# ExTaxsI: an exploration tool of biodiversity molecular data

Giulia Agostinetto[1], Anna Sandionigi[1,2,*], Adam Chahed[1], Alberto Brusati[1], Elena Parladori[2], Bachir Balech[3], Antonia Bruno[1], Dario Pescini[4], Maurizio Casiraghi[1]

**1 University of Milano-Bicocca, Department of Biotechnology and Biosciences, Milan, Italy**
**2 Quantia Consulting srl, Milan, Italy**
**3 Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies (CNR), via Amendola 165/A, 70126, Bari, Italy**
**4 University of Milano-Bicocca, Department of Statistics and Quantitative Methods, Milan, Italy**

*** anna.sandionigi@unimib.it**

## Abstract

**Background** The increasing availability of multi omics data is leading to continually revise estimates of existing biodiversity data. In particular, the molecular data enable to characterize novel species yet unknown and to increase the information linked to those already observed with new genomic data. For this reason, the management and visualization of existing molecular data, and their related metadata, through the implementation of easy to use IT tools have become a key point for the development of future research. The more users are able to access biodiversity related information, the greater the ability of the scientific community to expand the knowledge in this area. **Results** In our research we have focused on the development of ExTaxsI (Exploring Taxonomies Information), an IT tool able to retrieve biodiversity data stored in NCBI databases and provide a simple and explorable visualization. Through the three case studies presented here, we have shown how an efficient organization of the data already present can lead to obtaining new information that is fundamental as a starting point for new research. Our approach was also able to highlight the limits in the distribution data availability, a key factor to consider in the experimental design phase of broad spectrum studies, such as metagenomics. **Conclusions** ExTaxI can easily produce explorable visualization of molecular data and its metadata, with the aim to help researchers to improve experimental designs and highlight the main gaps in the coverage of available data.

**Keywords**: Biodiversity; Data visualization; Molecular data; Database; Data integration; Taxonomy gaps

## 1 Introduction

In recent years, studies investigating biodiversity at large scale started to create and incorporate molecular data. In particular, the spread of metagenomic studies (e.g. metabarcoding) have contributed to an exponential increase in genomic data availability. Thanks to this large amount of new information it is possible to expand our knowledge and enhance our scientific investigation capacity in many fields of research [47], ranging from macro-ecology and ecosystem monitoring, to food safety control, forensics

applications and microbiome identification [15, 47, 52]. Different groups of researchers emphasized the wealth of information collected in biological and molecular databases, with the aim to improve usefulness and reusability of data [21, 39, 57]. Therefore, building experimental designs that consider the totality of the data present in the biological databases could certainly increase the efficiency of these studies, and lead to more robust results [1, 41].

Biodiversity data retrieval and exploration has become a big data issue, forcing researchers to use Information Technologies (IT) tools to manage those data. In particular, the interpretation of results derived from metagenomic approaches requires computational pipelines and IT infrastructures that improve over time, but are strongly linked to the availability of pre-existing data stored in online databases (e.g. ENA - `www.ebi.ac.uk/ena`; and NCBI - `https://www.ncbi.nlm.nih.gov/`).

Visualization remains an effective strategy not only to aggregate and present the research results, but also to guide advanced investigations [22, 29]. At this moment, reference databases where molecular and taxonomic data are friendly explorable and punctually updated exist only for few molecular markers, such as SILVA for 16S and 18S genes [48], BOLD for animals and plants [50] or UNITE for Fungi domain [44]. However, these data resources are not representative of all the genomic and taxonomic diversity collected to date. On the other hand, although GenBank still resumes the majority of genetic data and their related metadata currently available [3, 5, 30], such information is not always easy to access without specific bioinformatics skills, which is a limiting factor to a large audience of scientists.

With the aim to help biologists to improve experimental designs and to encourage data exploration and exploitation, we have developed a tool, ExTaxsI (Exploring Taxonomies Information), to facilitate the molecular data integration with its associated metadata, eventually retrieved from heterogeneous sources. Moreover, its ease of use interface will help researchers and practitioners in the visualization phase. ExTaxsI can both query NCBI Nucleotide database for molecular data and accept data from an external source, exploiting the standard taxonomy notation. The tool is linked to NCBI taxonomy database [17] and ETE toolkit [26], in order to produce standard formats readable by most common software that deal with taxonomic information [4, 7, 8, 38, 51, 56], such as QIIME2 platform [7]. The tool is applicable to any marker, gene name or taxonomic group, so it is possible to create non-standard marker genes database usable in metagenomic/metabarcoding taxonomic assignment tools [7]. In addition, thanks to the integration of the NCBI query tool [11], ExTaxsI can reorganize personal datasets in a standardized format in order to easily describe taxonomic variability and geographic provenance of records.

## 2 ExTaxsI

ExTaxsI is a bioinformatic tool aimed to elaborate and visualize molecular and taxonomic information via a simple interface. This open-source user friendly instrument, developed in Python 3.7, starting from a list of taxa or gene name/s (as illustrated in Figure 1), allows i) the search of taxonomic, genetic and biogeographical data through NCBI databases, ii) the creation of local and formatted nucleotide sequences (FASTA format) dataset and iii) their related taxonomy classification paths/datasets, thanks to the integration of NCBI taxonomy data, iv) the creation of genetic markers lists coming from different studies and finally v) the production of interactive plots starting from NCBI query search results or directly from offline taxonomic files, including representative graphs for the exploration of taxonomy and refinement of biogeographical data, creating geographical maps with the locations of the species analyzed (Figure 1). It is important to note that ExTaxsI outputs are compatible with other tools for

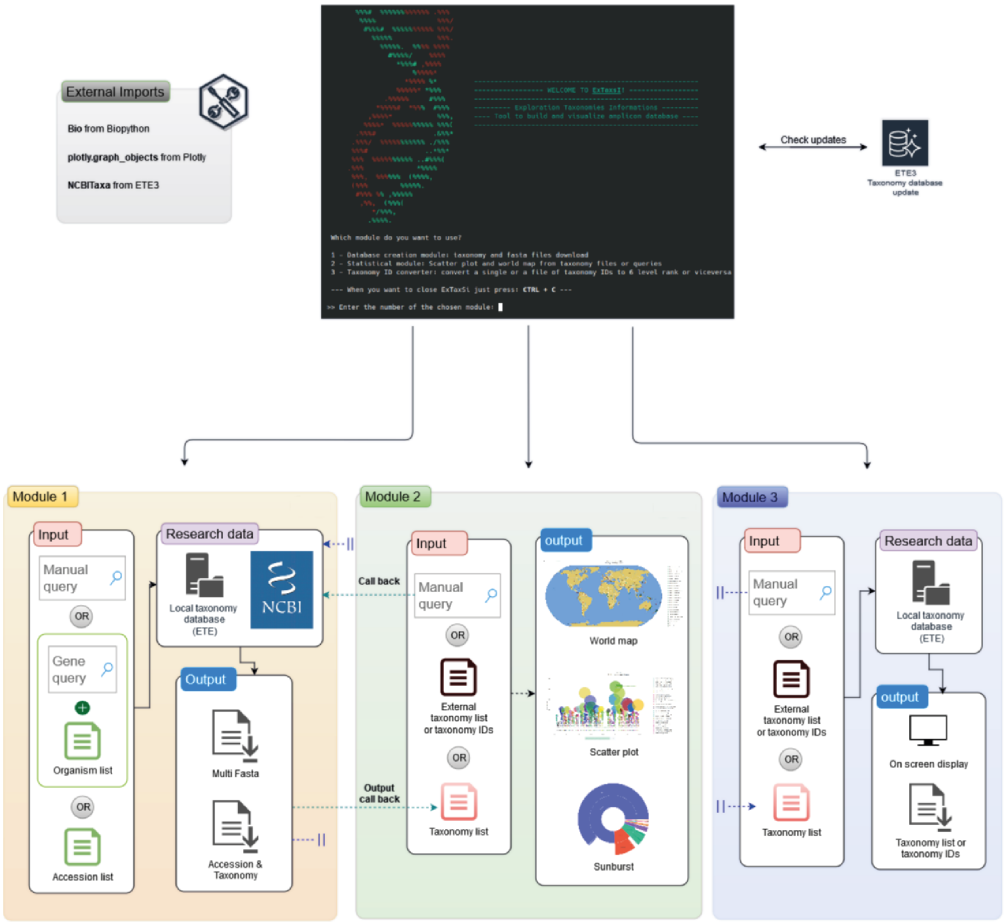



**Figure 1.** ExTaxsI pipeline: module 1 (orange) searches and creates files and databases; module 2 (green) processes georeferenced or taxonomic data for the creation of graphs and plots; module 3 (blue) converts taxonomic data into taxonomic ID (TaxID) and vice versa.

taxonomic assignment purposes such as QIIME2 platform [7].

The communication with NCBI server is mediated by the Entrez module [11], implemented in Biopython library [10], which allows to search, download and parse query results. To help NCBI interaction, when the requests are less than 2500, the search key is composed by a single query, otherwise the query will be splitted in groups of 2500 generating temporary files, which are then merged into single output file at the end of the process.

Regarding taxonomy handling, ETE toolkit was exploited [26]. In particular, ETE allows to create and maintain a local taxonomy database up to date by extrapolating the 6 main ranks (phylum, class, order, family, genus and species). If the organism is poorly described or it is an unknown species, the Taxonomy ID (i.e. TaxID) of its ancestor (known as parent TaxID) in ETE taxonomic tree is then used and converted into its scientific correspondent name. It is important to underline that all queries are carried out locally, avoiding unnecessary delays and allowing the integration of the tool in genomic and metagenomic pipelines.

