP12 chaperone) in the Alphaproteobacteria, and numerous photosynthesis-related genes within the Cyanobacteria (class *Oxyphotobacteria*).

Lineage-specific gene families can provide insights into the unique biology of their respective organisms. For example, eight lineage-specific Pfam and KO annotations were detected within the *Endozoicomonas* subtree, a clade of
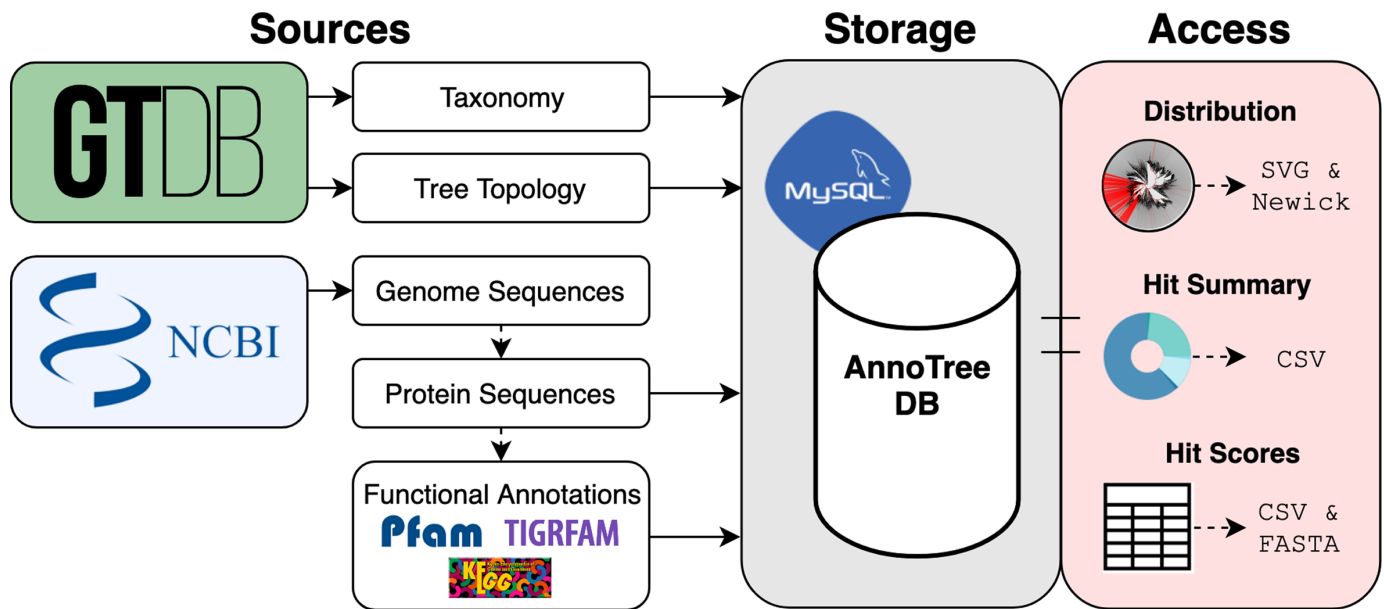
Figure 1. Data flow in the AnnoTree application. Raw values and computed features derived from data obtained from the GTDB is stored in a MySQL database that will be updated to match revisions made to the GTDB. Users can access data relevant to their queries in the form of figures and tables that are rendered in their browser. The figures themselves and the data used to generate them can be downloaded in various file formats from the AnnoTree interface.

endosymbiotic bacteria that inhabit numerous marine eukaryotic hosts (35). Consistent with possible utilization of host processes, the lineage-specific genes detected within this clade appear to be of eukaryotic origin and include genes involved in cytoskeletal organization (PF01302), eukaryotic cell–cell signaling (PF00812), apoptosis inhibition (K010343, K010344, K04725, PF07525) and eukaryotic proteolysis (K01378). Given the occurrence of numerous lineage-specific gene families in *Endozoicomonas*, we asked whether lineage-specific gene families may be overrepresented in certain taxa or branches of the bacterial tree. Indeed, lineage-specific genes were significantly enriched in specific taxonomic groups. Notable examples include 37 Pfam protein families within the *Bacillus_A* genus, and 19 Pfam protein families within the Actinobacteria that are largely composed of proteins of unknown function. We also observed an overrepresentation of lineage-specific gene families in numerous well-studied pathogens (e.g. *Bordetella*, *Helicobacter*, *Legionella* and *Vibrio*) (Supplementary Figures S5–S7; Supplementary Data File S1). This is in part due to the presence of lineage-specific virulence factors and toxins, but is also likely influenced by annotation bias towards organisms of biomedical interest (36).

