



HAL
open science

FreshOmics: A manually curated and standardized –omics database for investigating freshwater microbiomes

Corinne Biderre-Petit, Jean-christophe Charvy, Gisèle Bronner, Marina Chauvet, Didier Debroas, Hélène Gardon, Claire Hennequin, Isabelle Jouan-Dufournel, Anne Moné, Arthur Monjot, et al.

► **To cite this version:**

Corinne Biderre-Petit, Jean-christophe Charvy, Gisèle Bronner, Marina Chauvet, Didier Debroas, et al.. FreshOmics: A manually curated and standardized –omics database for investigating freshwater microbiomes. *Molecular Ecology Resources*, 2022, 10.1111/1755-0998.13692 . hal-03768857

HAL Id: hal-03768857

<https://hal.science/hal-03768857>

Submitted on 5 Sep 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

MOLECULAR ECOLOGY RESOURCES

FreshOmics: a manually curated and standardized -omics database for investigating freshwater microbiomes

Journal:	<i>Molecular Ecology Resources</i>
Manuscript ID	Draft
Manuscript Type:	Resource Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Bidierre-Petit, Corinne; Université Clermont Auvergne, LMGE UMR CNRS 6023 Charvy, Jean-Christophe; Université Clermont Auvergne, LMGE UMR CNRS 6023 Bronner, Gisele; Université Clermont Auvergne Chauvet, Marina; Université Clermont Auvergne, LMGE UMR CNRS 6023 Debroas, Didier; Université Clermont Auvergne Gardon, Helene; Université Clermont Auvergne, LMGE UMR CNRS 6023 Hennequin, Claire; Université Clermont Auvergne, LMGE UMR CNRS 6023 Jouan-Dufournel, Isabelle; Université Clermont Auvergne, CNRS, Laboratoire Microorganismes: Génome et Environnement Moné, Anne; Université Clermont Auvergne, LMGE UMR CNRS 6023 Monjot, Arthur; Université Clermont Auvergne Ravet, Viviane; Université Clermont Auvergne Vellet, Agnes; Université Clermont Auvergne Lepere, Cecile; Université Clermont Auvergne,
Keywords:	Freshwater, Microorganisms, -omics data, Database

1 **FreshOmics: a manually curated and standardized –omics database for investigating freshwater**
2 **microbiomes**

3

4 Running title: Microbial freshwater –omics database

5

6 Authors

7 Corinne Biderre-Petit^{1†*}, Jean-Christophe Charvy^{1†}, Gisèle Bronner¹, Marina Chauvet¹, Didier
8 Debroas¹, Hélène Gardon¹, Claire Hennequin¹, Isabelle Jouan-Dufournel¹, Anne Moné¹, Arthur
9 Monjot¹, Viviane Ravet¹, Agnès Vellet¹, Cécile Lepère¹

10

11 ¹CNRS, Laboratoire Microorganismes: Génome et Environnement, Université Clermont Auvergne,
12 Clermont-Ferrand, F-63000, France.

13 *For correspondence. E-mail corinne.petit@uca.fr; Tel+33 (0)473405139; Fax+33 (0)473407670.

14 †These authors contributed equally to this work.

15

16 **Abstract**

17 Freshwater is a critical resource for human survival but severely threatened by anthropic activities and
18 climate change. These changes also strongly impact the size and diversity of the microbial communities
19 hosted by freshwater ecosystems while they are key determinants of their functioning thanks to their
20 activities and stability. Although widely documented since the emergence of high-throughput
21 sequencing approaches, the information on these natural microbial communities is scattered among

22 thousands of papers and it is therefore difficult to investigate the temporal dynamics and the spatial
23 distribution of microbial taxa within or across ecosystems. To fill this gap, we built a manually curated
24 and standardized microbial freshwater –omics database (FreshOmics). Based on recognized ontologies
25 (ENVO, MIMICS, GO, ISO), FreshOmics describes 29 different types of freshwater ecosystems and uses
26 standardized attributes to depict biological samples, sequencing protocols and article attributes for
27 more than 2,487 geographical locations over 71 countries around the world. The database contains
28 24,808 run identifiers (mainly SRA, GSA and MG-RAST repositories) covering all sequence-based -omics
29 approaches used to investigate bacteria, archaea, microbial eukaryotes, and viruses. Therefore,
30 FreshOmics allows accurate and comprehensive analyses of microbial communities to answer
31 questions related to the microbial-driven roles in freshwater ecosystems functioning and resilience,
32 especially through meta-analysis studies. This collection also highlights strong discrepancies in
33 published works in the worldwide distribution of ecosystems studied, their types, and in the targeted
34 microorganisms.

35

36 **Keywords:** Freshwater, microorganisms, -omics data, database

37

38 **Introduction**

39 Covering less than 1% of Earth's surface, freshwater habitats are nevertheless critical for terrestrial life
40 (Dudgeon, 2019). Over 140,000 described species rely on freshwater habitats for their survival (IUCN,
41 <https://www.iucn.org/fr>). However, these ecosystems are severely threatened by the synergistic
42 effects of anthropogenic pressures and global change (dams, exploitation, pollution etc...) (Allan et al.,
43 2005; Jane et al., 2021). As a result, aquatic species face disproportionately higher extinction risks than
44 terrestrial or marine species (Webb & Mindel, 2015). Freshwater is therefore recognized as
45 conservation priorities (Tickner et al., 2020).

46 Amongst organisms hosted by freshwater habitats, microorganisms present an incredible diversity and
47 are crucial for the functioning and resilience of these ecosystems, underpinning all biogeochemical
48 cycles. In the last two decades, by coupling environmental DNA high-throughput sequencing with
49 bioinformatic methods, scientists have gained a better insight into the diversity and distribution of the
50 microbial populations in these ecosystems (Biessy et al., 2022; Debroas et al., 2017; Keck et al., 2020;
51 Ortiz-Álvarez et al., 2020; Pearmann, 2020; The Earth Microbiome Project Consortium et al., 2017).
52 Current microbiome analyses use multiple and complementary meta-omics approaches (e.g.
53 metabarcoding, metagenomics, metatranscriptomics), which generate an extensive volume of
54 sequencing data mostly stored in databases with public access. The most important is the Sequence
55 Read Archive (SRA), maintained by The National Center for Biotechnology Information (NCBI)
56 (Brooksbank et al., 2014; O'Leary et al., 2016). Thanks to these public repositories and the crescent
57 availability of -omics data, scientists can develop new working hypotheses and provide new insights in
58 many fields, especially through the combination of data from several studies (*i.e.* meta-analysis). This
59 powerful approach to produce novel discoveries is however, impeded by the laborious task of
60 extracting data of interest from different databases. Indeed, although public databases generally share
61 standardized data format, the effective interoperability of the data stored is reduced because of the
62 misuse (or absence of usage) of standardized vocabulary as proposed in biological ontologies (Jones et
63 al., 2015). Moreover, the stored data often present gaps, errors and contamination (Steinegger &
64 Salzberg, 2020). These problems have attracted much attention in the recent literature, leading to the
65 development of many tools to detect contaminants (Lupo et al., 2021; Tang, 2020) but mislabelled,
66 misleading, and missing data are hardly detectable.

67 To address these limitations, special-purpose databases have been multiplying these last few years,
68 covering a large variety of fields using domain-specific standardized metadata (see papers in the
69 Databases issues of Nucleic Acids Research (Rigden & Fernández, 2021, 2022), as well as papers in the
70 journal Databases, among other sources). In the case of environmental microbial studies, many

71 databases were also recently developed such as TerrestrialMetagenomeDB (TMDB) devoted to
72 terrestrial metagenomes (Corrêa et al., 2019), Planet Microbe devoted to oceanographic -omics data
73 (Ponsero et al., 2021), the MAR databases that compile data from marine microbial genomes
74 (Klemetsen et al., 2018), the Omics Database of Fermentative Microbes (ODFM) devoted to -omics
75 information for fermentative microorganisms (Whon et al., 2021) and MGnify developed to explore
76 microbiome data from a wide variety of biomes (Mitchell et al., 2019). Although some of these
77 databases contain freshwater data, they remain sparse and largely fragmented, which makes difficult
78 the exploration of novel biological hypotheses for this type of ecosystems especially through meta-
79 analysis studies.

