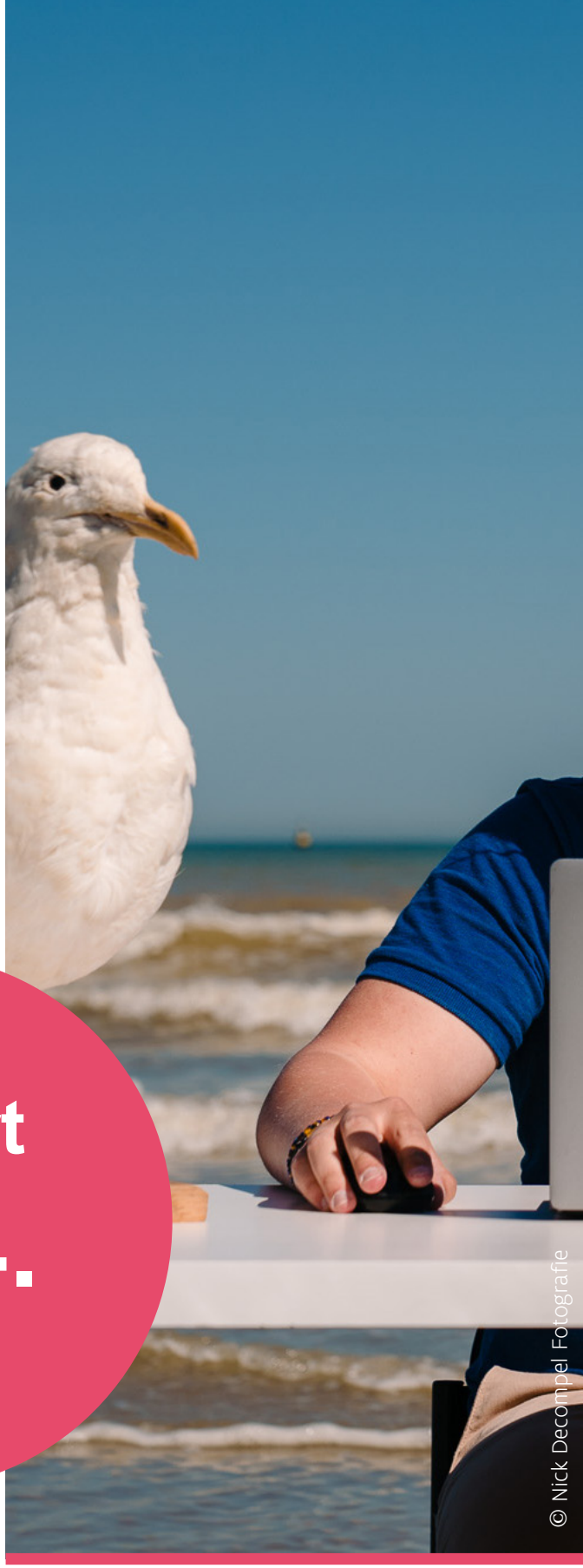
WP4 REPORT ON SURVEY OF AVAILABLE (ONLINE) TOOLS AND WORKFLOWS

Report
WP4.



Interreg North Sea Region
GEANS
European Regional Development Fund

EUROPEAN UNION

Genetic tools for Ecosystem health Assessment in the North Sea region

WP4 Report on survey of available (online) tools and work-flows

Version: 1.0

Project:
GEANS– Genetic tools for Ecosystem health Assessment in the North Sea region

Authors:
Pascal I. Hablützel, Rune Lagaisse, Marie-Catherine Bouquieaux

Date:
2023-04-07

Lead Partner:
VLIZ

Funding:
Funding for this research was received through the North Sea Program of the European Regional Development Fund of the European Union.

Main contact:
Pascal Hablützel, pascal.hablutzel@vliz.be

How to cite this report?
Hablützel P. I., Lagaisse R. & Bouquieaux M.-C. (2023)
WP4 Report on survey of available (online) tools and workflows. GEANS, Ostend, Belgium.

