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

Colbie J. Reed[1], Rémi Denise[1+], Jacob Hourihan[1], Jill Babor[1], Marshall Jaroch[1], Maria Martinelli[1&], Geoffrey Hutinet[2] and Valérie de Crécy-Lagard[132]*

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

[23]Department of Biology, Haverford College, 370 Lancaster Avenue, Haverford, PA 19041, USA

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

[+]APC Microbiome Ireland, Bioscience Building, University College Cork, Ireland, UK

[&]Burnett School of Biomedical Sciences, University of Central Florida, Orlando, Florida 32827

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

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 specific members of that said family. Using a previously gathered dataset of more than 280 references mentioning a member of the DUF34 (NIF3/Ngg1-interacting Factor 3), we evaluated the efficiency of different databases and search tools, and devised a workflow that experimentalists can use to capture the most published information on members of a protein family in the least amount of time. To complement this workflow, web-based platforms allowing for the exploration of protein family members across sequenced genomes or for the analysis of gene neighborhood information were reviewed for their versatility 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. 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 cloning design platforms as their 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 (e.g., MicrobesOnline, JGI-IMG) 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 the available 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 experimentalists 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. Additionally, we present a resource targeted towards novice bioinformatic tool users, the VDC-Lab Wiki, that compiles databases that we routinely use for research and teaching, doing so with an informed curatorial eye guided by 20 years of experience in 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 as a result of this work. The totality of these resources can be reviewed in the provided supplemental materials (**Data S2**). Venn diagrams were generated using the online bioinformatics tools of Gent University (https://bioinformatics.psb.ugent.be/webtools/Venn/).

Data Analysis, Figure Generation.

Microsoft Office Excel (Office16) 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.

Wiki website development and publishing, gathering and curating resource information.

The websites featured on the wiki originate from a list of websites amassed by Valérie de Crécy-Lagard over time, as well as from discoveries made by laboratory members throughout their research. The included websites were tested by at least one lab member, who then crafted a brief description for it. Subsequently, all the websites were categorized and incorporated into the wiki. The VDC lab wiki can be found at the following address: https://vdclab-wiki.herokuapp.com/

Results

Investigating workflows for capturing literature for all members of a protein family.

Comprehensively identifying literature pertaining to all members of a given protein family for the purposes of background review or hypothesis generation is often the first step in many biological studies. This task remains rife with challenges in an era defined by massive accumulations of biological data [9,17,18]. Most microbiologists depend on PubMed [19] to find literature, relying on its text-based search tools. Although efforts by the scientific community have been made within the last decade to popularize adherence to uniform data standards that prioritize the findability, accessibility interoperability and reusability (or 'FAIR') of information [20], these principles have yet to be systematically implemented among databases and publishing journals [21], and, as a result, the state of linking publications to the biomolecular entities (genes or proteins) that they describe remains suboptimal. The only journal to-date that has imposed such a 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 [22,23].

To both explore the challenges of finding all relevant literature of a protein family and propose potential solutions, a stepwise demonstration of the capture process was recapitulated using the conserved unknown protein family, DUF34, recently examined in Reed et al. [24]. 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). Additionally, "false positives" (i.e., resulting papers found to lack relevance to any DUF34 family member) were also flagged. The DUF34 family was selected for its high level of conservation across all three domains of life, its described within-family diversity, and its variable, pleiotropic functional associations observed between study organisms. This family was used as a shared target for different methods of literature capture described and compared in the subsequent sections of this work. In addition, a workflow using only

124 web-based bioinformatic tools was developed to guide experimentalists into capturing a high
125 proportion of published data pertinent to a protein family in a timely and efficient fashion (**Fig. 1**).
126 The DUF34 family, again, allows us to show examples of each of these recommended steps and
127 discuss their strengths and weaknesses in the process of reviewing and characterizing a protein
128 family of unknown function.

129 ***Orthology databases are useful for gathering an overview of protein family domain organization***
130 ***and taxonomic distributions.***

131 The first step in any protein family analysis requires the gathering of input data (e.g., a sequence
132 or an identifier) that will be used as seed information for queries (**Fig. 1 and Fig. S1-2)**. This process
133 generates two master lists: 1) a list of identifiers, gene/protein names; and 2) a list of representative
134 sequences. Protein family databases such as Pfam [25], InterPro [26], CDD [27], EggNOG [28] are
135 essential tools in generating these two lists.

136 Orthology databases, broad resources for examining proteins at the family level, are often pre-
137 computed by HMM, bi-directional best hits (BLAST), or motif signatures, and allow for swift analysis
138 of one target family at a time across a predetermined set of genomes (**Data S2a**). If a seed input
139 sequence is available, it can be used to directly query these databases and extract family names and
140 identifiers, in addition to sequences of other family members. It also provides topical insight into the
141 taxonomic and domain distribution of members of the family, which will guide subsequent queries.
142 For the DUF34 family, this step led to a list of ten most frequently used keywords among UniProtKB
143 entries for this protein family: NGG1 interacting factor 3, NIF3, NIF3L1, GTP Cyclohydrolase 1 type 2,
144 DUF34, YbgI, PF01784, COG0327, YqfO, and COG3323. Without a sequence, known keywords/aliases
145 must be used to acquire sequences from a general protein knowledge database (e.g., UniProtKB
146 [29], NCBI [30], JGI-IMG [31], BV-BRC [32]) (e.g., DUF34 protein family, **Data S1**) that can then be
147 used to query family databases (**Fig. S3b**). Together, these processes allow for populating a final list
148 of searchable identifiers/accessions/names (i.e., keywords; **Fig. 1**); this process is completed most
149 comprehensively for the DUF34 family in the context of the later-discussed "QCC (Query, Curate,
150 and Catalog) Cycle" method.

151 One of the challenges in family-level analyses is the uncertainty of a given family's domain
152 architectural diversity, as well as their corresponding taxonomic distributions. While orthology
153 databases provide a general view of these attributes, they can be incomplete or misleading as, they
154 erroneously combine or separate families. Examples of this can be seen when navigating the

155 clustered groups and hierarchical relations of the EggNOG (v6) Database. The DUF34 family COG
156 root cluster, LCOG0327, functional annotations include K22391, K07164, and K24730, all of which
157 are incorrectly attributed to this group, the former due to premature EC number assignment in *H.*
158 *pylori* [24] and the latter two due to DUF34 fusion sequences in bacteria and eukaryotes,
159 respectively (**Fig. S4**). The aggregation mechanism and presentation manner of annotations by
160 EggNOG implicitly suggests to users that all of these functions are performed by DUF34. However, in
161 truth, these annotations are merely *linked* to DUF34—and aggregated by EggNOG—as a result of
162 DUF34 member fusions with COG1579 (K07164) and CIAO (COG2319, K24730). While this implicit
163 suggestion can be dangerous for propagation, it can also be used to examine the functional
164 associations of DUF34 through comparative genomics, as fusions are often used as a form of guilt-
165 by-association evidence paired with physical clustering data in the prediction of protein function
166 [33,34]. Of note, EggNOG is one of few databases to offer a view feature for family paralog incidence
167 (**Fig. S4e**), knowledge that can be critical in correctly annotating protein subgroups [35–38].

168 ***Text-based query yields vary quantitatively and qualitatively with both the choice of input and***
169 ***search engine.***

170 After establishing and refining the two master lists of keywords and representative sequences
171 (**Fig. 1**, example), text-based queries can be pursued, which will result in the continued accumulation
172 of keywords and representative sequences. The choice of search engine used for text-based queries
173 is ultimately up to the user and examples of the commonly used platforms or "engines" include but
174 are not limited to PubMed, Google Scholar, and Europe PMC. To evaluate how the choices of input
175 keyword and search engine impacted the resulting literature yields and, therein, the information
176 retrieved relevant (or irrelevant) to a target protein family, 10 common keywords associated with
177 the DUF34 protein family were selected and used in iterative searches using 9 popular search
178 engines frequented for the retrieval of scientific literature (i.e., both specialized like PubMed and
179 more broad like Google) (**Table 1**). The results of this systematic survey demonstrated that the total
180 number of text-based hits for this example conserved unknown protein family were highly variable
181 between the distinct combinations of search tool and keyword (**Fig. 2**).

182 To investigate the influence of keyword and search engine choice on the occurrence of false-
183 positive paper yields, a more thorough examination of a subset of specialized search tools was
184 examined (i.e., PubMed, PubTator, Scinapse.io, and Europe PMC) (**Fig. 3; Data S3**). These results
185 indicated a notable presence of false-positive hits among iterations despite the specialized nature of
186 the selected tools. Most of the instances of false-positives were found to be the results of the

187 retrieved publications containing an unrelated scientific term that was identified by the tool as being
188 equivalent to one of the selected keywords (e.g., NiF-3, "Nickel Fluoride 3") (**Fig. S5**), the irrelevance
189 of which was only identified through manual curation by the user. In a related observation,
190 Pubtator's automated query adjustments (a common search engine subroutine intended to improve
191 search result totals by modifying the user-provided keyword, capturing hits for a more generalized
192 version of the term) were found to result in a substantial number of misleading, irrelevant results—
193 specifically for the keyword "GTP cyclohydrolase I type 2". In these cases, irrelevant results were
194 driven by the expansion of the query to "GTP cyclohydrolase I" without providing the user any
195 notification of the change (**Fig. 3**).

196 Although text-based search tools are widely used by experimentalists, they are vulnerable to
197 false-positives/false-negatives linked to human-computer language "mistranslations" or
198 "miscommunications". These disconnects between queryable identifiers, and the terms used in
199 publications can be broadly regarded as problems in identifier referenceability. In the case study of
200 DUF34, three major sources of poor referenceability were observed: 1) name/identifier multiplicity
201 (i.e., polyonymy/plurality); 2) mistaken identity (i.e., misleading homonymy/false synonymy); and 3)
202 the 'published and perished' phenomenon, discussed further below.

203 The first source of poor referenceability, name/identifier multiplicity, refers to the problems
204 generated by the many different aliases often assigned to biological entities like protein sequences,
205 which frequently ensure that searches based on any one specific term are always incomplete, as was
206 demonstrated in the DUF34 example (**Fig. 2a**). The second source of poor referenceability can be
207 described as 'false synonymy'. As gene/protein names are not designed as unique identifiers, the
208 same name can be used for distinct entities by mere coincidence, making it difficult to distinguish,
209 identify, retrieve, or sort them. The DUF34 family alias, "GTP cyclohydrolase I type 2", continues to
210 exemplify such problematic homonymy (**Fig. S6**). In this case, the alias, while widespread in
211 databases across bacterial member sequences, is frequently not recognized by text-based search
212 engine tagging systems/keyword libraries, resulting in a failed primary search that is often
213 automatically generalized by the engine to "GTP cyclohydrolase" in hit retrieval. Although a synonym
214 within reason for the original search term this change invariably retrieves only publications relevant
215 to FolE or RibA, GTP cyclohydrolases I and II. The final source, the so-called 'published and perished'
216 phenomenon [39], refers to aliases, descriptions, or characterizations that had been published in the
217 past but have since been overlooked and "re-discovered" by the work of one or more

218 contemporaries, resulting in the independent naming, describing, and/or characterizing of the same
219 entity (**Elaboration S1**).

