AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems

| Computational Biology | Resource Report

Annotating microbial functions with ProkFunFind

Keith Dufault-Thompson,[1] **Xiaofang Jiang**[1]

AUTHOR AFFILIATION See affiliation list on p. 9.

ABSTRACT Analyzing microbial genomes has become an essential part of microbiology research, giving valuable insights into the functions and evolution of microbial species. Identifying genes of interest and assigning putative annotations to those genes is a central task in genome analysis, and a plethora of tools and approaches have been developed for this task. The ProkFunFind tool was developed to bridge the gap between these various annotation approaches, providing a flexible and customizable search approach to annotate microbial functions. ProkFunFind is designed around hierarchical definitions of biological functions, where individual genes can be identified using heterogeneous search terms consisting of sequences, profile hidden Markov models, protein domains, and orthology groups. This flexible and customizable search approach allows for searches to be tailored to specific biological functions, and the search results are output in multiple formats to facilitate downstream analyses. The utility of the ProkFunFind search tool was demonstrated through its application in searching for bacterial flagella, which are complex organelles composed of multiple genes. Overall, ProkFunFind provides an accessible and flexible way to integrate multiple types of annotation and sequence data while annotating biological functions in microbial genomes.

IMPORTANCE Genome sequencing and analysis are increasingly important parts of microbiology, providing a way to predict metabolic functions, identify virulence factors, and understand the evolution of microbes. The expanded use of genome sequencing has also brought an abundance of search and annotation methods, but integrating the information from these different methods can be challenging and is often done through *ad hoc* approaches. To bridge the gap between different types of annotations, we developed ProkFunFind, a flexible and customizable search tool incorporating multiple search approaches and annotation types to annotate microbial functions. We demonstrated the utility of ProkFunFind by searching for gene clusters encoding flagellar genes using a combination of different annotation types and searches. Overall, ProkFunFind provides a reproducible and flexible way to identify gene clusters of interest, facilitating the meaningful analysis of new and existing microbial genomes.

KEYWORDS genome annotation, genomics, functional annotation

Sequence analysis has become a ubiquitous component of biological and biomedical research, providing information on the roles and evolution of microbes in their environments. While this has greatly expanded our collective knowledge about microbial diversity, it has also highlighted our limited understanding of many microbial functions (1). The large-scale analysis of microbial genomes has been used to study the evolution and distribution of multiple microbial pathways, providing valuable insights into the ecological importance of these pathways and their potential impacts on human health (2–5). This kind of analysis can be essential in biomedical research, where understanding the functions of the human microbiome is a key component in developing personalized medicine techniques (6) and probiotics (7).

Editor Samuel Chaffron, CNRS Delegation Bretagne et Pays de Loire, Nantes, France

Address correspondence to Xiaofang Jiang, xiaofang.jiang@nih.gov.

The authors declare no conflict of interest.

See the funding table on p. 9.

Received 10 January 2024
Accepted 22 January 2024
Published 16 February 2024

The identification of features of interest in microbial genomes relies on gene annotation, assigning putative functions to genes based on their sequences, predicted features, and similarity to other genes. Various tools have been developed to facilitate the annotation process, providing multiple ways to assign putative functions to new genes. Sequence and hidden Markov model (HMM)-based searches are common approaches and involve comparisons to other sequences or models using tools like BLAST (8) and HMMER (9). Tools like InterProScan can be used to functionally annotate proteins based on their similarity to curated databases of protein domains and families (10). Finally, grouping new sequences into larger groups of orthologs using tools like KofamScan (11) or eggNOG-mapper (12) can be an effective way to annotate new sequences in an evolutionary and functional context. These different approaches have distinct benefits and drawbacks, and in many cases, the annotation of multiple genes encoding biological pathways will rely on a combination of different approaches.

In most cases, the task of integrating different annotation data during genome analysis falls to individual researchers. This process often involves custom scripts and workflows, leading to a lack of reproducibility and standardization. Approaches that integrate different data types and facilitate reproducible and flexible searches are needed. To facilitate the efficient and customizable search for functions in genome data, we developed the ProkFunFind tool. ProkFunFind is a flexible search tool designed to facilitate the identification and characterization of different functions in microbial genomes. This tool incorporates multiple search and annotation approaches, allowing for the identification of functions based on sequences, HMM profiles, protein domains, and multiple common orthology definitions. ProkFunFind is based on searching for biological functions rather than single genes, providing a way to denote the relationship between different genes within larger functional definitions such as metabolic pathways. In addition, we have designed ProkFunFind to be modular and extensible, allowing for the incorporation of additional search approaches, annotation methods, and downstream analyses. Overall, ProkFunFind is an important step toward developing ways to analyze and interact with genome data that incorporate the multiple sources of information that are now easily accessible and commonly used.

