Beyond Blast: Enabling Microbiologists to Better Extract Literature, Taxonomic Distributions and Gene Neighborhood Information for Protein Families

Colbie Reed[1], Rémi Denise[1+], Jacob Hourihan[1], Jill Babor[1], Marshall Jaroch[1], Maria Martinelli[1&], Geoffrey Hutinet[1] and Valérie de Crécy-Lagard[1,2]*

[1]Department of Microbiology and Cell Science, University of Florida, Gainesville, FL 32611, USA

[2]University of Florida Genetics Institute, Gainesville, FL 32610, USA.

[+]APC Microbiome Ireland, Science Foundation Ireland Research Centre, Bioscience Building, School of Microbiology, University College Cork — National University of Ireland, Cork, Ireland

[&]University of Central Florida, 4000 Central Florida Blvd. Orlando, Florida, 32816, USA

*Correspondence should be sent to Valérie de Crécy-Lagard, vcrecy@ufl.edu

VDC: 0000-0002-9955-3785

CJR: 0000-0002-5039-2052

Keywords

Phylogenetic distribution, synteny, gene neighborhoods, orthology, UX/UI, bioinformatics, tools

Abstract

Capturing the published corpus of information on all members of a given protein family should be an essential step in any study focusing on any specific member of that said family. This step is often performed only superficially or partially by experimentalists as the most common approaches and tools to pursue this objective are far from optimal. Using a previously gathered dataset of 284 references mentioning a member of the DUF34 (NIF3/Ngg1-interacting Factor 3), we evaluated the productivity of different databases and search tools, and devised a workflow that can be used by experimentalists to capture the most information in less time. To complement this workflow, web-based platforms allowing for the exploration of member distributions for several protein families across sequenced genomes or for the capture of gene neighborhood information were reviewed for their versatility, completeness and ease of use. Recommendations that can be used for experimentalist users, as well as educators, are provided and integrated within a customized, publicly accessible Wiki.

Data summary

The authors confirm all supporting data, code, and protocols have been provided within the article or through supplementary data files. The complete set of supplementary data sheets may be accessed via FigShare.

Introduction

In the last 35 years, the field of microbiology has undergone a total revolution. The completion of the first whole genome sequence of a bacterium, *Haemophilus influenzae* RD40 in 1995 [1], changed the way bench scientists design and/or interpret their experiments: the analysis of sequences (gene, protein, whole genomes) has become an integral part of the whole process [2]. This led to the incredible success of the BLAST suite developed at NCBI by Altschul et al. [3] that allowed any scientist with an internet connection to ask whether his/her favorite gene/protein was similar to an already experimentally characterized one or whether a similar sequence was present in particular organisms. From 1995 to 2005, most microbiologists could get by with NCBI and digital cloning platforms for their limited bioinformatic toolboxes. The arrival of Next Generation Sequencing (NGS) technologies has made the sequencing of microbial genomes a routine procedure. Today, this technological advancement is feeding thousands of microbial genomes and metagenomes into GenBank [4] every week (or even every day) thus transforming many fields of microbiology, from ecology [5] to food microbiology [6], infectious diseases [7] and basic enzymology [8]. This 'deluge of data' [9] is making simple BLAST searches useless for most applications as, without specific filters, BLAST will just retrieve hundreds of sequences closely related to the input sequence. In an ideal world, every biologist would be trained in using command line and programming tools that would allow them to cope with this encumbrance of data [10]. This might be the case in a few years' time, but such a solution has yet to be realized and many researchers are likely to be left behind due to resource, access, and opportunity constraints. Fortunately, a plethora of databases have developed various programs with web-accessible Graphical User Interfaces (GUIs) that allow users with little to no programming experience to take full advantage of the information possible to be derived from the over 250K available complete microbial genome sequences [11].

Integrated microbial genome portals are the easiest entry points for accessing and analyzing data derived from microbial genomes. Many microbiologists become aware of these resources only when they need to annotate a genome sequenced in their own laboratories, as most offer user-friendly annotation pipelines [12–15]. These microbial genome web-portals are quite versatile and offer various tools that were recently extensively reviewed in a side-by-side comparison [16]. Some databases offer training through introductory workshops, which can be great gateways into these

resources, yet these tend to reach only a small audience and are often restricted to a specific platform. Tutorials are also available but—in our experience teaching the use of web-based tools to undergraduate, graduate, and post-graduate audiences, both, in formal classes and in workshops—we find that these are most useful when used to "refresh" the skills of seasoned users instead of being used to get a novice user started.

We have been using comparative genomic driven approaches using only web-based tools to link genes and functions for over 20 years, leading to the functional characterization of more than 65 gene families (**Table S1**). This work required the use of all the available microbial genome web-portals, learning the strengths and weaknesses of each in the process. Here, we address problems that routinely arise for biochemists/microbiologists interested in a specific protein family and show how they can be resolved using the web-portals as well as more specialized online tools. We focus on answering three specific questions. First, "what information has already been published for any member of a protein family?" Second, "how can one best analyze and visualize the taxonomic distribution for members of a protein family?" Finally, "how can physical clustering data for genes of a given family be gathered and visualized?" In answering these questions, we intend to showcase the different microbial web-portals, as well as identify and discuss their limitations. Moreover, we present a new resource targeted towards novice bioinformatic tool users, the VDC-Lab Wiki, that compiles databases we routinely use for research and teaching, doing so with an informed curatorial eye guided by 20 years of experience navigating biological databases.

Methods

Protein Family Case Study and Literature Review, Curation.

Process of retrieval described in detail in text; resulting accumulation of published keywords, identifiers and accessions is provided (**Data S1**). A list of tools, databases, and search engines were compiled for use in and a result of this work. The totality of these resources can be reviewed in the provided supplemental materials (**Data S2**).

Data Analysis, Figure Generation.

Microsoft Office Excel was used for tallying observations, query results, in addition to documenting the curation process and generating figures of curation results. Other figures and diagrams were created using Microsoft Office PowerPoint.

Results

Recommended workflow to capture literature for all members of a protein family.

Identifying all literature pertaining to all members of a given protein family remains a major challenge in an era defined by massive accumulations of biological data. Most microbiologists depend on PubMed [17] to find literature on a specific subject, relying on its text-based search tools. However, retrieval of published data for family-relevant homologs—often for the purposes of background review or hypothesis generation—remains a common objective rife with challenges even with the use of such corpus-centric tools. Indeed, efforts have been made in the last 5-10 years by the scientific community to adhere to a set of uniform data standards prioritizing the findability, accessibility interoperability and reusability (or 'FAIR') of information [18], but these principles have yet to be systematically implemented to optimally facilitate the processes linking publications to the biomolecular entities (genes or proteins) they describe. The only journal to-date that has imposed such a preemptive standard is Biochemistry: since 2018, it has required authors to complete a form providing UniProt entry information for the proteins described in the paper being submitted [19,20].

To both explore the challenges of finding all the literature linked to members of a protein family and explore potential solutions, a stepwise demonstration of the data capture process was recreated using the conserved unknown protein family, DUF34 (recently examined in Reed et al., Biomolecules 2021 [21]). With this, an approach framework constructed relying upon web-based (i.e., highly accessible) bioinformatic tools was developed with experimentalists (non-bioinformaticians, computational biologists) in mind. It is with the sharing and recommendation of this workflow that we hope to better guide researchers who find bioinformatics unapproachable or beyond their means.

Gathering initial lists of keywords and representative sequences is a required first step.

The first step in any protein family analysis requires the gathering of input data (e.g., a sequence or an identifier) that will be used as seed information for queries (**Fig. 1 and Fig. S1-2)**. This process will generate two master lists: 1) a list of identifiers, gene/protein names; and 2) a list of representative sequences. Protein family databases such as Pfam [22], InterPro [23], CDD [24], EggNOG [25] are essential tools in generating these two lists. If a seed input sequence is available, it can be used to directly query these databases and extract family names and identifiers, as well as sequences of other family members. It also provides early insight into the taxonomic distribution of members of the family, of which will guide subsequent queries. Without a sequence, known keywords/aliases must be used to acquire sequences from a general protein knowledge database (e.g., UniProtKB [26], NCBI [27], JGI/IMG [28], BV-BRC [29]) (e.g., DUF34 protein family, **Data S1**) that can then be used to query family databases (**Fig. S3b**). Together, these processes allow for populating a final list of searchable identifiers/accessions/names (i.e., keywords; **Fig. 1**).

Representative family members should reflect the potential functional subgroups defined by taxonomic distribution, subgroups of high similarity, and domain architectural diversity.