80 The lack of a specialist database for freshwater -omics data can be explained by the reduced amount
81 of data produced so far, especially, in comparison with large oceanic expeditions. However large-scale
82 initiatives are multiplying with, for example, studies on microbial biogeography at regional (Pyrenean
83 lakes) and European scale with the joint analysis of hundreds of lakes (Boenigk et al., 2018; Ortiz-
84 Álvarez et al., 2018; Ortiz-Álvarez et al., 2020) as well as long-term time series (Linz et al., 2017;
85 Mondav et al., 2020). We therefore developed a manually curated database focused on freshwater
86 ecosystems (FreshOmics), including all types of sequence-based -omics data from bacteria, archaea,
87 microbial eukaryotes, and viruses by compiling, and organizing most available data and knowledge
88 from published literatures and associated repositories. It is meant to promote the exploratory
89 possibilities of these data in a user-friendly web interface, and to encourage integrative analysis of
90 available public data. Our resource combines the -omics data present in most of the current
91 repositories including SRA/ENA, GSA, MG-RAST, JGI, DRYAD. This database will not only facilitate the
92 work of microbial ecology researchers but will also be of interest to people from a wide range of
93 domains working on -omics based research.

94

95 **Methods**

96 **Database construction**

97 **Database objective.** We aimed to develop FreshOmics for comprehensively, accurately, and rapidly
98 analysing microbial community diversity (*i.e.* bacteria, archaea, microbial eukaryotes, and viruses) in
99 freshwater ecosystems around the world from all -omics data (metabarcodes, metagenomes,
100 metatranscriptomes and single cell genomes). Here, we present the overall workflow of this database
101 in terms of how we curated and annotated the data, and how we designed and implemented the
102 database (Figure 1). Briefly, the actual database (January 2022) is built from 663 peer-reviewed
103 publications (published between early 2006 and 2022, Supplementary Figure S1A) by retrieving
104 metadata relating to i) the publication (author, title, journal, DOI, PMID), ii) the studied site (name,
105 GPS, ENVO-type), iii) the experimental design (-omics type, material, primers) and iv) data resource
106 (repository name, accessions) (Figure 1A). The selected attributes for these ecosystems were then
107 organised in a standardized metadata sheet according to bio-ontological standards (Figure 1B). Finally,
108 we implemented a web-application allowing the metadata exploration and retrieval (Figure 1C, 1D),
109 among which -omics data identifiers allowing further analyses.

110 **Data retrieval and non-freshwater data removal.** The data were collected from an extensive literature
111 search using various combinations of keywords. Several publication databases were used such as
112 PubMed, ScienceDirect, Scopus, HAL and BibCNRS (Figure 1A). For the collection, we created a glossary
113 of terms related to the microbial domains, freshwater ecosystems, sample material, -omics
114 approaches, barcode gene markers, and sequencing platforms (Supplementary Table S1). Research
115 articles lacking at least one of the information on these terms were excluded. Similarly, when -omics
116 data were cited in multiple papers, only the most informative paper was retained for FreshOmics.
117 Unlike other databases, such as HumanMetagenomeDB (Kasmanas et al., 2021) and TMDB (Corrêa et
118 al., 2020), no semi-automated approach for direct retrieval of metadata from repositories was

119 developed. Indeed, the comparison of the data found in the papers to those present in the associated
120 repositories (especially the SRA, the main repository for microbial -omics data) revealed a high
121 proportion of inconsistent information, corresponding to erroneous data records, preventing from
122 having a high quality database. As the source of these errors cannot be automatically resolved, we
123 systematically reviewed the publication including full texts, tables, figures and supplementary material
124 and associated repositories to identify the reliable data of interest. All information were inspected
125 manually and corrected when possible, to provide consistent annotation. For publications containing
126 both freshwater and non-freshwater samples (*e.g.* estuary and coastal samples of thalassic origin),
127 only freshwater data were kept (Supplementary Figure S1B).

128 **Standardization of attributes.** In both publications and repositories, most attributes of interest are
129 written in different ways, and their values are non-standardized. To circumvent this issue, we created
130 a list of synonyms and screened attribute names and their respective values. During data inspection,
131 we identified attributes related to the ecosystem, name, country of origin, geographical coordinates,
132 type, sample material and several other experimental characteristics listed in Supplementary Tables
133 S1 and S2. Whenever possible, we used ontologies and custom-built controlled vocabularies. Most
134 attributes and ontologies used were selected based on the MixS version 5.0 (<https://gensc.org/mixs/>)
135 compliant metadata lists for a metagenomic sample (MIMS) and a marker gene survey (MIMARKS).
136 The water ecosystem and sample material attributes were standardized according to the Environment
137 Ontology (EnvO; <https://sites.google.com/site/environmentontology/>). In few cases, the information
138 could not be standardized (*e.g.* ecosystem type not defined in ENVO) and was left as stated by the
139 original submitter. Freshwater ecosystems were standardized into 29 main categories (*e.g.* lake, pond,
140 spring, aquifer, etc...) and the sample material, into 12 categories (*e.g.* water, sediment, microbial mat,
141 etc...). Barcode gene markers were described according to the Gene Ontology (GO) terms
142 (<http://geneontology.org/>) and hierarchized into three main categories, *i.e.* phylogenetic, functional,
143 and viral markers. Finally, information regarding experiment and run identifiers were collected from

144 the repository metadata. Geographical coordinates were standardized to both decimal and
145 sexagesimal degrees formats. Country names and journal title abbreviations were manually labelled
146 based on the ISO 3166 and ISO 4 standard, respectively. PubMed ID, Digital Object Identifier (DOI) ID,
147 or both when available, were manually recovered for all articles. The different categories and the
148 complete set of standardized attributes can be found in Supplementary Tables S2 and S3, respectively.

149 **Data quality control.** As both data standardization and integration are manual and individual tasks,
150 some inconsistencies may marginally remain. Thus, an automated curation procedure was developed
151 to check metadata completeness and integrity post integration in the FreshOmics database. It works
152 as a search engine, looking for specific empty fields (*e.g.* absence of ecosystem type, author in
153 publication, run identifier, etc...) or lack of information (*e.g.* barcoding method specified but no
154 primers indicated, etc...), producing a report of all entries that need to be verified. A second run of
155 manual curation of spurious entries is then done. If the correction cannot be made because the
156 attribute identified as faulty is missing from the original publication or repository, two cases are
157 possible: i) the attribute is mandatory and therefore, the affected dataset is removed from the
158 database; ii) the attribute is not mandatory (*e.g.* marker region, primer sequences, etc...) and
159 therefore, the affected dataset is kept in the database and in memory for subsequent checks. To avoid
160 conversion errors, the validation form embeds an automatic GPS coordinate conversion process to
161 generate sexagesimal coordinates from decimal ones. Moreover, for publications with a PubMed ID,
162 author names and rank are automatically collected from the publication description page in 'PubMed'
163 format (previously named MEDLINE) and integrated directly into the database.

164 **Web application implementation and exploration.**

165 **Implementation.** The FreshOmics web interface was implemented using Apache HTTP server (version
166 2.4.6) integrated with PHP (version 7.2.34) and PostgreSQL (version 9.2.24). The home and informative
167 pages were built with Bludit CMS (version 3.13.1), a flat-file-based CMS with JSON format to store the

168 data. Map browsing was developed using the Leaflet package (version 1.7.1) to handle the maps
169 provided by the OpenStreetMap Project (<https://www.openstreetmap.org/>), together with the JQuery
170 library (version 3.2.1), to implement the interactivity of all pages. The PostgreSQL database was used
171 to process the data of the back-end.