Finally, the extracted data are visualized through the ScatterPlot and the SunBurst (Expansion Pie) for the taxonomy exploration, and ScatterGeo for the geographic

metadata plotting.

## 2.1 Use cases

Being ExTaxsI a taxonomy focused data exploration tool, we designed three possible scenarios of increasing complexity, to challenge it with increasing taxonomic variability and dimension of accession entries. The first scenario hypothesizes a query to explore data with i) low taxonomic variability and a high number of expected entries (1 species, more than 300,000 entries). The second scenario provides ii) a high taxonomic variability and a large expected number of entries (about 500 species, more than 300,000 entries). The third and more complex scenario explores a iii) complete case study with taxonomic input intersected by molecular data. As case studies of the first two scenarios, we focused on taxa of interest in marine fisheries: 1) the cod fish species (*Gadus morhua*), for which a worldwide economic interest exists, and 2) its taxonomic group at order level - the *Gadiformes* order - which supports long-standing commercial fisheries and aquaculture. These two case studies evaluate the capacity to explore data and to fill the geographic distribution of a species, prospecting also the available genes information to perform a genetic survey (e.g. DNA metabarcoding study). With the third use case, we aimed at demonstrating the flexibility of ExTaxsI in different contexts: a genetic exploration of the available data in NCBI associated to *SARS-CoV-2* virus - a very recent topic that involved many research groups, leading to huge amounts of data collected and deposited in public sources [6]. A large scale exploration of data related to this topic can potentially improve the reliability of results and can provide valuable evidence to inform decisions on public health protection, both now and most importantly in the future.

### 2.1.1 Insights into two taxonomic groups of commercial interest

The first scenario is the case of *Gadus morhua* species, also called Atlantic cod. In detail, *Gadus morhua* is a large, cold-adapted teleost fish that supports long-standing commercial fisheries and aquaculture [27, 28, 33, 34, 54].

ExTaxsI retrieved a total of 366,963 accessions using the taxonomy ID through the following query: "txid8049[ORGN]" (where 8049 is the specific *Gadus morhua* TaxID; 18 of June, 2020). Only 53,695 entries showed a 'gene' tag investigable by ExTaxsI. As a unique species, we decided to represent the results obtained from genes survey (Figure 2) and the world map plot (Figure 3). Regarding gene distribution, the most abundant gene is CYTB (with 985 accessions), followed by COI (434) and ND2 (311). The remaining most abundant genes are the other ND portions and Cytochrome Oxidase fragments (COIII and COII), belonging to the mitochondrial genome. These results show the increased effort in sequencing "standard" barcoding markers, while moderately sequencing whole mitochondrial genomes. The remaining genes in the retrieved list and their relative accession frequencies distribution (see the complete list in Additional file 1) demonstrate that the entire genome of this species was sequenced). These results are in line with those obtained by Knudsen and colleagues (2019), where they personally developed specific primers for CYTB amplification, as it is a widely used marker in fish molecular characterization.

Regarding the geographic area, the *Gadidae* family has a circumpolar distribution, comprising species occurring principally in northern and cool seas [28]. Further, as reported by Jorde and colleagues (2018), in Norway we can recognize four distinct stocks of the Atlantic cod: (1) the oceanic Northeast Arctic cod, (2) coastal cod north of 62°N, (3) coastal cod south of 62°N, and (4) a North Sea/Skagerrak stock, the most densely populated region in Norway [28]. This geographic distribution is partly visible

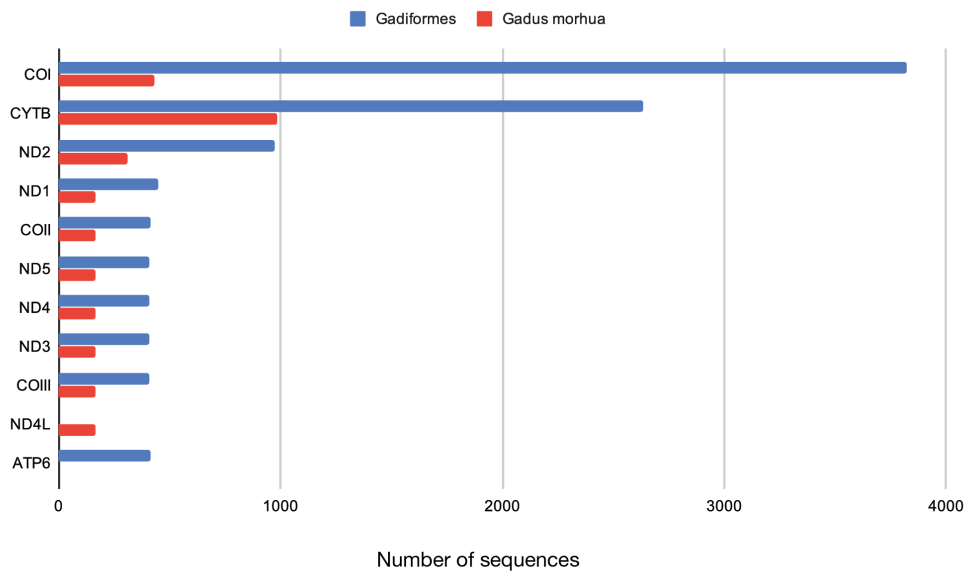



**Figure 2.** Gene distribution of accessions with complete 'gene' tag among *Gadus morhua* and *Gadiformes* taxons.

via the metadata extracted by ExTaxsI, as shown in the world map plot in Figure 3b (Additional file 2).

The second scenario takes as an example the *Gadiformes* Order (phylum: *Chordata*; class: *Actinopterygii*), a major group of organisms belonging to marine fisheries. It includes many important food fishes, variously marketed as cods, hakes, grenadiers, moras, moray cods, pelagic cods, codlets and eucla cods [43]. As a vast group, it comprises more than 500 species, which contribute to more than a quarter of the world's marine fish catch [13, 43].

Via ExTaxsI, this Order was explored using the following query "txid8043[ORGN]", yielding 388,603 accessions (where 8043 is the specific *Gadiformes* TaxID; 22 of June, 2020), where 60,703 showed the 'gene' tag. As a group spread on different taxonomic levels, both taxonomy and gene reports were created. In detail, in order to explore taxa distribution and accessions abundances across the entire Order, the tool created ScatterPlot and SunBurst HTML outputs. In Figure 3a genus abundances are documented in ScatterPlot modality, while SunBurst and entirely interactive plots are available in the Supplementary Material section (Additional files 3 and 4).

As it is shown, *Gadidae* is the most abundant family, considering the number of accessions available. In fact, a total of 380,658 accessions populate this group, followed by *Merlucciidae* (3,196) and *Macrouridae* (1,581) families. These results are in accordance with the literature, a *Gadidae* family is a primary marine, bottom-dwelling family of fishes in the Order of *Gadiformes* with great commercial power [33, 43].

Further, considering the ScatterPlot in Additional file 3, the interactive visualization allowed us to explore the taxonomy distribution among the accessions available, changing dynamically the rank that we want to explore. This feature allows us to disclose that the genus *Gadus* is the most abundant of the entire dataset, highlighting that *Gadus morhua* species composed 94,43% of all the data. This is an expected result, as *Gadus morhua* is documented to be a key species both in the North Atlantic ecosystem and commercial fisheries, with an increasing aquaculture production in