Because sequence-based queries provide a more direct path to capture the full diversity of sequence- and family-relevant literature, these types of queries should precede text-based searches. To date, several of these tools have been developed (e.g., Seq2Ref and Pubserver [30,31]) but, at the time of this publication's composition, only PaperBLAST [32] has remained functional and fully-maintained. Using the set of representative sequences selected in the initial gathering phase (**Fig. 1**), sequence-based queries using PaperBLAST can be performed to retrieve homolog-specific literature, with sequence-associated hits and respective crosslinked publications being called according to BLAST thresholds. From these results, "literature-sequence" associations can be reviewed for true-positive status, extracted, and compiled into a final collection organized by gene/protein identifier(s), organism, evidence, and PubMed identifier (PMID). Although the results of PaperBLAST can be viewed easily, exporting sets of hits along with their respective links and excerpts of cached descriptions is not as user-friendly, requiring users to manually copy-paste them in lieu of programmatic access (**Fig. S6**). To properly accommodate a family's taxonomic distributions and architectural diversity, it is recommended that several sequences from phylogenetically disparate organisms be used for completing sequence-based queries using PaperBLAST. It is also important that these sequences reflect the diversity in domain architectures of members of the family that can be determined as described in **Fig. S3** and **Fig. S4**. Like results typical of BLAST programs, performing multiple queries using members of the same protein family will lead to highs level of redundancy as observed for the sequence-based queries completed for DUF34 family members. Here, repeat PaperBLAST queries were completed using diverse sequences reflecting different superkingdoms and alternative domain architectures (**Fig. 2**). In this case study, publications were classified as being either "focal" (i.e., any family homolog being mentioned in the title or abstract) or "non-focal" (i.e., any family homolog being mentioned anywhere outside of the abstract or title, including supplementary materials), as well as whether the hit was determined to be a "false positive" that lacked relevance to any DUF34 family member (**Fig. 2**). On average, 26% of retrieved publications were observed to occur in less than half of the other query results, supporting the necessity of using diverse input sequences. As expected, repeat hits of the same publication were relatively frequent within individual query results (average redundancy rate of ~23%).

PaperBLAST it still relies upon a pre-existing or computed network of database cross-links, which include information and inferred associations extracted from accessible publications via text-mining [32]. This inherent feature makes them susceptible to many of the mistakes common to text-

based retrieval methods that will be discussed in the section below and require an important level of manual curation [33]. For example, false positives can be observed (average across queries, 8.4%; total across yields, 6.8%; **Fig. 2**). Nonetheless, due to its ease-of-use and powerful retrieval tools, PaperBLAST should be the first step when extracting published information on a given protein family. Unfortunately, the tool is not widely used (inferred by citation records). Only 45 publications are listed by PubMed to have cited the key reference paper for PaperBLAST since its publishing in 2017 [34] (https://pubmed.ncbi.nlm.nih.gov; accessed the 20th of December, 2022). PredictProtein [35] (https://predictprotein.org) is another sequence-based tool that remains active and more recently has added features allowing it to be used as an alternative to PaperBLAST; however, the literature search tool is a secondary feature of the primary functions and remains in the beta phase of development.

Performing text-based queries for a protein family is essential but rife with challenges.

After establishing and refining the two master lists of keywords and representative sequences (**Fig. 1**), text-based queries can be pursued, which will result in the continued accumulation of keywords and representative sequences. The choice of search engine used for text-based queries is ultimately up to the user's preference; however, some tools are more appropriate than others depending on the available time budget and overarching goals (i.e., comprehensive, or topical review of a family). Together, these two parameters—keyword(s) and search tool(s)—govern the ultimate productivity of text-based search efforts. Using select keywords common to DUF34 family member associations, a survey of the different search tools demonstrated that the total number of text-based hits are highly variable (**Fig. 3a**) and can depend on the keywords used (**Fig. 3b**). Further, many of these results in this case were found to be contaminated by false-positive hits due to publications containing an unrelated scientific term, NiF-3 ("Nickel Fluoride 3"; **Fig. S7**), only identified through manual curation by the user. Similarly, Pubtator's automated query adjustments can result in over-expanded, misleading results for the keyword, "GTP cyclohydrolase I type 2", by permitting additional irrelevant results for "GTP cyclohydrolase I" into the output (**Fig. 3b; Fig. 4**). It is unclear how this may impact the varied searches of other users.

Despite curation efforts, sequence-based search tools can fall victim to many of the same problems that encumber the text-based retrieval of literature.

Although text-based search tools are widely used by experimentalists, they are less direct than sequence-based tools, and their queries are vulnerable to false-positives/false-negatives linked to human-/computer-language "mistranslations". Any disconnect between queryable identifiers and synonymous terms used in publications can broadly be regarded as problems in *identifier referenceability*. In the case study of DUF34, three major sources of low or lax referenceability were

observed, each having the potential to influence that of the others: 1) *name* or *identifier multiplicity* (i.e., polyonymy/plurality); 2) mistaken identity (i.e., *problematic homonymy*); and 3) the *'published and perished'* phenomenon (discussed below).

The first refers to the problems generated by the many different aliases or terms of reference often assigned to biological entities like protein sequences. In the case of DUF34, numerous systematic identifiers (study system- and/or database-specific identifiers), for model and non-model organisms alike, were accumulated in the retrieval of publications (**Fig. S8**).

The second source of poor referenceability can also be described as *'problematic homonymy'* or *'false synonymy'*. When the same term is used for two distinct entities, either by intention or coincidence, it can make them difficult to distinguish, identify, retrieve, or sort. The DUF34 homolog, CT108 of *Chlamydia trachomatis* str. D/UW-3/CX, exemplified such problematic homonymy (**Fig. S9a-b**). In this case, PaperBLAST's retrieval of a false positive paper for the systematic identifier CT108 was the result of an author-assigned sample having the same name (a homonym), "CT108", as the putatively linked homolog, even though this sequence was never actually mentioned in the publication. Similarly, homolog BB0468, also picked up by PaperBLAST, resulted in an example of a true match false positive; that is, the match was linked to a gene/protein of the correct identifier, but the biological information within the context of the work contradicted other well-established published observations, suggesting that the functional attribution was likely mistaken (**Fig. S10**).

The final source, the so-called *'published and perished'* phenomenon [36], refers to aliases, descriptions, or characterizations that had been published in the past but have since been missed by the work of one or more contemporaries, resulting in the independent naming, describing, and/or characterizing of the same entity (**Elaboration S1**). In addition to these challenges, keywords containing family-level names or identifiers often returned the fewest number of hits, regardless of search engine (**Fig. 3**), which has a deleterious effect upon the findability of published members of a protein family. This is expected to reflect the lack of standardized, systematic family-level recognition across papers that mention or discuss protein sequences. Systems that employ such standards to be implemented by publishers are recommended to facilitate better database crosslinking and improve the findability of family homologs for which experimental data has been published. In summary, any output derived from text-based searched should also be checked by sequence for the expected protein family membership.

***Comprehensive literature search: capture through repeat iterations of queries in parallel with the accumulation of keywords and representative sequences*.**

Ideally, a comprehensive capture of all publications linked to all members of a protein family would require cycles of querying, curating, and cataloging ("QCC cycle", **Fig. 5**). However, there exists a point of diminishing returns with such a process, productivity exponentially decreasing as more time passes. Even for a dedicated biocuration expert, the total amount of non-redundant data retrieved per unit time exponentially decreases after a certain number of hours. Ultimately, the user will only retrieve duplicates or false positives with each new query cycle. In recognizing this phenomenon of diminishing returns, we suggest that when the ratio of new relevant papers retrieved across tools to the amount of time invested falls below one, the user should consider this the optimal stopping point for queries.

Phyletic patterning tools that examine the taxonomic distributions of protein families are essential components of comparative bioinformatic analyses.

Phyletic patterning or phylogenomic analyses are used to compare the occurrence of proteins/protein families across multiple genomes. Beyond just surveying the taxonomic distribution of a protein family that is an essential step for many basic bioinformatic tasks such as generating multiple alignments, phylogenetic trees and even literature searches as discussed above, phylogenomic analyses can survey the absence-presence of complete metabolic pathways, identifying co-occurring families, or families that match a specific taxonomic distribution pattern to identify missing enzymes or pathway holes [37,38]. We survey different types of webserver-based resources for phyletic patterning (**Data S2a**). These resources vary in their interoperability, explorability, and overall usefulness depending on a user's ultimate end goal and can be separated into several types: 1) general orthology databases (precomputed); 2) phyletic pattern generators/databases (often features within larger databases; most often precomputed). Of the latter type, there exists four subtypes defined by parametric restrictions and outputs: 1) custom genome selection with single target family selection; 2) tool-defined genome selection with single target family selection; 3) tool-defined genome selection with multiple target family selection; and the rarest of the subtypes, 4) custom genome selection with multiple target family selection.

