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.

MOLECULAR ECOLOGY
RESOURCES

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

Journal:	*Molecular Ecology Resources*
Manuscript ID	Draft
Manuscript Type:	Resource Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Biderre-Petit, Corinne; Université Clermont Auvergne, LMGE UMR CNRS 6023 Charvy, Jean-Christophe; Université Clermont Auvergne, LMGE UMR CNRS 6023 Bronner, Gisele; Universite Clermont Auvergne Chauvet, Marina; Université Clermont Auvergne, LMGE UMR CNRS 6023 Debroas, Didier; Universite 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; Universite Clermont Auvergne Lepere, Cecile; Universite Clermont Auvergne,
Keywords:	Freshwater, Microorganisms, -omics data, Database

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

Running title: Microbial freshwater –omics database

Authors

Corinne Biderre-Petit[1†*], Jean-Christophe Charvy[1†], Gisèle Bronner[1], Marina Chauvet[1], Didier Debroas[1], Hélène Gardon[1], Claire Hennequin[1], Isabelle Jouan-Dufournel[1], Anne Moné[1], Arthur Monjot[1], Viviane Ravet[1], Agnès Vellet[1], Cécile Lepère[1]

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

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

†These authors contributed equally to this work.

Abstract

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

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

Keywords: Freshwater, microorganisms, -omics data, database

Introduction

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

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

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

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

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

Methods

Database construction

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

***Data retrieval and non-freshwater data removal*.** The data were collected from an extensive literature search using various combinations of keywords. Several publication databases were used such as PubMed, ScienceDirect, Scopus, HAL and BibCNRS (Figure 1A). For the collection, we created a glossary of terms related to the microbial domains, freshwater ecosystems, sample material, -omics approaches, barcode gene markers, and sequencing platforms (Supplementary Table S1). Research articles lacking at least one of the information on these terms were excluded. Similarly, when -omics data were cited in multiple papers, only the most informative paper was retained for FreshOmics. Unlike other databases, such as HumanMetagenomeDB (Kasmanas et al., 2021) and TMDB (Corrêa et 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

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

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

Web application implementation and exploration.

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

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

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

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

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

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

Database content overview

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

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

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

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

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

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

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

Discussion

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

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

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

The -omics era has revolutionized microbial ecology and has led to the emergence of new knowledge for a better understanding of microbial diversity, interrelationships and function within an ecosystem. With FreshOmics, we highlight great discrepancies in the distribution of the data published in the literature for the freshwater ecosystems at the geographical, ecosystem type, and microbial domain scale. The largest amount of -omics data was so far produced by Asia, North America, and Europe (Figure 2). This gap between continents has probably multiple causes however it is likely due to the income inequality and technological inability to generate and analyze big data (*e.g*. data storage, processing). A strong imbalance is also observed at the country scale as, for example, for Asia where China encompasses 78% of all datasets from this part of the world and Russia only 6.6%, while almost twice as large in area and with the second largest amount of freshwater in the world after Brazil (Brazil: 8,233 Km^3; Russia: 4,508 Km^3; China: 2,840 Km^3, https://www.worldatlas.com/articles/countries-with-the-most-freshwater-resources.html). This can be explained by the lack of previous concerns and investments of Russia for the biological sciences, largely considerate as top-down approaches (Gronvall & Bland, 2020). A strong disparity is also observed between the types of ecosystems studied, with a large dominance of lakes (43% of all datasets) and overall, all lotic and lentic ecosystems (75.6% of all datasets), whereas they represent only 12.5% of the global volume of fresh water (75% are locked

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

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

Conclusion and future work

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

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

Funding

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

Acknowledgements

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

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

References

Alcamo, J. M., Vörösmarty, C. J., Naiman, R. J., Lettenmaier, D. P., & Pahl-Wostl, C. (2008). A grand challenge for freshwater research: understanding the global water system. *Environmental Research Letters*, *3*(1), 010202. https://doi.org/10.1088/1748-9326/3/1/010202

Allan, J. D., Abell, R., Hogan, Z., Revenga, C., Taylor, B. W., Welcomme, R. L., & Winemiller, K. (2005). Overfishing of Inland Waters. *BioScience*, *55*(12), 1041. https://doi.org/10.1641/0006-3568(2005)055[1041:OOIW]2.0.CO;2

Bailet, B., Bouchez, A., Franc, A., Frigerio, J.-M., Keck, F., Karjalainen, S.-M., & Kahlert, M. (2019). Molecular versus morphological data for benthic diatoms biomonitoring in Northern Europe freshwater and consequences for ecological status. *Metabarcoding and Metagenomics*, *3*, e34002. https://doi.org/10.3897/mbmg.3.34002

Beattie, R. E., Bandla, A., Swarup, S., & Hristova, K. R. (2020). Freshwater Sediment Microbial Communities Are Not Resilient to Disturbance From Agricultural Land Runoff. *Frontiers in Microbiology*, *11*, 539921. https://doi.org/10.3389/fmicb.2020.539921

Boenigk, J., Wodniok, S., Bock, C., Beisser, D., Hempel, C., Grossmann, L., & Jensen, M. (2018). Geographic distance and mountain ranges structure freshwater protist communities on a European scale. *Metabarcoding and Metagenomics*, *2*, e21519. https://doi.org/10.3897/mbmg.2.21519

Brooksbank, C., Bergman, M. T., Apweiler, R., Birney, E., & Thornton, J. (2014). The European Bioinformatics Institute's data resources 2014. *Nucleic Acids Research*, *42*(D1), D18–D25. https://doi.org/10.1093/nar/gkt1206