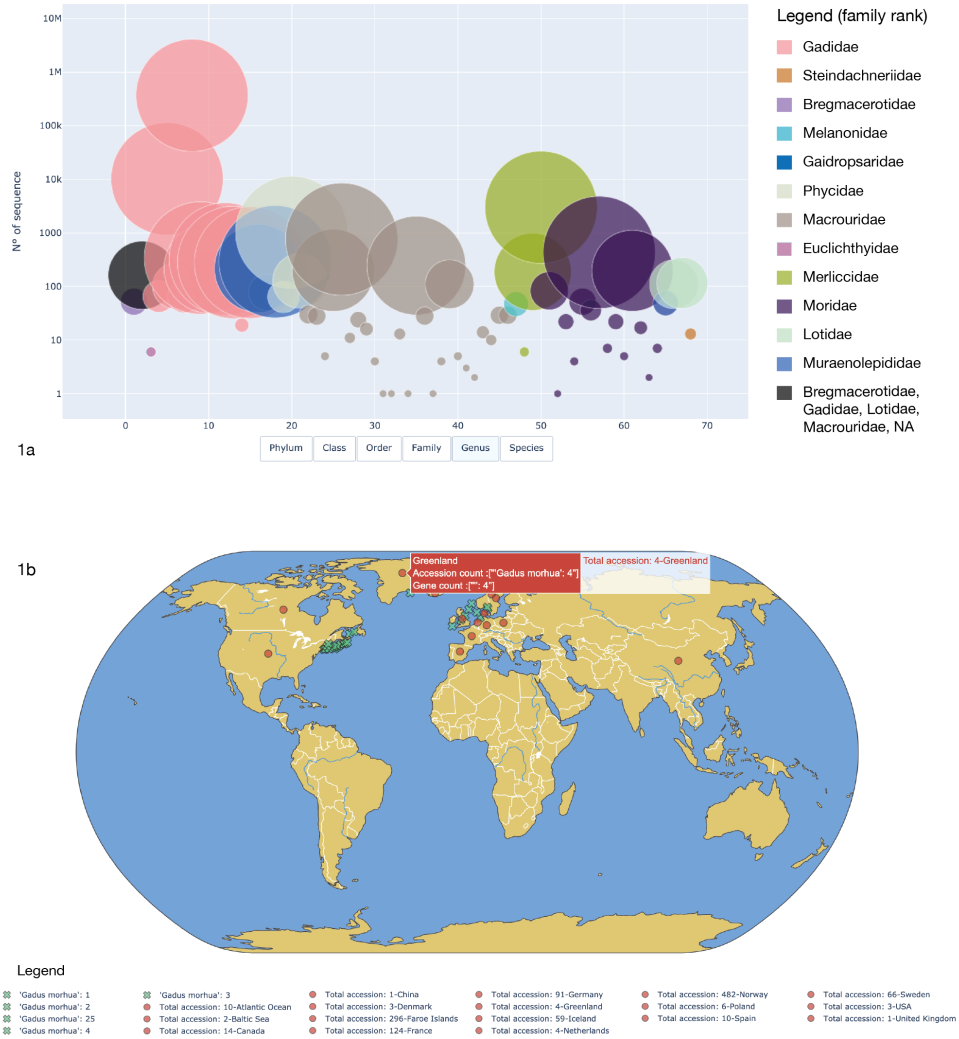



**Figure 3.** 3a) Scatter plot of *Gadiformes* accessions to represent sequence abundances among families; 3b) World map plot of *Gadus morhua* distribution considering geographic metadata extracted from records.

several countries [28]. Considering the genetic information reached by ExTaxsI, a total of 28,839 unique genes were found from the 60,703 completely tagged accessions. A classification of the most ten abundant genes is reported in Figure 2. As shown in the figure, at the first position we can find the COI gene, a widely used marker gene in metabarcoding projects (Knudsen et al. 2019), that deal mainly with animals detection [47], followed by CYTB and ND2 [47].

Concluding with these two case studies, the tool was able to accurately portrait the state of the art of the genetic information available in NCBI. Comparing the most abundant genes found among the records, it is possible to see a thin discrepancy between the two taxa explored (Figure 2), highlighting the disclosures that the survey can report. In general, the detection of mitochondrial genes, coding for Cytochrome Oxidase subunit I (COI) and Cytochrome B (CYTB), is in accordance with the

reliability of these DNA barcodes, principally used in the discrimination of animal species [23, 24, 42]. To date, considering the subjects of our case studies, diverse studies have used COI or CYTB barcoding to identify seafood products and explore broad patterns in fish mislabelling [9, 16, 18, 40, 49, 59, 61].

Regarding the extraction of geographic metadata from NCBI records, the completeness and collection of data can improve drastically the biogeographic and ecological research, allowing not only to explore sampling areas, but also to improve phylogeography investigations, biodiversity monitoring and environmental genomics strategies [12, 47].

The unbalance between the number of records and the number of genes explorable is in some cases due to the incompleteness of the 'gene' tag. In the very recent years genome sequences started playing a key role into public repositories, making sequences available for sharing and reuse. Submission process can be challenging and errors can affect the availability of the data. For this reason, there is a wide interest to integrate standardized procedures into the annotation process [19]. The promotion of FAIR principles and best practices can certainly avoid the error propagation in sequence databases [46, 58], making the data fully explorable in the future.

### 2.1.2 Explore biodiversity data in pandemic outbreak: the case of SARS-CoV-2

The severe acute respiratory syndrome coronavirus 2 (*SARS-CoV-2*) is an enveloped, positive-sense, single-stranded RNA virus that causes coronavirus disease 2019 (COVID-19). RNA and structural proteins are included into virus particles and mediate host cells invasion. After cell infection, RNA encodes structural proteins that make up virus particles. Virus assembly, transcription, replication and host control are mediated by nonstructural proteins [36]. The pandemic linked to *SARS-CoV-2* highlighted hidden virus reservoirs in wild animals and their potential to occasionally spillover into human populations [36]. A detailed understanding of this process is crucial to prevent future spillover events. As reported in the seminal paper of Andersen and colleagues (2020) [2], the risk of future re-emergence events increases if *SARS-CoV-2* pre-adapted in another animal species. *SARS-CoV-2* probably originated from *Rhinolophus affinis* bats, with pangolin (*Manis javanica*) as intermediate host [2]. Recently, other animal species were supposed to be possible intermediate hosts in between bats and humans. To date, ACE2 (Angiotensin-converting enzyme 2), the receptor which binds to the receptor-binding domain (RBD) of *SARS-CoV-2* S protein [35], is reported as crucial in host invasion.

To test our approach and explore the genetic information available in NCBI, we decided to extrapolate information of the ACE2 gene from the *Vertebrata* taxonomic group, with the following query: "txid7742[ORGN] AND ACE2[gene]" (where 7742 is the specific *Vertebrata* TaxID; 28 of June, 2020). The results show that the ACE2 gene is widely distributed throughout *Vertebrata*: we obtained a total of 1,189 accessions, distributed mainly among the *Mammalia* Class, with a high representation in *Actinopteri* and *Aves* groups (Figure 4; Additional files 5 and 6 are provided for an interactive exploration). In details, *Primates*, *Rodentia* and *Chiroptera* are the most represented, with 239, 132 and 108 accessions respectively. Supporting the exploitation of molecular data survey, Luan and colleague (2020) [37] analyzed the affinity to S protein of the 20 key residues in ACE2 from mammal, bird, turtle, and snake, and suggested that *Bovidae* and *Cricetidae* should be included in the screening of intermediate hosts for SARS-CoV-2. In addition, thanks to the analysis of spike glycoprotein sequences from different animals, the study of Dabravolski and Kavalionak [14] suggested that the human *SARS-CoV-2* could also come from yak as an intermediate host. ExTaxsI has the advantage to provide the complete list of taxa, allowing an exhaustive exploratory research. It allows to download all the sequences

available for the query input, generating in turn the input for downstream analyses, such as the calculating of sequence similarities among different taxa. Further, investigating shared features with other species can have important implications for understanding potential natural reservoirs, zoonotic transmission, and human-to-animal transmission. Noteworthy, the survey can give researchers an instrument to download data with a user-friendly approach, exploring interactively the data and program experiments.

Lastly, we explored the data available for *SARS-CoV-2* (Figure 4) using the following query "txid2697049" (where 2697049 is the specific severe acute respiratory syndrome coronavirus 2 TaxID; 29 of June, 2020). We obtained a total of 8,137 accessions. The top ten genes retrieved are shown in Figure 4c. In particular, the number of genes detected is quite similar among the top ten datasets and this is probably due to a high collection of genomes deposited into the database. The three most represented genes in the database are: ORF1AB (7892), followed by two important structural proteins: S (7829), the spike or surface glycoprotein, and N fragments (7817), the nucleocapsid protein. Considering the ORF1AB, several studies demonstrated its pivotal role among coronaviruses [55], providing a clinical target to break down *SARS-CoV-2* infection [31]. Regarding the second and third results, the nucleocapsid phosphoprotein is involved in packaging the RNA into virus particles and protects the viral genome. For these reasons, it has been suggested as an antiviral drug target [20, 60]. The spike glycoprotein, instead, is located outside the virus particle, mediating its attachment and promoting the entry into the host cell. It also gives viruses their crown-like appearance. In the very last research, the S protein was found as an important target for diagnostic antigen-based tests, antibody therapies and vaccine development [45, 53]. The entry of *SARS-CoV-2* is mediated by further processes, for example the activity of the protease TMPRSS2 [25]. Also in this case, the use of ExTaxsI can unearth similar proteases in possible intermediate hosts, revealing new insights into the mechanism of infection.

As also documented in Khailany et al., 2020 [31], the emergent and huge amounts of data collected in the last few months necessitates a large scale exploration of the data. The rapid increment of data releases may give some important insights about *SARS-CoV-2* behaviour in its host species, helping in improving not only our knowledge, but also models to predict COVID-19 outcomes and new drug targets.

# 3 Conclusions and future directions

ExTaxsI provides an easy-to-use standalone tool able to interact with NCBI databases and personal datasets, offering instruments to standardize taxonomy information and visualize vast quantities of data widespread on different taxonomic levels. It also provides interactive visualization plots, easily shareable through HTML formats.

The user-oriented interrogation of NCBI databases may help researchers involved in environmental genomics fields, from phylogeographic studies to DNA metabarcoding surveys, and also in projects related to the human health, as we demonstrated with the *SARS-CoV-2* case study.

With this work, we hope to meet the needs of a vast group of researchers, providing an instrument easy to install on common laptops and directly connected with NCBI databases. In our opinion, ExTaxsI data management ability with its visual interactive exploration can really improve the experimental design phase and the awareness of information available, facilitating the work and incentivizing data exploration and sharing.