220 ***Sequence-based searches stand as a more-than-adequate resource for initial retrieval and review***
221 ***of literature regarding a protein family.***

222 As discussed above, text-based searches are defined by their basic functions and reliance upon
223 the design of the engine and user-input preferences, each of which impact the proficiency of natural
224 language-to-computer information processing, searchable data/keyword indexing, and match
225 identification. An alternative to text-based tools are resources that use sequences and family-linked
226 HMMs to search for related publications. To date, several sequenced-based literature search tools
227 have been developed (e.g., Seq2Ref and Pubserver [40,41]) but, at the time of this study, only
228 PaperBLAST [42] remained functional and fully maintained. To determine the publication retrieval
229 productivity differences between these two search tool types, the results for a PaperBLAST query of
230 a single DUF34 homolog, YbgI of *E. coli* (UniProt: P0AFP6), were compared to those of PubMed text-
231 based searches for the same protein. For use in querying PubMed, 9 keywords (i.e., "YbgI",
232 "P0AFP6", "b0710", "JW0700", "PF01784", "NIF3", "Ngg1-interacting factor 3", "DUF34" and
233 "COG0327") were collected from the UniProt entry page for this homolog (UniProt: P0AFP6). Using
234 default search settings, a total of 47 unique publications were retrieved with PaperBLAST, of which
235 only 3 were determined to be false positives (6.4%; **Fig. 4**; **Data S4**). In contrast, while PubMed
236 returned a total of 19 unique publications, 6 of these were found to be false positives (32%).
237 Although PubMed's release announcements suggest nigh ubiquity of full-text search across their
238 database and increasingly more advanced search features
239 (https://www.ncbi.nlm.nih.gov/feed/rss.cgi?ChanKey=PubMedNews), the comparative results
240 between these two tools suggest that PubMed is not generally able to recognize search terms in a
241 publication if they are located outside of the title and/or abstract. This is made evident—at least in
242 the case of the keywords associated with this particular DUF34 homolog—by the recognition of
243 "non-focal" publications by PaperBLAST and the complete oversight of the same publications by
244 PubMed (**Fig. 4**), even though these 34 publications were all present in the full-text depository of
245 NCBI (PubMed Central, PMC) of which PubMed is designed to query (**Data S4d**).

246 To assess the different uses of PaperBLAST for the retrieval of publications relevant to many
247 members of a protein family, two separate search options were compared: via HMM or via single
248 sequence/sequence ID. These two methods were used to generate itemized lists of publications for
249 two family member sequences (*H. sapiens*, UniProt: Q9GZT8; *E. coli*, UniProt: P0AFP6; default

250 settings) and for one HMM classification identifier representing the same family (Pfam: PF01784).
251 False positives were still present among the results for both PaperBLAST retrieval methods (**Fig. 4**;
252 **Data S5-S6**). Despite querying the same database, the two single-sequence queries differed in their
253 results from, both, one another and compared to the HMM-based search. *H. sapiens* and *E. coli*
254 DUF34 homolog sequence results were found to share 9 and 7 unique publications with the HMM-
255 based results, respectively (**Fig. 4**; **Data S5-S6**). Only the retrieved set for the *E. coli* family member
256 sequence retrieved a publication (1) unique from both the *H. sapiens* homolog results and those of
257 the HMM-based search, suggesting two possibilities: 1) that the curation status of the latter two
258 have the greatest similarity in quality, which appears higher than that for the bacterial family
259 member sequence; and/or 2) that publications of the eukaryotic model system of *H. sapiens*
260 contribute the bulk of the overall family related literature accessioned by PaperBLAST's database.

261 Because of the differences observed in PaperBLAST outputs when using the *E. coli* or *H. sapiens*
262 outputs, we repeated PaperBLAST queries using seven diverse sequences from the DUF34 family
263 reflecting different superkingdoms and alternative domain architectures (**Fig. 4; Data S7-S9**). The
264 differences in output first observed with two sequences were confirmed with more sequences
265 tested. One set of results, those of the *B. cereus* DUF34 homolog (UniProt: Q818H0), produced
266 publications that were entirely unique among the seven queries, unshared with the results of other
267 single-sequence queries. Coincidentally, the homolog of *B. cereus* contains an inserted domain
268 distinguishing it and others like it as members of a putative functional subgroup of the DUF34 family
269 [24]. Therefore, these unique retrieved publications reinforce that an understanding of the
270 taxonomic distribution of protein family domain architecture diversities is important to develop
271 prior to selection of representatives for single-sequence-based literature retrieval via PaperBLAST.

272 In summary, the general lack of standards and guidelines across the community, of which could
273 be implemented by publishers [43], make extracting published information on all members of a
274 protein family a time-consuming and error-prone exercise, particularly when relying upon text-based
275 search tools. Sequence-based searches should be the starting point of any family level review, as
276 demonstrated above. This can be complemented by text-based searches, but any outputs derived
277 from these queries should also be checked by sequence for the expected protein family
278 membership.

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

281 Ideally, a comprehensive capture of all publications linked to all members of a protein family
282 would require: 1) both text-based and sequence-based search tools; and 2) iterative cycles of
283 querying, curating, and cataloging ("QCC cycle", **Fig. 5**). However, there exists an inevitable law of
284 diminishing returns with such a process (e.g., fewer total new relevant publications per hour), with
285 productivity gradually decreasing as more time passes until no new publications are retrieved with
286 continued iterations. Even for a dedicated biocuration expert, the total amount of non-redundant
287 data retrieved per unit time exponentially decreases after a certain number of hours. The amount of
288 time necessary to invest, however, will differ between protein families and the time investment
289 determined appropriate will vary between researchers performing the searches.

290 To optimize search time, an additional investigation was undertaken to better understand the
291 differences between the sequence- and HMM-based methods of PaperBLAST and the high-
292 investment approach presented by the "QCC cycle". The results of four different query result sets
293 henceforth referred to as four distinct "methods" of retrieval ({1} HMM-derived, {2} PubMed text-
294 derived, {3} QCC/Curated-derived, and {4} three separate sequence-derived sets: (A) one using only
295 *E. coli* DUF34 homolog sequence; (B) a second that was a merged pair derived from two sequences
296 (DUF34 homologs of *E. coli*, *H. sapiens*); and (C) a third constituted by merging seven sequence-
297 derived query result sets (DUF34 homologs of *E. coli*, *H. sapiens*, *B. cereus*, *M. jannaschii*, *M.*
298 *maripaludis*, *S. cerevisiae*, *S. pombe*, sequences chosen to represent the putative diversity of the
299 family) were compared for determining overlaps and method-unique yields (**Fig. 6a**; **Table S2; Data**
300 **S9**). Only 5 unique publications were shared between all three PaperBLAST-derived result sets
301 (methods {1}, {3} and {4, merged}). There were no results unique to any one of the single-sequence-
302 derived results (method {4}, lists A-C). A comparison of all different types of literature search
303 methods focusing on their relevant results (i.e., those classified as "non-focal" or "focal")
304 demonstrated that no single method, text- or sequence-based—regardless of the number of
305 sequences—can capture everything available, emphasizing the needs of these distinct methods of
306 search (**Fig. 6b**). Moreover, the sequence-based searches retrieved seven publications that were not
307 captured by the HMM-based results, the two single-sequence-based results (method {4}B), the one
308 single-sequence-based result (method {4}A), or the idealized "QCC cycle" method (**Fig. 6**; **Data S10**).
309 These data suggest that the HMM-based method was the most efficient approach of the two offered
310 by the PaperBLAST suite when seeking a more comprehensive family-level review of the corpus in
311 the least amount of time and use of only one resource. Finally, when comparing PaperBLAST as a
312 resource overall to the far more tedious "QCC cycle" approach, it was observed that PaperBLAST still
313 misses ~75% publications relevant to the example DUF34 family, so one cannot avoid the iterative

314 and semi-manual "QCC cycle" approach, for now, to capture the entirety of published knowledge
315 across the corpus for any given protein family.

316 The taxonomic distribution of protein families and their respective subgroups have been
317 demonstrated to impact the literature capture process through the availability heuristic [44,45] and
318 unique biases influencing researcher keywords and search engines of choice. That is, any given
319 researcher may have a different starting idea of a protein family and this beginning governs their
320 ultimate search results when attempting to broaden their understanding of the same protein family.
321 However, there are many methods one may use in microbiology to better understand proteins at
322 the family level before completing a literature search. In addition to the strategies addressed above,
323 the improvement of your starting conceptions of a family's diversity and putative subgroups can be
324 accomplished using a variety of phylogenomic and phylogenetic tools. These tools are discussed in
325 detail in the subsequent sections.

326 **Resources that examine the taxonomic distributions of protein families are essential components**
327 **of comparative bioinformatic analyses.**

328 Before comparing the tools important for family-level analyses, it is important to, first, define
329 some of the key terms used to describe them. Family-level bioinformatic tools are primarily split into
330 two categories: 1) phylogenomic and 2) phylogenetic. "Phylogenomic" describes analyses
331 considering whole genomes or large regions of genomes. Tools investigating gene synteny and
332 neighborhoods are often considered "phylogenomic", depending on the nature of their genomic
333 comparisons. Phylogenetic tools focus on the presence of individual gene/protein sequences
334 compared between genomes. For years, PubSEED [46] had been an ideal resource for examining
335 taxonomic distributions of protein families, as a user could rename member proteins and visualize
336 their distributions in sets of genomes or user-defined "Subsystems" with a color-coding system that
337 highlighted synteny and that could be used to gather both phylogenomic and phylogenetic data. If
338 PubSEED is still functional it is frozen at ~10,000 genomes and so cannot be considered a main
339 source for analyzing taxonomic distributions when over 200,000 complete genomes are now
340 available. Here, we surveyed different types of webserver-based resources designed for
341 phylogenomic and phylogenetic analyses (**Data S2a**). These resources can be separated into several
342 types: A) general orthology databases (precomputed phylogenetic distributions; can also include
343 protein family classification databases); B) synteny/gene neighborhood databases (precomputed
344 phylogenomic record data, distributions; often features within larger databases); and C)
345 phylogenetic pattern/profile databases (often features within larger databases; precomputed

346 phylogenetic distributions) (**Fig. 7**). These three types of analyses can be further defined by the
347 respective parametric restrictions of the tools used to execute them into four subtypes: 1) custom
348 genome selection with single target family selection; 2) tool-defined genome selection with single
349 target family selection; 3) tool-defined genome selection with multiple target family selection; and
350 the rarest of the subtypes, 4) custom genome selection with multiple target family selection. An
351 additional feature among these tools is sequence-based tool input or sequence similarity-based
352 thresholds in family identification across genomes (also noted in **Fig. 7** as a minor subcategory of
353 "Target Selection"). These four phylogenetic/phylogenomic tool subtypes are detailed further below,
354 and a subset we use most frequently for both teaching and research is given in **Table 2**. The
355 strengths and weaknesses of these tools are compared and discussed below using the specific
356 DUF34 family-linked clustered orthologous groups (COGs) identified previously: COG0327, COG1579,
357 and COG3323 [24].