Cahoon, A. B., Huffman, A. G., Krager, M. M., & Crowell, R. M. (2018). A meta-barcoding census of freshwater planktonic protists in Appalachia – Natural Tunnel State Park, Virginia, USA. *Metabarcoding and Metagenomics*, *2*, e26939. https://doi.org/10.3897/mbmg.2.26939

Corrêa, F. B., Saraiva, J. P., Stadler, P. F., & da Rocha, U. N. (2019). TerrestrialMetagenomeDB: a public repository of curated and standardized metadata for terrestrial metagenomes. *Nucleic Acids Research*, gkz994. https://doi.org/10.1093/nar/gkz994

Debroas, D., Domaizon, I., Humbert, J.-F., Jardillier, L., Lepère, C., Oudart, A., & Taïb, N. (2017). Overview of freshwater microbial eukaryotes diversity: a first analysis of publicly available metabarcoding data. *FEMS Microbiology Ecology*, *93*(4). https://doi.org/10.1093/femsec/fix023

Delmont, T. O., Gaia, M., Hinsinger, D. D., Fremont, P., Vanni, C., Guerra, A. F., & Jaillon, O. (2020). *Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics* [Preprint]. Microbiology. https://doi.org/10.1101/2020.10.15.341214

Dudgeon, D. (2019). Multiple threats imperil freshwater biodiversity in the Anthropocene. *Current Biology*, *29*(19), R960–R967. https://doi.org/10.1016/j.cub.2019.08.002

Edwards, R. A., Rodriguez-Brito, B., Wegley, L., Haynes, M., Breitbart, M., Peterson, D. M., & Rohwer, F. (2006). Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics*, *7*(1), 57. https://doi.org/10.1186/1471-2164-7-57

Gronvall, G. K., & Bland, B. (2020). Life-science research and biosecurity concerns in the Russian Federation. *The Nonproliferation Review*, *27*(4–6), 415–423. https://doi.org/10.1080/10736700.2020.1866323

Ingala, M. R., Werner, I. E., Fitzgerald, A. M., & Naro-Maciel, E. (2021). 18S rRNA amplicon sequence data (V1–V3) of the Bronx river estuary, New York. *Metabarcoding and Metagenomics*, *5*, e69691. https://doi.org/10.3897/mbmg.5.69691

Jane, S. F., Hansen, G. J. A., Kraemer, B. M., Leavitt, P. R., Mincer, J. L., North, R. L., & Rose, K. C. (2021). Widespread deoxygenation of temperate lakes. *Nature*, *594*(7861), 66–70. https://doi.org/10.1038/s41586-021-03550-y

Kasmanas, J. C., Bartholomäus, A., Corrêa, F. B., Tal, T., Jehmlich, N., Herberth, G., & da Rocha, N.U. (2021). HumanMetagenomeDB: a public repository of curated and standardized metadata for human metagenomes. *Nucleic Acids Research*, *49*(D1), D743–D750. https://doi.org/10.1093/nar/gkaa1031

Keck, F., Millet, L., Debroas, D., Etienne, D., Galop, D., Rius, D., & Domaizon, I. (2020). Assessing the response of micro-eukaryotic diversity to the Great Acceleration using lake sedimentary DNA. *Nature Communications*, *11*(1), 3831. https://doi.org/10.1038/s41467-020-17682-8

Kibona, D., Kidulile, G., & Rwabukambara, F. (2009). Environment, Climate Warming and Water Management. *Transition Studies Review*, *16*(2), 484–500. https://doi.org/10.1007/s11300-009-0084-z

Klemetsen, T., Raknes, I. A., Fu, J., Agafonov, A., Balasundaram, S. V., Tartari, G., & Willassen, N. P. (2018). The MAR databases: development and implementation of databases specific for marine metagenomics. *Nucleic Acids Research*, *46*(D1), D692–D699. https://doi.org/10.1093/nar/gkx1036

Linz, A. M., Crary, B. C., Shade, A., Owens, S., Gilbert, J. A., Knight, R., & McMahon, K. D. (2017). Bacterial Community Composition and Dynamics Spanning Five Years in Freshwater Bog Lakes. *MSphere*, *2*(3), e00169-17. https://doi.org/10.1128/mSphere.00169-17

Lupo, V., Van Vlierberghe, M., Vanderschuren, H., Kerff, F., Baurain, D., & Cornet, L. (2021). Contamination in Reference Sequence Databases: Time for Divide-and-Rule Tactics. *Frontiers in Microbiology*, *12*, 755101. https://doi.org/10.3389/fmicb.2021.755101

Matsuoka, S., Sugiyama, Y., Sato, H., Katano, I., Harada, K., & Doi, H. (2019). Spatial structure of fungal DNA assemblages revealed with eDNA metabarcoding in a forest river network in western Japan. *Metabarcoding and Metagenomics*, *3*, e36335. https://doi.org/10.3897/mbmg.3.36335

McDaniel, E. A., Wahl, S. A., Ishii, S., Pinto, A., Ziels, R., Nielsen, P. H., & Williams, R. B. H. (2021). Prospects for multi-omics in the microbial ecology of water engineering. *Water Research*, *205*, 117608. https://doi.org/10.1016/j.watres.2021.117608

Meredith, C., Hoffman, J., Trebitz, A., Pilgrim, E., Okum, S., Martinson, J., & Cameron, E. S. (2021). Evaluating the performance of DNA metabarcoding for assessment of zooplankton communities in Western Lake Superior using multiple markers. *Metabarcoding and Metagenomics*, *5*, e64735. https://doi.org/10.3897/mbmg.5.64735

Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., & Finn, R. D. (2019). MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research*, gkz1035. https://doi.org/10.1093/nar/gkz1035

Mondav, R., Bertilsson, S., Buck, M., Langenheder, S., Lindström, E. S., & Garcia, S. L. (2020). Streamlined and Abundant Bacterioplankton Thrive in Functional Cohorts. *MSystems*, *5*(5), e00316-20. https://doi.org/10.1128/mSystems.00316-20

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., & Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733-745. https://doi.org/10.1093/nar/gkv1189

Ortiz-Álvarez, R., Cáliz, J., Camarero, L., & Casamayor, E. O. (2020). Regional community assembly drivers and microbial environmental sources shaping bacterioplankton in an alpine lacustrine

district (Pyrenees, Spain). *Environmental Microbiology*, *22*(1), 297–309. https://doi.org/10.1111/1462-2920.14848

Ortiz-Álvarez, R., Fierer, N., de los Ríos, A., Casamayor, E. O., & Barberán, A. (2018). Consistent changes in the taxonomic structure and functional attributes of bacterial communities during primary succession. *The ISME Journal*, *12*(7), 1658–1667. https://doi.org/10.1038/s41396-018-0076-2

Ponsero, A. J., Bomhoff, M., Blumberg, K., Youens-Clark, K., Herz, N. M., Wood-Charlson, E. M., & Hurwitz, B. L. (2021). Planet Microbe: a platform for marine microbiology to discover and analyze interconnected 'omics and environmental data. *Nucleic Acids Research*, *49*(D1), D792–D802. https://doi.org/10.1093/nar/gkaa637

Rigden, D. J., & Fernández, X. M. (2021). The 2021 *Nucleic Acids Research* database issue and the online molecular biology database collection. *Nucleic Acids Research*, *49*(D1), D1–D9. https://doi.org/10.1093/nar/gkaa1216

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

Santillan, E., Seshan, H., Constancias, F., Drautz-Moses, D. I., & Wuertz, S. (2019). Frequency of disturbance alters diversity, function, and underlying assembly mechanisms of complex bacterial communities. *Npj Biofilms and Microbiomes*, *5*(1), 8. https://doi.org/10.1038/s41522-019-0079-4

Steinegger, M., & Salzberg, S. L. (2020). Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biology*, *21*(1), 115. https://doi.org/10.1186/s13059-020-02023-1

Tang, L. (2020). Contamination in sequence databases. *Nature Methods*, *17*(7), 654–654. https://doi.org/10.1038/s41592-020-0895-8

The Earth Microbiome Project Consortium, Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., & Knight, R. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*, *551*(7681), 457–463. https://doi.org/10.1038/nature24621

Tickner, D., Opperman, J. J., Abell, R., Acreman, M., Arthington, A. H., Bunn, S. E., & Young, L. (2020). Bending the Curve of Global Freshwater Biodiversity Loss: An Emergency Recovery Plan. *BioScience*, *70*(4), 330–342. https://doi.org/10.1093/biosci/biaa002

Valentin, V., Frédéric, R., Isabelle, D., Olivier, M., Yorick, R., & Agnès, B. (2019). Assessing pollution of aquatic environments with diatoms' DNA metabarcoding: experience and developments from France water framework directive networks. *Metabarcoding and Metagenomics*, *3*, e39646. https://doi.org/10.3897/mbmg.3.39646

Wani, G. A., Khan, M. A., Dar, M. A., Shah, M. A., & Reshi, Z. A. (2021). Next Generation High Throughput Sequencing to Assess Microbial Communities: An Application Based on Water Quality. *Bulletin of Environmental Contamination and Toxicology*, *106*(5), 727–733. https://doi.org/10.1007/s00128-021-03195-7

Webb, T. J., & Mindel, B. L. (2015). Global Patterns of Extinction Risk in Marine and Non-marine Systems. *Current Biology*, *25*(4), 506–511. https://doi.org/10.1016/j.cub.2014.12.023

Whon, T. W., Ahn, S. W., Yang, S., Kim, J. Y., Kim, Y. B., Kim, Y., & Roh, S. W. (2021). ODFM, an omics data resource from microorganisms associated with fermented foods. *Scientific Data*, *8*(1), 113. https://doi.org/10.1038/s41597-021-00895-x

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., & Eisen, J. A. (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature*, *462*(7276), 1056–1060. https://doi.org/10.1038/nature08656

Data availability statement

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

ORCID

Corinne Biderre-Petit https://orcid.org/0000-0001-7962-4171
Marina Chauvet https://orcid.org/0000-0002-6380-6162
Arthur Monjot https://orcid.org/0000-0002-6978-4785
Anne Moné https://orcid.org/0000-0002-8686-703X
Viviane Ravet https://orcid.org/0000-0001-7574-992X
Gisèle Bronner https://orcid.org/0000-0002-4863-4518?lang=en
Hélène Gardon https://orcid.org/0000-0001-8976-1372
Cécile Lepère https://orcid.org/0000-0003-4767-0477
Didier Debroas https://orcid.org/ 0000-0002-9915-1268

Conflict of interests

The authors declare that they have no known competing interests.