MATERIALS AND METHODS

Implementation

ProkFunFind is publicly available on GitHub (https://github.com/nlm-irp-jianglab/ProkFunFind.git) along with additional documentation describing the tool and a tutorial with working examples (https://prokfunfind.readthedocs.io/en/docs-and-tests/index.html). The tool is implemented in Python and has been designed to be modular and easily extended to incorporate new annotation formats and search approaches in the future. Depending on the types of search approaches and annotation features being used, ProkFunFind also incorporates multiple other tools, including BLAST (8), HMMER (9), InterProScan (10), KofamScan (11), eggNOG-mapper (12), Prokka (13), and Bakta (14). To perform a search, users can either download precomputed annotation information, like what is available for genomes from the MGnify database (15), prepare their own annotation files, or use the ProkFunAnnotate Snakemake pipeline (https://github.com/nlm-irp-jianglab/ProkFunAnnotate) to generate annotation files for new or existing genomes.

The ProkFunAnnotate pipeline performs a function similar to Prokka (13) and Bakta (14), with the additional step of providing annotations based on eggNOG-mapper and KofamScan. This Snakemake pipeline takes a user-provided collection of genomes and uses Prokka (version 1.14.5) (13) to predict genes and provide preliminary annotations and then produces KEGG Orthology (KO) Annotations and ortholog group annotations using KofamScan (version 1.3.0) (11) and eggNOG-mapper (version 2.1.12) (12). The ProkFunAnnotate pipeline can also be used to generate annotations for an already existing genome, using the genome sequence and a set of genes in the GenBank format file. The ProkFunFind tool uses the default output formats from these other tools to make setting up and performing searches on new genomes as straightforward as possible.

ProkFunFind performance

To assess the performance of ProkFunFind, searches were done using six different function definitions, each based on different annotation types, on a set of 5,000 randomly selected Genome Taxonomy Database (GTDB) representative genomes. The six function definitions are available as part of the ProkFunFind tutorial materials (https://github.com/nlm-irp-jianglab/prokfunfind-tutorial/tree/master/queries) and consist of different representations of 12 genes found in a gene cluster related to the production of the metabolite equol (2). Each search was run independently on a high-performance computing node with 20 central processing units and 6 GB of memory. These searches do not include generating the annotation files for each genome, which must be done before the search is performed. The average time to process a genome was then calculated for each search (Table S1).

Search for flagellar genes

To demonstrate a use case of ProkFunFind, a search for flagella gene clusters was performed using a mixture of HMM profiles, InterProScan predictions, and Clusters of Orthologous Genes (COGs). First, a function definition was designed to represent the essential genes that comprise the bacterial flagella (16). The definition includes 18 genes/gene families organized into 5 functional categories required to assemble the flagellar rod, hook, C ring, motor and MS ring, and export complexes. A combination of protein domain signatures from the Pfam (17), PANTHER (18), and TIGRFAMs (19) databases, hidden Markov models from the National Center for Biotechnology Information (NCBI) Protein Family Models database (20), and COGs (20, 21) was curated to search for the flagella-related genes in the Bacillota and Pseudomonadota phyla of bacteria (Table S2).

The microbial traits data set provided by Madin et al. (22) was used to identify bacterial species with and without flagella in the GTDB. The trait data were mapped to the GTDB representative species using the NCBI species taxonomy identifiers. Only species with their motility annotated as “flagella” or “no” in the trait database were included, giving 543 Bacillota species (68 with flagella and 475 without flagella) and 690 Pseudomonadota species (246 with flagella and 444 without flagella). The representative genomes from these species were then annotated with eggNOG-mapper using the ProkFunAnnotate pipeline and InterProScan (version 5.63–95.0).

The curation of the search profiles for the Bacillota and Pseudomonadota phyla was an iterative process that consisted of identifying search terms, comparing the predicted presence or absence of each gene to the ground truth data set, and adjusting the e-value threshold used to filter hits. For the identification of search terms, an initial annotation of the *Escherichia coli* K12 (U00096.3) and *Bacillus subtilis* (GCF_000009045.1) genomes was done using the ProkFunAnnotate pipeline and InterProScan (version 5.63–95.0). Search terms associated with each of the flagellar genes from these genomes were then identified based on the annotations of those genes from KofamScan, eggNOG-mapper, and InterProScan. The search terms were examined individually to identify terms that captured hits in the largest number of genomes annotated as having flagella possible, and e-value thresholds were chosen that minimized the number of hits in genomes annotated as lacking flagella. This process was repeated multiple times to generate function definitions that reasonably captured the presence of flagella in the Bacillota and Pseudomonadota phyla.