358 ***Subtype 1: custom genome selection, single target family selection.***

359 Phylogenetic and phylogenomic analyses, in general, are largely governed by two variables, as
360 alluded to in **Fig. 7**: 1) the number of targets (i.e., families, groups, neighborhoods)
361 viewable/analyzed at once, and 2) the number of genomes one can view these data across at once
362 (and whether those genomes can be custom selected). Many of the tools available via webserver are
363 restricted by one or the other, often both. Additionally, it is common for a webserver's workflow to
364 begin with a single sequence or family identifier, a paradigm common across phyletic tool subtypes
365 defined in this work. Divergences from this framework are discussed in later subsections (e.g., see
366 discussions of BV-BRC), but are also more commonplace in the first of our four defined subtypes.
367 One of the subtype 1 example resources for phylogenetic analyses, JGI-IMG, can begin with a single
368 sequence/family but can also begin at the level of user-selected genomes, taxonomic ranges. Most
369 subtype 1 tools are components or features within a larger suite or database, with or without an
370 account-linked workspace. Physical clustering is a key type of association-based inference derived
371 from genomic sequences and links genes to putative functions based on the annotations of their
372 encoded neighbors, given that strong conservation is observed [33]. Many gene neighborhood or
373 physically clustered gene viewers also fall within this subtype (**Fig. 7**). One example of a free-
374 standing subtype 1 gene neighborhood viewer is WebFlaGs [47] (https://server.atkinson-
375 lab.com/webflags) (**Fig. S7**), which permits the user to input many sequence identifiers (i.e., NCBI
376 protein accessions) within a protein family for the generation of a taxonomically clustered set of
377 gene neighborhoods.

378 Sequence similarity networks (SSN), while not classified as a major analysis type to be shown in
379 **Fig. 7**, are becoming more common across comparative genomics [48,49]. A popular SSN-generation
380 tool, EFI's Enzyme Similarity Tool (EST) [8,50] (https://efi.igb.illinois.edu/efi-est/) (**Fig. S8**), falls
381 within the subtype 1 group, specifically for its primary means of job submission ("Sequence BLAST"
382 and "Families"). Additionally, this suite provides options for gene neighborhood analyses, either
383 linked to submitted EST jobs or independently generated user-submitted SSNs. Because of the
384 numerous options made available within the EFI suite, this tool could also be considered subtype 3
385 or even subtype 4, depending upon a user's creativity in implementing the sequence list-based job
386 submission form and other features.

387 ***Subtype 2: tool-defined genome selection, single target family selection.***

388 Subtype 2 phylogenomic/phylogenetic tools are by far the most common and are often
389 embedded as part of a larger database (e.g., CDD, PANTHER, MBGD). Because they usually require
390 little computation or are precomputed, they are ideal for a first pass investigation of a protein
391 family.

392 A user-friendly subtype 2 tool for deriving family-level distribution information is the
393 phylogenetic tree viewer, Annotree [51] (http://annotree.uwaterloo.ca/annotree/). Building on the
394 protein family information derived from Pfam (now integrated into InterPro), TIGRFAM (no longer
395 maintained), or KEGG (KO families), Annotree provides a practical "first pass" in examining a protein
396 family's taxonomic distribution [51] (**Fig. S9**). Several output parameters can be actively modified by
397 the user in-browser (e.g., taxonomic ranges for tree branching and, separately, labeling of those
398 ranges). However, Annotree is restricted to bacterial and archaeal taxa and does not allow for the
399 examination of multiple target families.

400 Gene neighborhood and synteny tools that fall within subtype 2 are usually free-standing while
401 also being dependent upon the aggregation of information from other databases (e.g., COGNAT [52]
402 (https://depo.msu.ru/module/cognat) **(Fig. S10**)). An embedded feature of the KEGG Database
403 makes it possible to extract "Gene Cluster" (gene neighborhood) information for individual genes,
404 sometimes provided in parallel to neighborhood data of closely related organisms (that is, if the
405 majority of the Cluster is also conserved across those closely related target homologs). However, this
406 tool does not provide an option of viewing these data across taxa as a distribution per target
407 gene/protein family. Even more, this viewer is without an option for tabular export. Another subtype
408 2 comparative gene neighborhood tool that is also a component within another database is the

409 Genomic Neighborhood Comparison viewer subsection each *B. subtilis* gene entry within the recent
410 beta release ("December update") of SubtiWiki's "CoreWiki" (http://corewiki.uni-
411 goettingen.de/welcome) [53] (**Fig. S11**). This feature provides a swift overview of the homologous
412 gene neighborhoods of select model bacterial alongside that of the entry page's target gene and
413 corresponding neighborhood. The additional genomes featured in the viewer are fixed by the
414 database and the visualization is generated within-page using the beforementioned subtype 1 tool,
415 WebFlaGs.

416 ***Subtype 3: tool-defined genome selection, multiple target family selection.***

417 Phylogenetic/phylogenomic tools that are classified here as "subtype 3" allow for the selection
418 of multiple target gene/protein families but are restricted in the genomes across which those
419 families may be viewed or analyzed. Examples of precomputed phylogenetic databases of this
420 subtype include MicrobesOnline, STRING-DB, *fast.genomics*, and KEGG Orthology (KO).
421 MicrobesOnline, while being a multi-faceted sequence database, allows the user to choose a set of
422 input families using different types of systematic identifiers like COGs or Enzyme Commission (EC)
423 numbers for generating phylogenetic profiles, which are graphically produced and clustered
424 taxonomically with the absence-presence of the families/members across the database's benchmark
425 1,965 organisms (**Fig. S12**). This tool is notably user-friendly with different methods of family
426 member identification/filtration possible for selection per target; in addition to systematic identifier
427 annotations, options for these filters also include several BLAST cut-offs (**Fig. S13**). Users may also
428 view the precomputed phyletic profile for a single family via any gene entry's "Gene Info" tab (**Fig.
429 S14**). Unfortunately, MicrobesOnline is, to-date, frozen at a total of 3,707 genomes (retrieved
430 January 14th, 2022), and, further, these genomes are largely limited to bacterial organisms with only
431 94 archaea and 119 eukaryotes, the latter of which are mostly fungi.

432 While primarily an annotation network visualization tool, the STRING Database also features a
433 tool designed for the rapid survey of phylogenetic co-distribution of protein families (**Fig. S15**). While
434 useful for exploring annotations and hypothesis generation, the tool is a poor source of primary data
435 without the paired implementation of much more systematic, stringent analytic pipelines.

436 Because the KEGG database uses relatively stringent family relationships to create their
437 orthologous groups (i.e., "KOs" or K numbers [54]), we find that the KO database can be quite useful
438 to analyze the phylogenetic distribution of specific families. The tool produces a table of protein
439 distribution among all genomes present in the KEGG dataset using KO identifiers (**Fig. S16a**), which

440 the user provides in a space-separated list (**Fig. S16b**). However, not all protein families have been
441 assigned to a KO group and the genomes are organized in an order without clear reference to
442 taxonomic relations and the data shown is generated based upon the genomes in which at least one
443 of the submitted KOs occurs. Because of the latter feature, it is recommended that, with tools like
444 this, the user co-submit a positive control KO (i.e., a group that is known to be universally conserved
445 across database genomes) to ensure that all genomes benchmarked within KEGG are called in the
446 results. Further, export for this webserver output is not necessarily tabular or tabular-compatible
447 (i.e., HTML-embedded table) and, therefore, will require additional data tidying due to paralog-
448 related row duplications (i.e., duplicate rows lack names, which may be particularly troublesome for
449 tidying without specialized programmatic script development). Recently, KEGG Synteny database,
450 queries of which also use K numbers for input (two or more at a time), has been expanded to include
451 the entirety of genomes in KEGG.

452 A recently developed tool of particularly exciting functionality is *fast.genomics*, a tool built
453 within the suite of PaperBLAST (**Fig. 8**) [55] and that uses the genomes of MicrobesOnline. This tool
454 is one of very few that pairs the power of a sequence-based search in tandem with the inferential
455 value of clustering analysis for pairs of distinct protein families (**Fig. 4**). Several databases specifically
456 designed for biosynthetic gene cluster analyses also fall within tool subtype 3, allowing for multiple
457 target sequences/families as input while restricting genomes included in the analyses. An example of
458 this subtype is CAGECAT [56] (https://cagecat.bioinformatics.nl) (**Fig. S17**), which uses NCBI
459 sequence identifiers for input (multiple) chosen by the user. While users can also select specific
460 organisms or taxonomic clades (by Genus), the queries are somewhat limited and predefined by the
461 server's database. A final example of this tool subtype, FunCoup [57] (https://funcoup.org/search/)
462 (**Fig. S18**), takes a meta-analytical approach to family-level analyses, aggregating and summarizing
463 data supporting functional coupling between proteins across a limited number of genomes,
464 phylogenetic profiling via InParanoid [58] being one of several evidentiary criteria.

465 ***Subtype 4: custom genome selection, multiple target family selection.***

466 Subtype 4 tools are most commonly available in the form of online suites and custom account-
467 linked workspaces (e.g., Galaxy [59] (https://usegalaxy.org/), BV-BRC (formerly, PATRIC)). Like
468 Annotree, BV-BRC's Comparative Systems tool is restricted taxonomically to bacteria and archaea
469 [32] (https://www.bv-brc.org), but with the additional inclusion of viruses (**Fig. S19-S20**). The output
470 of this tool includes a searchable heatmap for all identified gene families across a custom selection
471 of genomes, the results of which can then be filtered using family identifiers (PGF IDs). MicroScope

472 (the microbial platform of GenoScope) also possesses a Gene Phyloprofile tool. Multiple genomes
473 (PkGDB, i.e., reannotated bacterial RefSeq genomes/proteomes) can be compared based on single
474 or multiple genes/proteins, in addition to whole genome-to-genome phylogenomics. The ultimate
475 result of this program is an output in the form of an HTML embedded table with each selected
476 genome represented in a separate column (not row). Finally, JGI-IMG provides a tool suite that
477 allows for the examination of custom genome lists with the use of many common systematic
478 identifiers, such as KOs, COGs, and Pfams (i.e., "Find Function" feature of the suite). Again, the
479 output for this tool is restricted to an HTML embedded table format but can be customized and
480 exported in tabular format. In general, if all these tools are quite user-friendly and useful for first
481 pass analyses, they are currently limited by the reliance on precomputed family annotations that can
482 be partial, too broad, or simply incorrect [60].

483 One phylogenetic analysis tool of subtype 4 that is not dependent upon a user account-linked
484 workspace is EggNOG's Phylogenetic Profile tool [61] (http://eggnog6.embl.de/app/phyloprofile/)
485 (**Fig. S21**). Uniquely, this browser-accessible resource allows for the input of multiple COGs and
486 multiple user-selected genomes, the latter input being taxonomic identifiers. The submitted job
487 results in a heatmap visualization showing absence-presence of selected orthologous groups across
488 the user's custom-selected genomes.

489 The Department of Energy Systems Biology Knowledgebase (KBase) has made analytic modules
490 and pipelines available for researchers who lack programming skills [62]. Any KBase user account
491 allows for browser-mediated access to complete suites of common bioinformatic analyses using
492 either publicly accessible or user-uploaded data. Resources provided by KBase map out specially
493 ordered "narratives" (i.e., an organized set of data objects and application queues within a digital
494 notebook) for completing phylogenomic analysis starting from species trees, but such a pipeline can
495 be unwieldy for novice users (**Fig. S22**; figure adapted from KBase 2020 phylogenetics narrative
496 diagram). It should also be noted that analyses can take many hours depending on the number of
497 genomes being analyzed, and such investments may be important timeline considerations for
498 experimentalists.

499 Because of those described and many other challenges to tool usage, resources providing
500 guidance appropriate for the microbiological field's diverse audiences are increasingly necessary.
501 Support of resources are often directly tied to the perceived use on part of the target community
502 (frequently measured by citations, link traffic when reported to supporting agencies, organizations).

503 Therefore, if we are to ensure the continued success and access to many high-quality resources, it is
504 sometimes necessary to assist in the effective accessibility of their use by diverse userbases.