For years, PubSEED [39] had been an ideal resource for examining taxonomic distributions of protein families, as a user could rename member proteins and visualize their distributions in sets of genomes or user-defined "Subsystems". If PubSEED is still functional at the time of this work's publishing, it will likely still be frozen at ~10,000 genomes and so cannot be considered a main source for analyzing taxonomic distributions. For most existing orthology-driven resources (**Data S2a**), the user relies on annotation- and family-propagation systems implemented by the specific databases. The

strengths and weaknesses of these tools are compared and discussed below using the specific DUF34 family linked clustered orthologous groups (COGs) identified previously: COG0327, COG1579, and COG3323 [21].

Orthology databases are useful as a first pass to gather an overview of the taxonomic distribution of members of a protein family.

Orthology databases, very broad resources for examining proteins at the family level, allow for fast but superficial views of one target family at a time across pre-computed sets of genomes. Pre-computed similarity-based groups or families, like those available through Pfam, InterPro, EggNOG and OrthoMCL, are useful but tend to be too superficial or lacking necessary specificity, especially in the case of smaller or clade specific families, like COG1579 (PF02591, IPR003743). These resources are limited most by the slow speed at which these databases are updated and are also not adapted to detect fusions. Further, more obscure functional subgroups can be missed. For example, when using the sequence-based search of the EggNOG database, a representative member of COG3323 will still retrieve multiple COGs varying in relevance to the query sequence and with the relationships between the different COGs left unspecified. We do find however, that even if these tools are very constrained in the genomes included, their simplicity and accessibility allow for the swift topical examination of protein families, and, in some cases, offer features not offered by more complex tools. An example of the latter includes a passive feature of EggNOG that permits a user access easy COG-dependent paralog retrieval through the “proteins” view of any given clustered group (**Fig. S11**).

Annotree: a more customizable “quick view” of single protein families in Archaea and Bacteria.

Building on the protein family information derived from Pfam (now integrated into InterPro), TIGRFAM (no longer maintained), or KEGG (KO families), Annotree provides a practical “first pass” in examining a protein family’s taxonomic distribution [40] (**Fig. S12**). Several output parameters can be actively modified by the user in-browser (e.g., taxonomic ranges for tree branching and, separately, labeling of those ranges). However, Annotree is restricted to bacterial and archaeal taxa and cannot allow for examination of multiple target families at once (**Fig. S12**). A separate tool, PhyloGene [41], provides another excellent example of a simple, fast phyletic patterning tool with an additional feature of determining putatively co-evolved genes; however, this tool, like many others, is restricted to eukaryotic homologs of humans as the primary query input(s).

Tools that allow analyzing multiple target families but restrict genome selection.

Phyletic patterning analysis is governed by two facets: 1) the number of families viewable at once (and how); and 2) the number of genomes one can view these data across at once (and whether

291 those genomes can be custom selected). Unfortunately, many of the tools available via webserver are
292 restricted by one or the other, even both, as precomputed outputs are far easier to manage and
293 retrieve upon external user queries. Examples of these kinds of precomputed phylogenomic databases
294 include MicrobesOnline, STRING-DB, and KEGG Orthology (KO). MicrobesOnline allows the user to
295 choose a set of input families using different types of systematic identifiers like COGs or Enzyme
296 Commission (EC) numbers for generating phyletic profiles, which are graphically produced clustered
297 taxonomically with the absence-presence of the families/members across the database's benchmark
298 1,965 organisms (**Fig. S13**). This tool is notably user-friendly with different methods of family member
299 identification/filtration possible for selection per target; in addition to systematic identifier
300 annotations, options for these filters also include several BLAST cut-offs (**Fig. S14**). Users may also view
301 the precomputed phyletic profile for a single family via any gene entry's "Gene Info" tab (**Fig. S15**).
302 Unfortunately, MicrobesOnline is, to-date, frozen at a total of 3,707 genomes (retrieved January 14th,
303 2022), and, further, these genomes are largely limited to bacterial organisms with only 94 archaea and
304 119 eukaryotes, the latter of which are mostly fungi.

305 The STRING database is a data aggregation-driven bio-entity network model database, and so
306 is constituted by a collection of edges and nodes propagated from other annotation knowledge bases.
307 More recently, the database has added a feature that allows for browsing of networks generated by
308 multiple user-provided clustered orthologous group identifiers (COGs), referred to as "families" on the
309 site (**Fig. S16a**). However, because of the underlying dependencies of the tool (i.e., the data aggregated
310 from various other annotation databases), the user-defined "settings" of the networks can lead to
311 misleading enrichments of node-node relationships, as well as losses of traceable data provenance
312 (i.e., the source and database history of a given annotation). Therefore, it is important that any outputs
313 are reviewed for their validity in consideration of the high potential for false positives (**Fig. S16b**).

314 Because the KEGG database uses relatively stringent family relationships to create their
315 orthologous groups (i.e., "KOs" or K numbers [42]), we find that the KO database can be quite useful.
316 The tool produces a table of protein distribution among all genomes present in the KEGG dataset using
317 KO identifiers (**Fig. S17a**), which the user provides in a space-separated list (**Fig. S17b**). However, the
318 genomes are organized in an order without clear reference to taxonomic relations and the data shown
319 is generated based upon the genomes in which at least one of the submitted KOs occurs. Because of
320 the latter feature, it is recommended that, with tools like this, the user co-submit a positive control KO
321 (i.e., a group that is known to be universally conserved across database genomes). Further, export for
322 this webserver output is not necessarily tabular or tabular-compatible (i.e., HTML-embedded table)
323 and, therefore, will require additional data tidying due to paralog-related row duplications (i.e.,

duplicate rows lack names, which may be particularly troublesome for tidying without specialized programmatic script development).

Tools that allow the choice of multiple target families with a custom selection of genomes.

Like Annotree, BV-BRC's Comparative Systems tool is restricted taxonomically to bacteria and archaea, but with the additional inclusion of viral pathogens (**Fig. S18**). The output of this tool includes a searchable heatmap for all identified gene families across a custom selection of genomes, the results of which can then be filtered using family identifiers (PGF IDs). MicroScope (the microbial platform of GenoScope) also possesses a Gene Phyloprofile tool. Multiple genomes (PkGDB, i.e., reannotated bacterial RefSeq genomes/proteomes) can be compared based on single or multiple genes/proteins, in addition to whole genome-to-genome phylogenomics. The ultimate result of this program is an output in the form of an HTML embedded table with each selected genome represented in a separate column (not row). Finally, JGI-IMG provides a tool suite that allow for the examination of custom genome lists with the use of many common systematic identifiers, such as KOs, COGs, and Pfams (i.e., "Find Function" feature of the suite). Again, the output for this tool is restricted to an HTML embedded table format but can be customized and exported in tabular format. In general, if all these tools are quite user-friendly and useful for a first pass analyses, they are currently limited by the reliance on precomputed family annotations that can be partial, too broad, or wrong [43].

KBase is one example of a tool suite designed to bridge the practical gaps between comparative analyses, setting a logical precedent for future development, integration.

The Department of Energy Systems Biology Knowledgebase (KBase) has made analytic modules and pipelines available for researchers that lack programming skills [44]. Any KBase user account allows for browser-mediated access and use of complete suites of common bioinformatic analyses using either publicly accessible data or data uploaded by the user. However, while the platform is relatively easy to use, it is currently missing the modules necessary to facilitate a complete phyletic patterning analysis pipeline from genome/genome set to visualization of recognized family members. Resources provided by KBase map out specially ordered "narratives" (i.e., an organized set of data objects and application queues within a digital notebook) for completing phylogenomic analysis starting from species trees, but such a pipeline can be challenging for novice users (**Fig. S20**; figure adapted from KBase 2020 phylogenetics narrative diagram). It should also be noted that analyses can take many hours depending on the number of genomes being analyzed, and such investments may be important timeline considerations for experimentalists.

KBase, along with the other tools discussed, have many useful elements but, unfortunately it seems that no single web-based tool possesses all the features in a direct, intuitive way interpretable by most experimentalists when addressing the problem of protein family absence-presence across genomes.