The ProkFunFind search was performed using search configuration files for the Bacillota and Pseudomonadota genomes (available as part of the ProkFunFind github repository: https://github.com/nlm-irp-jianglab/ProkFunFind.git). Hidden Markov models were compared to the genomes using HMMSCAN (version 3.3.2) (9), protein domain signatures were identified using the InterProScan results, and COGs were assigned based on the eggNOG-mapper annotations. The search results, indicating the presence or absence of each gene and the completeness of the overall flagella function, were compared to the trait database annotations and were mapped onto a subset of the GTDB reference phylogenetic tree using iTOL (version 6) (23).

RESULTS

ProkFunAnnotate overview

The ProkFunAnnotate pipeline was developed as a convenient way to annotate a genome of interest and produce a set of annotation files in a consistent format that can be used with ProkFunFind. ProkFunAnnotate is not a novel approach for gene calling or annotation but instead provides a self-contained pipeline to run multiple commonly used annotation programs. The ProkFunAnnotate pipeline starts with gene calling and preliminary annotation using Prokka, generating the associated genome files, including gene and protein fasta files, along with a preliminary annotation file generated by Prokka. Additional annotation files are then generated using KofamScan and eggNOG-mapper, providing a formatted collection of files that can be used with the ProkFunFind search tool. This collection of files provides users with everything needed to perform searches with ProkFunFind with any combination of protein sequences, HMM profiles, COGs, and KOs. These annotations also only need to be generated once, and any number of searches can then be performed, reusing the same data for the genomes.

ProkFunFind overview

The ProkFunFind tool is designed to facilitate searching for groups of genes in a genome that are related to a biological function (Fig. 1). ProkFunFind feature definitions are combinations of different kinds of queries that can be used to represent complex functions. This allows each search to be customized to address the problem of interest in the best way. For example, one gene in a pathway may only have a few characterized reference sequences, making a BLAST-based search a good starting point, while another gene may be a part of a well-defined protein family, allowing it to be identified based on an ortholog group assignment. This heterogeneous query structure can be refined by adding filtering parameters, providing a highly flexible and customizable search approach. The search output of ProkFunFind provides a basic summary of the presence and absence of the defined function and multiple output files that facilitate downstream analyses and visualization.

Search space

The search space for ProkFunFind consists of a genome or set of genomes with a standardized set of annotation information. The core information needed for each genome is a nucleotide fasta file containing the genome sequence, a GFF-formatted file with the gene coordinates, and a protein fasta file with the translated gene sequences. In addition to this basic genome information, ProkFunFind can utilize annotation information generated from general gene calling and annotation tools like Bakta (14) or Prokka (13), or from specialized annotation tools like eggNOG-mapper (12), KofamScan (11), and InterProScan (10), to perform searches for COGs, KO identifiers, and protein domain signatures, respectively. Only the annotation data associated with the search terms being used is required when running a search, allowing users to avoid generating and storing extra annotation files. The annotation files are expected to be in the standard output formats from these programs to make it as easy as possible to generate the files and to make it possible to run searches on genomes with precomputed annotation data from databases like the MGnify database (15). The ProkFunAnnotate Snakemake pipeline was also developed to provide a standardized way to generate the needed genome and annotation files for new or unannotated genomes. Having this collection of genome information facilitates flexible and customizable searches across large data sets without