505 **Creating a Wiki compiling a non-exhaustive list of web-based resources organized into pedagogical**
506 **modules for microbiologist.**

507 A persistent challenge exists within the bioinformatic community; that is, the ability to know
508 which tools are available and which are most suitable for fulfilling our data analysis and visualization
509 objectives. As time passes, more tools are published with others being decommissioned nearly at
510 much the same rate, the longevity of most tools maintaining uncertainty across their lifetime due to
511 funding instability [63]. In 2015, Attwood and colleagues found that after 18 years (1997-2015) over
512 60% of catalogued databases within DBcat (www.infobiogen.fr/services/dbcat) were "dead" (i.e.,
513 server was unresponsive, or search/other major functionality was not functioning) [63]. More
514 recently, Kern and colleagues found, in a survey of 2618 tools published between 2019 and 2020,
515 that 10% of them were unreachable, the rate of which they observed to increase linearly with earlier
516 dates of publication [64]. A few sites have been dedicated to the aggregation of the totality of useful
517 bioinformatics resources (e.g.: bio.tools [65], https://bio.tools; CNCB Database Commons [66],
518 https://ngdc.cncb.ac.cn/databasecommons/; Nucleic Acids Research's regular Database Issue [67]),
519 but—in addition to being understandably challenging to maintain—the lack of grassroots- or
520 leadership-level efforts to popularize some of these resources have left them of low findability and,
521 therein, in deficit of broad use by the community. Only more recently have sites like CNCB Database
522 Commons been recommended by the likes of *Cell Press* (https://marlin-
523 prod.literatumonline.com/pb-assets/journals/research/cellpress/data/RecommendRepositories.pdf)
524 or *Bioinformatics Advances* (https://academic.oup.com/bioinformaticsadvances/pages/instructions-
525 to-authors). As of November 29, 2023, CNCB Database Commons indicated that 5213 of 6380 (82%)
526 biological databases were annotated as being "alive".

527 In response to our own difficulties in navigating the ever-changing frontier of bioinformatic
528 tools, a wiki of webtools was established, initially, for our laboratory's in-house use, and, later, was
529 further developed with the intention of aiding other microbiologists (https://vdclab-
530 wiki.herokuapp.com/). With 15 years of instructional experience in bioinformatics specifically for
531 microbiologists, this resource was designed with our own graduate-level courses in mind, in addition
532 to featuring some of our lab's own commonly used bioinformatic workflows. The wiki was created
533 using the pedagogic modules of already well-established bioinformatic courses and their learning
534 objectives, which were used to model the website's subsets of information, keywords, tags, and

535 relationships between links. Additional pages provided in the main navigation like "VDC Favorites"
536 and "Recent Finds" contribute a personal touch—one curated and the other a chronological,
537 ongoing, and uncurated feed—to the website's suite of information. Ideally, this custom collection
538 of tools and workflows will aid other microbiological experimentalists that are less familiar with the
539 user-friendly bioinformatic resources available today.

540 Conclusions

541 Comprehensively extracting, synthesizing, and properly propagating scientific observations
542 among databases, all in a manner that adheres to and further fosters the use of FAIR data guidelines
543 remains a challenge [69]. Here, we examined the challenges pertaining to the capture of the
544 literature on whole protein families. The curation of published data, alongside the interrogation of
545 available tools common to this process, were surveyed and workflows incorporating different
546 iterations of them were compared to provide a minimal workflow that can be followed by users to
547 optimize search time (Fig. 1 and 5). Several potential pitfalls and stumbling blocks commonly
548 encountered by researchers during biological publishing were identified and described, and further
549 supplemented by examples of each using the case study of the DUF34 protein family. Importantly, it
550 was observed that the choice of keywords and search engines, though equally important, vary both
551 together and independently in how they influence published data capture results. Additionally, false
552 positives across search engine types illustrated the importance of thorough, well-informed curation
553 efforts, and the need for more stringent standards among publishers. Related and although tools
554 that allow this form of search are limited in number, sequence-based searches were determined to
555 be critical first steps of the data capture process at the protein family level. A comparison of merged
556 query result lists derived from varying numbers of single-sequence-based searches (i.e., 1-sequence,
557 2-sequences, and 7-sequences) to those of PaperBLAST's HMM-based search tool demonstrated
558 that, despite being drawn from the same publication-bioentity/sequence ID crosslink network, each
559 method—no matter the number of representative sequences used to generate a consolidated list
560 for the sequence-based results—provided yields dissimilar in total and quality. The greatest
561 distinctions between lists were observed between the sequence-based list derived from only one
562 family member sequence of *E. coli* and that of the HMM-based results. Interestingly, the yield
563 increase of single-sequence-based queries was shown to have largely plateaued between the lists
564 derived from one sequence to two sequences, with only a marginal increased yield of publications
565 between lists derived from two and seven sequences. These results implied that the most impactful
566 factor for improving single-sequence-based results was not necessarily the total number of

567 sequences used to compile the final results list. Moreover, and consistent with this hypothesis, the
568 member sequence representing the known most-divergent domain architecture of the DUF34
569 protein family, that of *B. cereus*, was also the sequence to derive the greatest number of
570 publications unique to its query when comparing the individual single-sequence-based yields.
571 Consideration of the QCC cycle approach in these comparisons further emphasized that, while no
572 single method appeared optimal, it was this more tedious method that was observed to outperform
573 all others in total unique publications and rate of false positives, the latter of which was largely
574 driven by the curation-based nature of the QCC cycle approach. As it relates to the interest of
575 increasing the efficiency with which a given experimentalist might capture a sufficient portion of
576 literature relevant to a protein family, these analyses suggested that the HMM-based method
577 available through PaperBLAST would likely best serve this purpose as it was the approach that
578 retrieved the most unique and relevant publications with the least amount of time and resource
579 investment.

580 Because of the importance of protein family-level information in guiding the published data
581 capture process, a survey of web-based phylogenomic and phylogenetic tools was performed
582 highlighting those of higher usability and interoperability. Mapping identifiers between databases is
583 central to comparative genomics approaches as users often must use a variety of resources with
584 distinct features and formats in order to accomplish their analytic objectives. However, the more
585 work and time necessary to transform and/or map data between resources, the less likely they are
586 to be used together or in tandem, regardless of how high-quality their visualizations or how well-
587 curated their data [70–74]. Although not thoroughly discussed, interoperability was observed to be
588 notably bereft between the phylogenetic/phylogenomic tools surveyed, as well as with those of
589 other types of biological databases. While it has been argued to simultaneously contribute to the
590 larger systemic problem, the growing diversity of tools also suggests that the challenges driven by
591 persisting interoperability deficits between resources have yet to be sufficiently ameliorated and
592 continue to obstruct analytic processes. That is, publishing behavior of biological tools suggests that
593 many researchers continue to resort to "reinventing the wheel" instead of developing tools to bridge
594 the gaps between different resources or, alternatively, work with existing resources to improve their
595 interoperability. To confront these trends in our recommendations to experimentalists, we designed
596 a pedagogical wiki of bioinformatic workflows using tools with long, well-tried histories of continued
597 support, adequate quality curation, and maintenance. Ultimately, it is our hope that this work
598 provides a framework with which experimental microbiologists can perform routine operations on
599 genes and proteins of a given family even as the genomic data size increases.

Figures and Tables

Table 1. DUF34 homolog search terms and search engines selected for investigating the quantitative impacts of user choice for each upon literature queries.

DUF34 Homologs	Literature Search Tools	
(Keywords/Search Phrases)	**Qualitative Search Category**	**Engine/Resource**
NIF3	*Specialized Scientific Literature Search*	PubMed
Ngg1-interacting Factor 3		Europe PMC
NIF3L1		PubTator
GTP cyclohydrolase 1 type 2		Scinapse.io
DUF34		BASE
Ybgl		
PF01784		
COG0372		
YqfO	*Broad Scientific Literature, Database Search*	ScienceResearch.org
COG3323		WorldWideScience.org
	Broad Scientific Literature Search	Microsoft Academic
		Google Scholar

Table 2. Preferred tools by phylogenomic/phylogenetic objective.

Tool	Feature Highlights
Distribution of Family Across Genomes *(Query Input: Target Sequence/Family)*	
fast.genomics	Exploratory; sequence-based; close-relative sequences only
OrthoMCL	Genome benchmarking; easy to use; limited genomes; independent protein classification system
MicrobesOnline	Genome benchmarking; easy to use; limited genomes
MBGD	Genome benchmarking; large number of tools; somewhat challenging to navigate
OrthoInspector	Genome benchmarking; limited genomes

Annotree	User-friendly; limited genomes
KEGG Orthology	Multiple families can be selected at once; large collection of benchmarked genomes; output is HTML embedded, making export difficult
Co-occurrence *(Query Input: Target Sequence/Family)*	
COGNAT	User-friendly; viewer is limited in scope; useful exploratory purposes, not for tabular analysis; single COG input, output highlights other COGs of high co-occurrence
fast.genomics	Sequence-based; hyperlinked table outputs
Taxonomic Distribution *(Query Input: Phyletic Pattern/Distribution)*	
MBGD	Genome benchmarking; large number of tools; somewhat challenging to navigate
Phylogenetic Profiler (JGI)	Choice of specific genomes
OrthoMCL	User-friendly; limited genomes
Physical Clustering/Synteny *(Query Input: Target Sequence/Family)*	
EFI GNT	Sequence similarity network-linked neighborhood analyses
BV-BRC (formerly, PATRIC)	Custom choice of genomes; personal account, workspace
fast.genomics	Exploratory, sequence-based; outputs somewhat taxonomically restricted
GeCoViz	Exploratory, custom choice of genomes (**Fig. S23**)
GizmoGene	Exploratory, custom sequence/genome selection; can work in tandem with BV-BRC (**Fig. S24**)