Gene families with patchy distributions

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 (37,38). To investigate this further, we ranked all Pfam and KEGG annotations according to their phylogenetic patchiness, determined by homoplasy score (total number of gains

and losses by parsimony) normalized by gene family size after filtering out traits with family size <50 (Supplementary Data File S2, see Materials and Methods). Next, we grouped KO terms into their higher-level functional categories for visual comparison of broader trends (Figure 3, Supplementary Data File S3). Not surprisingly, 'viral' (bacteriophage) genes ranked the highest in homoplasy in both Pfam and KEGG annotations, and therefore are the single most phylogenetically scattered class of genes in bacteria. In contrast, gene functions with extremely low homoplasy include sporulation, photosynthesis, and core processes such as transcription, replication and protein synthesis (Figure 3). Highly scattered genes showed significant overrepresentation among specific taxonomic groups such as the genera *Pseudomonas_E, Streptomyces,* and *Mycobacterium* (Supplementary Data Files S4 and S5), suggesting that these taxa may be taxonomic 'hotspots' of HGT.

We then examined in more detail the top 100 gene families that showed the most scattered distributions across the bacterial tree. Not surprisingly, this list of gene families is dominated by transposases, CRISPR- and bacteriophage-associated gene families (Supplementary Data File S2). Numerous gene families of unknown function were included among the most patchy gene families, but further examination revealed that most of these genes are likely bacteriophage-derived. The extreme phylogenetic patchiness of bacteriophage and CRISPR genes is not only consistent with their known evolutionary dynamics but could also reflect the ongoing 'arms race' between these two opposing biological forces (phage infection versus phage defense). Other biologically relevant members of the 1% most highly scattered KO genes include: K19057-K19059 (*merC*, *merD*, and *merR* of the *mer* operon) for mercury resistance; K19155 and K19156, components of a toxin-antitoxin sys-
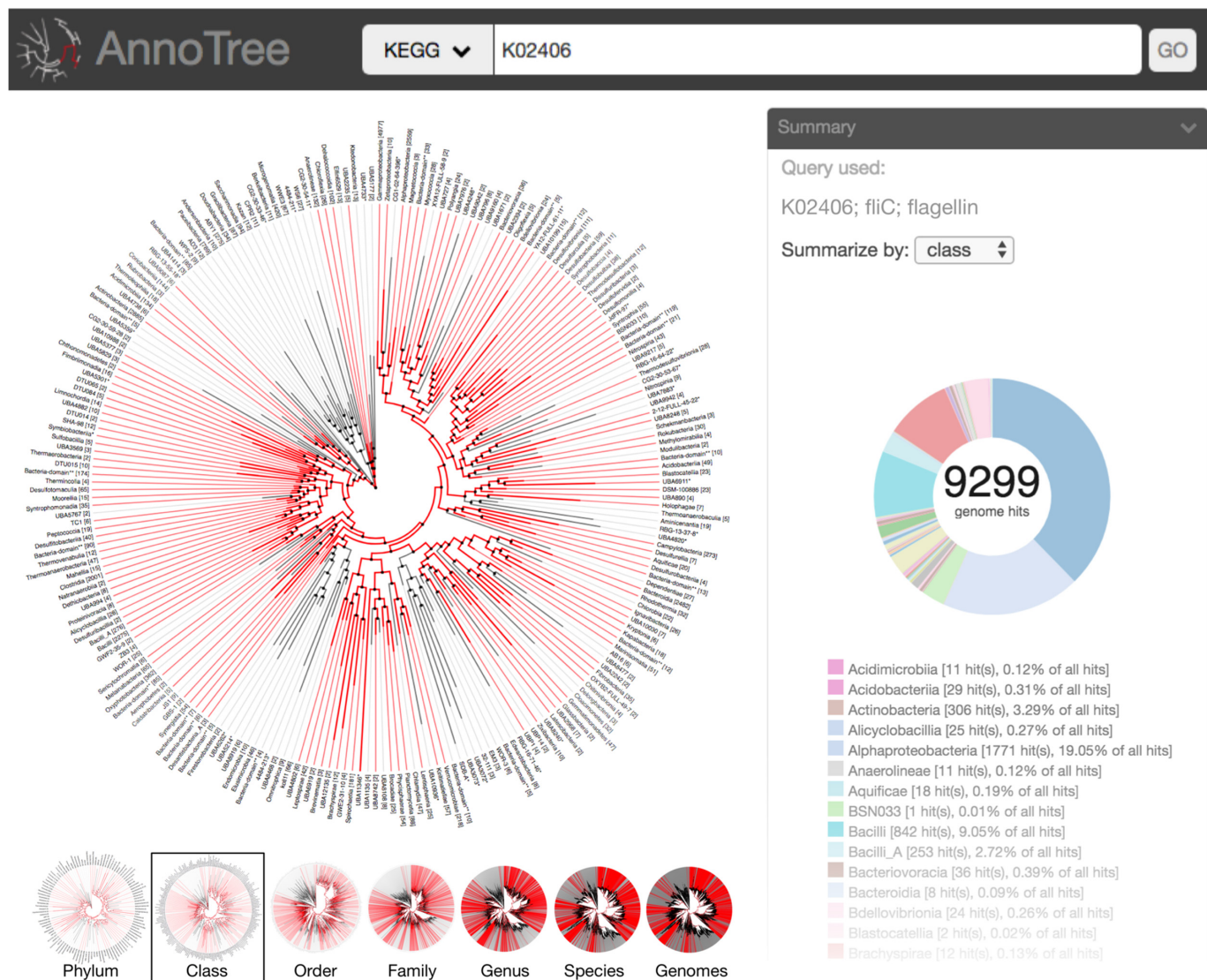
Figure 2. AnnoTree interface overview. AnnoTree can be queried with any number of KO identifiers, Pfam families, Tigrfam families, or NCBI taxon identification numbers to display a mapping of those traits on the GTDB tree at any resolution. Lineages containing at least one genome with the query annotation(s) are highlighted in red. A circle chart displays a taxonomic summary of the genomes containing the flagellin gene (KO identifier: K02406) at a chosen taxonomic level. Smaller trees below show the interactive view when different taxonomic levels are selected by the user. When a highlighted node is clicked, a window appears (not shown in figure) displaying basic taxonomic information, zooming options, and annotation confidence scores.