Table of Contents

- Summary

DNA sequences must undergo a suite of bioinformatic analyses and conversion into biologically relevant outcomes, necessitating the utilization of computational techniques to process the numerous raw DNA sequences produced through DNA metabarcoding. As part of the GEANS project extension, our objective was to assist stakeholders in their quest for a standardized analysis method. To achieve this, we compiled a comprehensive list of appropriate analytical approaches currently employed by both the GEANS consortium and external sources. Through conducting a questionnaire of accessible (online) tools and workflows, we generated a report that highlights pressing issues in computational analysis of environmental DNA data.

To get a view of bioinformatic workflows used within and outside of the GEANS consortium, we launched two surveys on bioinformatics and biodiversity informatics with the aim of identifying steps in workflows where experts are struggling with, that are too time consuming or that are prone to errors. The surveys concerned data pipelines on metabarcoding and metagenomics. The first section targeted processing of raw data, where we inquired about the techniques and sequencing platforms used and the aims to get an overall view. After we asked participants to describe their current preferred pipeline and the main motivation for using it, what experience is needed, the type of output it produces, where parameter settings are defined in the bioinformatic pipeline, if it includes provenance tracking, what the most time consuming step is and what pitfalls are experienced. The second section of the surveys covered data analysis, where we again first tried to get a general view of what the user is interested in, how taxonomy is assigned and how accuracy is assured and what reference database is used. The last section of the surveys was on archiving data, where we asked what long-term repositories are used to store data and whether these are public, if participants quality check data before archiving, if standardized nomenclature is followed, how metadata is published, what sequence data are published (Amplicon Sequence Variants, ASVs, or Operational Taxonomic Units, OTUs), where script/code and parameter settings and version numbers are published, under what license data and code are published, and if data reuse is prevalent.

From the survey results we gathered that automatization of pipelines and provenance tracking could be improved by the use of input and output files for scripts and that ASV and OTUs are often not published in data aggregators, restricting their accessibility and usability for broader scientific investigations, facilitating more robust and comprehensive analyses. To further standardize bioinformatic workflows and make these pipelines and their output more FAIR (Findable, Accessible, Interoperable and Reusable), we build a custom R tool that standardizes DADA2 output, as this pipeline is used by multiple GEANS partners, in Darwin Core Archive format. The tool standardizes metadata input and output of the pipeline as well as provenance tracking while reformatting data in standardized format for long-term archiving in biodiversity repositories, like OBIS or GBIF. The tool is freely available on GitHub, including a dummy dataset, and can be easily adapted to fit to alternative metabarcoding pipelines. Because of a lack of agreement among experts on which fields should be required and/or highly recommended we could only incorporate a number of fields in the tool. Another obstacle is that there is currently no field that can host the bootstrapping values of the Bayesian classifier of DADA2 limiting the provenance tracking. So, while the tool is a good step to making bioinformatic genetic analysis more standardized and FAIR, there are still a number of issues that need to be addressed in the field.

- Survey within the GEANS consortium

We obtained eight responses from the following institutes: the Naturalis Biodiversity Center (The Netherlands), the Institute for Agriculture and Fisheries Research (Belgium), Aarhus University (Denmark), Wageningen University & Research (The Netherlands) and the Senckenberg Society for Nature Research (Germany).

- PROCESSING

The most frequently used technique among participants was metabarcoding (8), followed by quantitative PCR (qPCR, 5) and metagenomics (2) and to lesser extent metatranscriptomics (1). The main results aimed for are species lists (8), biodiversity indices (7), ecological indicators (6), and to a lesser extent impact assessments (1), quantification of species (1) and phylogenomics (1). The most commonly used sequencing platform was Illumina (7), followed by Oxford Nanopore (3).