607

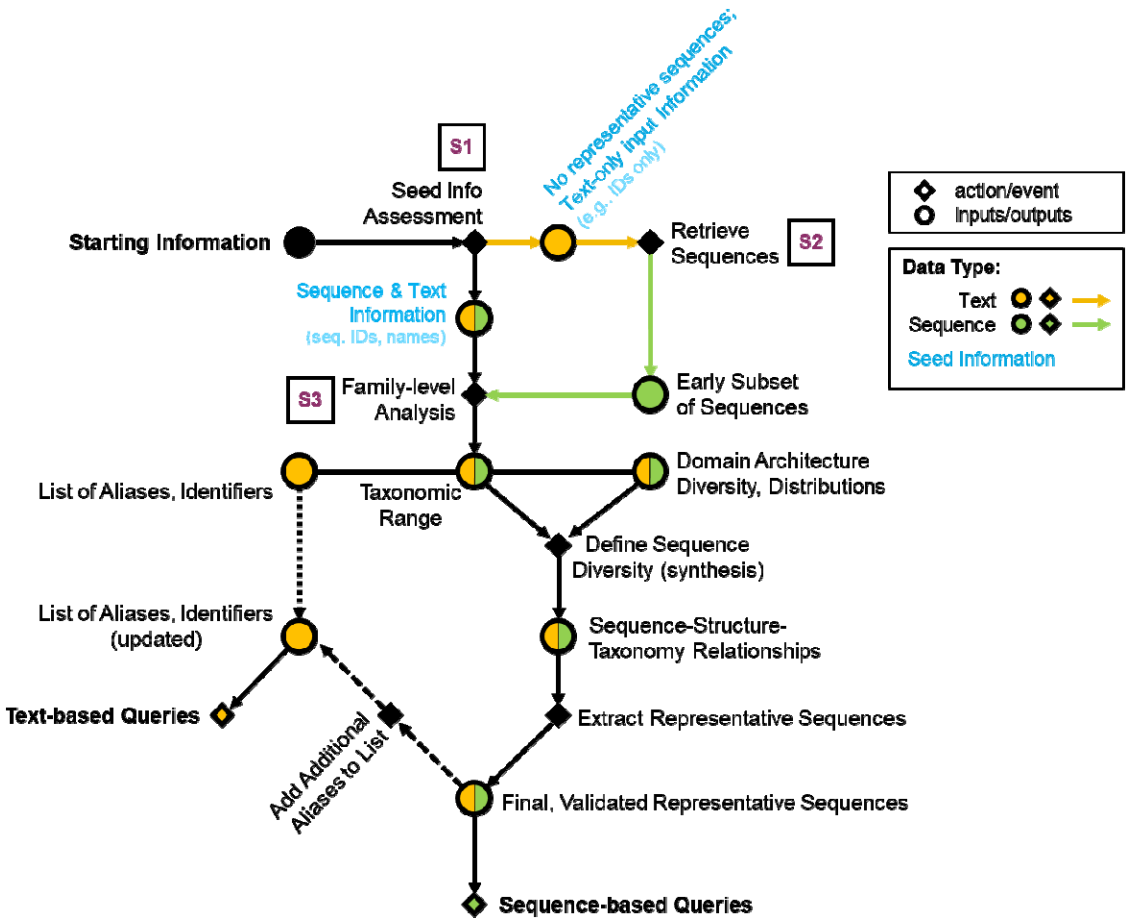


Figure 1. Workflow diagram recommended for capturing published data of a protein family. Supplemental figures were generated for examples (**Figures S1-S3**). Accompanying supplemental figures are boxed in the diagram and shown in purple text. Nodes of the directed network diagram are distinguished by the shape and color. Circular nodes indicate inputs/outputs (data), while diamond-shaped nodes denote an action or event. The colors of circular nodes convey the type of input/output data: green for sequences and yellow for text. Circular nodes regarding data of both types are split in two (proportions not to scale, do not reflect any guaranteed experience observed with the applied workflow).

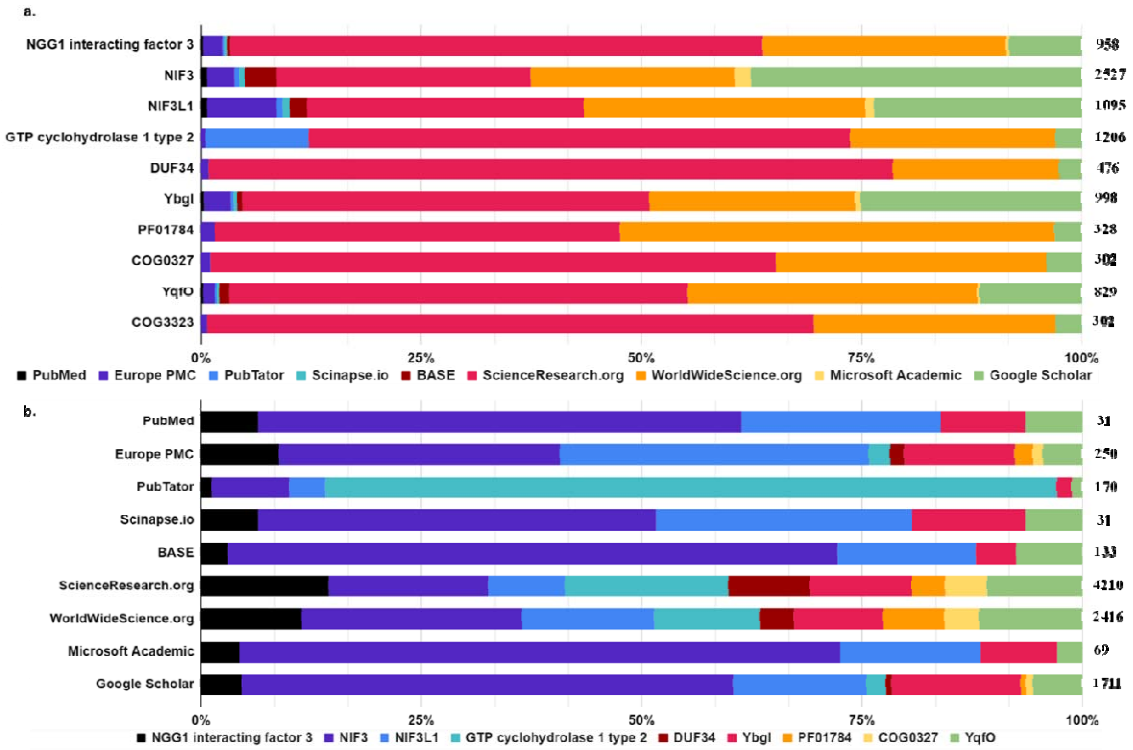


Figure 2. Query yield distributions per search tool as a function of keyword (a) and per keyword as a function of search tool (b). A subset of nine keywords most frequently associated with the target protein family, DUF34, was organized and used to compare the query results (i.e., total hits) across 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.

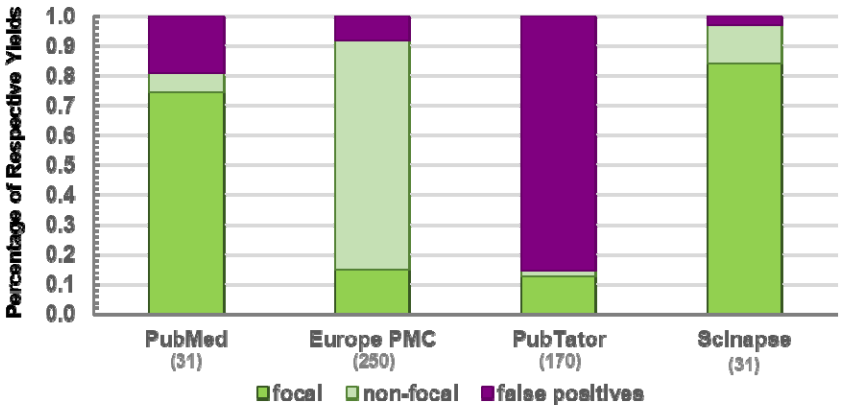


Figure 3. Query yield quality of text-based searches using more research-centered, conservative search tools. Focal publications were labeled as such if they featured relevant keywords specific to the protein family/family homolog in either the title or abstract of the publication. Non-focal publications were labeled if the keywords occurred in any other section of the paper. False positives were manually curated on a case-by-case basis.

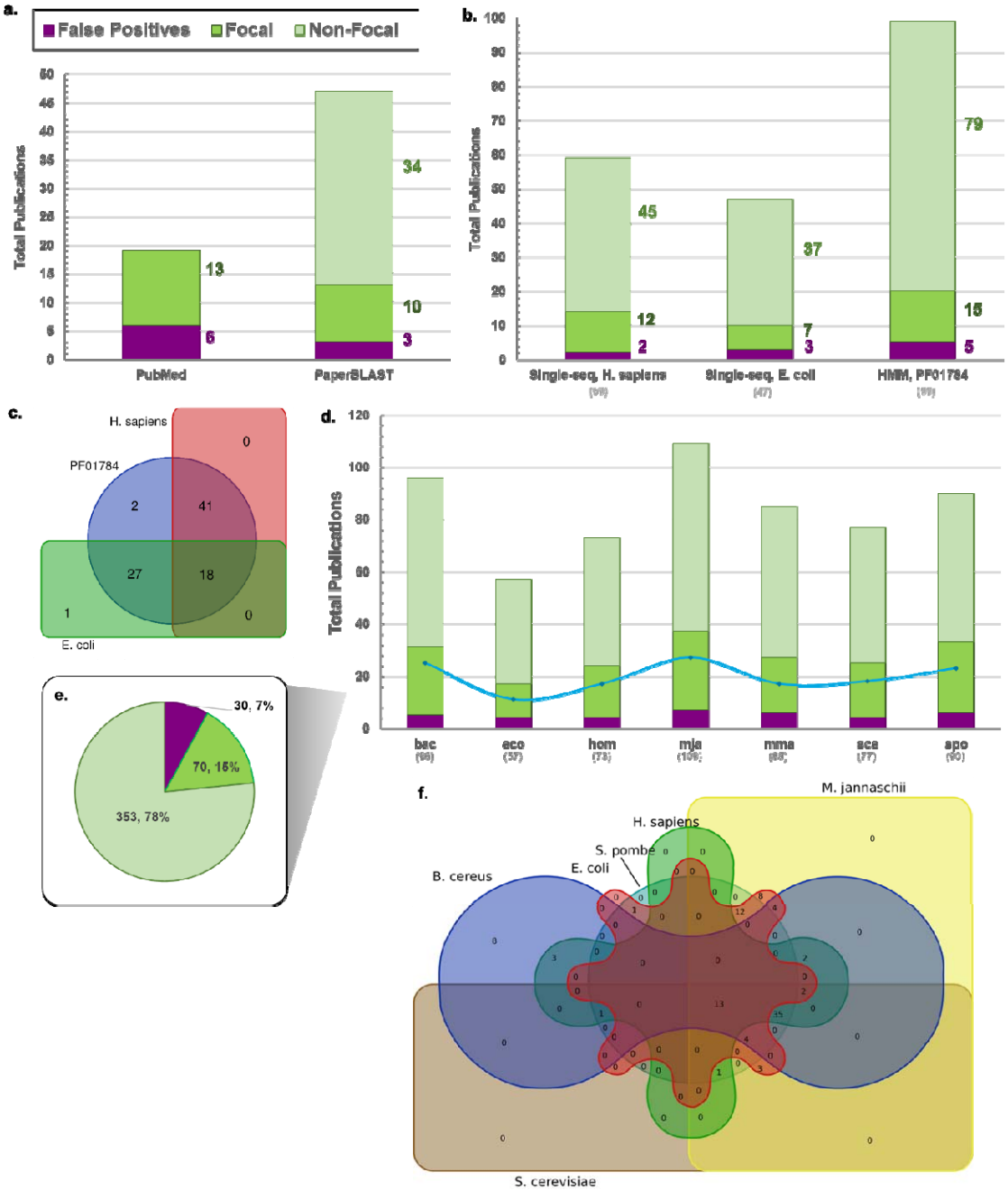
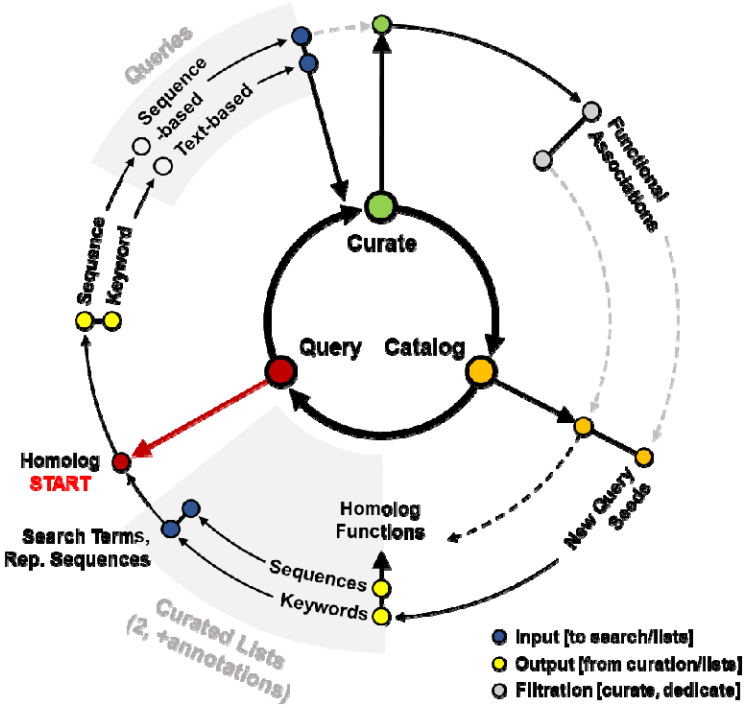