Author contributions

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

Figure titles and legends

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

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

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

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

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

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

FreshOmics: a database dedicated to freshwater microbiomes

Metadata retrieval

A

Pubmed
HAL
BibCNRS

Published data

Not relevant articles (i)

Relevant articles

Non-freshwater samples (ii)

Curated data

Repository data

Standardization and Merging

B

Attribute mining

- Article data
  - DOI
  - PMID
  - Title
  - Authors
  - Mail
- Ecosystem data
  - Country
  - GPS
  - Type
  - ...
  - Material
- Microbial data
  - Fraction
  - Microbe
  - Molecule
  - Marker
  - Region
- Sequencing data
  - Platform
  - Method
  - Repository
  - Archive number
  - Run number

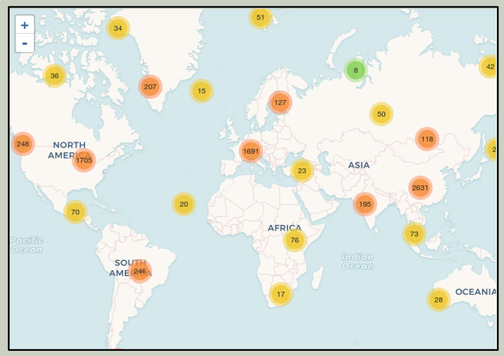
Standardization **Standardization** **Standardization** **Standardization**

Onthologies by:
ENVO, MIxS.v5 (MINS and MIMARKS), GO, ISO

Merging dataset

Web Application

C

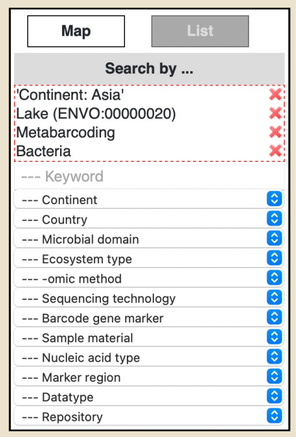


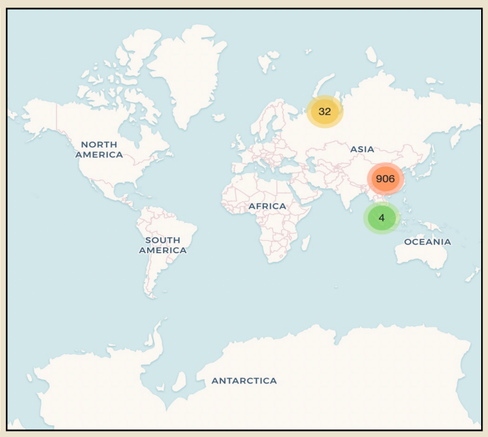
Summary file for dataset

Ecosystem	Lake (ENVO:00000020) / Lake Ward Hunt
Country	Antarctica
Latitude	N83°04'53.504"
Longitude	W74°09'11.987"
Method	Metabarcoding
Sequencing	Miseq
Marker	16S
Sample type	Mat
Experimentation number	8
Publication	*Multiple strategies for light-harvesting, photoprotection, and carbon flow in high latitude microbial mats*
Authors	Vigneron A, Cruaud P, Mohit V, Martineau MJ, Culley AI, Lovejoy C, Vicent MF, Front Microbiol, 2018, 9:2881
Contact mail	warwick.vincent_AT_bio.ulaval.ca