Physical clustering and synteny analyses

Physical clustering is a key type of association-based inference derived from genomic sequences and links genes to putative functions based on the annotations of their encoded neighbors, given strong conservation is observed [45]. Webtools available for curating and exploring the physical clustering among genes within and between genomes are more limited than those available for phyletic patterning. Often, these analyses are optional subroutines within other webservers and are sometimes a combined product of synteny analyses. To better understand the diversity of tool features, utility, and manners of presentation, a survey of publicly available webtools for the objective of physical clustering analysis was completed (**Data S2b**).

While synteny tools and pangenome viewers are plentiful (**Data S2c**), we will focus here tools that allow to capture localized physical clustering across genomes (**Data S2c**). Neighborhood viewers, the most basic of physical clustering analyses, are often a feature embedded within database entries or are optional features of other, larger tools (e.g., EFI Tools [46]), and, in cases of the former, neighborhoods are typically shown one at a time relative to the respective webpage's contents. For comparative genomics, however, it is often necessary to view multiple neighborhoods at once aligned relative to a specified target gene or genes. Single-neighborhood viewers were not of interest to this work and have not been included in any of the survey data shown. Analytically, physical clustering tools can be defined by several factors: 1) the number of genomes possible to consider at once; 2) annotation quality of genomes (i.e., CDSs, regulatory sites, operons); and 3) visualization objectives (e.g., region sizes, labeling, color coding). In more practical terms, however, these tools can be more easily classified according to how they allow users to interact with data (i.e., "data interaction types"), which can be either exploratory in nature (e.g., PubSEED [39]) or a more customizable experience through user-defined *ab initio* analyses (e.g., Genomic Context Visualizer [47], EFI-GNT, GeCont3 [48], COGNAT [49], FlaGs [50], GizmoGene (http://www.gizmogene.com); **Fig. 7**, **Fig. S21-S22**). Tools allowing for the customization of graphical outputs with the goal of creating publication quality figures can be considered a third variety of utility or data interaction type. Tools suited to this third category can be a post-production feature of some analytic suites, but, more often, are developed as stand-alone applications (e.g., Gene Graphics [51]).

These tools can be further described by their data filtration and, therein, their relationship to other tools. Some tools, such as the EFI Gene Neighborhood Tool, can be used as either a stand-alone method or in-tandem with a separate analysis (i.e., EFI's Enzyme Similarity Tool) to then be influenced by those preceding results. Likewise, GizmoGene is interoperable with BV-BRC database [29], providing easy access to its vast collection of microbial genomes (>700,000 as of February 2023), orthologous protein families, and comparative genome analysis tools. In particular, this simplifies the creation of user-defined genome groups for GizmoGene input and facilitates the downstream analysis of generated neighborhoods.

Such interoperability, however, is rare among physical clustering tools. Even rarer are features allowing for assessments of a generated collection of neighborhoods (i.e., annotation assessments, e.g., the frequency of occurrence of different neighboring genes, etc.). Gene Context Tool NG (GeCont3) is one of the few tools that provide any form of annotation assessment or summary of generated neighborhoods (i.e., "Phylo" and "Category" tabs accompanying the tool's outputs; **Fig. S21**), although the protein neighbor annotation networks generated by EFI-GNT may be considered a likened visual summary**.**

Creating a Wiki compiling a non-exhaustive list of web-based resources for the common microbiologist.

A persistent challenge exists within the bioinformatic community; that is, the ability to know which tools are available and which are most suitable for fulfilling our data and visualization objectives. As time passes, more tools are published with others being decommissioned nearly at a matched rate, the longevity of most tools not often surpassing any more than a year [52]. Many of the webtools designed for biologists' use are scattered across the internet, often without crosslinking to more centralized, well-known resources. A few sites have been dedicated to the aggregation of the totality of useful bioinformatics resources (e.g.: bio.tools [53], https://bio.tools; CNCB Database Commons [54], https://ngdc.cncb.ac.cn/databasecommons/; Nucleic Acids Research regular Database Issue [55]), but—in addition to being understandably challenging to maintain—the lack of grassroots- or leadership-level efforts to popularize some of these resources have left them of low findability and, therein, a deficit of broad use by the community. Only more recently have sites like CNCB Database Commons been recommended by the likes of *Cell Press* (https://marlin-prod.literatumonline.com/pb-assets/journals/research/cellpress/data/RecommendRepositories.pdf) or *Bioinformatics Advances* (https://academic.oup.com/bioinformaticsadvances/pages/instructions-to-authors).

In response to our own difficulties in navigating the ever-changing frontier of bioinformatic tools, a wiki of webtools was established, initially, for our laboratory's in-house use, and, later, was further developed with the intention of aiding other microbiologists. With 15 years of instructional experience in bioinformatics for microbiologists, this resource was designed with our own graduate-level courses in mind, as well as common bioinformatic workflows of familiar to more interdisciplinary microbiologists, a guide informing the organization and presentation of the site and its contents. It was also of interest to create a digital space where users could share feedback on tool use and performance with others. The importance of community discussion around the use, decommissioning, and creation of tools allows for a more complete documented history of the shifts and pivots of the broader analytic spheres of computational and comparative biology. Curricula of bioinformatics courses should represent the foremost and well-proven approaches in the field [56], and, in this, was deemed ideal for guiding our collection of resources. With this, the wiki was created using the pedagogic modules of our courses and their respective learning objectives to model the subsets of information, keywords, tags, and relationships between links. The list of webtools provided in the customized "VDC Favorites" page represent the closest reimagining of these course materials, as well as being the most user-friendly, interoperable, and accessible of the greater collection of resources. Ideally, this presentation of preferred tools and workflows aids other experimentalists less familiar with bioinformatic resources.

Figures and tables

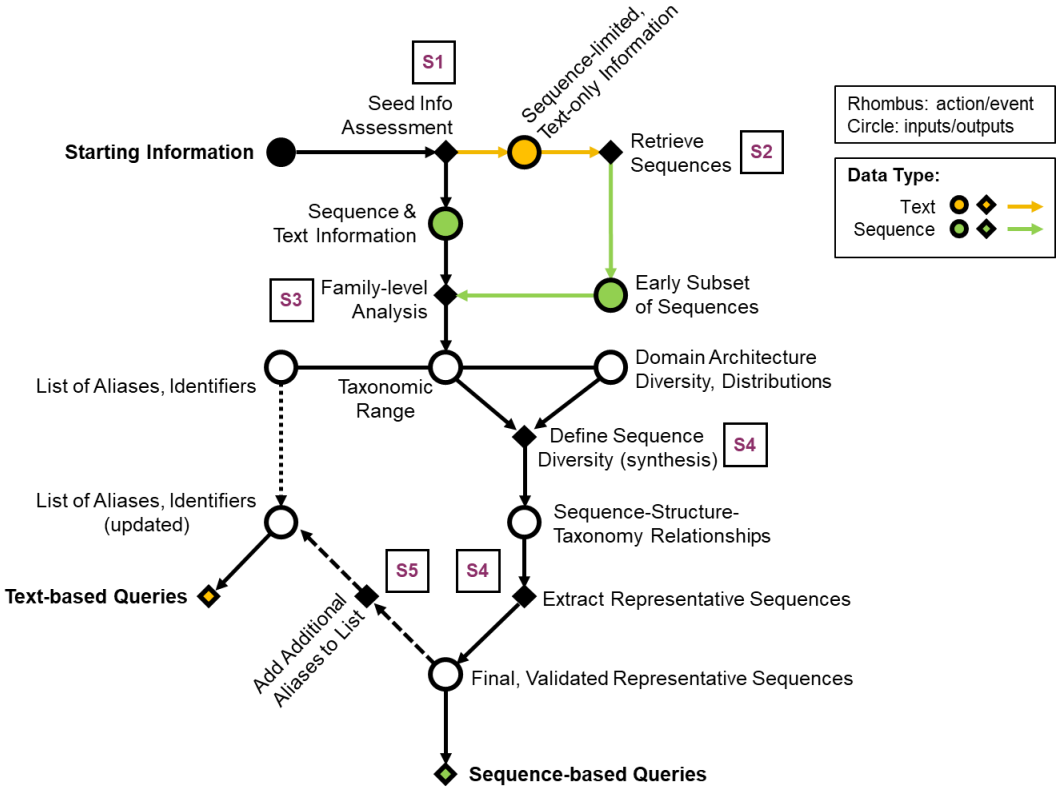
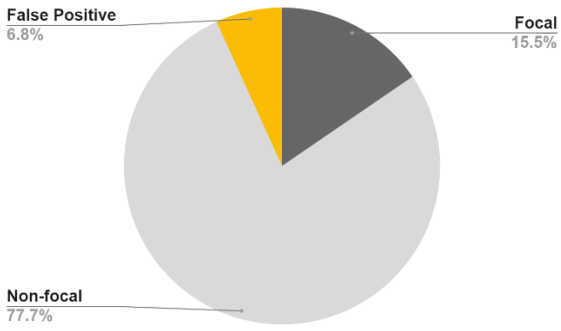


Figure 1. Workflow diagram recommended for capturing published data of a protein family. Supplemental figures were generated for examples (**Figures S1-S5**). Accompanying supplemental figures are boxed in the diagram and shown in purple text.

a,



b.