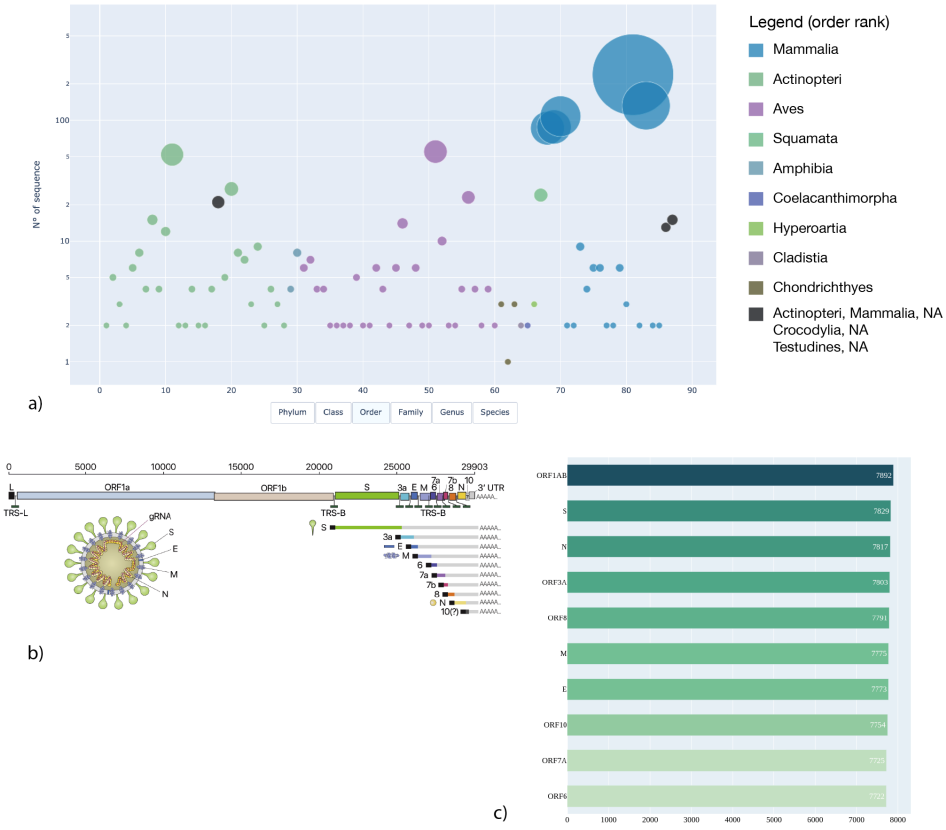



**Figure 4.** 4a) Scatter plot of ACE2 accessions representing sequence abundances among taxa at order level; 4b) *SARS-CoV-2* representation, from [32]; 4c) gene distribution across accessions with complete 'gene' tag of *SARS-CoV-2* data.

# 4 Implementation

No specific system requirements are needed for the installation of ExTaxsI, however for the correct functioning of the software we suggest a minimum of 4GB of RAM. Moreover, to successfully run ExTaxsI, the following python libraries must be installed: NumPy, SciPy, Matplotlib, ipython, Pandas, SymPy, nose, genutils and Plotly, in addition to ETE toolkit [26]. To install all the dependencies compatible versions, we provide a requirement list at the GitHub page ( `https://github.com/qLSLab/extaxsi`), with a detailed guideline to set directly a conda environment.

Regarding the organization of the tool, ExTaxsI is designed in separate modules, albeit interconnected, in order to work directly from different points of its workflow and to allow greater simplicity in the integration of additional modules in the future.

## 4.1 Database creation module

The module 'Database' allows users to create multi FASTA files composed of nucleotide sequences, taxonomic lists, genes names and their related accessions, starting from either a single query or a batch mode using csv/tsv files (Figure 1). After indicating the

type of input, the tool asks, with the exception of the file accession, whether or not the user wants to integrate the query with one or more gene name/s (or other details). This step allows the user to restrict the research in NCBI if needed. In general, the output formats are i) a multi-FASTA file (widely used format for molecular sequences) and ii) text file in TSV format, with two columns composed by the accessions code followed by the taxonomy path of each accession at the six main levels separated by semicolons: phylum, class, order, family, genus and species. When requested by the user, the output file of genes names is in TSV format consisting of a table with two columns, one with the list of genes and the other with the frequency values of the respective genes found in the analyzed records. The tool also provides a summary table containing the most frequent genes from a list of taxid, accessions or organisms. In addition, it is possible to create a barplot with the top ten of this summary table, downloadable as a PNG file.

## 4.2 Visualization and statistical module

The module 'Visualization' allows users to create interactive plots, starting from Database module output or external sources (Additional files 3, 4 and 5) containing taxonomic lists. Before producing the plots, a dialogue box will ask the user to choose a filter value on the data based on the frequency. If the chosen filter value is 0, the tool processes all the data. Otherwise, all the taxonomic units that have not reached the minimum value are inserted into an additional text file, specifically created with a name containing the filter used. The available plots generated by ExTaxsI are i) ScatterPlot (Additional file 3), ii) SunBurst (Additional file 4) and iii) world map plot (Additional file 5). All figures created by the Visualization module can be downloaded as HTML format files. In detail, ScatterPlot uses taxonomy as input to produce a graph that indicates the quantity of each individual taxonomic unit; the interactive plot enables the user to: i) choose the taxonomic level to be displayed using the buttons located under the graph; ii) hover over points to show details, such as the number of records within taxa, names of selected taxa and name of the higher taxon from which they derives. The plot allows also to compare more data on mouse-over, highlight an area of interest with zoom function and view a specific group or remove taxa from the graph. SunBurst, instead, from a taxonomy input creates an Expansion Pie that allows to explore taxonomy by clicking on the taxonomic group of interest and showing the underlying taxa within a new SunBurst. Also in this case, hovering over points shows the number of records within taxa. Regarding world map plot, the initial input is processed in order to obtain geographic data. The tool exploits the 'Country' metadata stored in the NCBI records to produce a map indicating the position of each entry. In this step, based on the type of geographic data obtained, ExTaxsI divides results into two different arrays: i) a specific array of coordinates (if the coordinates are present in the record) or ii) a specific array of states names (if the coordinates are not present in the record). Also external sources can be processed and added to the map. In each map created, coordinates are indicated by green X signs, while States by red circles. Thinking of multiple taxa plotting, each symbol can have a legend that summarizes the data downloaded with the same country or coordinates description. Further, it is possible to see both genes and counts available among the accessions represented.

## 4.3 Taxonomy ID converter module

This module allows to convert TaxID to the main six ranks taxonomy and vice versa (phylum, class, order, family, genus and species); it can convert single manual inputs or multiple inputs from a tsv/csv file complete of a TaxIDs list.

# 5 Availability of source code and requirements

- Project name: ExTaxsI
- Project home page: `https://github.com/qLSLab/extaxsi`
- Operating system(s): Platform independent
- Programming language: Python
- License: GNU GPL version 3

## 5.1 Additional Files

**Additional file 1**: Gene list in TSV format obtained through ExTaxsI for the species *Gadus morhua*. Gene counts were extracted by 366,963 accessions (query: "txid8049[ORGN]"; 18 of June, 2020).

**Additional file 2**: World map plot in HTML format created via ExTaxsI extracting the values of 'Country' tag contained into 366,963 accessions of *Gadus morhua* (query: "txid8049[ORGN]"; 18 of June, 2020). Coordinates are indicated by green X signs, while States by red circles.

**Additional file 3**: ScatterPlot in HTML format created via ExTaxsI extracting the taxonomy of 388,603 accessions of *Gadiformes* Order (txid8043[ORGN]"; 22 of June, 2020).

**Additional file 4**: SunBurst Plot in HTML format created via ExTaxsI extracting the taxonomy of 388,603 accessions of *Gadiformes* Order (txid8043[ORGN]"; 22 of June, 2020).

**Additional file 5**: ScatterPlot in HTML format created via ExTaxsI extracting the taxonomy related to 1,189 accessions of ACE2 genes belonging to the *Vertebrata* taxonomic group (query: "txid7742[ORGN] AND ACE2[gene]"; 28 of June, 2020).

**Additional file 6**: SunBurst Plot in HTML format created via ExTaxsI extracting the taxonomy related to 1,189 accessions of ACE2 genes belonging to the *Vertebrata* taxonomic group (query: "txid7742[ORGN] AND ACE2[gene]"; 28 of June, 2020).

# 6 Declarations

## 6.1 List of abbreviations

SILVA: High quality ribosomal RNA databases; BOLD: Barcode of Life Data System; UNITE: Database and sequence management environment centered on the eukaryotic nuclear ribosomal ITS region; ETE: Environment for Tree Exploration; QIIME2: Quantitative Insights Into Microbial Ecology; FASTA: Text-based format for representing either nucleotide sequences or peptide sequences; TAXID: Taxonomy ID; HTML: Hyper-Text Markup Language; COI: Cytochrome Oxidase I; COI: Cytochrome Oxidase II; COIII: Cytochrome Oxidase III; CYTB: Cytochrome B; ND2: NADH dehydrogenase 2; ACE2: Angiotensin-Converting enzyme 2; RBD: Receptor-Binding Domain; PNG: Portable Network Graphics; NCBI: National Center for Biotechnology Information; ENA: European Nucleotide Archive

## 6.2 Author's Contributions

Giulia Agostinetto: Conceptualization, Investigation, Software development, Visualization, Original Draft Preparation, Review & Editing. Anna Sandionigi: Conceptualization, Original Draft Preparation, Review & Editing, Supervision, Project

Administration. Adam Chahed: Software development, Visualization. Alberto Brusati: Investigation, Software development, Visualization, Review & Editing. Elena Parladori: Software development, Visualization. Bachir Balech: Review & Editing, Validation. Antonia Bruno: Review & Editing, Validation. Dario Pescini: Review & Editing, Supervision. Maurizio Casiraghi: Funding Acquisition, Supervision. All authors read and approved the final manuscript, contributing critically important comments.