Figure 4. Quantitative and qualitative analyses of sequence-based literature search using PaperBLAST. (a) The yields of a PubMed text-based searches of 9 UniProt entry-derived keywords (source UniProt entry: *E. coli* DUF34 homolog, P0AFP6; keywords: "Ybgl", "P0AFP6", "b0710", "JW0700", "PF01784", "NIF3", "Ngg1-interacting factor 3", "DUF34", "COG0327") compared to those of PaperBLAST using one single-sequence-based query of the same source UniProt entry sequence. Bar color key for "false positives" (purple), "focal" (green), and "non-focal" (light green) designated publications applies to this panel (a) and panels (b), (d), and (e). "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

635 supplemental materials. (b) Comparison of publication retrieval methods available through PaperBLAST (i.e., HMM or Pfam
636 identifier-input-based query compared to single-sequence-based queries). Two disparately related DUF34 family homolog
637 sequences (one used in the preceding comparison of text- and sequence-based queries (a) and the other of a model
638 organism, *H. sapiens*) were chosen for comparison to the family's linked HMM profile identifier extracted from Pfam
639 (PF01784). (c) Visual comparison (Venn diagram) of unique yields derived from the prior sequence- and HMM-based
640 literature searches using PaperBLAST. The results for the homolog of *E. coli* (UniProt: P0AFP6) are shown in green area and
641 those for the *H. sapiens* homolog (UniProt: Q9GZT8) are shown in red. The results of the HMM-based query using the
642 family Pfam (PF01784) are shown in blue. (d) Query quality of PaperBLAST hits per DUF34 protein family member
643 sequence, one query sequence per organism. Organisms selected by tentative domain architectural subclasses and
644 taxonomic distribution of members. A blue line marks the occurrence of redundant results within a single query (avg. ~23%
645 of hits per query). All selected query sequences have been independently described in a scientific publication [24]. Total
646 hits per sequence/query are shown below the x-axis labels. (e) Overall query quality across all representative sequences
647 used to query PaperBLAST shown in (d). Total of represented hits equals 453. (f) Visual comparison (Venn diagram) of all
648 methods of literature retrieval examined here. Yields shown are 6 of the 7 sets of single-sequence PaperBLAST results;
649 diagram generating tool could not suit the total of 7 lists. With this, the results for the *M. maripaludis* DUF34 homolog
650 were not observed to have provided any novel results not retrieved by the other 6 sequences and so its yield list was not
651 included in the generation of this figure.

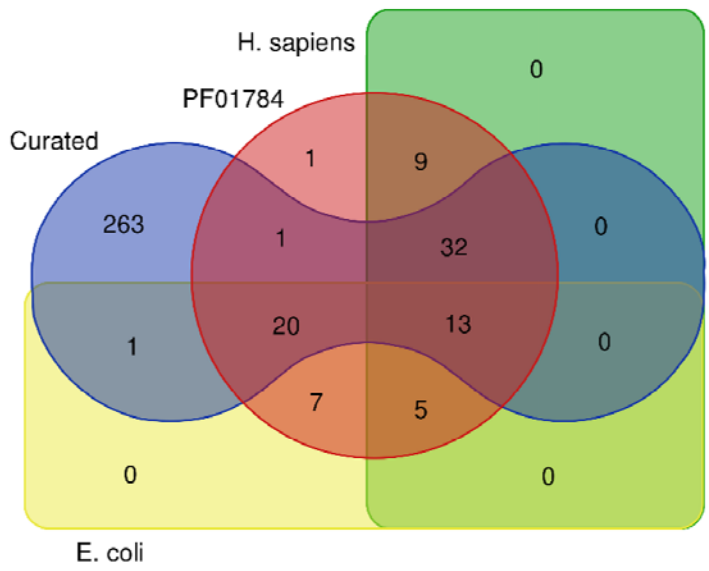
652



653 **Figure 5.** Idealized cycle of accumulating keywords, homolog annotations representative sequences, and publications
654 necessary to optimize the capture of all published data relating to a protein family. Query phase refers to the process
655 (often multiple in parallel) of using various webservers to retrieve literature relevant to target homolog(s). The Curate
656 phase is distinguished by its filtration and review of retrieved information, and, frequently, the identification of
657 experimentally validated functional associations to be noted for select homologs in the subsequent phase. The final phase
658 of the cycle, the Catalog phase, is defined by the multiple diverging paths along which the different identified information
659 will be channeled. These distinct paths of information included the two curated lists of keywords and representative
660 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 flow, whereas the black, solid arrows indicate explicit information flow. The dashed, black arrow denotes explicit information flow out* of the QCC cycle.

a.



b.

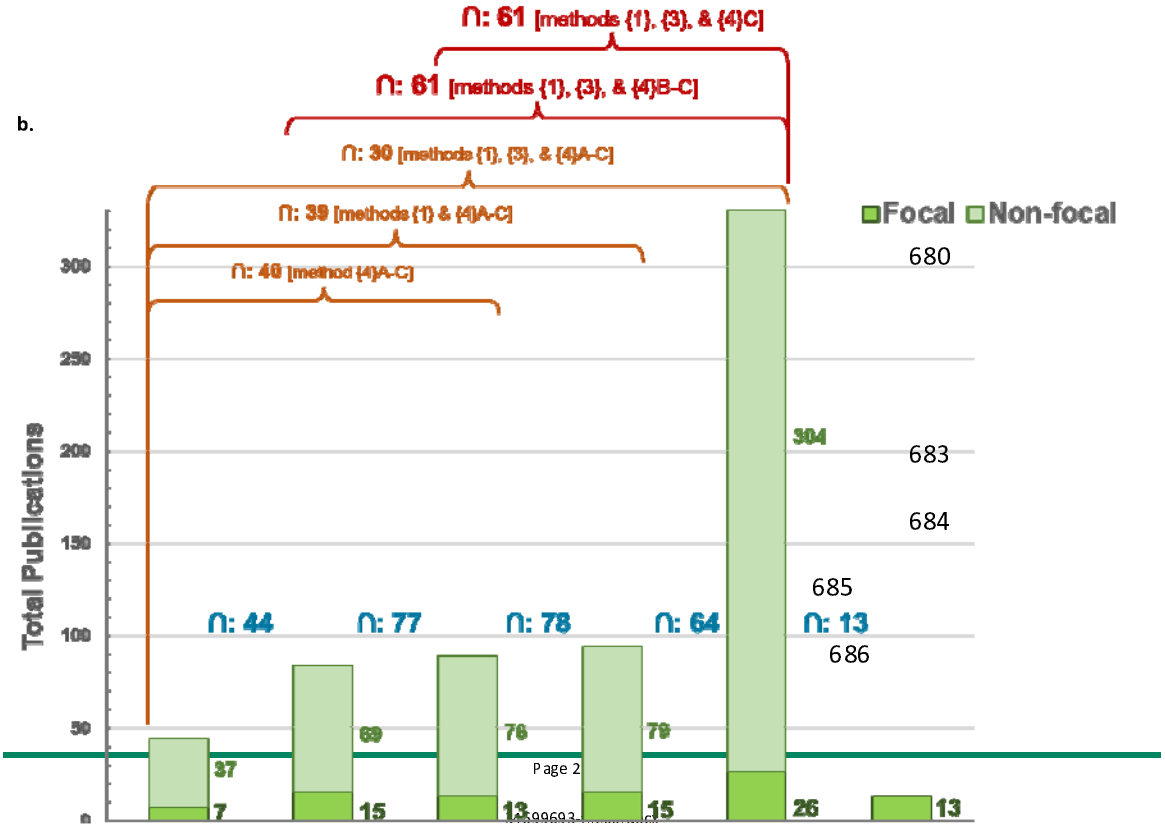
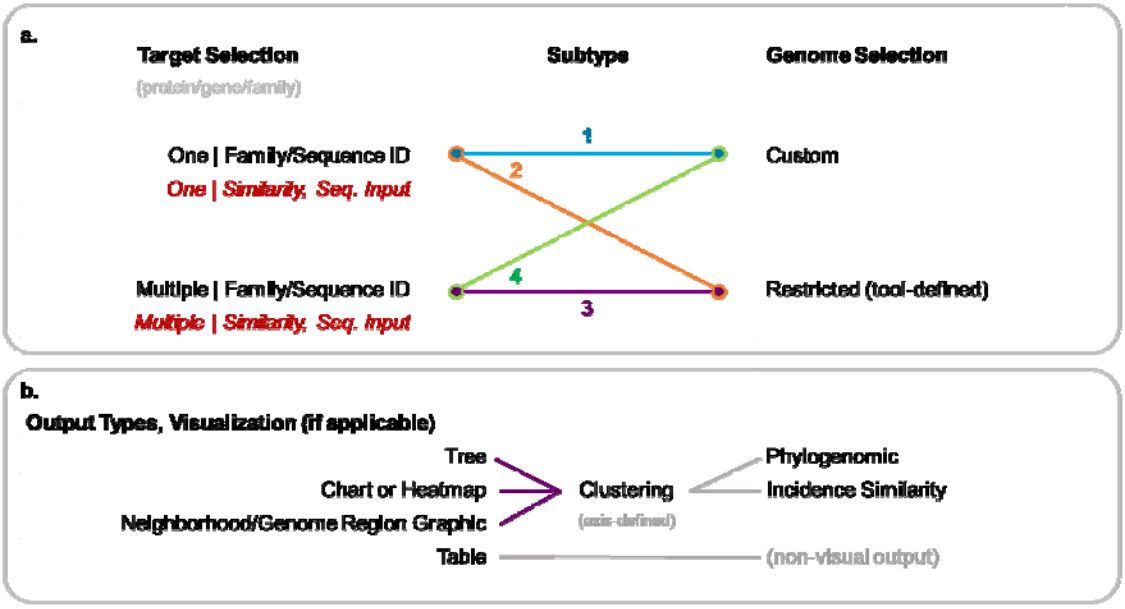


Figure 6. (a) Comparison of literature yields for selected PaperBLAST retrievals and the QCC cycle method. Asymmetric Venn diagram illustrates the distinct hits for each method. "H. sapiens Seq" (green area) denotes the unique results of single-sequence PaperBLAST output for *H sapiens* DUF34 homolog sequence (UniProt: Q9GZT8). "E. coli Seq" (yellow area) denotes the unique results of single-sequence PaperBLAST output for *E. coli* DUF34 homolog sequence (UniProt: P0AFP6). "Curated" (blue area) denotes the unique results of the idealized "QCC" Cycle output for the comprehensive investigation of the DUF34 family. "PF01784" (red area) denotes the unique results of the HMM-based PaperBLAST query output (HMM/Pfam: PF01784). (b) Focal and non-focal publications captured across all methods. Stacked bar plot of unique focal and non-focal results of all distinct literature retrieval methods. Intersection of yields between each pair of methods is shown in blue. Orange labels denote the "method" group, as do the red labels, the color of which were changed from those of the orange variety to emphasize the data patterns described in the text.