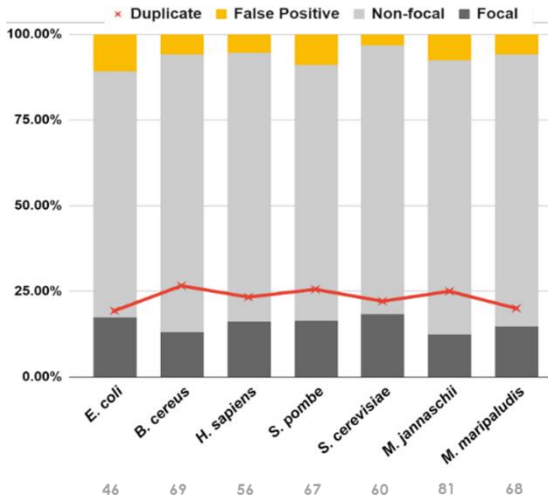


Figure 2. (a) Overall query quality across all representative sequences used to query PaperBLAST. "Focal" publications were defined as those having mentions of respective homologs in the abstract and/or the title, while "non-focal" were those with homolog mentions anywhere else in the publication and/or supplemental materials. Total represented hits equals 584 (false positives= 30; focal= 68; non-focal= 349). (b) Query quality of PaperBLAST hits per DUF34 protein family member sequence, one query sequence per organism. Organisms selected by tentative domain architectural subclasses and taxonomic distribution of members. Red line with 'x's marks the occurrence of redundant results within a single query (avg. ~23% per query). All selected query sequences have been independently described in a scientific publication [21]. Total hits per sequence/query are shown below the x-axis labels.

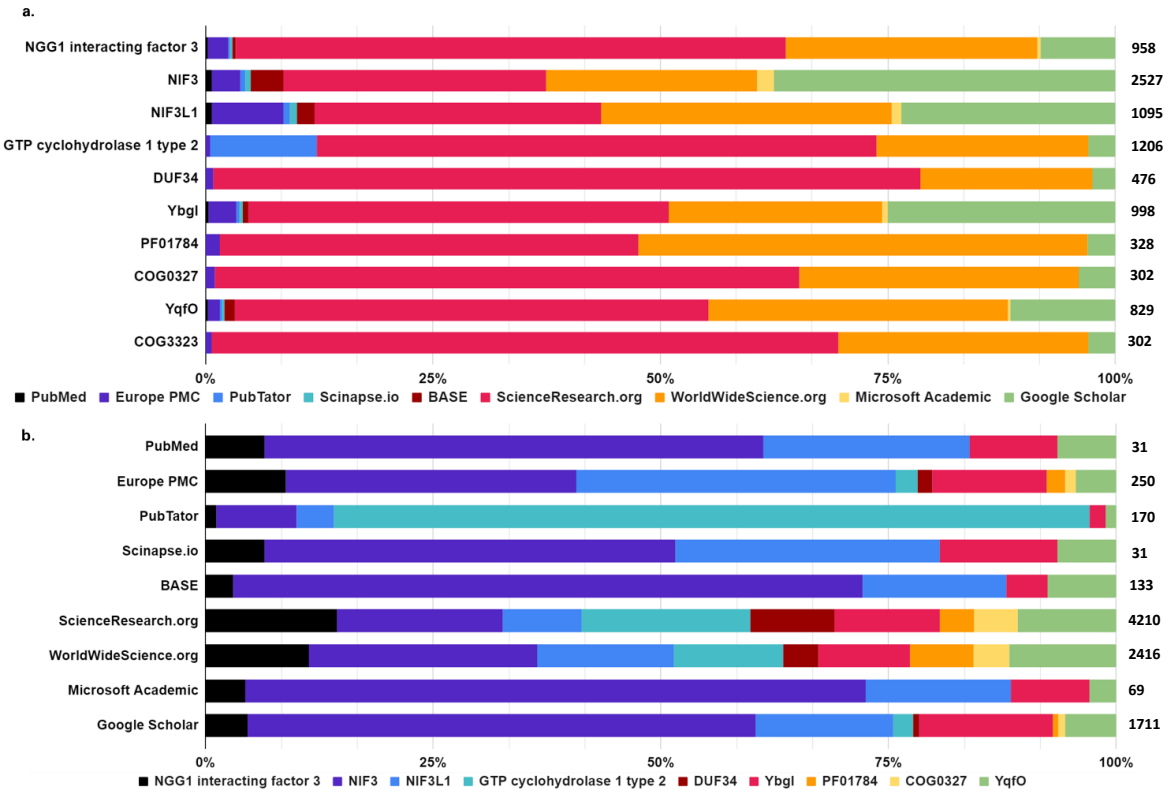


Figure 3. Query yield distributions per search tool as a function of keyword (a) and per keyword as a function of search tool (b). A subset of nine keywords most frequently associated with the target protein family, DUF34, was organized and used to compare the query results (i.e., total hits) across nine distinct search engines commonly used in published data retrieval for scientific research. Totals of each row are shown on the right axis of each figure.

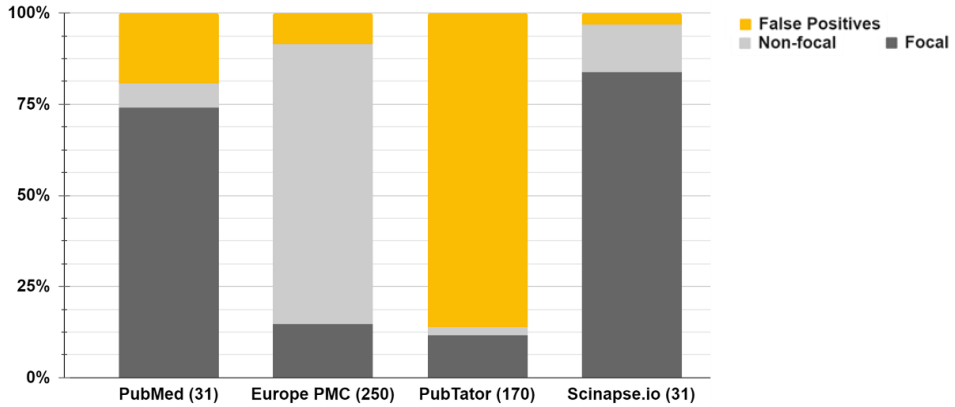


Figure 4. Query yield quality of more research-centered, conservative search tools. Focal publications were labeled as such if they featured relevant keywords specific to the protein family/family homolog in either the title or abstract of the publication. Non-focal publications were labeled if the keywords occurred in any other section of the paper. False positives were manually curated on a case-by-case basis.