172 **Exploration.** The application was designed with a tab layout. The FreshOmics 'Home' page
173 (<http://freshomics.lmge.uca.fr/>) gives an overview of the resource. The 'Browse' section constitutes
174 the main search interface of FreshOmics data, either through attributes filtering or map navigation.
175 The 'Interactive map' locates all ecosystems for which an -omics study was published, based on their
176 latitude/longitude coordinates. The 'List' mode allows user to search in FreshOmics using various
177 descriptors or keywords through the 'Search by' box. Once the selection step finished, all the datasets
178 are displayed in a final cart list. For each dataset in this list, a short description of the sample is available
179 through the 'See more' option (Figure 1C). Further, the 'Statistics' section offers a graphical
180 representation of different metadata present in the database. Finally, the 'Help' section includes an
181 overview of the database organization and its practical uses, while the 'About' page provides details
182 about the database and the contributors.

183 **Metadata download.** Users can browse the data and view their search results, however, they will need
184 to create an account to download the datasets which are available in the cart. Data can be downloaded
185 as JSON file | YAML file | XML file | HTML file | TXT file | CSV file. When opening a CSV file, we
186 recommend using the UTF-8 encoding to ensure a better display of special characters. Thereafter, the
187 sequencing data of interest must be downloaded from the repositories where they are stored, either
188 by using the run identifiers (Run_ID) for the SRA and GSA repositories or the experiment identifiers
189 (Exp_ID) for the MG-RAST repository, which can be found in the downloaded metadata file. To
190 facilitate the download, we provided a script called 'freshcart_downloader.py' that takes as input the

191 downloaded JSON-file from our database. Also, a tutorial video on how to use the script is available in
192 the tutorial playlist.

193 **Private access.** A private access to the web server allows data management by curators. Possible tasks
194 include adding metadata sheets to the database, data integrity checking and correction (see data
195 quality control), author access management, or attribute standard updating.

196

197 **Results**

198 **The manual curation of microbial -omics freshwater-related data from the literature**

199 After keyword queries and selection of thousands of research articles, a quick manual sort allowed to
200 remove papers which did not reach the criteria of the database (*e.g.* non-high throughput sequencing
201 data, not freshwater body, aquatic animal microbiomes, microcosm, etc...) resulting in the selection of
202 792 papers containing non-redundant -omics data (January 2022). Although these articles were
203 published in a wide variety of journals (138 in total), most of them were found in few specific ones, *i.e.*
204 *Frontiers in Microbiology* (13%), followed by *FEMS Microbiology Ecology* (6%), *PLoS One* (6%), *ISME*
205 *Journal* (5%) and *Scientific Reports* (5%) (Figure S1A). A second and more in-depth curation resulted in
206 the supplemental removal of more than 16.2% of these articles, as some mandatory descriptors were
207 lacking and non-recoverable.

208 Among the 663 remaining articles, data for more than 20% of them had to be modified prior to their
209 incorporation into FreshOmics, either to remove non-freshwater (*e.g.* soil, rock, saline water) and
210 microcosm samples (1.4% of the articles) or to correct misannotated descriptors (>18.7%) (Figure S1B,
211 Table S4). Only the multiplexed -omics data could not be modified in input data, and thus induce a
212 slight redundancy in the database in terms of Exp_ID and Run_ID, several samples and, therefore
213 several datasets, being identified with the same Exp_ID and Run_ID. Although the rectifiable errors

214 were mostly related to GPS coordinates, all fields of the metadata were affected (*e.g.* annotation as
215 RNAseq instead of DNAseq or *vice versa*, amplicon/metagenome, isolate/community,
216 bacteria/eukaryote, but also errors in technology used, marker, marker region, sample material,
217 inversion of data between samples, etc...). In summary, >32% of the articles initially selected had data
218 with misannotations and misleadings, whether or not they were correctable (Figure S1B, Table S4).

219 Finally, the 663 articles kept covered data collected from 2,487 unique locations distributed around
220 the world. They correspond to 29 distinct ecosystem types where lakes are the main water bodies
221 represented (50.9% of all locations; among these lakes 5.6% are saline with athalassic origin), followed
222 by hot springs (9.4%), rivers (6.8%), ponds (6.5%), wastewaters (5.3%) and streams (5%) (Table S5).

223

224 **Database content overview**

225 **Dataset location and distribution.** From the 2,487 ecosystems identified in this study, a total of 6,912
226 datasets were generated (January 2022), as multiple datasets may be available for some of them.
227 Although the ecosystems span all seven continents, most datasets were obtained from Asia (37.7% of
228 datasets), North America (28.9%) and Europe (24.9%) (Figure 2A). A large disparity in distribution was
229 also observed between the different countries inside these continents, some countries being over-
230 represented compared to others (Figure 2B). Finally, annotated datasets span 71 countries with China
231 being the most represented country (29.6% of datasets), followed by the United States of America
232 (USA; 16.3%), Canada (8.1%) and Spain (5.5%). Consistent with the previous result for ecosystem
233 abundance, most of the datasets came from lakes with 42.9% (2,968 datasets), followed by rivers with
234 17.8% (1,232), streams with 8.9% (615) and hot springs with 6.2% (431) (Figure 3A). Moreover,
235 FreshOmics contains 12 distinct sample materials, the most documented being water (>64% of the
236 datasets) followed by sediment (>22%) (Figure 3B).

237 **Sequencing data overview.** Most of the datasets included in the FreshOmics database derived from
238 the SRA repository (94.7 % of datasets), followed by MG-RAST (2%) and GSA (1.6%). These datasets
239 cover 16 years of experiments, since the first freshwater metagenome submitted to MG-RAST in 2006
240 (Edwards et al., 2006). Most of the datasets contain metabarcoding data (86.4%), followed by
241 metagenome (12.5%, of which 13.8% specific to virus and called metavirome in the present study),
242 metatranscriptome (0.9%) and single cell genome (0.2%) (Figure 4A). Regarding the metabarcoding
243 approach, as expected, the most present markers in FreshOmics are 16S (72.5% of metabarcoding
244 datasets) and 18S rDNAs (17.7%) followed by the ITS region (5.5%) (Figure 4B). Furthermore, the most
245 represented functional markers in datasets are the genes encoding the large subunit of the ribulose-
246 1,5-bisphosphate carboxylase/oxygenase enzyme (*rbcL*; 1.1% of the datasets) and the ammonia
247 monooxygenase enzyme (*amoA*; 1%), respectively. Most of the samples were sequenced using Illumina
248 technology (79.6% of datasets) (Figure 5A), mostly using the Miseq system (79.5% of Illumina
249 sequencing). However, the analysis of published data shows an evolution in the different sequencing
250 technologies, with a decrease of the use of Roche 454 technology starting in 2016 and the beginning
251 of the use of third-generation sequencing technologies in 2019 (Figure 5B). Considering the SRA and
252 GSA repositories, all datasets cover 24,518 run numbers (99.2% are stored in the former) distributed
253 as follows: 85.8% for metabarcoding, 8.1% for metagenomics, 2.7% for metatranscriptomics, 1.7% for
254 single cell genomics and 1.6% for metaviromics. For MG-RAST, the 136 datasets correspond in reality
255 to 290 Exp_ID of which 51.4% for gene amplicon data and 48.6% for metagenomes (of which 9.2% are
256 metaviromes).