Participants were then asked to describe their preferred pipeline for bioinformatic analysis, whether it is publicly available and where the analysis is run. Majority of participants used DADA2 (Callahan et al., 2016) with or without individual adjustments (6), either run locally, on a server or on a supercomputer. One participant used Galaxy with both pre-existing and custom made tools and run on a local cloud. One participant used QIIME2 (Bolyen et al., 2019, VSEARCH (Rognes et al. 2016), and a DADA2 plugin for QIIME run on a server. One participant used a Decona pipeline developed in-house (https://github.com/Saskia-Oosterbroek/decona) and run locally or on HPC. In a follow-up question participants were asked about their main motivation for using the described workflow. Three participants indicated that they preferred to work with ASVs as opposed to implementing similarity thresholds for OTUs, another participant said they preferred the use of the learnErrors method, the Bayesian classifier and the retained quality profile of DADA2, one participant mentioned it is convenient to run code in a language they are familiar with, while another one mentioned a preference for a graphical user interface (GUI) for their preferred pipeline (hence omitting the need for command line code), the last participant mentioned 'open access, the community to help with' as a reason for preference. The majority of the mentioned workflows require coding experience (6), a few required setting up computing environments (3), one workflow had a Graphical User Interface (GUI) and one workflow required a single line of code to be adapted to the specifics of the data to be analyzed. Main coding languages required for the workflows were R (5) and shell scripting (5), and to a lesser extent Python (1) and C++ (1). Main outputs of the mentioned workflows are taxonomically annotated sequences (8) and abundance tables (8), followed by ASV tables (7), OUT tables (2), consensus sequences (2) and .fasta files from ASV sequences sorted into higher taxonomic level. Participants set parameter settings either manually in scripts (62.5 %), followed by manual entry in the GUI (1), preconfigured code with default parameters (1) or using separate configuration or input files (1). Average run time of a workflow takes around a few hours to a full day, depending on the amount of data. According to users, the most time-consuming step in the workflows was running the analyses (6), transforming input and output data (2) and quality control (2), adjusting the code for new study (1), documenting the analysis (1) and error determining (1). When asked where they think they could save time, participants mentioned adjusting the code or making the code more automated (4), improving taxonomic assignment to reduce time spent on quality control (1) and correcting controls and combining replicates (1). The main issues and pitfalls people

are experiencing are related to reference databases and taxonomic assignment (3), including the World Register of Marine Species (WoRMS, WoRMS Editorial Board, 2023) not recognizing species matches from National Centre for Biotechnology Information (NCBI) better annotation from Decona data output (1) and getting DADA2 to work in the Galaxy environment (1).

- **DATA ANALYSIS**

Participants indicated to be mainly interested in biodiversity analysis (8), ecological interpretation (7) and presence/absence of species (7). 50 % of participants work with ASVs, with 12.5 % of participants switching from OUTs to ASV at the moment of the survey, and 37.5 % participants work with species names.

All participants used the GEANS reference database (8), followed by NCBI (6), BoLD (6), Silva (Yilmaz et al., 2014)(4), MIDORI (Machida et al., 2017) (4), custom databases (4), PR2 (Guillou et al., 2013) (2) and UNITE, R-Syst (Nilsson et al., 2016) (1). When asked how they assure accuracy of taxonomic assignment, most participants mentioned they use Bayesian or maximum likelihood approaches with or without bootstrapping (5), similarity thresholds (of either 97, 98, or 99 %) (3), grade of percentage identity and query cover (1), or a least common ancestor based algorithm in MEGAN (Huson et al., 2007) (1). All people use R for statistical analysis, one participant also uses Python.

- **ARCHIVING**

The majority of participants publish their data on NCBI (5), in the Marine Data Archive (MDA) (2), European Nucleotide Archive (ENA, 1), Barcode of Life Data system (BoLD, 1), one participant only stores their data internally. All participants submit data to a quality control check before storage. Main standard nomenclatures used are the WoRMS (60 %), NCBI (20 %) or are project dependent (20 %). 62.5 % of participants sometimes fill in optional fields, 12.5 % rarely do this, 25 % of participants never fill in optional fields. 50 % of participants publish metadata as a separate file in a repository, 37.5 % publish metadata in the methods section of a publication, 12.5 % in the supplement of the publication. More than half of the participants publish raw data with minimal processing (57.1 %), 28.6 % publish raw data, 14.3 % publish results. ASVs and OTUs are mainly published as supplement to the publication (3) or on NCBI/Dryad (1). Majority of participants put their code and scripts on GitHub (4), on GitLab (2) or in the supplement of the research article (4). Version numbers and parameter settings are stored mainly in scripts (2), in the methods section of the research article (2) or in internal reports (1). Most participants publish their data open access (3), one participant indicated it depends on the project, one participant mentioned Nagoya, and one participant mentioned that they are unfamiliar with the subject. 71.4 % of participants have reused data from other researchers, 57.1 % of participants have had their data be reused by other researchers.