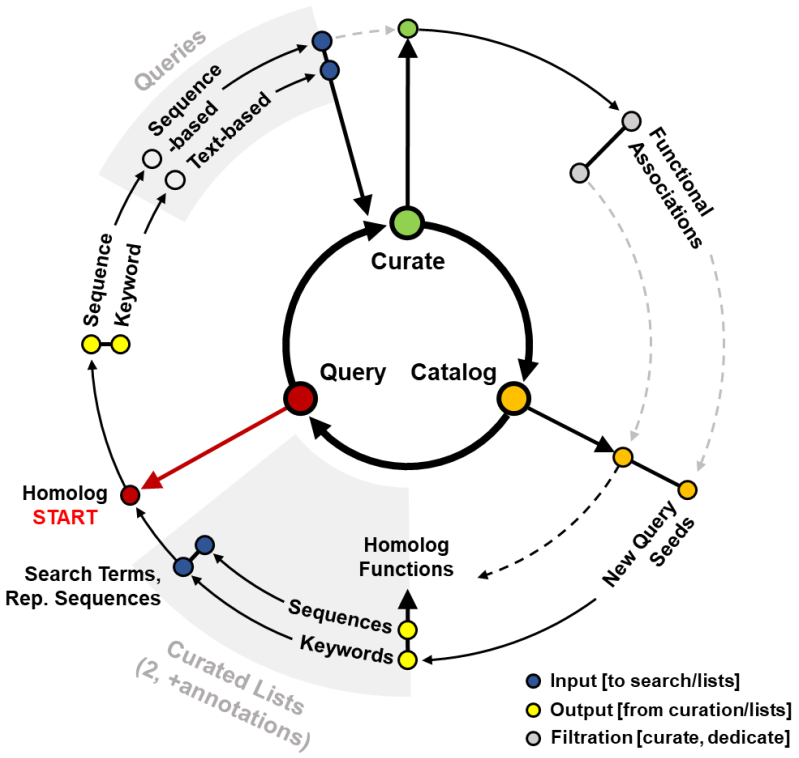


Figure 5. Idealized cycle of accumulating keywords, homolog annotations representative sequences, and publications necessary to optimize the capture of all published data relating to a protein family. Query phase refers to the process (often multiple in parallel) of using various webservers to retrieve literature relevant to target homolog(s). The Curate phase is distinguished by its filtration and review of retrieved information, and, frequently, the identification of experimentally validated functional associations to be noted for select homologs in the subsequent phase. The final phase of the cycle, the Catalog phase is defined by the multiple diverging paths which the different, identified information will be dedicated, which includes the two curated lists of keywords and representative sequences, as well as a collection of experimentally validated functional annotations of select homologs (publications cited). Multiple nodes within a single radial location indicates a split or merge, depending on the direction of the respectively linked arrow. Light gray dashed arrows indicate implicit information

472 flow, whereas the black, solid arrows indicate explicit information flow. The dashed, black arrow denotes explicit information
473 flow out* of the QCC cycle.

474

475

476

477

478

479

480

481

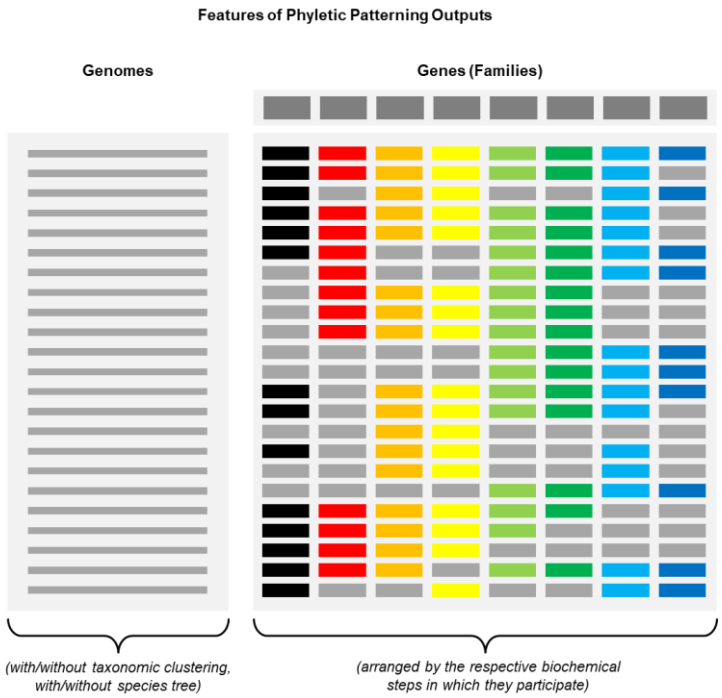
482

483

484

485

486



487

488 **Figure 6.** Outputs of phyletic patterning tools, visualized with options of arrangements noted.

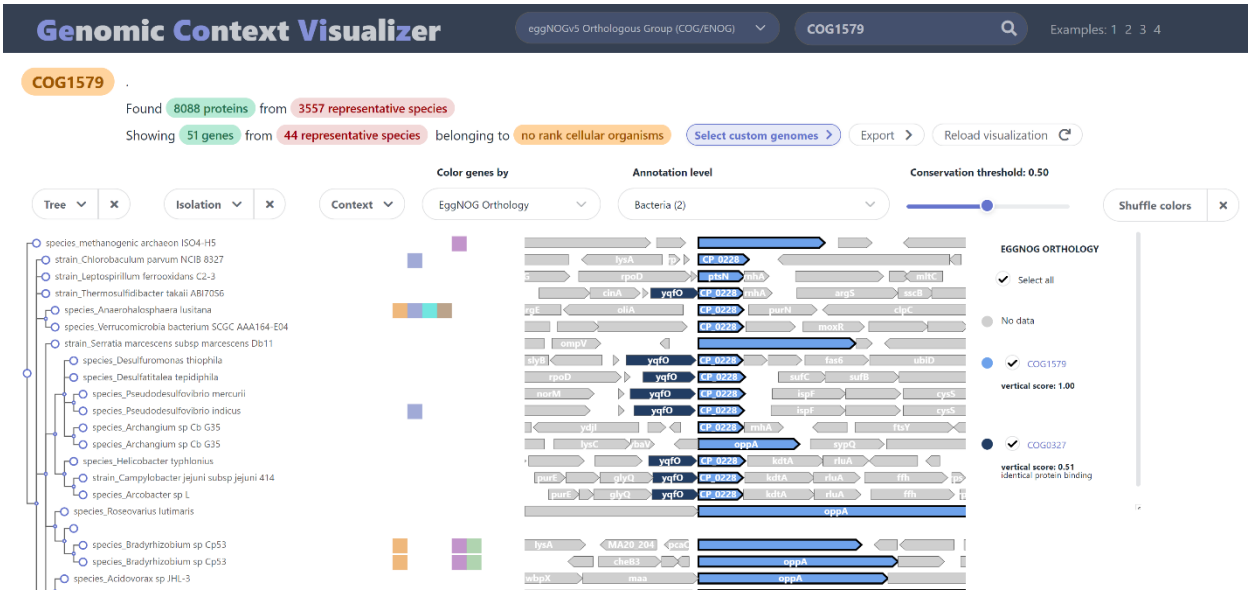


Figure 7. Exemplary physical clustering tool output of GeCoViz (Genomic Context Visualizer) using COG1579 as input. GeCoViz is one of few physical clustering tools that is generated accompanied by a taxonomic tree.

Discussion

To date, various fields of biology continue to be weighed down by the responsibilities and challenges of comprehensively extracting, synthesizing, and properly propagating scientific observations among databases, all in a manner that adheres to and further fosters the use of FAIR data guidelines [57]. Here, we examined these challenges of the literature review process in the context of whole protein families. The curation of published data, alongside the interrogation of available tools common to this process, were surveyed and workflows incorporating different iterations of them were compared. Several potential pitfalls and stumbling blocks for researchers found to be common to biological publishing were identified and described, supplemented by examples of each observed with the case study of the DUF34 protein family. Importantly, it was observed that the choice of keywords and search engines, though equally important, vary both together and independently in how they influence published data capture results. Additionally, false positives across search engine types illustrated the importance of thorough, well-informed curation efforts, or more stringent publishing standards among publishers. On the contrary, sequence-based searches were determined to be critical first steps of the data capture process at the protein family level. Unfortunately, it was also noted that sequence-based query tools like PaperBLAST are limited in number. Additionally, these tools rely

heavily on curated cross-linking networks of annotation data and publications, which obligates active, regular manual curation of cross-linked observations and biological "entities" across these networks.

As anticipated after a thorough investigation, it was determined that the process of accomplishing an 'ideal' level of completeness or comprehensiveness (i.e., requiring a repetitive, cyclic processes plagued by exponentially decreasing efficiency as greater completeness is achieved) is wholly impractical for the experimental microbiologist, given that completeness persists as a stubborn function of time, totality of keywords, and different tools invested in the retrieval of published works relevant to the target protein family. These factors, together with the aforementioned difficulties inherent to publishing, drive a pressure of diminishing returns on the retrieval process. Therefore, it was necessary to also describe an optimal strategy—a far more practical one—that could be easily employed by experimentalists aiming to improve the comprehensiveness of published data capture when completing such reviews at the family level. This recommended process was defined in a series of steps summarized within a decision tree framed by the necessity of sequences- or text-based queries, and responsibly informed by the overarching diversities and distributions of the target protein family.

In response to the importance of family-level information in guiding a comprehensive data capture process, a survey of web-based (i.e., casual-user accessible) bioinformatic tools was performed to assist in accomplishing the pedagogical objectives of this work. Beginning with phyletic patterning tools, the types of tools were reviewed and those of higher usability and interoperability were highlighted. Physical clustering tools were examined second. Precomputed databases (e.g., orthology databases) were dominant among both categories of analysis. Deficits of analytic completeness and output interoperability with other tools was also emphasized for both surveyed analysis types.