tem characterized in *E. coli*; K15943, K15945, and K16411 for polyketide antibiotic biosynthesis; and K19173-K19175 for DNA backbone S-modification (phosphorothioation) (Supplementary Data File S2).

Reductive dehalogenases

As a case study for the hypothesis generation and data mining strengths of AnnoTree, we selected a gene family of significant biological interest that ranked among the top percentile of homoplasy scores: *pcpC*; tetrachloro-p-hydroquinone reductive dehalogenase (K15241) Supplementary Data File S2). As key enzymes in bioremediation of chlorinated solvents, there has been extensive characterization of the diversity and phylogenomic distribution of reductive dehalogenases (Rdhs) and organohalide respiring organisms (39). Using AnnoTree, we compiled a dataset of

Rdh genes and associated taxa using Pfam query PF13486. Our analysis produced a comprehensive dataset of 1,299 putative Rdh genes from 385 genera and 38 phyla (Supplementary Table S1, Figures S8, S9), which not only recapitulates the known diversity of Rdh-associated phyla, but significantly expands it. In comparison, a manually-curated Rdh-specific database contains 264 Rdh genes from only 19 genera and 6 phyla (39), less than 15% of the total diversity identified by AnnoTree (Supplementary Table S1). The AnnoTree-derived dataset includes several newly predicted *rdh*-encoding taxa discovered from metagenome-assembled genomes (Supplementary Table S2), including the candidate phyla KSB1 (4 of 6 genomes, *rdh* copy number = 1) and UBP10 (7 of 14 genomes, *rdh* copy number = 1), as well as *Rhodospirillales* UBA2165 (*rdh* copy number = 13) and *Acidobacterium* UBA2161 (*rdh* copy number = 8) (Supplemen-
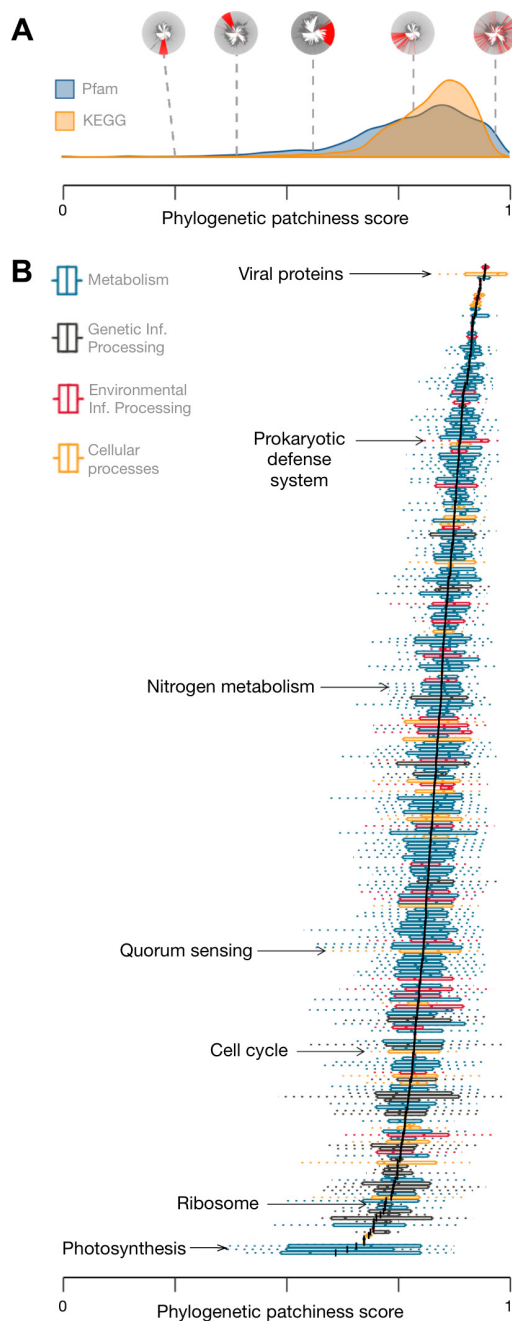
Figure 3. Phylogenetic patchiness of annotations inferred using AnnoTree. Phylogenetic patchiness was computed for each KEGG KO identifier and Pfam protein family using the consistency index (CI), a common homoplasy metric representing the inverse of the minimum possible number of state changes (trait gain or loss) given the tree topology. The final phylogenetic patchiness score is equal to $-\log(CI)/\log(\text{family size})$ where family size is the total number of genomes containing the trait. (**A**) Density plot showing the distribution of phylogenetic patchiness scores of Pfam protein families and KO identifiers with different visual examples of varying patchiness (red = present; gray = absent). The phylogenetic distribution plots are, from left to right: K10922 (transmembrane