Detailed answers of the survey can be found in the Annex.

From the survey we gathered that automatization of pipelines and provenance tracking could be improved by the use of input and output files for scripts. We propose to archive data in Darwin Core Archive format (DwC-A) for archiving to biodiversity repositories, like OBIS (Ocean Biodiversity Information System, 2023) or GBIF (Global Biodiversity Information Facility, 2023). We proposed to build a custom tool for DADA2, the pipelines used by most consortium members, that standardizes analysis and data input and output as well as provenance tracking, and formats data in standardized format for long-term archiving in biodiversity repositories.

-

- Bioinformatic survey outside the GEANS consortium

A slightly more concise version of the bioinformatic survey was launched outside the consortium through a news article on the https://northsearegion.eu/geans website, and through the GEANS twitter https://twitter.com/GEANS_Interreg. Consortium members were asked to spread the survey within their network. Unfortunately, the survey only had three participants, all belonging to academia or research.

- PROCESSING

All participants indicated they use metabarcoding (3), two participants also use metagenomics, one participant also uses metatranscriptomics. The aim for all participants is biodiversity indices, ecological indicators and species list (3), one participant is also interested in species function. Three participants use Illumina sequencing, two participants also use Oxford Nanopore as a sequencing platform. One participant uses the DADA2 pipeline for amplicon data and a custom local pipeline for metagenomes, another participant uses Cascabel, a snakemake in-house developed pipelines, runs with Pear, Qiime and DADA2, the last participant uses Anvia for metagenomes, run local and on HPC, or ORP for metatranscriptomes on HPC server or Galaxy pipeline with USEARCH (Edgar, RC (2010)) on a cloud. The main motivation for using the described workflows is because they are reliable and tested (1), easy in use (1), complete in-house control and availability and support in-house (1).

Coding experience needed for the mentioned workflows is R (1), Python (1), GUI (1), setting up computing environments (1) or little experience needed (1). The main workflows products are taxonomic annotated sequences (3), abundance tables (3), ASV (3) and OUT (3) tables, and predicted protein annotations (1). Parameters are either manually entered in a GUI (1), in code script (1) or in a separate configuration or input file (1). Two participants do provenance tracking, a third does not do this automatically. Running the analysis was the most consuming part of the workflow for 2 participants, transforming input and output data was most consuming for the third participant. When asked about pitfalls experienced, one participant mentioned that metatranscriptomics and metagenomics analysis are not easy to use as established metabarcoding pipelines, another participant mentioned RAM, a third participant experiences no issues at the moment. All participants use the NCBI database for taxonomic assignment, some also use Silva (2), BoLD (1), MIDORI (1) and a custom database (1). When asked how they assure accuracy of taxonomic assignment, participants use similarity threshold (98% species level matches), phylogenetic placement, LCA, Diamond workflow, Phlyotree, nucleotide BLAST, and Bayesian approach (Rdp classifier).

- **ARCHIVING**

  All participants submit their data to quality controls stages before storage. Most participants publish results (e.g. ASVs) (3), followed by raw data (2) and raw data with minimal processing (1). All participants published their data on ENA (3), followed by NCBI (2) and keeping internal copies (2), and to a lesser extent also published on GenBank (1) and Barcode of Life Data system (BoLD, 1). Two participants use NCBI/GenBank/ENA standard nomenclature, one participant uses Darwin Core Archive standard. Two participants indicated that they sometimes put effort filling in optional fields, one participant always fills in optional fields. All participants have published metadata in a separate file in the repository (3) and in a supplement to the publication (3), two participants have published metadata in the methods section of a publication. Participants either publish code on GitHub (1), in a publication supplement (1) or in the methods section of a publication (1). Two participants published version numbers and parameter settings in the methods section of a publication, or in the publication supplement (1), in the published script (1) or they don't share this metadata (1). Code and data have been published under the following licenses: MIT, open access, or depending on publisher – as free as possible.

Detailed answers of the survey can be found in the Annex.

Participants here also mention that transforming input and output data is time consuming, they also archive in long-term repositories, in standardized formats, and put effort into filling in optional fields but don't archive to long-term biodiversity repositories, like OBIS or GBIF. Parameter settings and version numbers are not shared or in sections related to the publication or in the script.