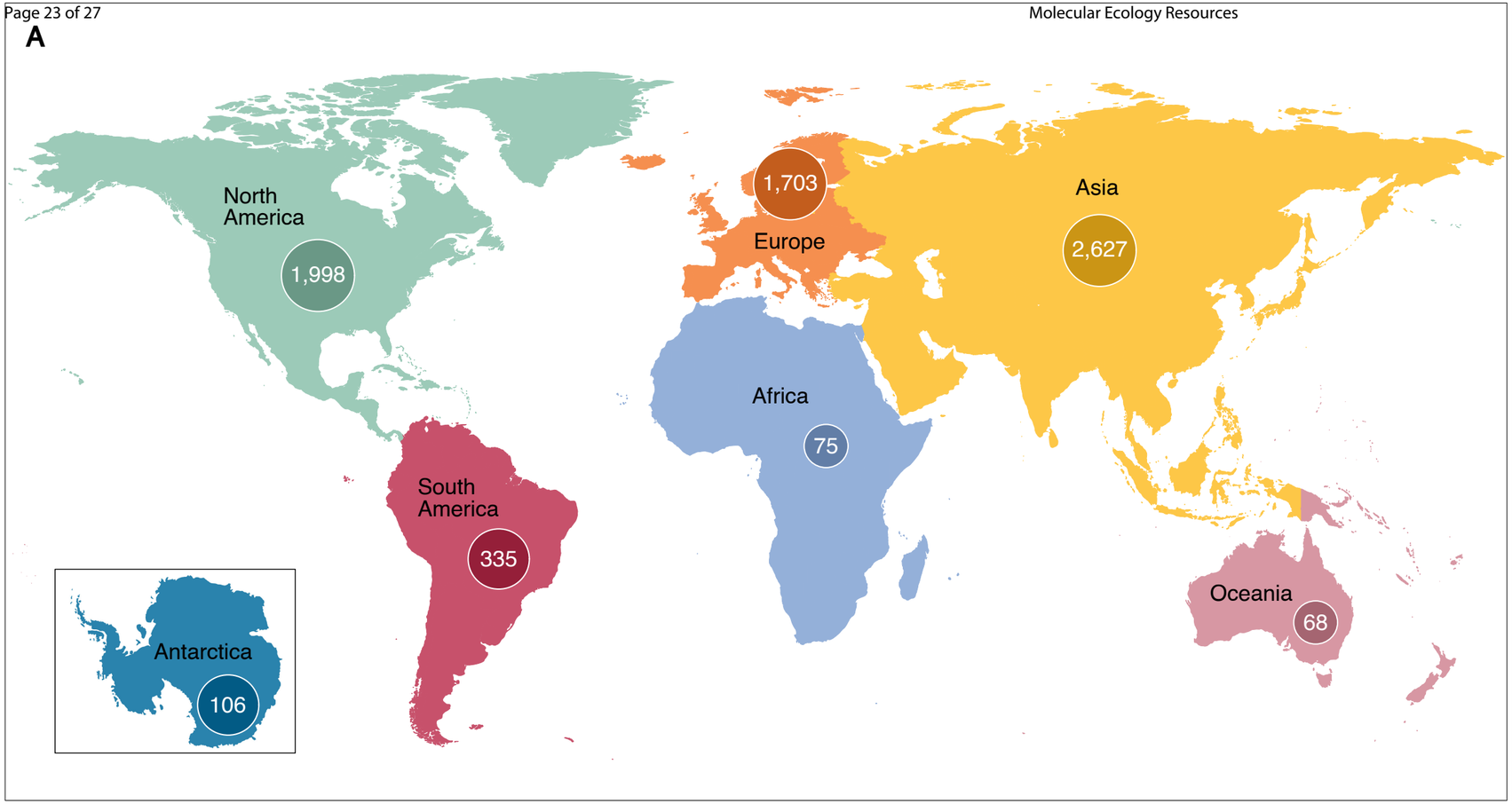
Metadata download

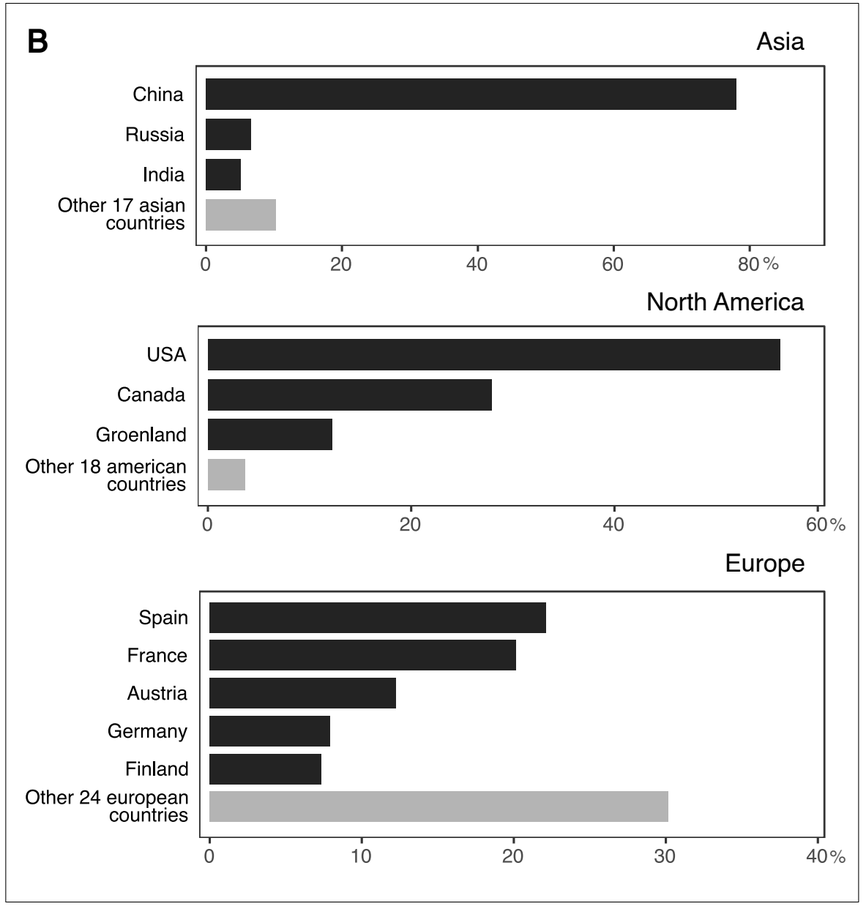
D



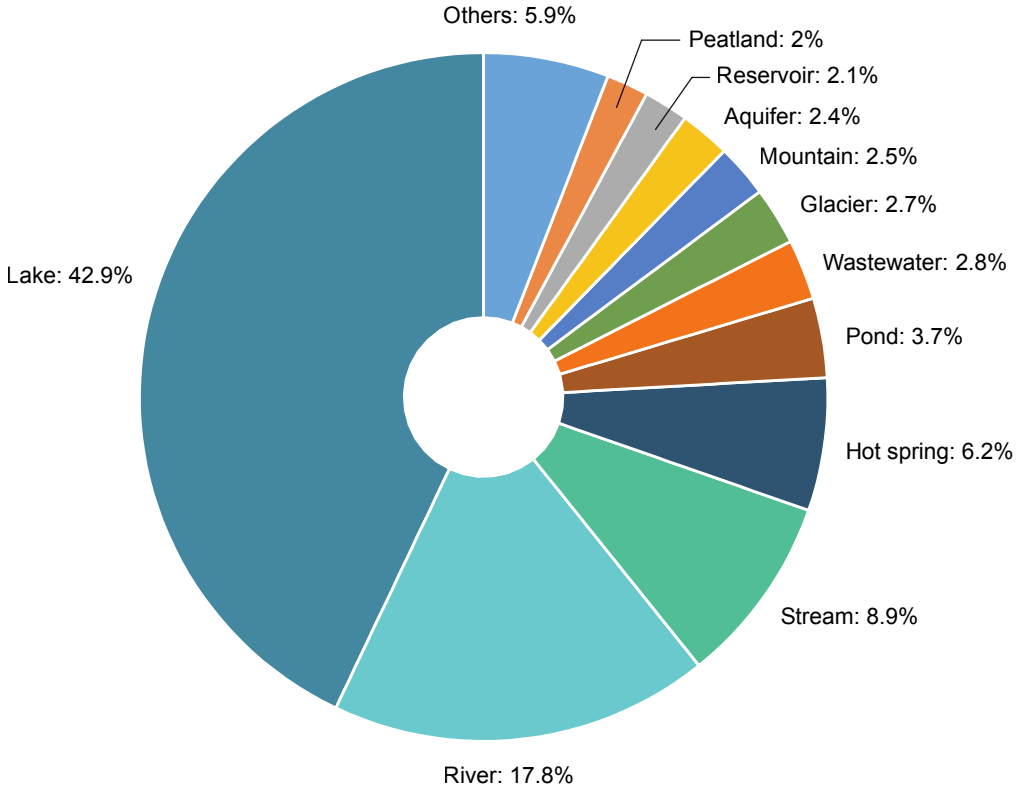


Selected Metadata Download