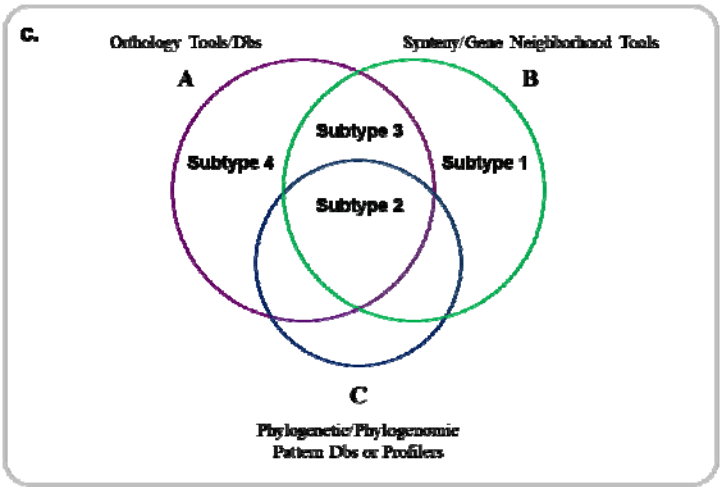


Figure 7. Categorization of phylogenomic, phylogenetic analyses and the diversity of tools available for performing them. (a) A list of three major categories of phylogenetic/phylogenomic analyses are listed first, A-C. (b) These major categories of analyses are further broken down into subtypes of tools used to perform those analyses. (c) All tools have a variety of outputs and result visualizations, and are, below the prior, examined in the context of the root output type, as well as any kernel-based clustering and corresponding type of clustering relative to axis/data-type being clustered (if applicable). (d) A summary diagram that illustrates the effective overlap of tool subtypes relative to the three major analysis types.

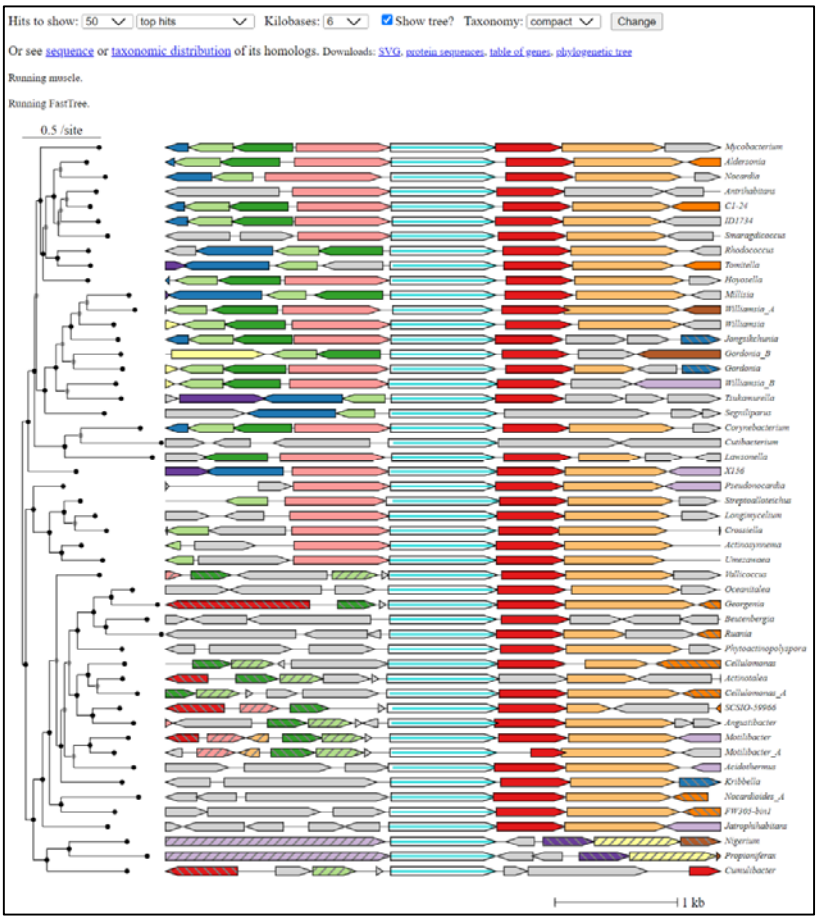


Figure 8. Graphical output of *fast.genomics* for one protein, Rv2230c (DUF34 family homolog, COG0327, P9WFM1 of *Mycobacterium tuberculosis* str. ATCC 25618 / H37Rv). A tree clustered view of physical clustering for closely related homologs is shown.

Author statements

Author contributions

Conceptualization: Valérie de Crécy-Lagard; Data Curation: Colbie J. Reed; Formal Analysis: Colbie J. Reed; Funding Acquisition: Valérie de Crécy-Lagard; Investigation: Colbie J. Reed; Methodology: Colbie J. Reed, Valérie de Crécy-Lagard; Project Administration: Valérie de Crécy-Lagard, Colbie J. Reed; Resources: Valérie de Crécy-Lagard; Software: Remi Denise; Supervision: Valérie de Crécy-Lagard; Validation: Colbie J. Reed, Valérie de Crécy-Lagard; Visualization: Colbie J. Reed; Writing – Original Draft: Colbie J. Reed, Valérie de Crécy-Lagard; Writing – Review & Editing: Colbie J. Reed, Valérie de Crécy-Lagard, Geoffrey Hutinet, Jill Babor, Rémi Denise; Tool/Resource Curation, Review

for Website Collections: Colbie J. Reed, Rémi Denise, Jacob Hourihan, Jill Babor, Marshall Jaroch, Maria Martinelli, Geoffrey Hutinet, and Valérie de Crécy-Lagard.

Conflicts of interest

The author(s) declare that there are no conflicts 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 (1979)* **1995**, *269*, 496–512, doi:10.1126/science.7542800.

2. Bansal, A.K. Bioinformatics in Microbial Biotechnology – a Mini Review. *Microb Cell Fact* **2005**, *4*, 19, doi:10.1186/1475-2859-4-19.

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.

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

762 9. Klimke, W.; O'Donovan, C.; White, O.; Brister, J.R.; Clark, K.; Fedorov, B.; Mizrachi, I.; Pruitt,
763 K.D.; Tatusova, T. Solving the Problem: Genome Annotation Standards before the Data
764 Deluge. *Stand Genomic Sci* **2011**, *5*, 168–193, doi:10.4056/sigs.2084864.

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

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

769 12. Vallenet, D.; Calteau, A.; Dubois, M.; Amours, P.; Bazin, A.; Beuvin, M.; Burlot, L.; Bussell, X.;
770 Fouteau, S.; Gautreau, G.; et al. MicroScope: An Integrated Platform for the Annotation and
771 Exploration of Microbial Gene Functions through Genomic, Pangenomic and Metabolic
772 Comparative Analysis. *Nucleic Acids Res* **2020**, doi:10.1093/nar/gkz926.

773 13. Chen, I.-M.A.; Chu, K.; Palaniappan, K.; Pillay, M.; Ratner, A.; Huang, J.; Huntemann, M.;
774 Varghese, N.; White, J.R.; Seshadri, R.; et al. IMG/M v.5.0: An Integrated Data Management
775 and Comparative Analysis System for Microbial Genomes and Microbiomes. *Nucleic Acids Res*
776 **2019**, *47*, D666–D677, doi:10.1093/nar/gky901.

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

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

784 16. Karp, P.D.; Ivanova, N.; Krummenacker, M.; Kyrpides, N.; Latendresse, M.; Midford, P.; Ong,
785 W.K.; Paley, S.; Seshadri, R. A Comparison of Microbial Genome Web Portals. *Front Microbiol*
786 **2019**, *10*, 208, doi:10.3389/fmicb.2019.00208.

787 17. Borda, S. If Data Is Used in the Forest and No-One Is Around to Hear It, Did It Happen? A
788 Citation Count Investigation. *Int J Digit Curation* **2023**, *17*, 14, doi:10.2218/ijdc.v17i1.830.

789 18. Blake, J.A.; Bult, C.J. Beyond the Data Deluge: Data Integration and Bio-Ontologies. *J Biomed*
790 *Inform* **2006**, *39*, 314–320, doi:10.1016/j.jbi.2006.01.003.

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

793 20. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, Ij.J.; Appleton, G.; Axton, M.; Baak, A.;
794 Blomberg, N.; Boiten, J.-W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding
795 Principles for Scientific Data Management and Stewardship. *Sci Data* **2016**, *3*, 160018,
796 doi:10.1038/sdata.2016.18.

797 21. Wei, C.-H.; Allot, A.; Leaman, R.; Lu, Z. PubTator Central: Automated Concept Annotation for
798 Biomedical Full Text Articles. *Nucleic Acids Res* **2019**, *47*, W587–W593,
799 doi:10.1093/nar/gkz389.

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

802 23. Wang, Y.; Wang, Q.; Huang, H.; Huang, W.; Chen, Y.; McGarvey, P.B.; Wu, C.H.; Arighi, C.N. A
803 Crowdsourcing Open Platform for Literature Curation in UniProt. *PLoS Biol* **2021**, *19*,
804 e3001464, doi:10.1371/journal.pbio.3001464.

805 24. Reed, C.J.; Hutinet, G.; de Crécy-Lagard, V. Comparative Genomic Analysis of the DUF34
806 Protein Family Suggests Role as a Metal Ion Chaperone or Insertase. *Biomolecules* **2021**, *11*,
807 1282, doi:10.3390/biom11091282.

808 25. Finn, R.D.; Mistry, J.; Tate, J.; Coggill, P.; Heger, A.; Pollington, J.E.; Gavin, O.L.; Gunasekaran,
809 P.; Ceric, G.; Forslund, K.; et al. The Pfam Protein Families Database. *Nucleic Acids Res* **2009**,
810 *38*, 211–222, doi:10.1093/nar/gkp985.

811 26. Blum, M.; Chang, H.-Y.; Chuguransky, S.; Grego, T.; Kandasaamy, S.; Mitchell, A.; Nuka, G.;
812 Paysan-Lafosse, T.; Qureshi, M.; Raj, S.; et al. The InterPro Protein Families and Domains
813 Database: 20 Years On. *Nucleic Acids Res* **2021**, *49*, D344–D354, doi:10.1093/nar/gkaa977.

814 27. Lu, S.; Wang, J.; Chitsaz, F.; Derbyshire, M.K.; Geer, R.C.; Gonzales, N.R.; Gwadz, M.; Hurwitz,
815 D.I.; Marchler, G.H.; Song, J.S.; et al. CDD/SPARCLE: The Conserved Domain Database in 2020.
816 *Nucleic Acids Res* **2020**, *48*, D265–D268, doi:10.1093/nar/gkz991.

817 28. Huerta-Cepas, J.; Szklarczyk, D.; Heller, D.; Hernández-Plaza, A.; Forslund, S.K.; Cook, H.;
818 Mende, D.R.; Letunic, I.; Rattei, T.; Jensen, L.J.; et al. EggNOG 5.0: A Hierarchical, Functionally
819 and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms and 2502
820 Viruses. *Nucleic Acids Res* **2019**, *47*, D309–D314, doi:10.1093/nar/gky1085.

821 29. Bateman, A. UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Res* **2019**, *47*,
822 D506–D515, doi:10.1093/nar/gky1049.