257 **Usage and functionalities.** The FreshOmics user interface is divided in two main search modes, *i.e.* a
258 'List' mode that allows to select -omics data of interest through a set of descriptors that can be
259 combined, and an 'Interactive map' mode that provides a more intuitive way of selecting data directly
260 from the world map. Both give access to the full content of the current version of the database as only
261 metadata with valid GPS coordinates were included in FreshOmics database.

262 *List mode.* Each dataset in the list can be selected by checking them individually. Moreover, all datasets
263 can also be selected or de-selected in one click using the buttons on the top 'Check all/check none'.
264 The attribute 'See more' on the bottom left of each dataset provides a short overview of the attributes
265 associated with the dataset. For filtering, a set of 12 descriptors is provided in a search box. By pushing
266 each descriptor, the dashboard is expanded to show all available filters. Further, the 'Keyword' field
267 allows user to perform a simple search for any attribute not proposed in the search box descriptors.
268 Complex queries are therefore possible using one or more conditional descriptors, combined or not to
269 a specific keyword. A panel is fixed on the top displaying the current number of filtered -omics datasets,
270 so user can keep track of how each filtering step shapes the data. Once the selection step is finished,
271 all marked datasets can be added to the cart by pushing the button 'Add selected datasets to cart'. If
272 no filter is applied, the whole dataset (the full FreshOmics data) can be added to the cart using the
273 button 'Add list datasets to cart'. Pushing the panel 'View cart' allows user to access the selection. At
274 this step, the selection can still be deleted by using the function 'Clear your cart' or its content modified
275 by eliminating unwanted datasets individually. Once finalized, the selection can be downloaded under
276 six different formats.

277 *Interactive map.* The 'Interactive map' tab allows to interactively explore the world map and select -
278 omics data from all around the world. By zooming in and out on the map, the user can select different
279 parts of the globe and collect the corresponding datasets using the button 'Add map datasets to cart'.
280 This selection can be performed multiple times in various regions of the map, adding datasets to the
281 cart for each search. The button 'Reset' allows user to return to the complete map. Individual points
282 in the map may indicate several samples collected in the same coordinates or one sample analyzed
283 through multiple approaches. For each point, a short label shows the ecosystem type, the method
284 analysis but also the marker if appropriate. For a search on specific criteria, the descriptors of interest
285 must be selected in the search box as described previously. Once the selection step is done, all marked
286 datasets can be added to the cart by pushing the button 'Add map datasets to cart' and user can

287 download its selection as described for the 'List' mode. To provide practical examples, we made two
288 video tutorials about the usage of the web application. A link to the tutorials can be found in the item
289 3 of the 'Help' section in the web application.

290 *Downloading -omics data of interest.* FreshOmics is not a repository of sequencing data. Once datasets
291 selected, the corresponding sequencing data must be downloaded from their original repository. To
292 facilitate the download, we provided a script called 'freshcart_downloader.py' available in a tutorial
293 and in the item 5 of the 'Help' section in the web application.

294

295 **Discussion**

296 Freshwater ecosystems are a vital natural resource (drinking water, animal habitats, etc...), also
297 needed for many uses (agriculture, energy production, recreation, manufacturing, etc...). These uses
298 put a huge pressure on this resource, stresses that are likely to be exacerbated by climate change
299 (Alcamo et al., 2008; Kibona et al., 2009). Microbial communities within these freshwater ecosystems
300 support critical function due to their high abundance, diversity and stability but are also subjected to
301 the wide variety of anthropogenic disturbances (Beattie et al., 2020; Santillan et al., 2019). Microbial
302 community composition and function both in pristine and disturbed natural environments are far from
303 being well known, however more and more studies are undertaken facilitated by the technological
304 advances in microbiology (McDaniel et al., 2021; Wani et al., 2021). Nonetheless, the information on
305 these natural communities is scattered among thousands of papers making it difficult to collect,
306 compare, and integrate.

307 FreshOmics is a unique resource in its purpose and content. It centralizes and standardizes metadata
308 for freshwater -omics data present in publications and the most used repositories. Our effort to
309 establish FreshOmics was prompted by the need of researchers to access a comprehensive and reliable
310 dataset, ever growing since the emergence of high-throughput sequencing approaches. Indeed, with

311 the continuous improvement of these technologies, accuracy, sequencing depth and price, the volume
312 of data increased tremendously. Prior FreshOmics, only a few databases provided information related
313 to sequencing data for microorganisms in freshwaters (*e.g.* TMDB, MGnify), but with limited features
314 and often restricted to a specific -omics method (*e.g.* metagenome) or repository (*e.g.* SRA, MG-RAST).
315 Moreover, their utilization to obtain valuable information regarding verified data was also not
316 satisfactory because these data were retrieved from repositories containing many errors. We believe
317 that our database fills therefore a long-standing gap in the freshwater and microbial research fields
318 and will hopefully help answering important questions related to the diverse roles of microorganisms
319 in ecosystem functioning and resilience.

320 The -omics era has revolutionized microbial ecology and has led to the emergence of new knowledge
321 for a better understanding of microbial diversity, interrelationships and function within an ecosystem.
322 With FreshOmics, we highlight great discrepancies in the distribution of the data published in the
323 literature for the freshwater ecosystems at the geographical, ecosystem type, and microbial domain
324 scale. The largest amount of -omics data was so far produced by Asia, North America, and Europe
325 (Figure 2). This gap between continents has probably multiple causes however it is likely due to the
326 income inequality and technological inability to generate and analyze big data (*e.g.* data storage,
327 processing). A strong imbalance is also observed at the country scale as, for example, for Asia where
328 China encompasses 78% of all datasets from this part of the world and Russia only 6.6%, while almost
329 twice as large in area and with the second largest amount of freshwater in the world after Brazil (Brazil:
330 8,233 Km³; Russia: 4,508 Km³; China: 2,840 Km³, [https://www.worldatlas.com/articles/countries-with-](https://www.worldatlas.com/articles/countries-with-the-most-freshwater-resources.html)
331 [the-most-freshwater-resources.html](https://www.worldatlas.com/articles/countries-with-the-most-freshwater-resources.html)). This can be explained by the lack of previous concerns and
332 investments of Russia for the biological sciences, largely considerate as top-down approaches
333 (Gronvall & Bland, 2020). A strong disparity is also observed between the types of ecosystems studied,
334 with a large dominance of lakes (43% of all datasets) and overall, all lotic and lentic ecosystems (75.6%
335 of all datasets), whereas they represent only 12.5% of the global volume of fresh water (75% are locked

336 up in ice) on earth. Indeed, these ecosystems are of major concerns for humans as the main source of
337 drinking water but also for their recreative and economic uses. However, we can also notice a country-
338 dependent distribution bias for at least the four main represented ecosystems in datasets (*i.e.* hot
339 spring, stream, river, and lake; Figure 6), which may reflect differences in scientific policies and
340 priorities implemented at the country or continental level. The proportion of datasets dedicated to
341 these four ecosystems is also variable across continents, ranging from 59.7 to 84.4% for the North
342 America and Asia, suggesting more diverse freshwater ecosystem models in North America (Figure 6).

343 The proportion of the different microbial domains also varies in the datasets, with an over-
344 representation of bacteria compared to the other domains, especially microbial eukaryotes (for
345 example, 73.1% and 21.4% of the metabarcoding datasets, respectively). This is because the vast
346 majority of microbial biodiversity -omics studies have overwhelmingly focused on bacterial
347 communities. The best example is the launching of the 'Genomic Encyclopedia of Bacteria and Archaea'
348 (Wu et al., 2009) which encourages the use of phylogenomics to enhance our knowledge on these
349 microorganisms, without any equivalent for microbial eukaryotes. Indeed, even though the microbial
350 eukaryote diversity is more studied now especially with the use of metabarcoding, their study through
351 other approaches such as metagenomics and single cell genomics is less developed especially due the
352 complexity of their genomes (*e.g.* large intergenic regions, introns, repetitive DNA) making them very
353 difficult to analyze. However, this imbalance will hopefully tend to diminish with the high number of
354 ongoing projects targeting this group (*e.g.* the PELAGICS project covering 70 European lakes and
355 targeting both bacteria and eukaryotes) and the advance in the analysis of eukaryotic genomics data
356 (Bailet et al., 2019; Cahoon et al., 2018; Delmont et al., 2020; Ingala et al., 2021; Matsuoka et al., 2019;
357 Meredith et al., 2021; Valentin et al., 2019).