File Formats: JSON, YAML, XML, HTML, TXT and CSV

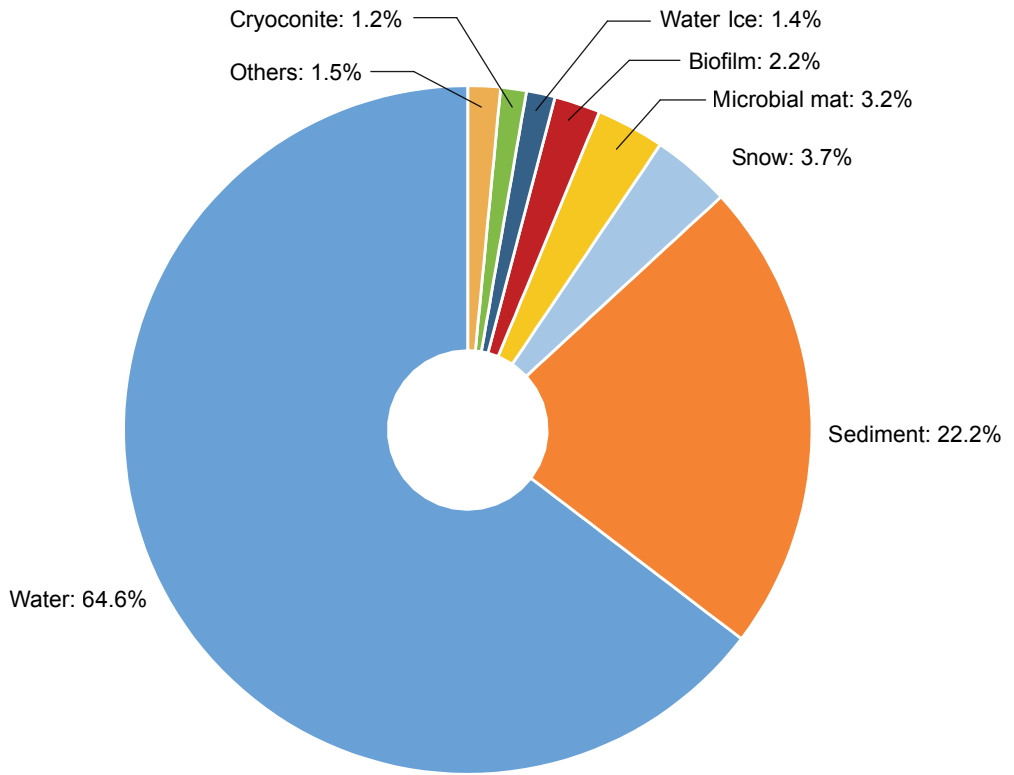

A
North America
1,998
Europe
1,703
Asia
2,627
Africa
75
South America
335
Oceania
68
Antarctica
106