## 7 Acknowledgements

Many thanks are due to ELIXIR Biodiversity community for all the support.

## References

1. A. Almeida, A. L. Mitchell, M. Boland, S. C. Forster, G. B. Gloor, A. Tarkowska, T. D. Lawley, and R. D. Finn. A new genomic blueprint of the human gut microbiota. *Nature*, 568(7753):499–504, 2019.

2. K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, and R. F. Garry. The proximal origin of sars-cov-2. *Nature medicine*, 26(4):450–452, 2020.

3. M. J. Ankenbrand, A. Keller, M. Wolf, J. Schultz, and F. Förster. Its2 database v: Twice as much. *Molecular Biology and Evolution*, 32(11):3030–3032, 2015.

4. J. Bengtsson-Palme, M. Hartmann, K. M. Eriksson, C. Pal, K. Thorell, D. G. J. Larsson, and R. H. Nilsson. Metaxa2: improved identification and taxonomic classification of small and large subunit rrna in metagenomic data. *Molecular ecology resources*, 15(6):1403–1414, 2015.

5. D. Benson, I. Karsch-Mizrachi, D. Lipman, J. Ostell, and D. Wheeler. Genbank nucleic acids res. *jan*, 1:33, 2008.

6. N. Blomberg and K. B. Lauer. Connecting data, tools and people across europe: Elixir's response to the covid-19 pandemic. *European Journal of Human Genetics*, pages 1–5, 2020.

7. E. Bolyen, J. R. Rideout, M. R. Dillon, N. A. Bokulich, C. C. Abnet, G. A. Al-Ghalith, H. Alexander, E. J. Alm, M. Arumugam, F. Asnicar, et al. Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. *Nature biotechnology*, 37(8):852–857, 2019.

8. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. Blast+: architecture and applications. *BMC bioinformatics*, 10(1):421, 2009.

9. E. Cline. Marketplace substitution of atlantic salmon for pacific salmon in washington state detected by dna barcoding. *Food Research International*, 45(1):388–393, 2012.

10. P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

11. N. R. Coordinators. Database resources of the national center for biotechnology information. *Nucleic acids research*, 42(D1):D7–D17, 2014.

12. T. Cordier, L. Alonso-Sáez, L. Apothéloz-Perret-Gentil, E. Aylagas, D. A. Bohan, A. Bouchez, A. Chariton, S. Creer, L. Frühe, F. Keck, et al. Ecosystems monitoring powered by environmental genomics: a review of current strategies with an implementation roadmap. *Molecular Ecology*, 2020.

13. M. J. Costello, P. Bouchet, G. Boxshall, K. Fauchald, D. Gordon, B. W. Hoeksema, G. C. Poore, R. W. van Soest, S. Stöhr, T. C. Walter, et al. Global coordination and standardisation in marine biodiversity through the world register of marine species (worms) and related databases. *PloS one*, 8(1):e51629, 2013.

14. S. A. Dabravolski and Y. K. Kavalionak. Sars-cov-2: Structural diversity, phylogeny, and potential animal host identification of spike glycoprotein. *Journal of medical virology*, 2020.

15. K. Deiner, H. M. Bik, E. Mächler, M. Seymour, A. Lacoursière-Roussel, F. Altermatt, S. Creer, I. Bista, D. M. Lodge, N. De Vere, et al. Environmental dna metabarcoding: Transforming how we survey animal and plant communities. *Molecular ecology*, 26(21):5872–5895, 2017.

16. A. Di Pinto, P. Di Pinto, V. Terio, G. Bozzo, E. Bonerba, E. Ceci, and G. Tantillo. Dna barcoding for detecting market substitution in salted cod fillets and battered cod chunks. *Food chemistry*, 141(3):1757–1762, 2013.

17. S. Federhen. The ncbi taxonomy database. *Nucleic acids research*, 40(D1):D136–D143, 2012.

18. T. J. Fernandes, J. Costa, M. B. P. Oliveira, and I. Mafra. Dna barcoding coupled to hrm analysis as a new and simple tool for the authentication of gadidae fish species. *Food Chemistry*, 230:49–57, 2017.

19. S. M. Geib, B. Hall, T. Derego, F. T. Bremer, K. Cannoles, and S. B. Sim. Genome annotation generator: a simple tool for generating and correcting wgs annotation tables for ncbi submission. *GigaScience*, 7(4):giy018, 2018.

20. D. E. Gordon, G. M. Jang, M. Bouhaddou, J. Xu, K. Obernier, K. M. White, M. J. O'Meara, V. V. Rezelj, J. Z. Guo, D. L. Swaney, et al. A sars-cov-2 protein interaction map reveals targets for drug repurposing. *Nature*, pages 1–13, 2020.

21. S. E. Hampton, M. B. Jones, L. A. Wasser, M. P. Schildhauer, S. R. Supp, J. Brun, R. R. Hernandez, C. Boettiger, S. L. Collins, L. J. Gross, et al. Skills and knowledge for data-intensive environmental research. *BioScience*, 67(6):546–557, 2017.

22. A. Hardisty, D. Roberts, et al. A decadal view of biodiversity informatics: challenges and priorities. *BMC ecology*, 13(1):16, 2013.

23. P. D. Hebert, S. Ratnasingham, and J. R. De Waard. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(suppl_1):S96–S99, 2003.

24. R. S. Hellberg, M. D. Kawalek, K. T. Van, Y. Shen, and D. M. Williams-Hill. Comparison of dna extraction and pcr setup methods for use in high-throughput dna barcoding of fish species. *Food analytical methods*, 7(10):1950–1959, 2014.

25. M. Hoffmann, H. Kleine-Weber, S. Schroeder, N. Krüger, T. Herrler, S. Erichsen, T. S. Schiergens, G. Herrler, N.-H. Wu, A. Nitsche, et al. Sars-cov-2 cell entry depends on ace2 and tmprss2 and is blocked by a clinically proven protease inhibitor. *Cell*, 2020.

26. J. Huerta-Cepas, F. Serra, and P. Bork. Ete 3: reconstruction, analysis, and visualization of phylogenomic data. *Molecular biology and evolution*, 33(6):1635–1638, 2016.

27. S. D. Johansen, D. H. Coucheron, M. Andreassen, B. O. Karlsen, T. Furmanek, T. E. Jørgensen, Å. Emblem, R. Breines, J. T. Nordeide, T. Moum, et al. Large-scale sequence analyses of atlantic cod. *New Biotechnology*, 25(5):263–271, 2009.

28. P. E. Jorde, A. R. Kleiven, M. Sodeland, E. M. Olsen, K. Ferter, S. Jentoft, and H. Knutsen. Who is fishing on what stock: population-of-origin of individual cod (gadus morhua) in commercial and recreational fisheries. *ICES Journal of Marine Science*, 75(6):2153–2162, 2018.

29. P. Kaur, F. Klan, and B. König-Ries. Issues and suggestions for the development of a biodiversity data visualization support tool. In *EuroVis (Short Papers)*, pages 73–77, 2018.

30. A. Keller, S. Hohlfeld, A. Kolter, J. Schultz, B. Gemeinholzer, and M. J. Ankenbrand. Bcdatabaser: on-the-fly reference database creation for (meta-) barcoding. *Bioinformatics*, 36(8):2630–2631, 2020.

31. R. A. Khailany, M. Safdar, and M. Ozaslan. Genomic characterization of a novel sars-cov-2. *Gene reports*, page 100682, 2020.

32. D. Kim, J.-Y. Lee, J.-S. Yang, J. W. Kim, V. N. Kim, and H. Chang. The architecture of sars-cov-2 transcriptome. *Cell*, 2020.

33. S. W. Knudsen, R. B. Ebert, M. Hesselsøe, F. Kuntke, J. Hassingboe, P. B. Mortensen, P. F. Thomsen, E. E. Sigsgaard, B. K. Hansen, E. E. Nielsen, et al. Species-specific detection and quantification of environmental dna from marine fishes in the baltic sea. *Journal of experimental marine biology and ecology*, 510:31–45, 2019.

34. M. Kurlansky and R. M. Davidson. *Cod: a Biography of the Fish that Changed the world*. Phoenix Books, 2006.

35. M. Letko, A. Marzi, and V. Munster. Functional assessment of cell entry and receptor usage for sars-cov-2 and other lineage b betacoronaviruses. *Nature microbiology*, 5(4):562–569, 2020.

36. R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet*, 395(10224):565–574, 2020.

37. J. Luan, X. Jin, Y. Lu, and L. Zhang. Sars-cov-2 spike protein favors ace2 from bovidae and cricetidae. *Journal of medical virology*, 2020.

38. F. Mahé, T. Rognes, C. Quince, C. de Vargas, and M. Dunthorn. Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ*, 3:e1420, 2015.

39. W. K. Michener and M. B. Jones. Ecoinformatics: supporting ecology as a data-intensive science. *Trends in ecology & evolution*, 27(2):85–93, 2012.

40. D. D. Miller and S. Mariani. Smoke, mirrors, and mislabeled cod: poor transparency in the european seafood industry. *Frontiers in Ecology and the Environment*, 8(10):517–521, 2010.

41. A. L. Mitchell, A. Almeida, M. Beracochea, M. Boland, J. Burgin, G. Cochrane, M. R. Crusoe, V. Kale, S. C. Potter, L. J. Richardson, et al. Mgnify: the microbiome analysis resource in 2020. *Nucleic acids research*, 48(D1):D570–D578, 2020.