358

359 **Conclusion and future work**

360 Having overcome the technical difficulties imposed by the diversity in quality, methodology and
361 presentation of published data, we created a reliable manually curated resource. FreshOmics is
362 intended to benefit the scientific community by providing (i) a freely accessible, easy-to-use, efficiently
363 organized resource with relevant -omics data from all types of freshwater and microorganisms, (ii) a
364 tool allowing to facilitate some routine and time-consuming computer tasks for people working on
365 freshwater microorganisms (*e.g.* data selection, recovering, curation) and (iii) a user-friendly and
366 beneficial application to a broader community of researchers by enabling the investigation of the
367 temporal dynamics and the spatial distribution of microbial taxa within or across ecosystems. These
368 characteristics may allow to test hypotheses about how microbial communities are structured and how
369 they respond to environmental change. In addition, FreshOmics provides a tool to quickly and easily
370 visualize the distribution of -omics data around the world and highlights existing gaps in the published
371 studies of freshwater microbiomes that future work could fill.

372 FreshOmics has been thoroughly tested and will be periodically updated by the core curator team.
373 Researchers in the field are also invited to contribute by providing references of their studies such as
374 the DOI and the archive name (or numbers) of the datasets they produce through the mail address
375 freshomics.lmge_AT_listes.uca.fr or by filling the form found in the 'Contribute' section of the
376 website. Next releases will mainly focus on enriching the database with the data from publications that
377 are currently pending and from those that will be published in the future and on integrating other –
378 omics data types such as proteomics, metabolomics when available.

379 **Funding**

380 A M, H G and M C are supported by PhD fellowships of the French National Fund for Scientific Research
381 (Ministère de l'Enseignement Supérieur de la Recherche et de l'Innovation; MESRI).

382

383 **Acknowledgements**

384 Computational resources have been provided by the Direction Opérationnelle des Systèmes
385 d'Information (DOSI) of the University Clermont-Auvergne (UCA). The authors warmly thank Alexis

386 OZWALD for the provision of the development server and his help in its maintenance. The authors are
387 also grateful to Jean-Louis RENAUD for the supply of the production hosting needed for the web server
388 and the database and to Eric TOURAILLE and Alexis OZWALD for the firewall configuration.

389

390 References

- 391 Alcamo, J. M., Vörösmarty, C. J., Naiman, R. J., Lettenmaier, D. P., & Pahl-Wostl, C. (2008). A grand
392 challenge for freshwater research: understanding the global water system. *Environmental*
393 *Research Letters*, 3(1), 010202. <https://doi.org/10.1088/1748-9326/3/1/010202>
- 394 Allan, J. D., Abell, R., Hogan, Z., Revenga, C., Taylor, B. W., Welcomme, R. L., & Winemiller, K. (2005).
395 Overfishing of Inland Waters. *BioScience*, 55(12), 1041. [https://doi.org/10.1641/0006-](https://doi.org/10.1641/0006-3568(2005)055[1041:OOIW]2.0.CO;2)
396 [3568\(2005\)055\[1041:OOIW\]2.0.CO;2](https://doi.org/10.1641/0006-3568(2005)055[1041:OOIW]2.0.CO;2)
- 397 Bailet, B., Bouchez, A., Franc, A., Frigerio, J.-M., Keck, F., Karjalainen, S.-M., & Kahlert, M. (2019).
398 Molecular versus morphological data for benthic diatoms biomonitoring in Northern Europe
399 freshwater and consequences for ecological status. *Metabarcoding and Metagenomics*, 3,
400 e34002. <https://doi.org/10.3897/mbmg.3.34002>
- 401 Beattie, R. E., Bandla, A., Swarup, S., & Hristova, K. R. (2020). Freshwater Sediment Microbial
402 Communities Are Not Resilient to Disturbance From Agricultural Land Runoff. *Frontiers in*
403 *Microbiology*, 11, 539921. <https://doi.org/10.3389/fmicb.2020.539921>
- 404 Boenigk, J., Wodniok, S., Bock, C., Beisser, D., Hempel, C., Grossmann, L., & Jensen, M. (2018).
405 Geographic distance and mountain ranges structure freshwater protist communities on a
406 European scale. *Metabarcoding and Metagenomics*, 2, e21519.
407 <https://doi.org/10.3897/mbmg.2.21519>
- 408 Brooksbank, C., Bergman, M. T., Apweiler, R., Birney, E., & Thornton, J. (2014). The European
409 Bioinformatics Institute's data resources 2014. *Nucleic Acids Research*, 42(D1), D18–D25.
410 <https://doi.org/10.1093/nar/gkt1206>
- 411 Cahoon, A. B., Huffman, A. G., Krager, M. M., & Crowell, R. M. (2018). A meta-barcoding census of
412 freshwater planktonic protists in Appalachia – Natural Tunnel State Park, Virginia, USA.
413 *Metabarcoding and Metagenomics*, 2, e26939. <https://doi.org/10.3897/mbmg.2.26939>
- 414 Corrêa, F. B., Saraiva, J. P., Stadler, P. F., & da Rocha, U. N. (2019). TerrestrialMetagenomeDB: a
415 public repository of curated and standardized metadata for terrestrial metagenomes. *Nucleic*
416 *Acids Research*, gkz994. <https://doi.org/10.1093/nar/gkz994>
- 417 Debroyas, D., Domaizon, I., Humbert, J.-F., Jardillier, L., Lepère, C., Oudart, A., & Taïb, N. (2017).
418 Overview of freshwater microbial eukaryotes diversity: a first analysis of publicly available
419 metabarcoding data. *FEMS Microbiology Ecology*, 93(4).
420 <https://doi.org/10.1093/femsec/fix023>
- 421 Delmont, T. O., Gaia, M., Hingsinger, D. D., Fremont, P., Vanni, C., Guerra, A. F., & Jaillon, O. (2020).
422 *Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed*
423 *by genome-resolved metagenomics* [Preprint]. *Microbiology*.
424 <https://doi.org/10.1101/2020.10.15.341214>
- 425 Dudgeon, D. (2019). Multiple threats imperil freshwater biodiversity in the Anthropocene. *Current*
426 *Biology*, 29(19), R960–R967. <https://doi.org/10.1016/j.cub.2019.08.002>
- 427 Edwards, R. A., Rodriguez-Brito, B., Wegley, L., Haynes, M., Breitbart, M., Peterson, D. M., & Rohwer,
428 F. (2006). Using pyrosequencing to shed light on deep mine microbial ecology. *BMC*
429 *Genomics*, 7(1), 57. <https://doi.org/10.1186/1471-2164-7-57>
- 430 Gronvall, G. K., & Bland, B. (2020). Life-science research and biosecurity concerns in the Russian
431 Federation. *The Nonproliferation Review*, 27(4–6), 415–423.
432 <https://doi.org/10.1080/10736700.2020.1866323>