regulatory protein ToxS), K18955 (WhiB transcriptional regulator), PF01848 (ATP12 chaperone), PF01848 (Hok/Sok antitoxin system), and K07495 (putative transposase). (**B**) Mean-sorted box plots containing phylogenetic patchiness scores of KO identifiers in their respective KEGG pathways and KEGG BRITE categories. The mean patchiness score of a set of KO identifiers in a KEGG pathway or KEGG BRITE category is indicated by a black line.

tary Figure S9, Table S2). The novel organisms with high *rdh* copy numbers are potential obligate organohalide respirers and may be valuable for remediation efforts. By revealing both known and potentially novel groups of organohalide respiring bacteria, the Rdh case study highlights the ability of AnnoTree to capture a broad and complete taxonomic diversity of a gene family, with accompanying hypothesis generation around the evolution and ecology of a function of interest.

DISCUSSION

Ultimately, by combining functional annotation data with evolutionary data, AnnoTree provides an automated framework for users to explore the distribution of function across the bacterial and archaeal phylogenies. These visualizations allow users to investigate a wide variety of research questions concerning their genes and functions of interest. As starting points for future analyses, we have assessed bacterial Pfam and KEGG annotations based on phylogenetic conservation, homoplasy, and lineage-specificity. However, while AnnoTree provides a snapshot of gene occurrence, additional sequence and phylogenetic analyses are required to validate many of these predictions. The AnnoTree database will also be continuously and automatically updated to reflect revisions of the GTDB taxonomy as the data become available. We anticipate that AnnoTree will become a valuable resource for exploring the evolution and phylogenomic distribution of genes and functional traits across the tree of life.

DATA AVAILABILITY

The AnnoTree application is available at http://annotree.uwaterloo.ca. All software and data used within AnnoTree can be downloaded at: http://annotree.uwaterloo.ca/downloads.html, and source code can be downloaded at: https://bitbucket.org/account/user/doxeylabcrew/projects/AN. Documentation for AnnoTree, including instructions on use of custom trees and datasets, is located at https://annotree-docs.readthedocs.io. Additional data for the genomes and taxonomy derived from the GTDB can be found at: http://gtdb.ecogenomic.org/downloads.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