42. S. Mueller, S. M. Handy, J. R. Deeds, G. O. George, W. J. Broadhead, S. E. Pugh, and S. D. Garrett. Development of a cox1 based pcr-rflp method for fish species identification. *Food Control*, 55:39–42, 2015.

43. J. S. Nelson, T. C. Grande, and M. V. Wilson. *Fishes of the World*. John Wiley & Sons, 2016.

44. R. H. Nilsson, K.-H. Larsson, A. F. S. Taylor, J. Bengtsson-Palme, T. S. Jeppesen, D. Schigel, P. Kennedy, K. Picard, F. O. Glöckner, L. Tedersoo, et al. The unite database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic acids research*, 47(D1):D259–D264, 2019.

45. T. S. Pillay. Gene of the month: the 2019-ncov/sars-cov-2 novel coronavirus spike protein. *Journal of Clinical Pathology*, 2020.

46. W. Pirovano, M. Boetzer, M. F. Derks, and S. Smit. Ncbi-compliant genome submissions: tips and tricks to save time and money. *Briefings in Bioinformatics*, 18(2):179–182, 2017.

47. T. M. Porter and M. Hajibabaei. Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Molecular ecology*, 27(2):313–338, 2018.

48. E. Pruesse, C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies, and F. O. Glöckner. Silva: a comprehensive online resource for quality checked and aligned ribosomal rna sequence data compatible with arb. *Nucleic acids research*, 35(21):7188–7196, 2007.

49. R. S. Rasmussen and M. T. Morrissey. Dna-based methods for the identification of commercial fish and seafood species. *Comprehensive reviews in food science and food safety*, 7(3):280–295, 2008.

50. S. Ratnasingham and P. D. Hebert. Bold: The barcode of life data system (http://www. barcodinglife. org). *Molecular ecology notes*, 7(3):355–364, 2007.

51. T. Rognes, T. Flouri, B. Nichols, C. Quince, and F. Mahé. Vsearch: a versatile open source tool for metagenomics. *PeerJ*, 4:e2584, 2016.

52. K. M. Ruppert, R. J. Kline, and M. S. Rahman. Past, present, and future perspectives of environmental dna (edna) metabarcoding: A systematic review in methods, monitoring, and applications of global edna. *Global Ecology and Conservation*, 17:e00547, 2019.

53. G. Salvatori, L. Luberto, M. Maffei, L. Aurisicchio, G. Roscilli, F. Palombo, and E. Marra. Sars-cov-2 spike protein: an optimal immunological target for vaccines. *Journal of Translational Medicine*, 18:1–3, 2020.

54. B. Star, A. J. Nederbragt, S. Jentoft, U. Grimholt, M. Malmstrøm, T. F. Gregers, T. B. Rounge, J. Paulsen, M. H. Solbakken, A. Sharma, et al. The genome sequence of atlantic cod reveals a unique immune system. *Nature*, 477(7363):207–210, 2011.

55. Y. Wan, J. Shang, R. Graham, R. S. Baric, and F. Li. Receptor recognition by the novel coronavirus from wuhan: an analysis based on decade-long structural studies of sars coronavirus. *Journal of virology*, 94(7), 2020.

56. Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole. Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, 73(16):5261–5267, 2007.

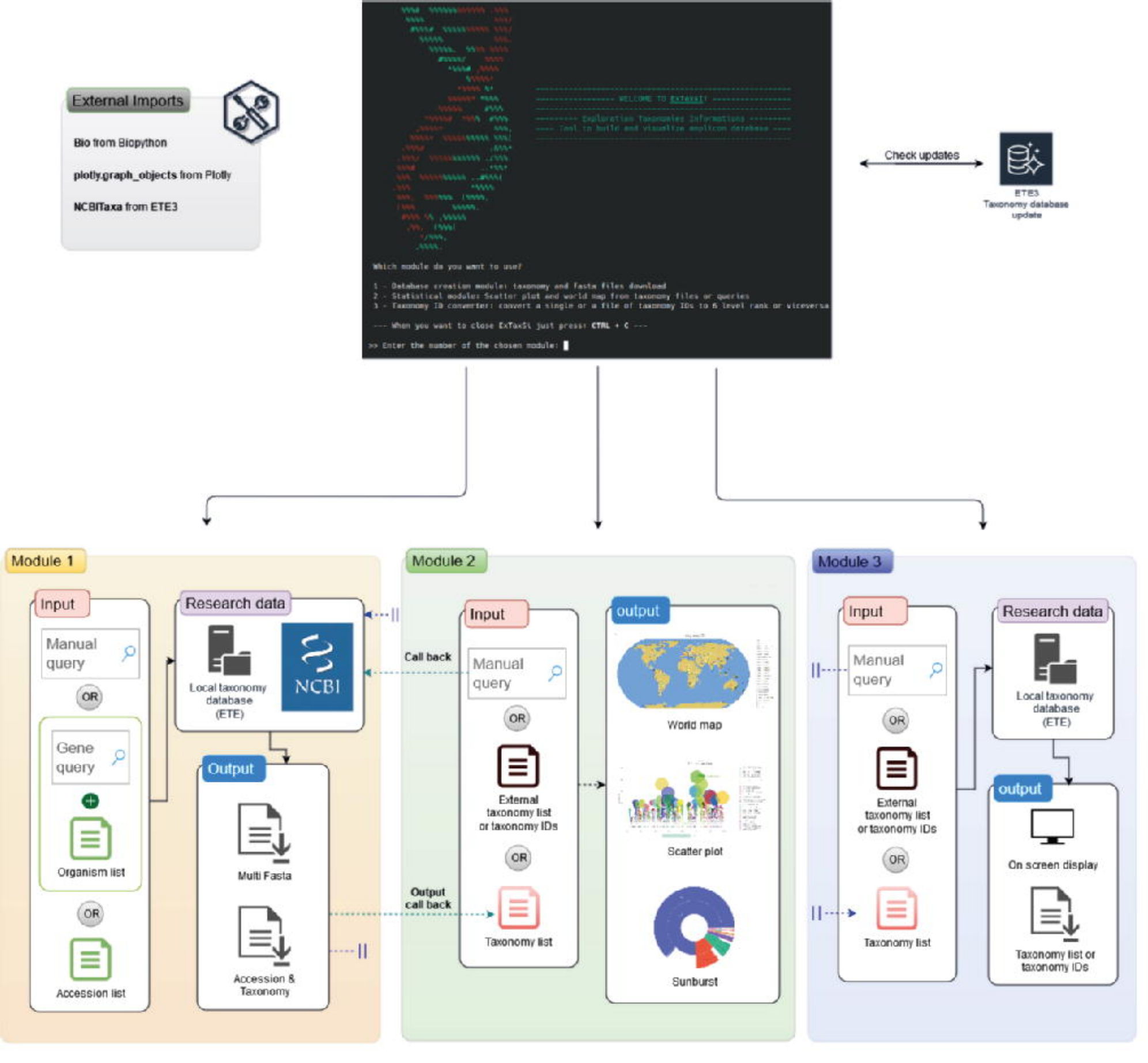
57. E. P. White, E. Baldridge, Z. T. Brym, K. J. Locey, D. J. McGlinn, and S. R. Supp. Nine simple ways to make it easier to (re) use your data. *Ideas in Ecology and Evolution*, 6(2), 2013.

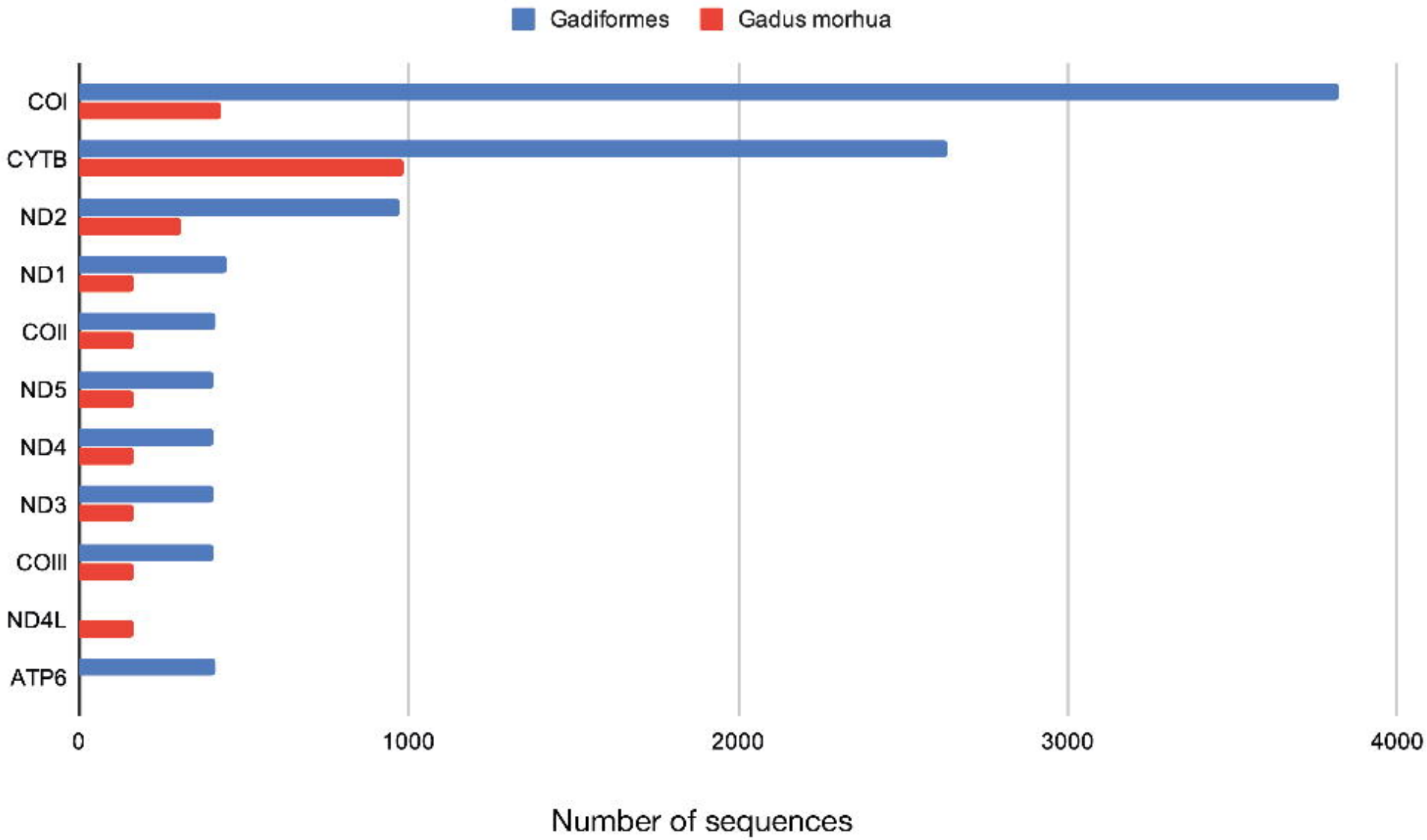
58. M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