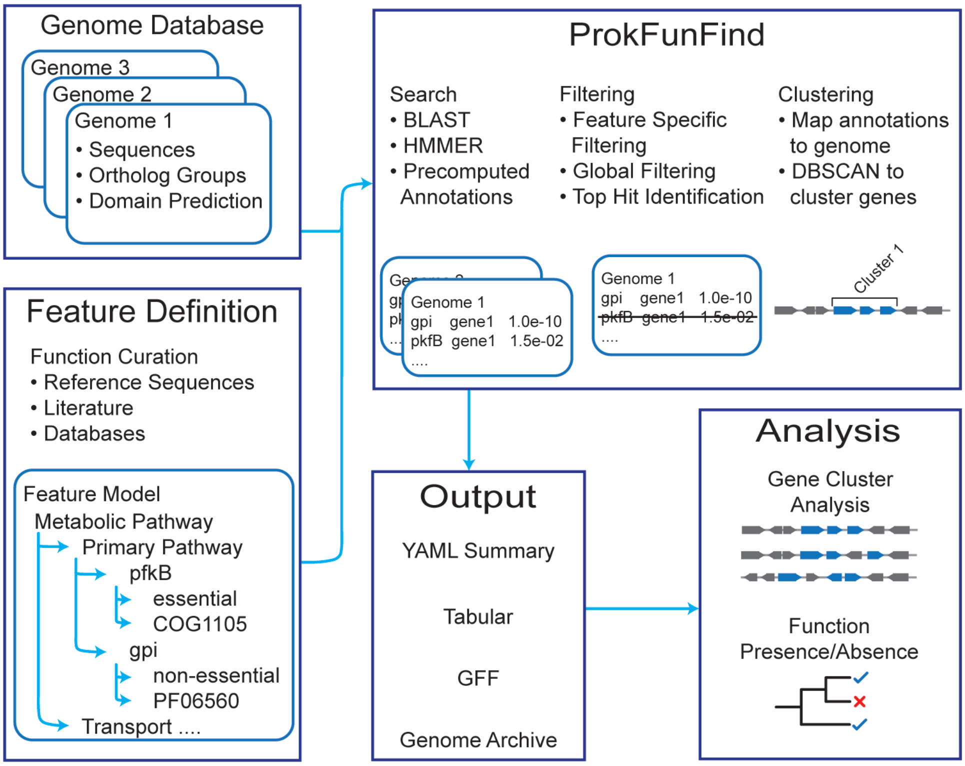

FIG 1 A diagram showing the ProkFunFind pipeline components and workflow.

having to rely on *ad hoc* approaches to generating, parsing, and filtering the data associated with each genome.

Definition of a biological function

ProkFunFind is designed around searching for a collection of genes related to a biological function. In contrast to most search approaches, this emphasizes the biological function and how its components are related to each other. The biological function is represented by a feature model, which is a hierarchical definition of the function, its different components, the associated genes, and the search terms that can be used to identify those genes. For example, the metabolism of a specific compound would be an overall function, with the components being discrete functions (e.g., transport, metabolic enzymes, and regulation) and each component would be associated with one or more distinct genes and search terms. This organization allows for the flexible definition of different kinds of functions in a way that best fits the biological problem. The function definition is formatted in a user-friendly and readable YAML format, making it straightforward to write and edit.

The lowest level of the function definition is the genes that are being searched for. Each of these genes is associated with one or more search terms that define how the search is performed. These search terms can be amino acid sequences, profile HMMs, KEGG or COG orthology IDs, or protein domain signatures. The genes can be associated with one or multiple search terms of the same or different types. To further customize the search, each search term can be filtered using the relevant parameters, including e-value, bit score, or percent identity, allowing for different genes to be identified using more or less strict parameters. Each component and gene can also be designated as being essential or non-essential to the overall function, allowing for additional accessory genes that are not core parts of the function but still of biological interest to be searched for.

Detection of candidate genes

The detection of putative genes associated with a function is done by examining the search and annotation results and applying a set of filters to identify the best hits for each component of the function. If sequences or profile HMMs are used in the search, then a search using *BLASTp* or *hmmscan* is performed first. The results of the searches and the annotation results from InterProScan, KofamScan, and eggNOG-mapper are then parsed to identify any genes that match the search terms defined in the function definition. A set of user-defined thresholds are then applied to each gene to remove low-quality hits from the results. These thresholds can include cutoffs for the e-values, percent identity for sequence-based searches, and annotation scores for KofamScan. User-defined thresholds can be customized for each search term in the function, or defaults can be set for any terms using that search approach. All putative hits for each component of the function are then summarized and added to a central genome object for subsequent analysis.

Output and performance

ProkFunFind searches return hits to multi-component functions, and understanding if these putative hits are located in similar regions of the genome can be biologically informative. To capture these relationships between the identified genes, ProkFunFind utilizes the DBSCAN algorithm to detect groups of hits that are located near each other in the genome. The clustering settings are adjustable, allowing for the fine-tuning of what qualifies as a “gene cluster” to suit different biological questions best. After this clustering step is performed, the annotations and cluster information are reported to the user. To mirror the flexibility of the search input, the output of ProkFunFind was designed to provide the putative hits and annotation information in multiple formats to facilitate downstream analyses. These include an easily parsable tab-separated table format, a GFF-formatted file containing information about each putative hit and their locations on the genomes, a pickle-formatted archive that can be interacted within Python, and a YAML-formatted summary of each component’s presence and absence. These output formats can be loaded into other analysis tools, allowing for additional downstream analysis and visualization.