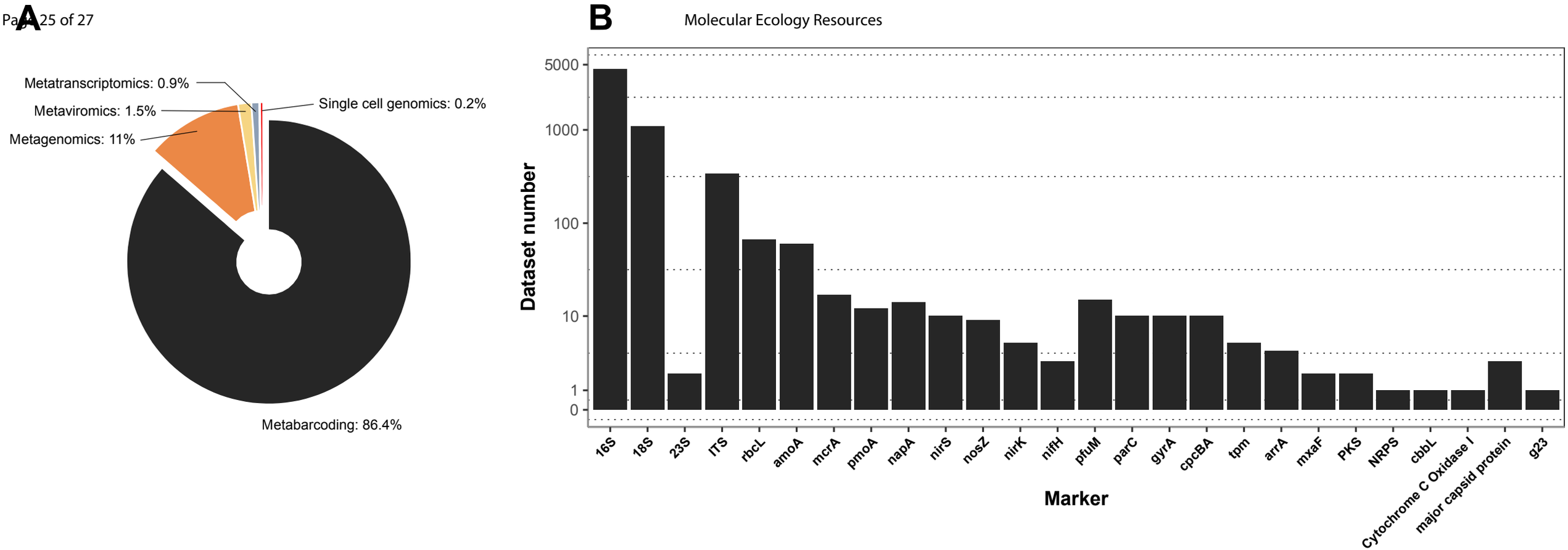

B
Asia
China
Russia
India
Other 17 asian countries
0
20
40
60
80 %
North America
USA
Canada
Groenland
Other 18 american countries
0
20
40
60 %
Europe
Spain
France
Austria
Germany
Finland
Other 24 european countries
0
10
20
30
40 %

A

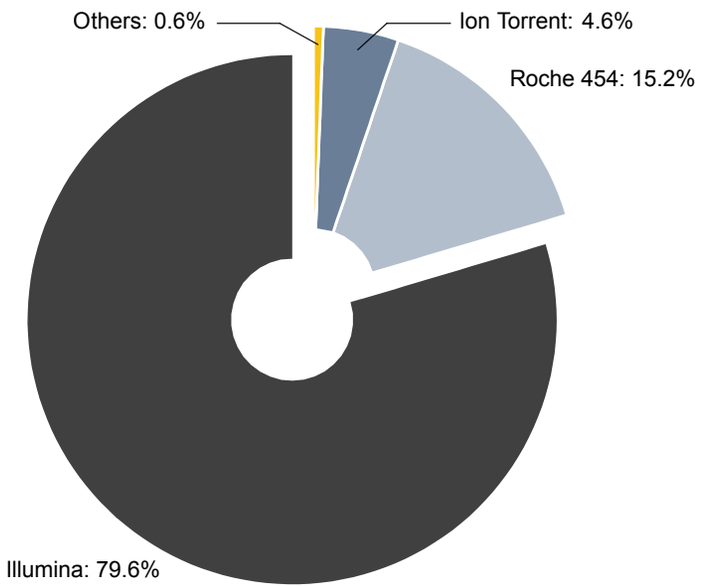

Others: 5.9%
Peatland: 2%
Reservoir: 2.1%
Aquifer: 2.4%
Mountain: 2.5%
Glacier: 2.7%
Wastewater: 2.8%
Pond: 3.7%
Hot spring: 6.2%
Stream: 8.9%
River: 17.8%
Lake: 42.9%