REFERENCES

1. Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,Ja., Wu,D., Paulsen,I., Nelson,K.E., Nelson,W. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.
2. Demuth,J. and Hahn,M. (2009) The life and death of gene families. *Bioessays*, **31**, 29–39.
3. Andersson,J.O., Hirt,R.P., Foster,P.G. and Roger,A.J. (2006) Evolution of four gene families with patchy phylogenetic distributions: influx of genes into protist genomes. *BMC Evol. Biol.*, **6**, 27.
4. Ravenhall,M., Skunca,N., Lassalle,F. and Dessimoz,C. (2015) Inferring horizontal gene transfer. *PLoS Comput. Biol.*, **11**, e1004095.
5. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST plus: architecture and applications. *BMC Bioinformatics*, **10**, 1.
6. 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 Res.*, **44**, W242–W245.
7. 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.
8. Yang,S. and Bourne,P.E. (2009) The Evolutionary History of Protein Domains Viewed by Species Phylogeny. *PLoS One*, **4**, e8378.
9. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.
10. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
11. 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 Struct. Funct. Bioinforma*, **85**, 72–77.
12. 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. *Nat. Biotechnol.*, **36**, 996–1004.
13. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
14. Bromberg,R., Grishin,N.V. and Otwinowski,Z. (2016) Phylogeny reconstruction with alignment-free method that corrects for horizontal gene transfer. *PLoS Comput. Biol.*, **12**, e1004985.
15. 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. *et al.* (2016) A new view of the tree of life. *Nat. Microbiol.*, **1**, 16048.
16. 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.
17. Tripp,H.J., Sutton,G., White,O., Wortman,J., Pati,A., Mikhailova,N., Ovchinnikova,G., Payne,S.H., Kyrpides,N.C. and Ivanova,N. (2015) Toward a standard in structural genome annotation for prokaryotes. *Stand. Genomic Sci.*, **10**, 45.
18. Haft,D.H. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.
19. 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.
20. Eddy,S.R. (2011) A new generation of homology search tools based on probabilistic inference. *Genome Inform.*, **23**, 205–211.
21. Buchfink,B., Xie,C. and Huson,D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
22. Martiny,A.C., Treseder,K. and Pusch,G. (2013) Phylogenetic conservatism of functional traits in microorganisms. *ISME J.*, **7**, 830–838.
23. Louca,S. and Doebeli,M. (2018) Efficient comparative phylogenetics on large trees. *Bioinformatics*, **34**, 1053–1055.
24. Schliep,K.P. (2011) phangorn: Phylogenetic analysis in R. *Bioinformatics*, **27**, 592–593.
25. Meier,R., Kores,P. and Darwin,S. (1991) Homoplasy slope ratio: A better measurement of observed homoplasy in cladistic analyses. *Syst. Biol.*, **40**, 74–88.
26. Nasir,A., Kim,K.M. 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 Evol. Biol.*, **12**, 156.
27. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.
28. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
29. 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.
30. Daims,H., Lebedeva,E.V., Pjevac,P., Han,P., Herbold,C., Albertsen,M., Jehmlich,N., Palatinszky,M., Vierheilig,J., Bulaev,A. *et al.* (2015) Complete nitrification by Nitrospira bacteria. *Nature*, **528**, 504–509.
31. Mansfield,M.J., Adams,J.B. and Doxey,A.C. (2015) Botulinum neurotoxin homologs in non-Clostridium species. *FEBS Lett.*, **589**, 342–348.
32. Mansfield,M.J., Wentz,T.G., Zhang,S., Lee,E.J., Dong,M., Sharma,S.K. and Doxey,A.C. (2019) Bioinformatic discovery of a toxin family in Chryseobacterium piperi with sequence similarity to botulinum neurotoxins. *Sci. Rep.*, **9**, 1634.
33. 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 Lett.*, **592**, 2693–2705.
34. McDonald,B.R. and Curriea,C.R. (2017) Lateral gene transfer dynamics in the ancient bacterial genus Streptomyces. *MBio*, **8**, e00644-17.
35. 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. *Appl. Microbiol. Biotechnol.*, **100**, 8315–8324.
36. Haynes,W.A., Tomczak,A. and Khatri,P. (2018) Gene annotation bias impedes biomedical research. *Sci. Rep.*, **8**, 1362.
37. Ochman,H., Lawrence,J.G. and Grolsman,E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.
38. Eisen,J.A. (2000) Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Curr. Opin. Genet. Dev.*, **10**, 606–611.
39. Hug,L.A., Maphosa,F., Leys,D., Löffler,F.E., Smidt,H., Edwards,E.A. and Adrian,L. (2013) Overview of organohalide-respiring bacteria and a proposal for a classification system for reductive dehalogenases. *Philos. Trans. R. Soc. B Biol. Sci.*, **368**, 20120322.