The performance of ProkFunFind was evaluated using a collection of search configurations using different query types in a search against 5,000 randomly selected GTDB representative genomes. On average, the ProkFunFind searches took between 4.7 and 8.5 s per genome across the different search approaches (Table S1). The sequence and HMM profile-based searches were all relatively quick, while the other searches, which require parsing through the precomputed annotation files, took slightly longer. Overall, ProkFunFind searches are relatively quick and lightweight, allowing users to efficiently scan collections of genomes for features of interest.

Using ProkFunFind to detect flagellar genes

To demonstrate a use case for ProkFunFind, a search was performed for flagellar genes in a collection of genomes from the Bacillota and Pseudomonadota phyla of bacteria. Flagella are complex organelles that are involved in bacterial motility (24). Flagella synthesis and activity involve dozens of distinct genes, with a core set of 21 genes typically being conserved across all motile bacteria (16). Feature models representing these core genes, grouped into five functional categories, were developed for both phyla using a combination of HMM profiles, protein domain signatures, and COGs. The ProkFunFind search results were compared to motility trait annotations in a microbial traits database (22) to assess if the search captured known flagellated bacteria.

For the Bacillota phylum, the ProkFunFind search identified all of the core flagella genes in 66 of the 68 genomes annotated as having flagella in the trait database (Fig. 2; Fig. S1). The Bacillales order contained a majority (38) of the species annotated as having flagella, with the ProkFunFind search identifying flagellar genes in 37 of these