- 433 Ingala, M. R., Werner, I. E., Fitzgerald, A. M., & Naro-Maciel, E. (2021). 18S rRNA amplicon sequence
434 data (V1–V3) of the Bronx river estuary, New York. *Metabarcoding and Metagenomics*, *5*,
435 e69691. <https://doi.org/10.3897/mbmg.5.69691>
- 436 Jane, S. F., Hansen, G. J. A., Kraemer, B. M., Leavitt, P. R., Mincer, J. L., North, R. L., & Rose, K. C.
437 (2021). Widespread deoxygenation of temperate lakes. *Nature*, *594*(7861), 66–70.
438 <https://doi.org/10.1038/s41586-021-03550-y>
- 439 Kasmanas, J. C., Bartholomäus, A., Corrêa, F. B., Tal, T., Jehmlich, N., Herberth, G., & da Rocha, N.U.
440 (2021). HumanMetagenomeDB: a public repository of curated and standardized metadata
441 for human metagenomes. *Nucleic Acids Research*, *49*(D1), D743–D750.
442 <https://doi.org/10.1093/nar/gkaa1031>
- 443 Keck, F., Millet, L., Debroas, D., Etienne, D., Galop, D., Rius, D., & Domaizon, I. (2020). Assessing the
444 response of micro-eukaryotic diversity to the Great Acceleration using lake sedimentary
445 DNA. *Nature Communications*, *11*(1), 3831. <https://doi.org/10.1038/s41467-020-17682-8>
- 446 Kibona, D., Kidulile, G., & Rwabukambara, F. (2009). Environment, Climate Warming and Water
447 Management. *Transition Studies Review*, *16*(2), 484–500. <https://doi.org/10.1007/s11300-009-0084-z>
- 449 Klemetsen, T., Raknes, I. A., Fu, J., Agafonov, A., Balasundaram, S. V., Tartari, G., & Willassen, N. P.
450 (2018). The MAR databases: development and implementation of databases specific for
451 marine metagenomics. *Nucleic Acids Research*, *46*(D1), D692–D699.
452 <https://doi.org/10.1093/nar/gkx1036>
- 453 Linz, A. M., Crary, B. C., Shade, A., Owens, S., Gilbert, J. A., Knight, R., & McMahon, K. D. (2017).
454 Bacterial Community Composition and Dynamics Spanning Five Years in Freshwater Bog
455 Lakes. *MSphere*, *2*(3), e00169-17. <https://doi.org/10.1128/mSphere.00169-17>
- 456 Lupo, V., Van Vlierberghe, M., Vanderschuren, H., Kerff, F., Baurain, D., & Cornet, L. (2021).
457 Contamination in Reference Sequence Databases: Time for Divide-and-Rule Tactics. *Frontiers*
458 *in Microbiology*, *12*, 755101. <https://doi.org/10.3389/fmicb.2021.755101>
- 459 Matsuoka, S., Sugiyama, Y., Sato, H., Katano, I., Harada, K., & Doi, H. (2019). Spatial structure of
460 fungal DNA assemblages revealed with eDNA metabarcoding in a forest river network in
461 western Japan. *Metabarcoding and Metagenomics*, *3*, e36335.
462 <https://doi.org/10.3897/mbmg.3.36335>
- 463 McDaniel, E. A., Wahl, S. A., Ishii, S., Pinto, A., Ziels, R., Nielsen, P. H., & Williams, R. B. H. (2021).
464 Prospects for multi-omics in the microbial ecology of water engineering. *Water Research*,
465 *205*, 117608. <https://doi.org/10.1016/j.watres.2021.117608>
- 466 Meredith, C., Hoffman, J., Trebitz, A., Pilgrim, E., Okum, S., Martinson, J., & Cameron, E. S. (2021).
467 Evaluating the performance of DNA metabarcoding for assessment of zooplankton
468 communities in Western Lake Superior using multiple markers. *Metabarcoding and*
469 *Metagenomics*, *5*, e64735. <https://doi.org/10.3897/mbmg.5.64735>
- 470 Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., & Finn, R. D. (2019).
471 MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research*, gkz1035.
472 <https://doi.org/10.1093/nar/gkz1035>
- 473 Mondav, R., Bertilsson, S., Buck, M., Langenheder, S., Lindström, E. S., & Garcia, S. L. (2020).
474 Streamlined and Abundant Bacterioplankton Thrive in Functional Cohorts. *MSystems*, *5*(5),
475 e00316-20. <https://doi.org/10.1128/mSystems.00316-20>
- 476 O’Leary, N. A., Wright, M. W., Brister, J. R., Ciufu, S., Haddad, D., McVeigh, R., & Pruitt, K. D. (2016).
477 Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and
478 functional annotation. *Nucleic Acids Research*, *44*(D1), D733-745.
479 <https://doi.org/10.1093/nar/gkv1189>
- 480 Ortiz-Álvarez, R., Cáliz, J., Camarero, L., & Casamayor, E. O. (2020). Regional community assembly
481 drivers and microbial environmental sources shaping bacterioplankton in an alpine lacustrine

- 482 district (Pyrenees, Spain). *Environmental Microbiology*, 22(1), 297–309.
483 <https://doi.org/10.1111/1462-2920.14848>
- 484 Ortiz-Álvarez, R., Fierer, N., de los Ríos, A., Casamayor, E. O., & Barberán, A. (2018). Consistent
485 changes in the taxonomic structure and functional attributes of bacterial communities during
486 primary succession. *The ISME Journal*, 12(7), 1658–1667. [https://doi.org/10.1038/s41396-](https://doi.org/10.1038/s41396-018-0076-2)
487 [018-0076-2](https://doi.org/10.1038/s41396-018-0076-2)
- 488 Ponsoero, A. J., Bomhoff, M., Blumberg, K., Youens-Clark, K., Herz, N. M., Wood-Charlson, E. M., &
489 Hurwitz, B. L. (2021). Planet Microbe: a platform for marine microbiology to discover and
490 analyze interconnected ‘omics and environmental data. *Nucleic Acids Research*, 49(D1),
491 D792–D802. <https://doi.org/10.1093/nar/gkaa637>
- 492 Rigden, D. J., & Fernández, X. M. (2021). The 2021 *Nucleic Acids Research* database issue and the
493 online molecular biology database collection. *Nucleic Acids Research*, 49(D1), D1–D9.
494 <https://doi.org/10.1093/nar/gkaa1216>
- 495 Rigden, D. J., & Fernández, X. M. (2022). The 2022 *Nucleic Acids Research* database issue and the
496 online molecular biology database collection. *Nucleic Acids Research*, 50(D1), D1–D10.
497 <https://doi.org/10.1093/nar/gkab1195>
- 498 Santillan, E., Seshan, H., Constancias, F., Drautz-Moses, D. I., & Wuertz, S. (2019). Frequency of
499 disturbance alters diversity, function, and underlying assembly mechanisms of complex
500 bacterial communities. *Npj Biofilms and Microbiomes*, 5(1), 8.
501 <https://doi.org/10.1038/s41522-019-0079-4>
- 502 Steinegger, M., & Salzberg, S. L. (2020). Terminating contamination: large-scale search identifies
503 more than 2,000,000 contaminated entries in GenBank. *Genome Biology*, 21(1), 115.
504 <https://doi.org/10.1186/s13059-020-02023-1>
- 505 Tang, L. (2020). Contamination in sequence databases. *Nature Methods*, 17(7), 654–654.
506 <https://doi.org/10.1038/s41592-020-0895-8>
- 507 The Earth Microbiome Project Consortium, Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A.,
508 Ladau, J., & Knight, R. (2017). A communal catalogue reveals Earth’s multiscale microbial
509 diversity. *Nature*, 551(7681), 457–463. <https://doi.org/10.1038/nature24621>
- 510 Tickner, D., Opperman, J. J., Abell, R., Acreman, M., Arthington, A. H., Bunn, S. E., & Young, L. (2020).
511 Bending the Curve of Global Freshwater Biodiversity Loss: An Emergency Recovery Plan.
512 *BioScience*, 70(4), 330–342. <https://doi.org/10.1093/biosci/biaa002>
- 513 Valentin, V., Frédéric, R., Isabelle, D., Olivier, M., Yorick, R., & Agnès, B. (2019). Assessing pollution of
514 aquatic environments with diatoms’ DNA metabarcoding: experience and developments
515 from France water framework directive networks. *Metabarcoding and Metagenomics*, 3,
516 e39646. <https://doi.org/10.3897/mbmg.3.39646>
- 517 Wani, G. A., Khan, M. A., Dar, M. A., Shah, M. A., & Reshi, Z. A. (2021). Next Generation High
518 Throughput Sequencing to Assess Microbial Communities: An Application Based on Water
519 Quality. *Bulletin of Environmental Contamination and Toxicology*, 106(5), 727–733.
520 <https://doi.org/10.1007/s00128-021-03195-7>
- 521 Webb, T. J., & Mindel, B. L. (2015). Global Patterns of Extinction Risk in Marine and Non-marine
522 Systems. *Current Biology*, 25(4), 506–511. <https://doi.org/10.1016/j.cub.2014.12.023>
- 523 Whon, T. W., Ahn, S. W., Yang, S., Kim, J. Y., Kim, Y. B., Kim, Y., & Roh, S. W. (2021). ODFM, an omics
524 data resource from microorganisms associated with fermented foods. *Scientific Data*, 8(1),
525 113. <https://doi.org/10.1038/s41597-021-00895-x>
- 526 Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., & Eisen, J. A. (2009). A
527 phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature*, 462(7276), 1056–
528 1060. <https://doi.org/10.1038/nature08656>