- Standardized workflow to prepare metabarcoding data for submission to biodiversity data aggregators

After identifying pain points in genetic workflow used by experts in- and outside the GEANS consortium, we saw the opportunity for improving automatization and provenance tracking through standardizing in- and output of bioinformatic pipelines and reformatting output data to standardized nomenclature for biodiversity archiving, in order to increase harmonization among scientists. To this end, we created a computational tool to process Illumina-generated metabarcoding data in the programming language R that converts the output of a metabarcoding bioinformatics pipeline to a format that can be submitted to international biodiversity data aggregators. As the DADA2 pipeline proved to be most popular in use amongst all participants of the surveys, we decided to focus on this pipeline as an example. However, only few changes are necessary to adapt the tool to alternative metabarcoding pipelines.

A Darwin Core Archive file for metabarcoding data requires three separate input files. The first is a metadata table, describing the relevant context of the observations, the second is the actual

occurrence table and the third contains information on the amplicon sequence variants (ASVs) that were registered.

While the base of the metadata table must be created manually by the user, the two other files will be created by this tool. The column heads of the metadata table should be identical to the ones presented in the metadata table example provided in the GitHub repository. Those names are based on the field names in the Darwin core archive and MIx standards. It is also mandatory that the values of one field (samp_name) in the metadata table correspond with the names of the samples in the DADA2 analysis. Once created, this metadata table can be used as an input for the tool, as well as the path to the folder containing the fastq.gz files. The tool then combines the content of the metadata table with the native output of the DADA2 workflow, namely the sequences, read counts and taxonomic assignments. Finally, this content is distributed over the required input files for the Darwin Core Archive (metadata table, occurrence table and DNA extension table).

A reproducible minimal example (including example sequence data and reference database) of the usage of the tool is available on GitHub (https://github.com/pascalhabluetzel/GEANS_WP4_dwca). Currently, only a limited number of Darwin core archive and MIxS fields are incorporated in the tool. The main reason for this is that there is no agreement among experts and data holders on which fields are required or (highly) recommended and what the values of these fields should exactly be. There is currently also no field that can accommodate the bootstraps values of the Bayesian analysis of DADA2 and this function is therefore not used in the example. A detailed guide of how the tool should be used can be found in the aforementioned GitHub repository. The idea that bioinformatics pipelines produce data that can be directly uploaded to biodiversity data aggregators seems to be a promising avenue for biodiversity informatics. Despite the current limitations we already publish the code for doing so on output of DADA2 in the hope it will be a useful guide for DADA2 users or bioinformatics and biodiversity informatics tool developers.

- Conclusions and recommendations

The field of bioinformatics relies heavily on the utilization of a heterogeneous set of workflows to analyze and interpret DNA sequence data. While there is a wide range of bioinformatics pipelines available, we noted that most users tend to rely on published and widely used workflows. This trend can be attributed to several factors, including familiarity, reliability, and the perceived validation associated with established pipelines. However, there are certain limitations and challenges that need to be addressed in order to enhance the efficiency and effectiveness of bioinformatics pipelines.

One of the prominent issues faced by users is the time-consuming process of adjusting code for new analyses. The customization of existing workflows to accommodate specific research needs can be a tedious and resource-intensive task. Therefore, there is a pressing need for more automation in bioinformatics workflows. This would allow scientists to focus more on the biological interpretation of results rather than spending excessive time on technical adjustments.

Another critical challenge in bioinformatics pipelines is the insufficient compatibility between different tools. The interoperability of tools is crucial for seamless data flow and integration of results from various analysis steps. However, the lack of standardization and compatibility protocols often leads to data compatibility issues. For instance, the taxonomic backbones of widely used databases like WoRMS and NCBI are not fully compatible, which can hinder accurate taxonomic assignments and data

integration. It is essential for the bioinformatics community to collaborate and establish standardized formats and protocols to improve compatibility between different tools and databases.