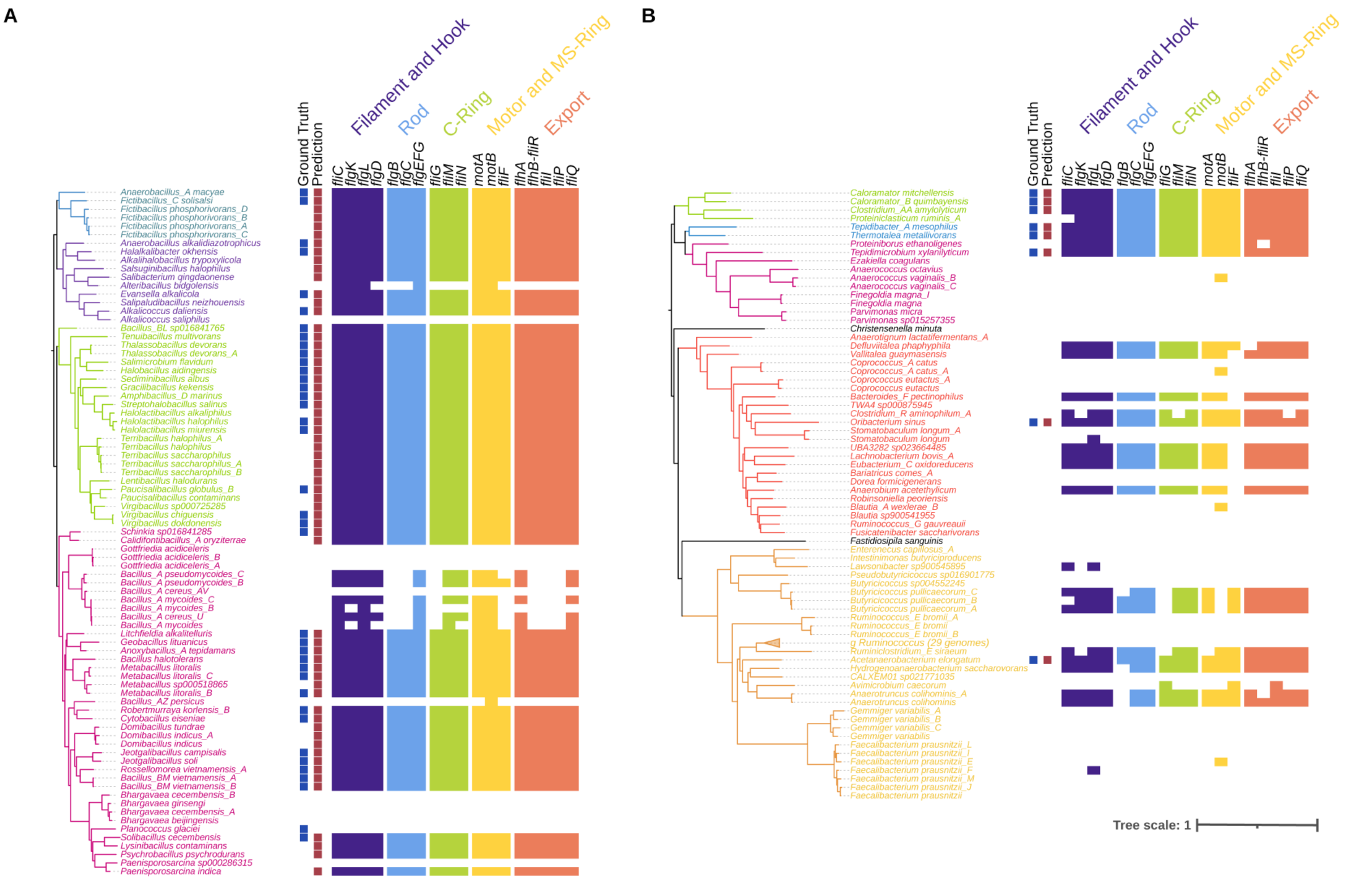

FIG 2 Subset phylogenetic trees showing the detected presence of flagellar genes in the Bacillales orders (Bacillales, Bacillales D, Bacillales G, and Bacillales H in GTDB) (A) and Clostridia class (B). Blue and red boxes next to the tree indicate if the species were annotated as having flagella in the microbial traits database (Ground True, blue box) and predicted to have all core flagellar genes (Prediction, red box). Colored boxes for each gene are used to indicate the predicted presence of the genes. Species on the trees are colored by order.

species (Fig. 2A). Gene clusters were identified in an additional 25 of the species that had been annotated as not having flagella in the trait database, but these species were almost always closely related to other species that are known to have flagella. In contrast, the Clostridia class contained only a few species annotated as having flagella, with all eight of them being identified in the ProkFunFind search (Fig. 2B). The search in the Pseudomonadota phylum was more variable, with 189 out of the 246 flagella annotated species being identified by the ProkFunFind search and flagellar genes being identified in an additional 88 species that were not annotated as having flagella in the trait database (Fig. S2).

The search for flagella resulted in the identification of flagellar genes in 81% of the 314 Bacillota and Pseudomonadota species that were annotated as having flagella in the microbial traits database. An additional 125 species had putative flagellar gene clusters but were annotated as not being motile in the traits databases. The differences between the trait database and the ProkFunFind predictions were further examined for the Bacillota genomes, which had 2 species that were falsely predicted to have incomplete sets of flagellar genes and 37 species that were falsely predicted to have flagellar genes (Table S3). For 10 of these species, there were differences between the trait database annotation and what was reported in the literature, with the literature either disagreeing with the trait database annotation or there being reports of strain-level differences in motility for that species. Three additional species were observed to have strain-level heterogeneity within the GTDB species cluster, with the representative genome having putative flagellar genes, while other genomes from that species were missing them. This highlights a shortcoming of this ground truth data set, where the microbial trait annotations are reported as a consensus annotation at a species level, not at the individual strain or genome level (22), resulting in the disagreements that are seen when looking at just the representative genomes for each species in GTDB.

Twenty-three of the falsely predicted Bacillota species had no clear explanation, with the search identifying reasonable flagellar gene clusters. These cases may partly be explained by the complex evolutionary history of flagellar genes. Flagellar gene clusters are subject to frequent mutation events (25), leading to significant heterogeneity in flagellar gene presence, the pseudogenization or genes, and resulting in partial flagellar gene clusters in some strains. Additionally, components of the flagellar complex are homologous to other systems, notably the Type 3 secretion system (26, 27), which may explain some of the partial sets of genes detected in many of the Pseudomonadota genomes. Overall, the flagellar search demonstrates the utility of the ProkFunFind search approach, using a heterogeneous set of search terms to identify complex sets of genes in diverse genomes.

DISCUSSION

ProkFunFind has been developed to bridge the gap between the numerous and varied annotation approaches that are commonly used in genome analysis, providing a flexible platform to perform searches with heterogeneous data. The rapid and continued development of new annotation approaches and the sequencing of new genomes have led to a flood of data. However, this abundance of data can lead to analytical difficulties, as combining information from multiple sources is often left to individual researchers using *ad hoc* approaches. ProkFunFind provides a tool to combine different types of annotations and facilitate meaningful data exploration using different features.

One of the central concepts of ProkFunFind is the ability to perform searches using heterogeneous data types. Biological functions are often complex, involving multiple gene products that have different importance to the overall function. Similar concepts have been applied in the MacSyFinder and ConJScan tools, where collections of HMM profiles are used to search for complex systems of genes like those related to conjugation and secretion systems, providing sensitive ways to search for the functions (28, 29). The challenge that comes from searching for genes associated with complex functions is that one type of search is often not sufficient to capture each of the genes. Some genes may have well-defined orthology definitions in commonly used databases like KEGG, while others may only have one or two representative sequences and would be best searched for using a direct sequence search approach. ProkFunFind's function definition allows for a mixture of these search terms, allowing each search to be tailored to the function of interest. This flexible and customizable approach can lead to more meaningful searches and biological insights.