530 Data availability statement

531 FreshOmics database files have been made available at <http://freshomics.lmge.uca.fr/>

532

533 **ORCID**

534 Corinne Biderre-Petit <https://orcid.org/0000-0001-7962-4171>

535 Marina Chauvet <https://orcid.org/0000-0002-6380-6162>

536 Arthur Monjot <https://orcid.org/0000-0002-6978-4785>

537 Anne Moné <https://orcid.org/0000-0002-8686-703X>

538 Viviane Ravet <https://orcid.org/0000-0001-7574-992X>

539 Gisèle Bronner <https://orcid.org/0000-0002-4863-4518?lang=en>

540 Hélène Gardon <https://orcid.org/0000-0001-8976-1372>

541 Cécile Lepère <https://orcid.org/0000-0003-4767-0477>

542 Didier Debroas <https://orcid.org/0000-0002-9915-1268>

543

544 **Conflict of interests**

545 The authors declare that they have no known competing interests.

546

547 **Author contributions**

548 CBP, JCC and DD designed the research. CBP, GB, CH, IJ, AMné, VR, AG, DD collected data from the
549 literature. MC, HG, AMjot conceived figures and pictograms for the web site. JCC designed the web
550 site application. CBP, JCC and CL wrote the paper with input from all co-authors.

551

552 **Figure titles and legends**

553 **Figure 1: Overview of the FreshOmics web interface (Home page) and construction method.** (A)
554 Metadata retrieval from the literature and repositories following two successive steps. The removal of
555 non-relevant papers (i) and the removal of non-freshwater samples (ii) were carried out as described
556 in the text. (B) Attributes mining, standardization and merging. The standardization of the attributes
557 was based on the ENVO, MIxS v.5, GO and ISO ontologies. (C) FreshOmics was made available online
558 through a user-friendly web application developed out as described in the text. (D) When selected, the
559 metadata can be downloaded under six different formats.

560

561 **Figure 2: Overall distribution of the datasets around the world.** (A) Distribution per continent. (B)
562 Distribution of -omics datasets per country for the most represented continents (*i.e.* Asia, North
563 America, and Europe).

564

565 **Figure 3: Distribution of the datasets (A) per ecosystem type and (B) per material type.**

566

567 **Figure 4: Distribution of the datasets per method (A) and for metabarcoding, per marker (B).**

568

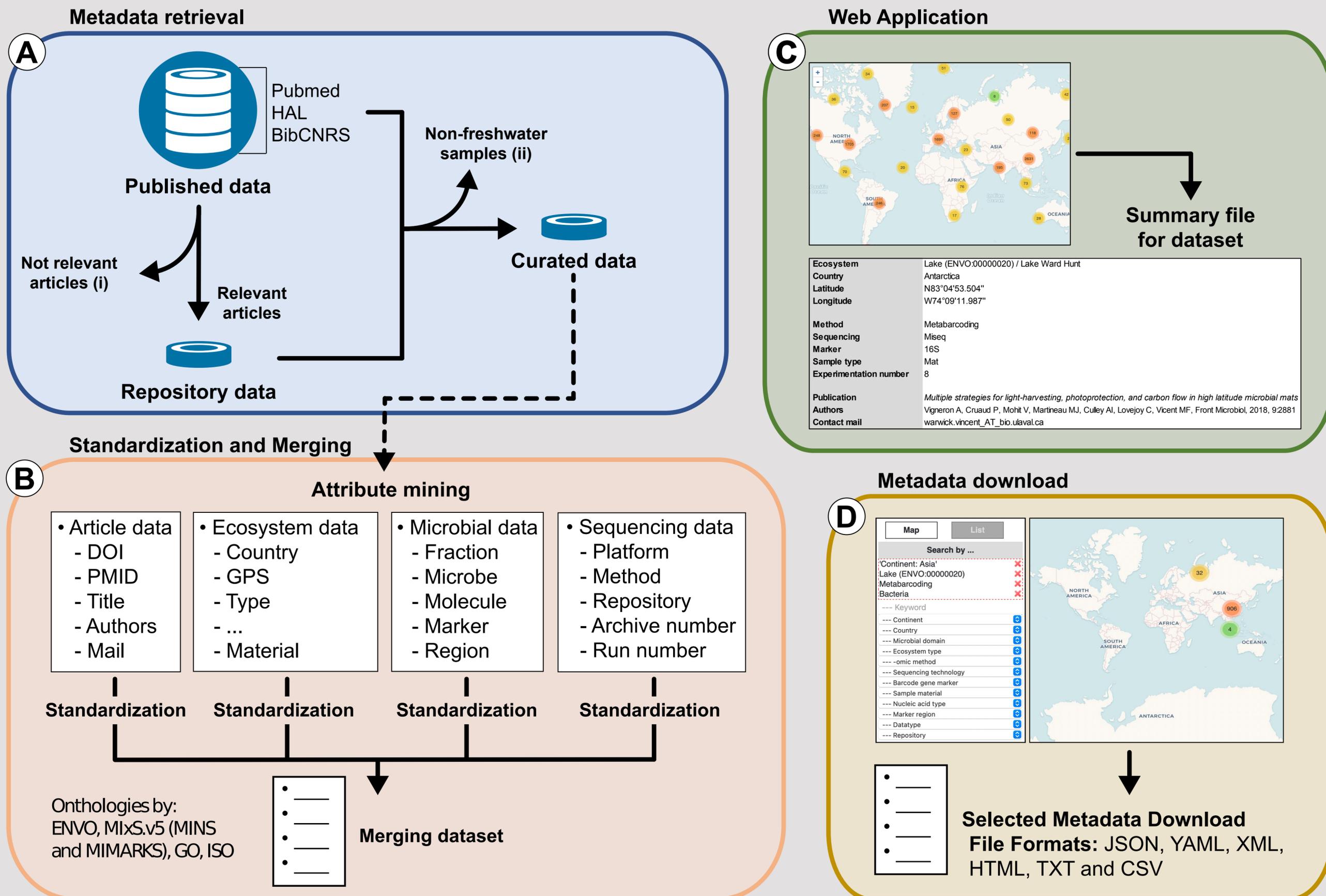
569 **Figure 5: Distribution of the datasets per sequencing technology (A) and evolution of technology**
570 **usage along the years (B).**