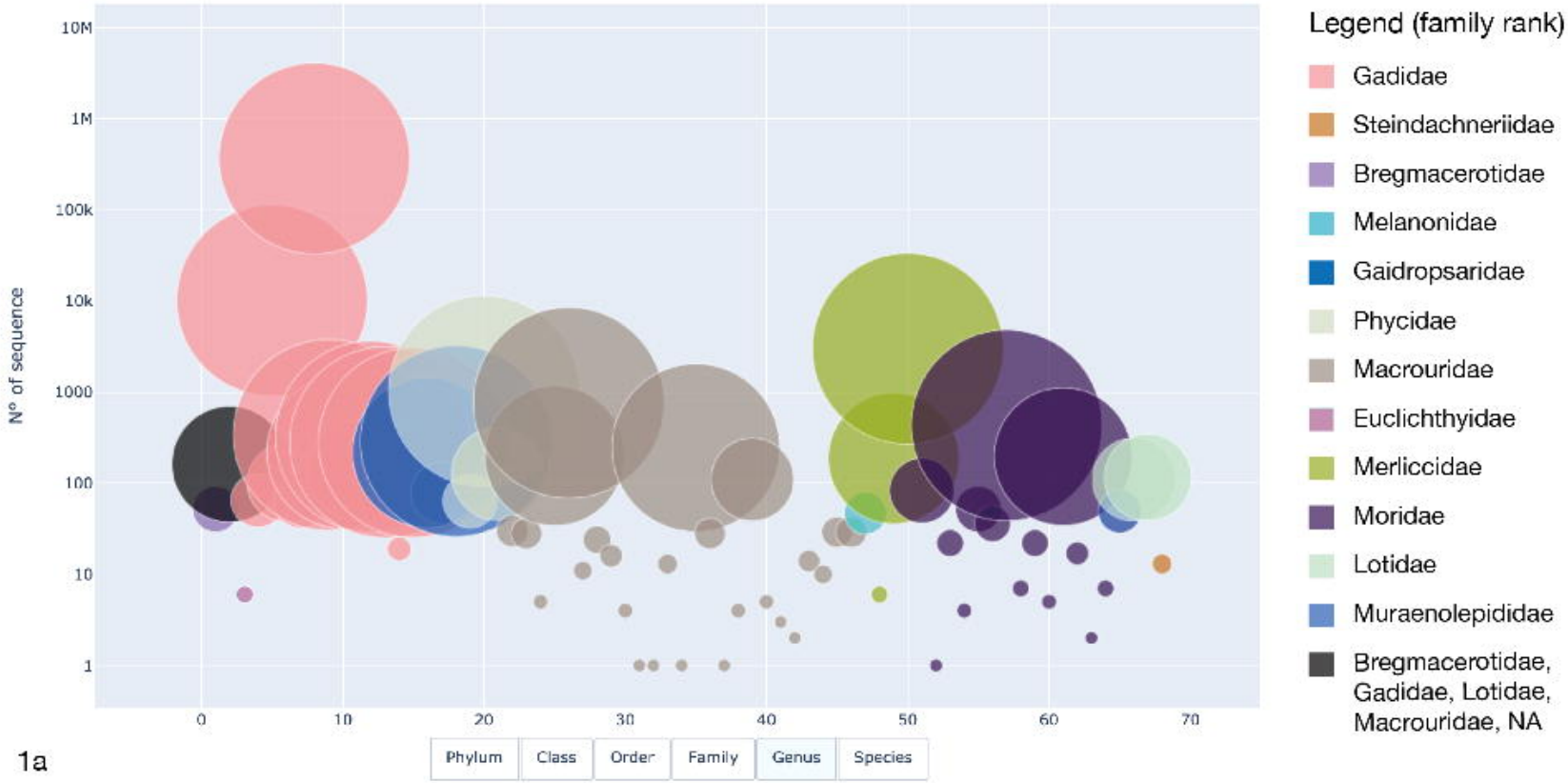
59. E. H.-K. Wong and R. H. Hanner. Dna barcoding detects market substitution in north american seafood. *Food Research International*, 41(8):828–837, 2008.

60. F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, et al. A new coronavirus associated with human respiratory disease in china. *Nature*, 579(7798):265–269, 2020.

61. H. F. Yancy, T. S. Zemlak, J. A. Mason, J. D. Washington, B. J. Tenge, N.-L. T. Nguyen, J. D. Barnett, W. E. Savary, W. E. Hill, M. M. Moore, et al. Potential use of dna barcodes in regulatory science: applications of the regulatory fish encyclopedia. *Journal of Food Protection*, 71(1):210–217, 2008.

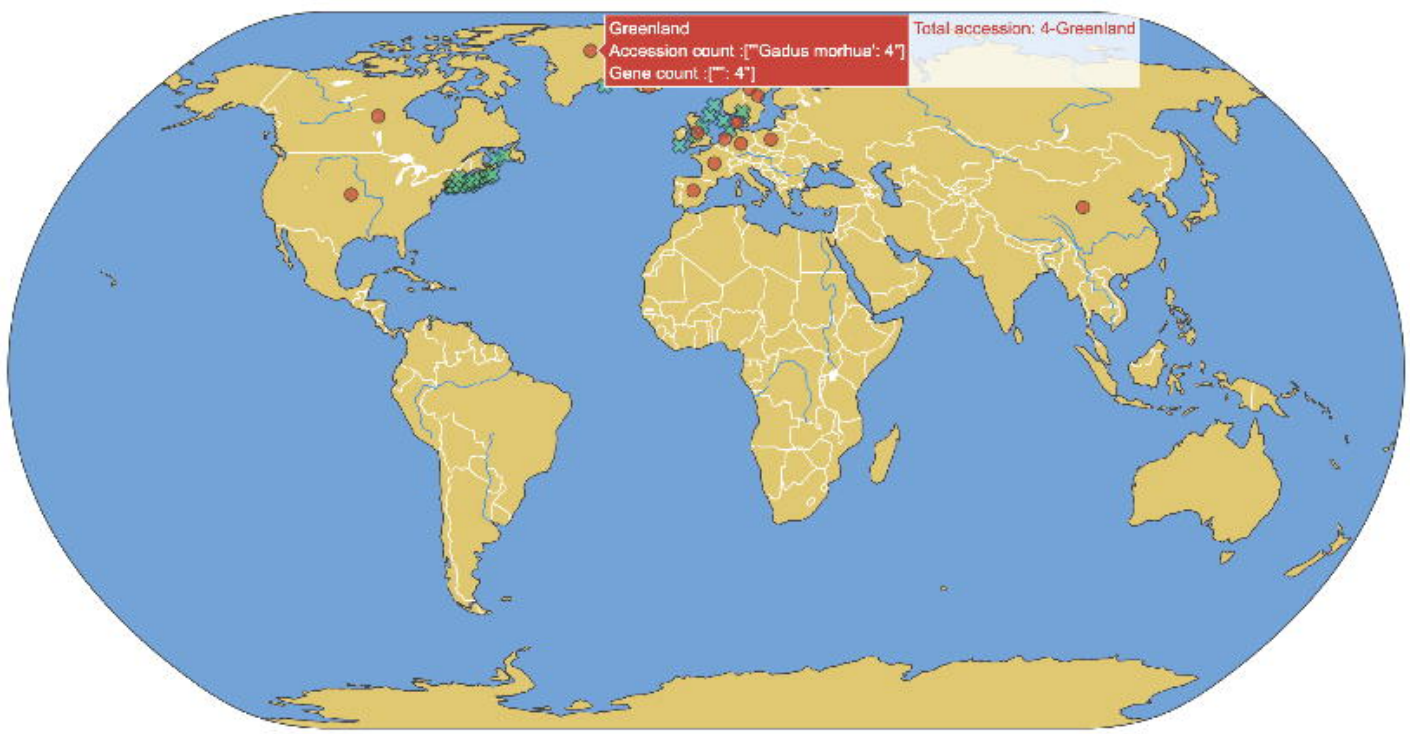

External Imports
Bio from Biopython
plotly.graph_objects from Plotly
NCBITaxa from ETE3
Which module do you want to use?
1 - Database creation module: taxonomy and fasta files download
2 - Statistical module: Scatter plot and world map from taxonomy files or queries
3 - Taxonomy ID converter: convert a single or a file of taxonomy IDs to 6 level rank or viceversa
--- When you want to close ExTaxSi just press: CTRL + C ---
>> Enter the number of the chosen module:
Check updates
ETE3
Taxonomy database
update
Module 1
Input
Manual query
OR
Gene query
Organism list
OR
Accession list
Research data
Local taxonomy database (ETE)
NCBI
Output
Multi Fasta
Accession & Taxonomy
Module 2
Input
Call back
Manual query
OR
External taxonomy list or taxonomy IDs
OR
Taxonomy list
Output call back
output
World map
Scatter plot
Sunburst
Module 3
Input
Manual query
OR
External taxonomy list or taxonomy IDs
OR
Taxonomy list
Research data
Local taxonomy database (ETE)
output
On screen display
Taxonomy list or taxonomy IDs