Ultimately, it is our hope that this work provides a framework with which experimental microbiologists might more easily approach the published data capture and the bioinformatic processes they would, without this guide, otherwise not be willing to explore.

Author contributions

Colbie Reed was responsible for data curation, formal analysis, investigation, methodology, validation, visualization, and writing (original draft, review and editing). Rémi Denise also contributed to writing (original draft), as well as software, resources, and creating the Heroku-supported wiki. Jacob Hourihan, Jill Babor, Marshall Jaroch, Maria Martinelli and Geoffrey Hutinet contributed significantly to data curation for the forementioned wiki. Additionally, Geoffrey Hutinet was responsible for editing

the final manuscript. Valérie de Crécy-Lagard was responsible for conceptualization, funding acquisition, methodology, project administration, supervision, and writing (review and editing).

Conflicts of interest

The authors declare no conflict of interest.

Funding information

Grant GM70641 to V dC-L.

Acknowledgements

We thank Marc Chevrette and Svetlana Gerdes for critical reading and suggestions and all students and participants in the training and courses given by the Crécy lab in the last 20 years for spurring the idea of this manuscript.

References

1. Fleischmann, R.D.; Adams, M.D.; White, O.; Clayton, R.A.; Kirkness, E.F.; Kerlavage, A.R.; Bult, C.J.; Tomb, J.-F.; Dougherty, B.A.; Merrick, J.M.; et al. Whole-Genome Random Sequencing and Assembly of Haemophilus influenzae Rd. *Science (80-.).* **1995**, *269*, 496–512, doi:10.1126/science.7542800.

2. Bansal, A.K. Bioinformatics in microbial biotechnology - A mini review. *Microb. Cell Fact.* 2005.

3. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410.

4. Benson, D.A.; Karsch-mizrachi, I.; Lipman, D.J.; Ostell, J.; Rapp, B.A.; Wheeler, D.L. GenBank. **2000**, *28*, 15–18.

5. 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. A new view of the tree of life. *Nat. Microbiol.* **2016**, *1*, 16048, doi:10.1038/nmicrobiol.2016.48.

6. Jagadeesan, B.; Gerner-Smidt, P.; Allard, M.W.; Leuillet, S.; Winkler, A.; Xiao, Y.; Chaffron, S.; Van Der Vossen, J.; Tang, S.; Katase, M.; et al. The use of next generation sequencing for improving food safety: Translation into practice. *Food Microbiol.* **2019**, *79*, 96–115, doi:https://doi.org/10.1016/j.fm.2018.11.005.

7. Quainoo, S.; Coolen, J.P.M.; van Hijum, S.A.F.T.; Huynen, M.A.; Melchers, W.J.G.; van Schaik, W.; Wertheim, H.F.L. Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. *Clin. Microbiol. Rev.* **2017**, *30*, 1015 LP – 1063, doi:10.1128/CMR.00016-17.

8. Zallot, R.; Oberg, N.; Gerlt, J.A. The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. *Biochemistry* **2019**, *58*, 4169–4182, doi:10.1021/acs.biochem.9b00735.

9. Klimke, W.; O'Donovan, C.; White, O.; Brister, J.R.; Clark, K.; Fedorov, B.; Mizrachi, I.; Pruitt, K.D.; Tatusova, T. Solving the problem: Genome annotation standards before the data deluge. *Stand. Genomic Sci.* **2011**, *5*, 168–193, doi:10.4056/sigs.2084864.

10. Shade, A.; Teal, T.K. Computing Workflows for Biologists: A Roadmap. *PLoS Biol.* **2015**, *13*, e1002303–e1002303, doi:10.1371/journal.pbio.1002303.

11. Zhulin, I.B. Databases for Microbiologists. *J. Bacteriol.* **2015**, *197*, 2458–2467, doi:10.1128/JB.00330-15.

12. Vallenet, D.; Calteau, A.; Dubois, M.; Amours, P.; Bazin, A.; Beuvin, M.; Burlot, L.; Bussell, X.; Fouteau, S.; Gautreau, G.; et al. MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. *Nucleic Acids Res.* **2020**, doi:10.1093/nar/gkz926.

13. Chen, I.-M.A.; Chu, K.; Palaniappan, K.; Pillay, M.; Ratner, A.; Huang, J.; Huntemann, M.; Varghese, N.; White, J.R.; Seshadri, R.; et al. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* **2019**, *47*, D666–D677, doi:10.1093/nar/gky901.

14. Arkin, A.P.; Cottingham, R.W.; Henry, C.S.; Harris, N.L.; Stevens, R.L.; Maslov, S.; Dehal, P.; Ware, D.; Perez, F.; Canon, S.; et al. KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat. Biotechnol.* **2018**, *36*, 566.

15. Davis, J.J.; Wattam, A.R.; Aziz, R.K.; Brettin, T.; Butler, R.; Butler, R.M.; Chlenski, P.; Conrad, N.; Dickerman, A.; Dietrich, E.M.; et al. The PATRIC Bioinformatics Resource Center: Expanding data and analysis capabilities. *Nucleic Acids Res.* **2020**, *48*, D606–D612, doi:10.1093/nar/gkz943.

16. Karp, P.D.; Ivanova, N.; Krummenacker, M.; Kyrpides, N.; Latendresse, M.; Midford, P.; Ong, W.K.; Paley, S.; Seshadri, R. A Comparison of Microbial Genome Web Portals. *Front. Microbiol.* **2019**, *10*, 208, doi:10.3389/fmicb.2019.00208.

17. White, J. PubMed 2.0. *Med. Ref. Serv. Q.* **2020**, *39*, 382–387, doi:10.1080/02763869.2020.1826228.

18. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, Ij.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018, doi:10.1038/sdata.2016.18.

19. Gerlt, J.A. The Need for Manuscripts To Include Database Identifiers for Proteins. *Biochemistry* **2018**, *57*, 4239–4240, doi:10.1021/acs.biochem.8b00705.

20. Wang, Y.; Wang, Q.; Huang, H.; Huang, W.; Chen, Y.; McGarvey, P.B.; Wu, C.H.; Arighi, C.N. A crowdsourcing open platform for literature curation in UniProt. *PLOS Biol.* **2021**, *19*, e3001464, doi:10.1371/journal.pbio.3001464.

21. Reed, C.J.; Hutinet, G.; de Crécy-Lagard, V. Comparative Genomic Analysis of the DUF34 Protein Family Suggests Role as a Metal Ion Chaperone or Insertase. *Biomolecules* **2021**, *11*, 1282, doi:10.3390/biom11091282.

22. Finn, R.D.; Mistry, J.; Tate, J.; Coggill, P.; Heger, A.; Pollington, J.E.; Gavin, O.L.; Gunasekaran, P.; Ceric, G.; Forslund, K.; et al. The Pfam protein families database. *Nucleic Acids Res.* **2009**, *38*, 211–222, doi:10.1093/nar/gkp985.

23. Blum, M.; Chang, H.-Y.; Chuguransky, S.; Grego, T.; Kandasaamy, S.; Mitchell, A.; Nuka, G.; Paysan-Lafosse, T.; Qureshi, M.; Raj, S.; et al. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **2021**, *49*, D344–D354, doi:10.1093/nar/gkaa977.

24. Lu, S.; Wang, J.; Chitsaz, F.; Derbyshire, M.K.; Geer, R.C.; Gonzales, N.R.; Gwadz, M.; Hurwitz, D.I.; Marchler, G.H.; Song, J.S.; et al. CDD/SPARCLE: the conserved domain database in 2020.

Nucleic Acids Res. **2020**, *48*, D265–D268, doi:10.1093/nar/gkz991.

25. Huerta-Cepas, J.; Szklarczyk, D.; Heller, D.; Hernández-Plaza, A.; Forslund, S.K.; Cook, H.; Mende, D.R.; Letunic, I.; Rattei, T.; Jensen, L.J.; et al. EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **2019**, *47*, D309–D314, doi:10.1093/nar/gky1085.

26. Bateman, A. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **2019**, *47*, D506–D515, doi:10.1093/nar/gky1049.

27. Bethesda (MD): National Library of Medicine (US), N.C. for B.I. National Center for Biotechnology Information (NCBI)[Internet] Available online: https://www.ncbi.nlm.nih.gov/.