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

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

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

832 33. Overbeek, R.; Fonstein, M.; D'Souza, M.; Pusch, G.D.; Maltsev, N. The Use of Gene Clusters to
833 Infer Functional Coupling. *Proceedings of the National Academy of Sciences* **1999**, *96*, 2896–
834 2901, doi:10.1073/pnas.96.6.2896.

835 34. Pejaver, V.R.; Lee, H.; Kim, S. Gene Cluster Prediction and Its Application to Genome
836 Annotation. In *Protein Function Prediction for Omics Era*; Kihara, D., Ed.; Springer
837 Netherlands: Dordrecht, 2011; pp. 35–54 ISBN 978-94-007-0881-5.

838 35. Altenhoff, A.M.; Glover, N.M.; Dessimoz, C. Inferring Orthology and Paralogy. In *Methods in*
839 *Molecular Biology*; Humana Press: New York, New York, 2019; Vol. 1910, pp. 149–175 ISBN
840 9781493990740.

841 36. Gurska, D.; Jentzsch, I.M.V.; Panfilio, K.A. Mutual Regulation Underlies Paralogue Functional
842 Diversification. *bioRxiv* **2019**, 427245, doi:10.1101/427245.

843 37. Mirny, L.A.; Gelfand, M.S. Using Orthologous and Paralogous Proteins to Identify Specificity
844 Determining Residues. *Genome Biol* **2002**, *3*, doi:10.1186/gb-2002-3-3-preprint0002.

845 38. Zallot, R.; Harrison, K.; Kolaczkowski, B.; de Crécy-Lagard, V. Functional Annotations of
846 Paralogs: A Blessing and a Curse. *Life* **2016**, *6*, 39, doi:10.3390/life6030039.

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

850 40. Li, W.; Cong, Q.; Kinch, L.N.; Grishin, N. V. Seq2Ref: A Web Server to Facilitate Functional
851 Interpretation. *BMC Bioinformatics* **2013**, *14*, 1, doi:10.1186/1471-2105-14-30.

852 41. Jaroszewski, L.; Koska, L.; Sedova, M.; Godzik, A. PubServer: Literature Searches by Homology.
853 *Nucleic Acids Res* **2014**, *42*, W430–W435, doi:10.1093/nar/gku450.

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

856 43. de Crécy-lagard, V.; Amorin de Hegedus, R.; Arighi, C.; Babor, J.; Bateman, A.; Blaby, I.; Blaby-
857 Haas, C.; Bridge, A.J.; Burley, S.K.; Cleveland, S.; et al. A Roadmap for the Functional
858 Annotation of Protein Families: A Community Perspective. *Database* **2022**, *2022*, 1–16,
859 doi:10.1093/database/baac062.

860 44. Novin, A.; Meyers, E. Making Sense of Conflicting Science Information. In Proceedings of the
861 Proceedings of the 2017 Conference on Conference Human Information Interaction and
862 Retrieval; ACM: New York, NY, USA, March 7 2017; pp. 175–184.

863 45. MENG, S. Availability Heuristic Will Affect Decision-Making and Result in Bias. *DEStech*
864 *Transactions on Social Science, Education and Human Science* **2017**, 267–272,
865 doi:10.12783/dtssehs/msie2017/15448.

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

868 47. Saha, C.K.; Sanches Pires, R.; Brolin, H.; Delannoy, M.; Atkinson, G.C. FlaGs and WebFlaGs:
869 Discovering Novel Biology through the Analysis of Gene Neighbourhood Conservation.
870 *Bioinformatics* **2021**, *37*, 1312–1314, doi:10.1093/bioinformatics/btaa788.

871 48. Knox, H.L.; Allen, K.N. Expanding the Viewpoint: Leveraging Sequence Information in
872 Enzymology. *Curr Opin Chem Biol* **2023**, *72*, 102246, doi:10.1016/j.cbpa.2022.102246.

873 49. Copp, J.N.; Anderson, D.W.; Akiva, E.; Babbitt, P.C.; Tokuriki, N. Exploring the Sequence,
874 Function, and Evolutionary Space of Protein Superfamilies Using Sequence Similarity
875 Networks and Phylogenetic Reconstructions. In *Methods in enzymology*; Elsevier Inc., 2019;
876 Vol. 620, pp. 315–347 ISBN 9780128168295.

877 50. Oberg, N.; Zallot, R.; Gerlt, J.A. EFI-EST, EFI-GNT, and EFI-CGFP: Enzyme Function Initiative
878 (EFI) Web Resource for Genomic Enzymology Tools. *J Mol Biol* **2023**, *435*, 168018,
879 doi:10.1016/j.jmb.2023.168018.

880 51. Mendler, K.; Chen, H.; Parks, D.H.; Lobb, B.; Hug, L.A.; Doxey, A.C. AnnoTree: Visualization
881 and Exploration of a Functionally Annotated Microbial Tree of Life. *Nucleic Acids Res* **2019**,
882 *47*, 4442–4448, doi:10.1093/nar/gkz246.

883 52. Klimchuk, O.I.; Konovalov, K.A.; Perekhvatov, V. V.; Skulachev, K. V.; Dibrova, D. V.;
884 Mulkidjanian, A.Y. COGNAT: A Web Server for Comparative Analysis of Genomic
885 Neighborhoods. *Biol Direct* **2017**, *12*, 26, doi:10.1186/s13062-017-0196-z.

886 53. Pedreira, T.; Elfmann, C.; Stülke, J. The Current State of Subti Wiki, the Database for the
887 Model Organism Bacillus Subtilis. *Nucleic Acids Res* **2022**, *50*, D875–D882,
888 doi:10.1093/nar/gkab943.

889 54. Kanehisa, M.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. KEGG as a Reference
890 Resource for Gene and Protein Annotation. *Nucleic Acids Res* **2016**, *44*, D457–D462,
891 doi:10.1093/nar/gkv1070.

892 55. Price, M.N.; Arkin, A.P. A Fast Comparative Genome Browser for Diverse Bacteria and
893 Archaea. *bioRxiv* **2023**, 1–17.

894 56. Gilchrist, C.L.M.; Chooi, Y.-H. Clinker & Clustermap.Js: Automatic Generation of Gene
895 Cluster Comparison Figures. *Bioinformatics* **2021**, *37*, 2473–2475,
896 doi:10.1093/bioinformatics/btab007.

897 57. Persson, E.; Castresana-Aguirre, M.; Buzzao, D.; Guala, D.; Sonnhammer, E.L.L. FunCoup 5:
898 Functional Association Networks in All Domains of Life, Supporting Directed Links and Tissue-
899 Specificity. *J Mol Biol* **2021**, *433*, 166835, doi:10.1016/j.jmb.2021.166835.

900 58. Sonnhammer, E.L.L.; Östlund, G. InParanoid 8: Orthology Analysis between 273 Proteomes,
901 Mostly Eukaryotic. *Nucleic Acids Res* **2015**, *43*, D234–D239, doi:10.1093/nar/gku1203.

902 59. Jalili, V.; Afgan, E.; Gu, Q.; Clements, D.; Blankenberg, D.; Goecks, J.; Taylor, J.; Nekrutenko, A.
903 The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses:
904 2020 Update. *Nucleic Acids Res* **2020**, *48*, W395–W402, doi:10.1093/nar/gkaa434.

905 60. Kasif, S.; Roberts, R.J. We Need to Keep a Reproducible Trace of Facts, Predictions, and
906 Hypotheses from Gene to Function in the Era of Big Data. *PLoS Biol* **2020**, *18*, e3000999,
907 doi:10.1371/journal.pbio.3000999.

908 61. Hernández-Plaza, A.; Szklarczyk, D.; Botas, J.; Cantalapiedra, C.P.; Giner-Lamia, J.; Mende,
909 D.R.; Kirsch, R.; Rattei, T.; Letunic, I.; Jensen, L.J.; et al. EggNOG 6.0: Enabling Comparative
910 Genomics across 12 535 Organisms. *Nucleic Acids Res* **2023**, *51*, D389–D394,
911 doi:10.1093/nar/gkac1022.

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

915 63. Attwood, T.K.; Agit, B.; Ellis, L.B.M. Longevity of Biological Databases. *EMBnet J* **2015**, *21*, 1–8,
916 doi:10.14806/ej.21.0.803.

917 64. Kern, F.; Fehlmann, T.; Keller, A. On the Lifetime of Bioinformatics Web Services. *Nucleic Acids*
918 *Res* **2020**, *48*, 12523–12533, doi:10.1093/nar/gkaa1125.

919 65. Ison, J.; Rapacki, K.; Ménager, H.; Kalaš, M.; Rydza, E.; Chmura, P.; Anthon, C.; Beard, N.;
920 Berka, K.; Bolser, D.; et al. Tools and Data Services Registry: A Community Effort to Document
921 Bioinformatics Resources. *Nucleic Acids Res* **2016**, *44*, D38–D47, doi:10.1093/nar/gkv1116.

922 66. Ma, L.; Zou, D.; Liu, L.; Shireen, H.; Abbasi, A.A.; Bateman, A.; Xiao, J.; Zhao, W.; Bao, Y.;
923 Zhang, Z. Database Commons: A Catalog of Worldwide Biological Databases. *Genomics*
924 *Proteomics Bioinformatics* **2022**, doi:10.1016/j.gpb.2022.12.004.

925 67. Rigden, D.J.; Fernández, X.M. The 2022 Nucleic Acids Research Database Issue and the Online
926 Molecular Biology Database Collection. *Nucleic Acids Res* **2022**, *50*, D1–D10,
927 doi:10.1093/nar/gkab1195.

928 68. Mulder, N.; Schwartz, R.; Brazas, M.D.; Brooksbank, C.; Gaeta, B.; Morgan, S.L.; Pauley, M.A.;
929 Rosenwald, A.; Rustici, G.; Sierk, M.; et al. The Development and Application of Bioinformatics
930 Core Competencies to Improve Bioinformatics Training and Education. *PLoS Comput Biol*
931 **2018**, *14*, 1–14, doi:10.1371/journal.pcbi.1005772.

932 69. Sansone, S.-A.; McQuilton, P.; Rocca-Serra, P.; Gonzalez-Beltran, A.; Izzo, M.; Lister, A.L.;
933 Thurston, M. FAIRsharing as a Community Approach to Standards, Repositories and Policies.
934 *Nat Biotechnol* **2019**, *37*, 358–367, doi:10.1038/s41587-019-0080-8.

935 70. Mathers, B.J.; L'Hours, H. Increasing the Reuse of Data through FAIR-Enabling the
936 Certification of Trustworthy Digital Repositories. *Int J Digit Curation* **1970**, *17*, 5,
937 doi:10.2218/ijdc.v17i1.852.

938 71. Zhao, M.; Yan, E.; Li, K. Data Set Mentions and Citations: A Content Analysis of Full-Text
939 Publications. *J Assoc Inf Sci Technol* **2018**, *69*, 32–46, doi:10.1002/asi.23919.

940 72. Silvello, G. Theory and Practice of Data Citation. *J Assoc Inf Sci Technol* **2017**,
941 doi:10.1002/asi.23917.

942 73. Kafkas, Ş.; Kim, J.-H.; McEntyre, J.R. Database Citation in Full Text Biomedical Articles. *PLoS*
943 *One* **2013**, *8*, e63184, doi:10.1371/journal.pone.0063184.

944 74. Piwowar, H.A.; Vision, T.J. Data Reuse and the Open Data Citation Advantage. *PeerJ* **2013**, *1*,
945 e175, doi:10.7717/peerj.175.

946