B

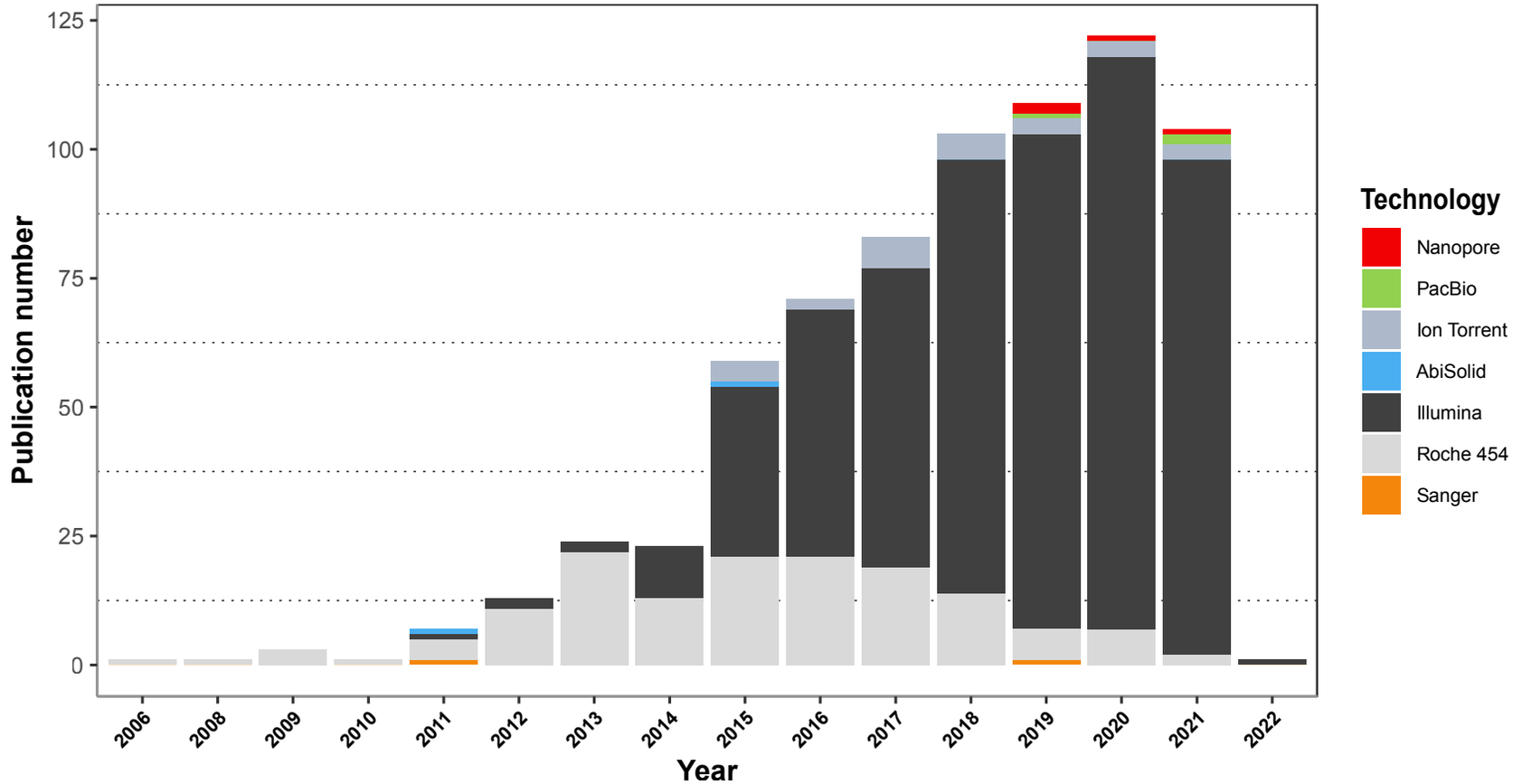

Cryoconite: 1.2%
Water Ice: 1.4%
Others: 1.5%
Biofilm: 2.2%
Microbial mat: 3.2%
Snow: 3.7%
Sediment: 22.2%
Water: 64.6%

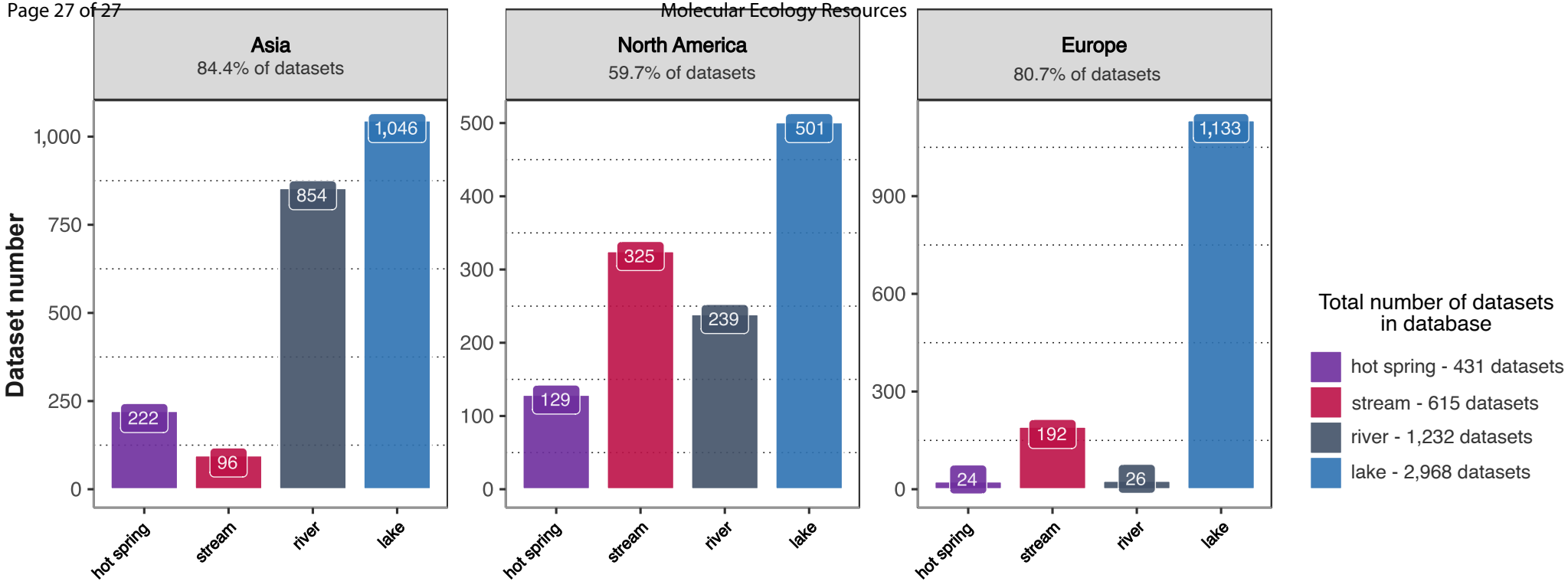

A
Metatranscriptomics: 0.9%
Metaviromics: 1.5%
Metagenomics: 11%
Single cell genomics: 0.2%
Metabarcoding: 86.4%
B
Dataset number
5000
1000
100
10
1
0
16S
18S
23S
ITS
rbcL
amoA
mcrA
pmoA
napA
nirS
nosZ
nirK
nifH
pfuM
parC
gyrA
cpcBA
tpm
arrA
mxaF
PKS
NRPS
cbbL
Cytochrome C Oxidase I
major capsid protein
g23
Marker

A


Others: 0.6%
Ion Torrent: 4.6%
Roche 454: 15.2%
Illumina: 79.6%

B


Publication number
125
100
75
50
25
0
2006
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Year
Technology
Nanopore
PacBio
Ion Torrent
AbiSolid
Illumina
Roche 454
Sanger


Asia
84.4% of datasets
North America
59.7% of datasets
Europe
80.7% of datasets
Dataset number
1,000
750
500
250
0
500
400
300
200
100
0
900
600
300
0
222
96
854
1,046
129
325
239
501
24
192
26
1,133
hot spring
stream
river
lake
Total number of datasets in database
hot spring - 431 datasets
stream - 615 datasets
river - 1,232 datasets
lake - 2,968 datasets