571

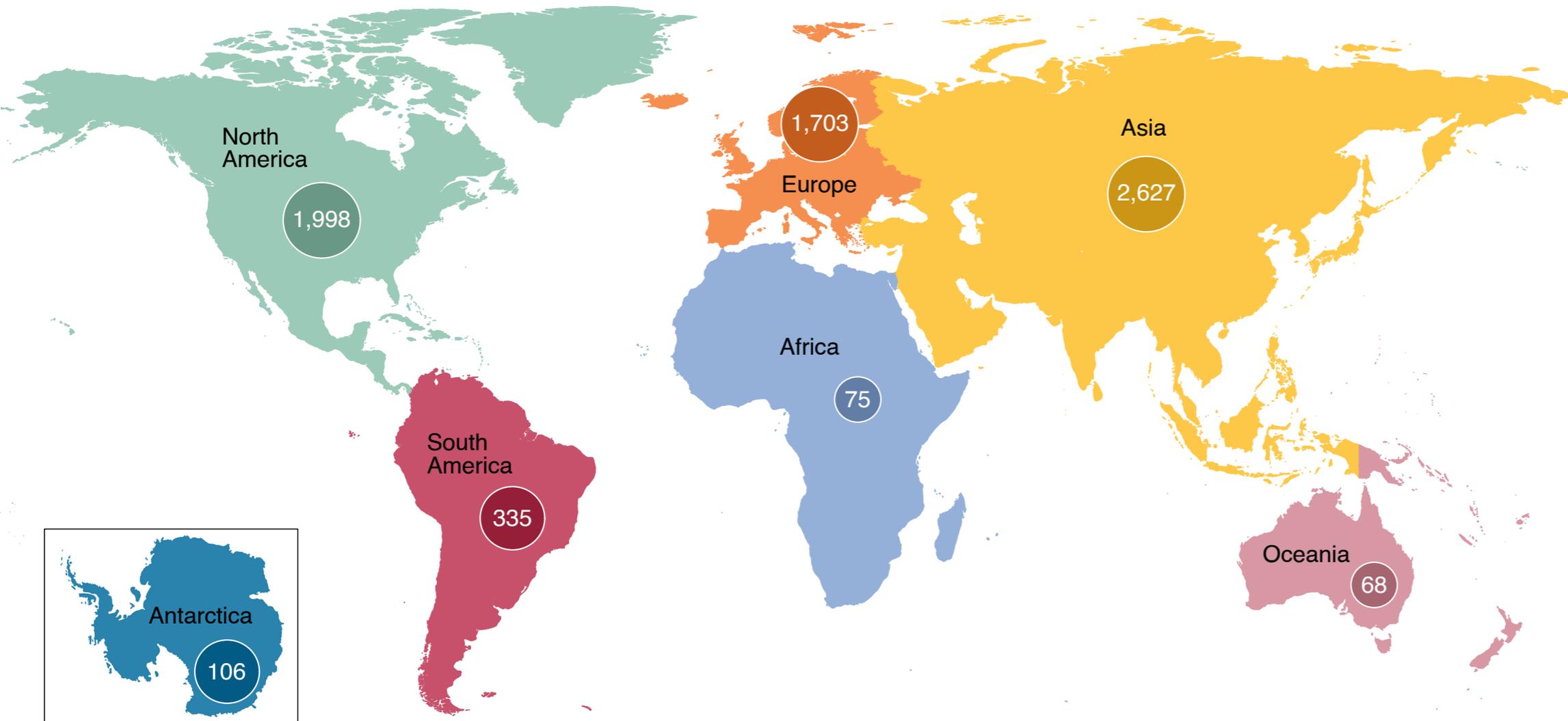
572 **Figure 6: Relative abundance of the datasets for the four main ecosystems (hot spring, stream, river**
573 **and lake) in Asia, Europe and North America.** The total percent of datasets devoted to these four
574 ecosystems for each continent is indicated in the box above each histogram.

For Review Only

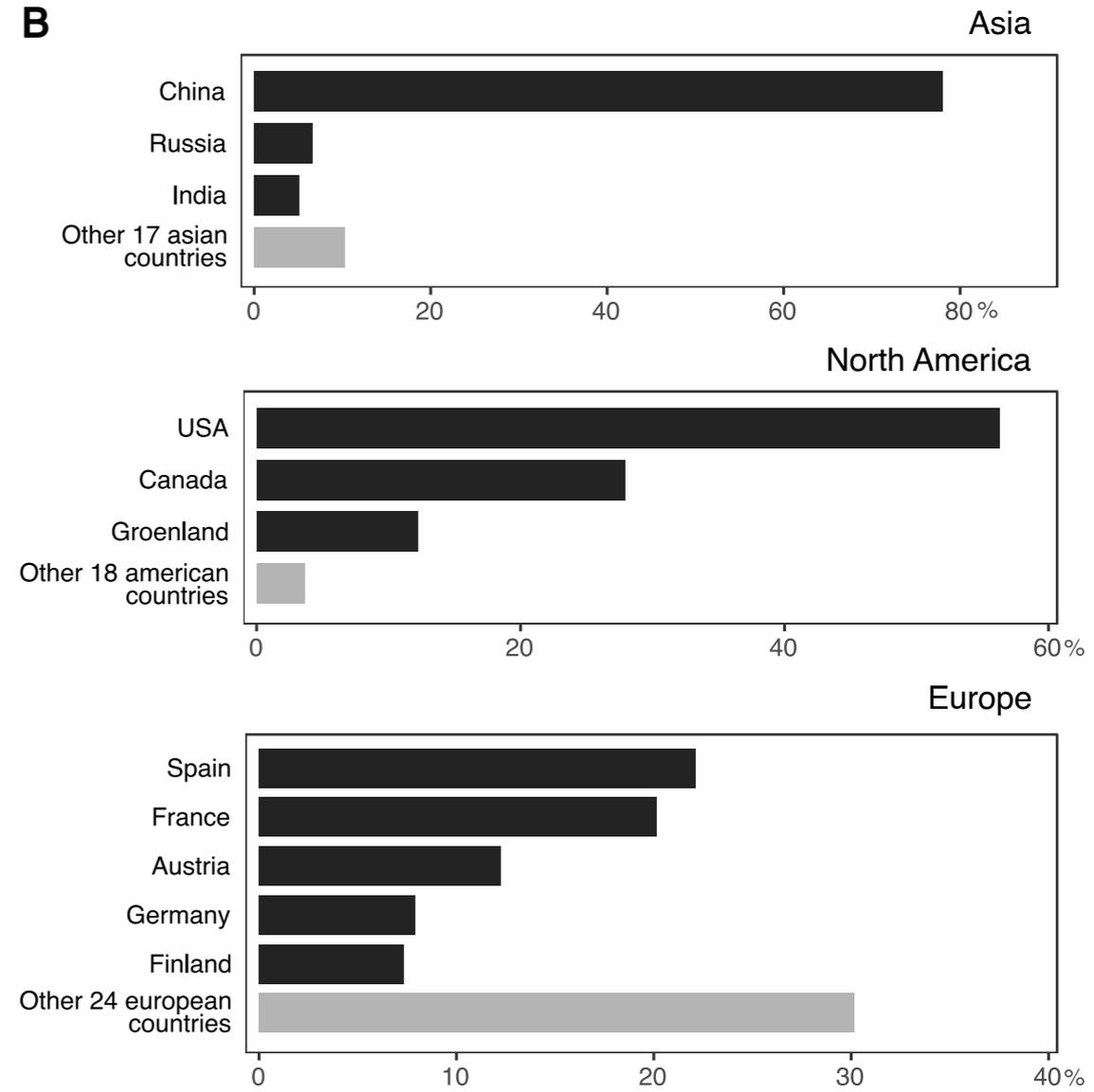
FreshOmics: a database dedicated to freshwater microbiomes



A



B



A**B**