Additionally, the limited integration of Alternative Sequence Variants (ASVs) and Operational Taxonomic Units (OTUs) into data aggregators is a notable concern. ASVs and OTUs are widely used in metagenomic studies to identify and classify biological communities. However, the lack of their inclusion in data aggregators restricts their accessibility and usability for broader scientific investigations. Therefore, efforts should be made to encourage the integration of ASVs and OTUs into existing data repositories and aggregators. This would promote the sharing of comprehensive and diverse datasets, facilitating more robust and comprehensive analyses. We took a step in this direction by writing a tool that does transform the output of the DADA2 pipeline into a format that is suitable for submission to biodiversity data aggregators. However, this tool is currently limited in its capabilities and is rather a proof of principle and further development is needed before it can be applied in real data analysis.

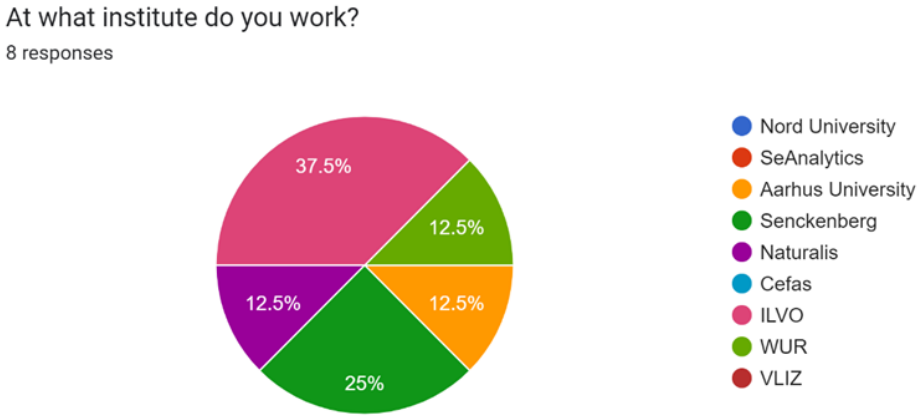
Furthermore, the issue of data and code reproducibility and interoperability needs to be addressed to ensure the FAIR (Findable, Accessible, Interoperable, and Reusable) principles are followed. While data is usually shared upon publications, ensuring that the associated code is well-documented and easily interpretable remains a challenge. The lack of interoperability and standardization of code can impede the reproducibility and transparency of research findings. To tackle these challenges, the bioinformatics community should actively encourage the use of code sharing platforms, such as GitHub, where researchers can share their pipelines and make them accessible to the wider scientific community. In addition, efforts should be directed towards developing standardized documentation practices and promoting the use of metadata standards, enabling better interoperability and reproducibility of data and code.

In conclusion, the field of bioinformatics relies on a variety of pipelines, but the dominance of published and widely used workflows limits the exploration of new approaches. To overcome this, there is a need for more automation in pipelines, allowing users to easily adapt to new analyses. Additionally, improving compatibility between different tools and databases will enhance the efficiency and reliability of bioinformatics workflows. The integration of ASVs and OTUs into data aggregators should be encouraged to broaden the accessibility and usability of metagenomic studies. Lastly, addressing the challenges of data and code interoperability and reproducibility will promote the adoption of FAIR- data principles (Findable, Accessible, Interoperable, Reusable) in bioinformatics research. By taking these recommendations into account, we can strive towards more efficient, reliable, and accessible pipelines.