28. Nordberg, H.; Cantor, M.; Dusheyko, S.; Hua, S.; Poliakov, A.; Shabalov, I.; Smirnova, T.; Grigoriev, I. V.; Dubchak, I. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res.* **2014**, *42*, 26–31, doi:10.1093/nar/gkt1069.

29. Olson, R.D.; Assaf, R.; Brettin, T.; Conrad, N.; Cucinell, C.; Davis, J.J.; Dempsey, D.M.; Dickerman, A.; Dietrich, E.M.; Kenyon, R.W.; et al. Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Res.* **2023**, *51*, D678–D689, doi:10.1093/nar/gkac1003.

30. Li, W.; Cong, Q.; Kinch, L.N.; Grishin, N. V. Seq2Ref: A web server to facilitate functional interpretation. *BMC Bioinformatics* **2013**, *14*, 1, doi:10.1186/1471-2105-14-30.

31. Jaroszewski, L.; Koska, L.; Sedova, M.; Godzik, A. PubServer: literature searches by homology. *Nucleic Acids Res.* **2014**, *42*, W430–W435, doi:10.1093/nar/gku450.

32. Price, M.N.; Arkin, A.P. PaperBLAST: Text Mining Papers for Information about Homologs. *mSystems* **2017**, *2*, 1–10, doi:10.1128/mSystems.00039-17.

33. Poux, S.; Magrane, M.; Arighi, C.N.; Bridge, A.; O'Donovan, C.; Laiho, K. Expert curation in UniProtKB: A case study on dealing with conflicting and erroneous data. *Database* **2014**, *2014*, 1–9, doi:10.1093/database/bau016.

34. Price, M.N.; Arkin, A.P. PaperBLAST: Text Mining Papers for Information about Homologs. *mSystems* **2017**, *2*, 1–10, doi:10.1128/mSystems.00039-17.

35. Bernhofer, M.; Dallago, C.; Karl, T.; Satagopam, V.; Heinzinger, M.; Littmann, M.; Olenyi, T.; Qiu, J.; Schütze, K.; Yachdav, G.; et al. PredictProtein - Predicting Protein Structure and Function for 29 Years. *Nucleic Acids Res.* **2021**, *49*, W535–W540, doi:10.1093/nar/gkab354.

36. Griss, J.; Côté, R.G.; Gerner, C.; Hermjakob, H.; Vizcaíno, J.A. Published and Perished? The Influence of the Searched Protein Database on the Long-Term Storage of Proteomics Data. *Mol. Cell. Proteomics* **2011**, *10*, M111.008490, doi:10.1074/mcp.M111.008490.

37. Hanson, A.D.; Pribat, A.; Waller, J.C.; Crécy-Lagard, V. de 'Unknown' proteins and 'orphan' enzymes: the missing half of the engineering parts list – and how to find it. *Biochem. J.* **2010**, *425*, 1–11, doi:10.1042/BJ20091328.

38. Osterman, A. Missing genes in metabolic pathways: a comparative genomics approach. *Curr. Opin. Chem. Biol.* **2003**, *7*, 238–251, doi:10.1016/S1367-5931(03)00027-9.

39. Overbeek, R. The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes. *Nucleic Acids Res.* **2005**, *33*, 5691–5702, doi:10.1093/nar/gki866.

40. Mendler, K.; Chen, H.; Parks, D.H.; Lobb, B.; Hug, L.A.; Doxey, A.C. AnnoTree: visualization and

exploration of a functionally annotated microbial tree of life. *Nucleic Acids Res.* **2019**, *47*, 4442–4448, doi:10.1093/nar/gkz246.

41. Sadreyev, I.R.; Ji, F.; Cohen, E.; Ruvkun, G.; Tabach, Y. PhyloGene server for identification and visualization of co-evolving proteins using normalized phylogenetic profiles. *Nucleic Acids Res.* **2015**, *43*, W154–W159, doi:10.1093/nar/gkv452.

42. Kanehisa, M.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **2016**, *44*, D457–D462, doi:10.1093/nar/gkv1070.

43. Kasif, S.; Roberts, R.J. We need to keep a reproducible trace of facts, predictions, and hypotheses from gene to function in the era of big data. *PLOS Biol.* **2020**, *18*, e3000999, doi:10.1371/journal.pbio.3000999.

44. Arkin, A.P.; Cottingham, R.W.; Henry, C.S.; Harris, N.L.; Stevens, R.L.; Maslov, S.; Dehal, P.; Ware, D.; Perez, F.; Canon, S.; et al. KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat. Biotechnol.* **2018**, *36*, 566–569, doi:10.1038/nbt.4163.

45. Overbeek, R.; Fonstein, M.; D'Souza, M.; Pusch, G.D.; Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci.* **1999**, *96*, 2896–2901, doi:10.1073/pnas.96.6.2896.

46. Gerlt, J.A.; Bouvier, J.T.; Davidson, D.B.; Imker, H.J.; Sadkhin, B.; Slater, D.R.; Whalen, K.L. Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochim. Biophys. Acta - Proteins Proteomics* **2015**, *1854*, 1019–1037, doi:10.1016/j.bbapap.2015.04.015.

47. Botas, J.; Rodríguez del Río, Á.; Giner-Lamia, J.; Huerta-Cepas, J. GeCoViz: genomic context visualisation of prokaryotic genes from a functional and evolutionary perspective. *Nucleic Acids Res.* **2022**, *50*, W352–W357, doi:10.1093/nar/gkac367.

48. Martinez-Guerrero, C.E.; Ciria, R.; Abreu-Goodger, C.; Moreno-Hagelsieb, G.; Merino, E. GeConT 2: gene context analysis for orthologous proteins, conserved domains and metabolic pathways. *Nucleic Acids Res.* **2008**, *36*, 176–180, doi:10.1093/nar/gkn330.

49. Klimchuk, O.I.; Konovalov, K.A.; Perekhvatov, V. V.; Skulachev, K. V.; Dibrova, D. V.; Mulkidjanian, A.Y. COGNAT: a web server for comparative analysis of genomic neighborhoods. *Biol. Direct* **2017**, *12*, 26, doi:10.1186/s13062-017-0196-z.

50. Saha, C.K.; Sanches Pires, R.; Brolin, H.; Delannoy, M.; Atkinson, G.C. FlaGs and webFlaGs: discovering novel biology through the analysis of gene neighbourhood conservation. *Bioinformatics* **2021**, *37*, 1312–1314, doi:10.1093/bioinformatics/btaa788.

51. Harrison, K.J.; Crécy-Lagard, V. De; Zallot, R. Gene Graphics: a genomic neighborhood data visualization web application. *Bioinformatics* **2018**, *34*, 1406–1408, doi:10.1093/bioinformatics/btx793.

52. Kern, F.; Fehlmann, T.; Keller, A. On the lifetime of bioinformatics web services. *Nucleic Acids Res.* **2020**, *48*, 12523–12533, doi:10.1093/nar/gkaa1125.

53. Ison, J.; Rapacki, K.; Ménager, H.; Kalaš, M.; Rydza, E.; Chmura, P.; Anthon, C.; Beard, N.; Berka, K.; Bolser, D.; et al. Tools and data services registry: a community effort to document bioinformatics resources. *Nucleic Acids Res.* **2016**, *44*, D38–D47, doi:10.1093/nar/gkv1116.

54. Ma, L.; Zou, D.; Liu, L.; Shireen, H.; Abbasi, A.A.; Bateman, A.; Xiao, J.; Zhao, W.; Bao, Y.; Zhang,

Z. Database Commons: A Catalog of Worldwide Biological Databases. *Genomics. Proteomics Bioinformatics* **2022**, doi:10.1016/j.gpb.2022.12.004.

55. Rigden, D.J.; Fernández, X.M. The 2022 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Res.* **2022**, *50*, D1–D10, doi:10.1093/nar/gkab1195.

56. Mulder, N.; Schwartz, R.; Brazas, M.D.; Brooksbank, C.; Gaeta, B.; Morgan, S.L.; Pauley, M.A.; Rosenwald, A.; Rustici, G.; Sierk, M.; et al. The development and application of bioinformatics core competencies to improve bioinformatics training and education. *PLoS Comput. Biol.* **2018**, *14*, 1–14, doi:10.1371/journal.pcbi.1005772.

57. Sansone, S.-A.; McQuilton, P.; Rocca-Serra, P.; Gonzalez-Beltran, A.; Izzo, M.; Lister, A.L.; Thurston, M. FAIRsharing as a community approach to standards, repositories and policies. *Nat. Biotechnol.* **2019**, *37*, 358–367, doi:10.1038/s41587-019-0080-8.