The other central concept in the ProkFunFind approach emphasizes searching for biological features rather than single genes. Many biological features are not defined by the presence or absence of a single gene but instead, involve multiple gene products and may have multiple other genes that serve accessory roles. To assess if a function is putatively present or not in a given genome requires searching for multiple genes in a way that accounts for how these genes are related to each other. The ProkFunFind query was designed to represent a collection of genes that are related to a function in a flexible way that can account for the relationships between different gene products and the essentiality of different components to the overall function. The utility of this approach to searching for functions was highlighted through the application of ProkFunFind to identify putative flagellar genes, which are part of a complex and multi-gene system. The benefit of searching for sets of genes related to a function and allowing for mixed types of search terms is highlighted by this example, where the flagellar genes have complex evolutionary histories, and identifying homologs of single genes is not informative in relation to the overall function.

Overall, ProkFunFind provides a useful tool that can aid in the analysis of biological systems. Genome analysis is becoming a central component of many biological studies, and methods that provide accessible ways to interact with genome data will be important moving forward. The flexibility and customizability of ProkFunFind allow it to be used when searching for simple and complex microbial functions while providing an interface to interact with multiple types of annotation data at the same time.

ACKNOWLEDGMENTS

This work utilized the computational resources of the NIH HPC Biowulf cluster. (http://hpc.nih.gov).

K.D. and X.J. are supported by the Intramural Research Program of the NIH, National Library of Medicine.

AUTHOR AFFILIATION

[1]National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA

AUTHOR ORCIDs

Keith Dufault-Thompson http://orcid.org/0000-0002-0991-2255
Xiaofang Jiang http://orcid.org/0000-0002-0955-8284

FUNDING

Funder	Grant(s)	Author(s)
HHS | National Institutes of Health (NIH)		Xiaofang Jiang Keith Dufault-Thompson

AUTHOR CONTRIBUTIONS

Keith Dufault-Thompson, Conceptualization, Formal analysis, Software, Writing – original draft, Writing – review and editing | Xiaofang Jiang, Conceptualization, Formal analysis, Software, Supervision, Writing – original draft, Writing – review and editing

DATA AVAILABILITY

ProkFunFind is available on GitHub (https://github.com/nlm-irp-jianglab/ProkFunFind) alongside the ProkFunAnnotate pipeline (https://github.com/nlm-irp-jianglab/ProkFunAnnotate). All genome sequences used in the analysis are available as part of GTDB (https://gtdb.ecogenomic.org/).

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Fig. S1 (mSystems00036-24-s0001.pdf). Presence of flagella in Bacillota.
Fig. S2 (mSystems00036-24-s0002.pdf). Presence of flagella in Pseudomonadota.
Legends (mSystems00036-24-s0003.docx). Supplemental material legends.
Supplemental Tables (mSystems00036-24-s0004.xlsx). Tables S1-S3.

REFERENCES

1. Jiao J-Y, Liu L, Hua Z-S, Fang B-Z, Zhou E-M, Salam N, Hedlund BP, Li W-J. 2021. Microbial dark matter coming to light: challenges and opportunities. Natl Sci Rev 8:nwaa280. https://doi.org/10.1093/nsr/nwaa280
2. Dufault-Thompson K, Hall B, Jiang X. 2022. Taxonomic distribution and evolutionary analysis of the equol biosynthesis gene cluster. BMC Genomics 23:182. https://doi.org/10.1186/s12864-022-08426-7
3. Braccia DJ, Jiang X, Pop M, Hall AB. 2021. The capacity to produce hydrogen sulfide (HS) via cysteine degradation is ubiquitous in the

human gut microbiome. Front Microbiol 12:705583. https://doi.org/10.3389/fmicb.2021.705583