Gadiformes
Gadus morhua
COI
CYTB
ND2
ND1
COII
ND5
ND4
ND3
COIII
ND4L
ATP6
0
1000
2000
3000
4000
Number of sequences



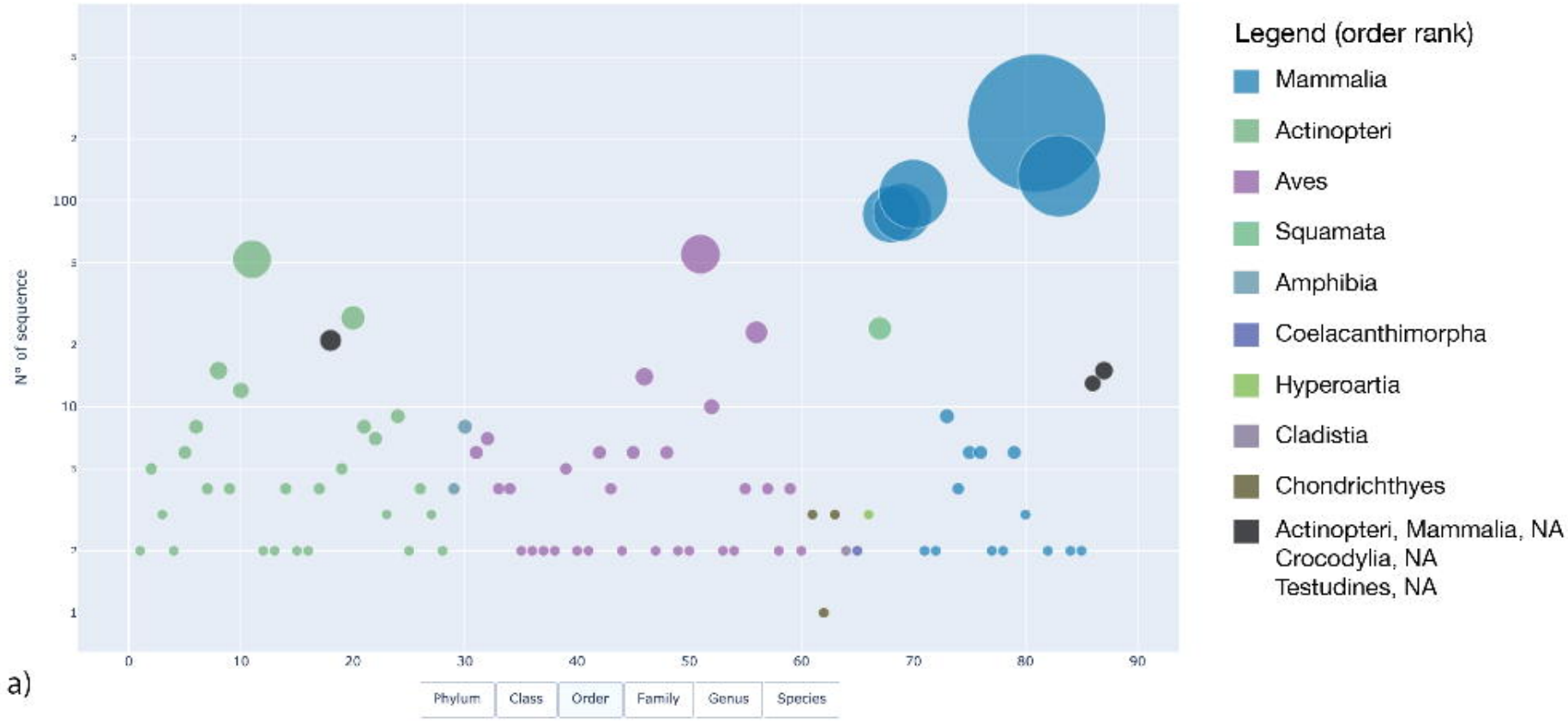

1a

1b

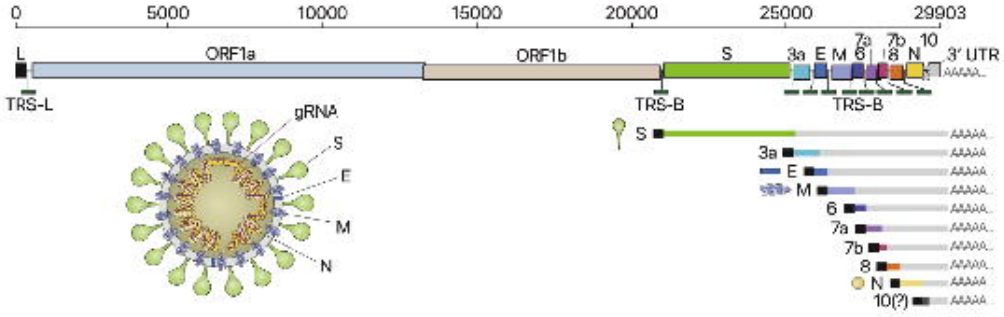



Legend

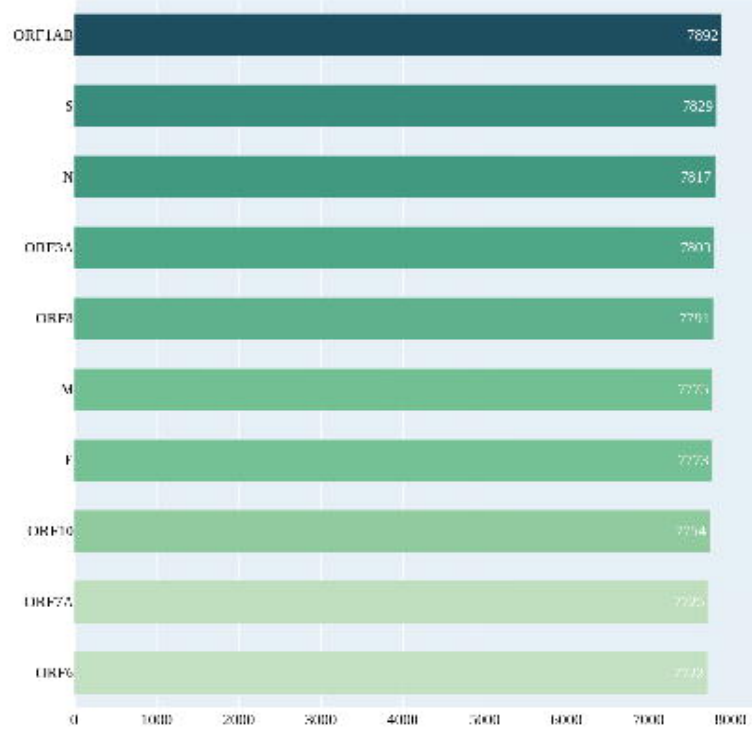
'Gadus morhua': 1
'Gadus morhua': 2
'Gadus morhua': 25
'Gadus morhua': 4
'Gadus morhua': 3
Total accession: 10-Atlantic Ocean
Total accession: 2-Baltic Sea
Total accession: 14-Canada
Total accession: 1-China
Total accession: 3-Denmark
Total accession: 296-Faroe Islands
Total accession: 124-France
Total accession: 91-Germany
Total accession: 4-Greenland
Total accession: 59-Iceland
Total accession: 4-Netherlands
Total accession: 482-Norway
Total accession: 6-Poland
Total accession: 10-Spain
Total accession: 66-Sweden
Total accession: 3-USA
Total accession: 1-United Kingdom


Nº of sequence
0
10
20
30
40
50
60
70
80
90
Phylum
Class
Order
Family
Genus
Species
Legend (order rank)
Mammalia
Actinopteri
Aves
Squamata
Amphibia
Coelacanthimorpha
Hyperoartia
Cladistia
Chondrichthyes
Actinopteri, Mammalia, NA
Crocodylia, NA
Testudines, NA

a)


0
5000
10000
15000
20000
25000
29903
L
ORF1a
ORF1b
S
3a
E
M
6
7a
7b
8
N
10
3' UTR
TRS-L
TRS-B
TRS-B
gRNA
S
E
M
N
S
3a
E
M
6
7a
7b
8
N
10(?)

b)


7892
7829
7817

c)