4. Mou Z, Yang Y, Hall AB, Jiang X. 2021. The taxonomic distribution of histamine-secreting bacteria in the human gut microbiome. BMC Genomics 22:695. https://doi.org/10.1186/s12864-021-08004-3
5. Hall B, Levy S, Dufault-Thompson K, Arp G, Zhong A, Ndjite GM, Weiss A, Braccia D, Jenkins C, Grant MR, Abeysinghe S, Yang Y, Jermain MD, Wu CH, Ma B, Jiang X. 2024. BilR is a gut microbial enzyme that reduces bilirubin to urobilinogen. Nat Microbiol 9:173–184. https://doi.org/10.1038/s41564-023-01549-x
6. Kashyap PC, Chia N, Nelson H, Segal E, Elinav E. 2017. Microbiome at the frontier of personalized medicine. Mayo Clin Proc 92:1855–1864. https://doi.org/10.1016/j.mayocp.2017.10.004
7. Seol D, Jhang SY, Kim H, Kim S-Y, Kwak H-S, Kim SH, Lee W, Park S, Kim H, Cho S, Kwak W. 2019. Accurate and strict identification of probiotic species based on coverage of whole-metagenome shotgun sequencing data. Front Microbiol 10:1683. https://doi.org/10.3389/fmicb.2019.01683
8. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
9. Eddy SR, Pearson WR. 2011. Accelerated profile HMM searches. PLoS Comput Biol 7:e1002195. https://doi.org/10.1371/journal.pcbi.1002195
10. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240. https://doi.org/10.1093/bioinformatics/btu031
11. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, Ogata H. 2020. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics 36:2251–2252. https://doi.org/10.1093/bioinformatics/btz859
12. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper V2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol 38:5825–5829. https://doi.org/10.1093/molbev/msab293
13. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153
14. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. 2021. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microb Genom 7:000685. https://doi.org/10.1099/mgen.0.000685
15. Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G, Crusoe MR, Kale V, Potter SC, Richardson LJ, Sakharova E, Scheremetjew M, Korobeynikov A, Shlemov A, Kunyavskaya O, Lapidus A, Finn RD. 2020. MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res 48:D570–D578. https://doi.org/10.1093/nar/gkz1035
16. Liu R, Ochman H. 2007. Stepwise formation of the bacterial flagellar system. Proc Natl Acad Sci U S A 104:7116–7121. https://doi.org/10.1073/pnas.0700266104
17. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. 2021. Pfam: the protein families database in 2021. Nucleic Acids Res 49:D412–D419. https://doi.org/10.1093/nar/gkaa913
18. Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou L-P, Mi H. 2022. PANTHER: making genome-scale phylogenetics accessible to all. Protein Sci 31:8–22. https://doi.org/10.1002/pro.4218
19. Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, White O. 2001. TIGRFAMs: A protein family resource for the functional identification of proteins. Nucleic Acids Res 29:41–43. https://doi.org/10.1093/nar/29.1.41
20. Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, Badretdin A, Coulouris G, Chitsaz F, Derbyshire MK, Durkin AS, Gonzales NR, Gwadz M, Lanczycki CJ, Song JS, Thanki N, Wang J, Yamashita RA, Yang M, Zheng C, Marchler-Bauer A, Thibaud-Nissen F. 2021. RefSeq: expanding the prokaryotic genome annotation pipeline reach with protein family model curation. Nucleic Acids Res 49:D1020–D1028. https://doi.org/10.1093/nar/gkaa1105
21. Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28:33–36. https://doi.org/10.1093/nar/28.1.33
22. Madin JS, Nielsen DA, Brbic M, Corkrey R, Danko D, Edwards K, Engqvist MKM, Fierer N, Geoghegan JL, Gillings M, et al. 2020. A synthesis of bacterial and archaeal phenotypic trait data. Sci Data 7:170. https://doi.org/10.1038/s41597-020-0497-4
23. Letunic I, Bork P. 2021. Interactive tree of life (iTOL) V5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 49:W293–W296. https://doi.org/10.1093/nar/gkab301
24. Nakamura S, Minamino T. 2019. Flagella-driven motility of bacteria. Biomolecules 9:279. https://doi.org/10.3390/biom9070279
25. Liu R, Ochman H. 2007. Origins of flagellar gene operons and secondary flagellar systems. J Bacteriol 189:7098–7104. https://doi.org/10.1128/JB.00643-07
26. Diepold A, Armitage JP. 2015. Type III secretion systems: the bacterial flagellum and the Injectisome. Philos Trans R Soc Lond B Biol Sci 370:20150020. https://doi.org/10.1098/rstb.2015.0020
27. Bergeron JR. 2016. Structural modeling of the flagellum MS ring protein FliF reveals similarities to the type III secretion system and sporulation complex. PeerJ 4:e1718. https://doi.org/10.7717/peerj.1718
28. Abby SS, Néron B, Ménager H, Touchon M, Rocha EPC. 2014. MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. PLoS One 9:e110726. https://doi.org/10.1371/journal.pone.0110726
29. Cury J, Abby SS, Doppelt-Azeroual O, Néron B, Rocha EPC. 2020. Identifying conjugative plasmids and integrative conjugative elements with CONJscan. Methods Mol Biol 2075:265–283. https://doi.org/10.1007/978-1-4939-9877-7_19