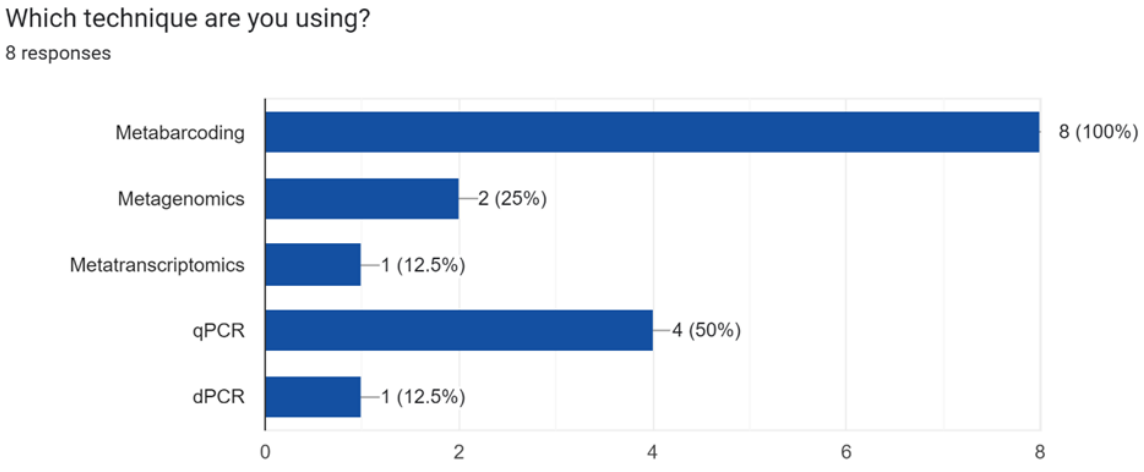
- Annex

- WP4 - BIOINFORMATICS SURVEY INSIDE GEANS CONSORTIUM

As part of WP4: Harmonization and standardization of (genetic) protocols and tools, we would like to get a view of bioinformatic pipelines used in GEANS partner institutes. Therefore, we would like you all to fill in the following survey on bioinformatics and biodiversity informatics. This information will help us to further consolidate protocols and standardize our workflows. Importantly, the goal of this survey is not to identify the best pipeline, but to help us to identify steps in workflows where even experts are struggling with, loose time or which are prone to introduce errors. This questionnaire concerns only metabarcoding and metagenomics (i.e. only techniques that produce nucleotide sequences).

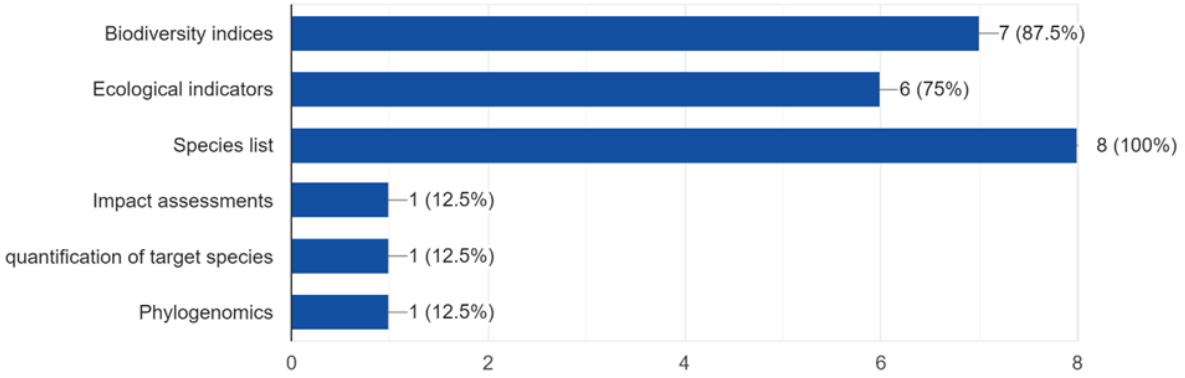


Section 1 – Processing



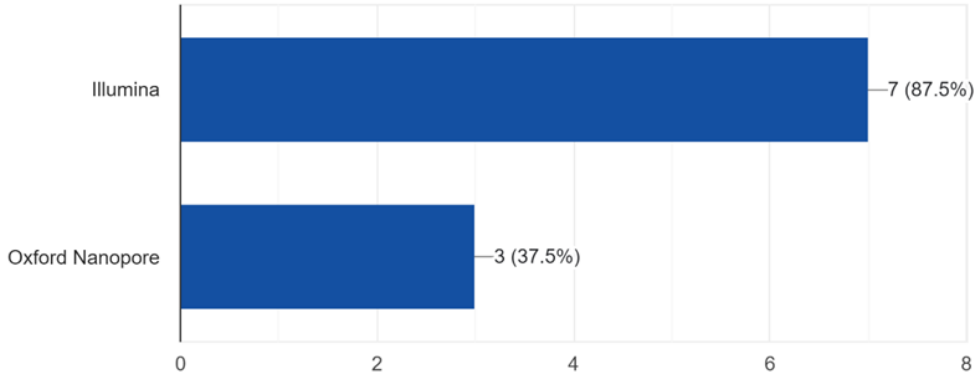
What type of result do you aim for?

8 responses



Which sequencing platform are you using?

8 responses



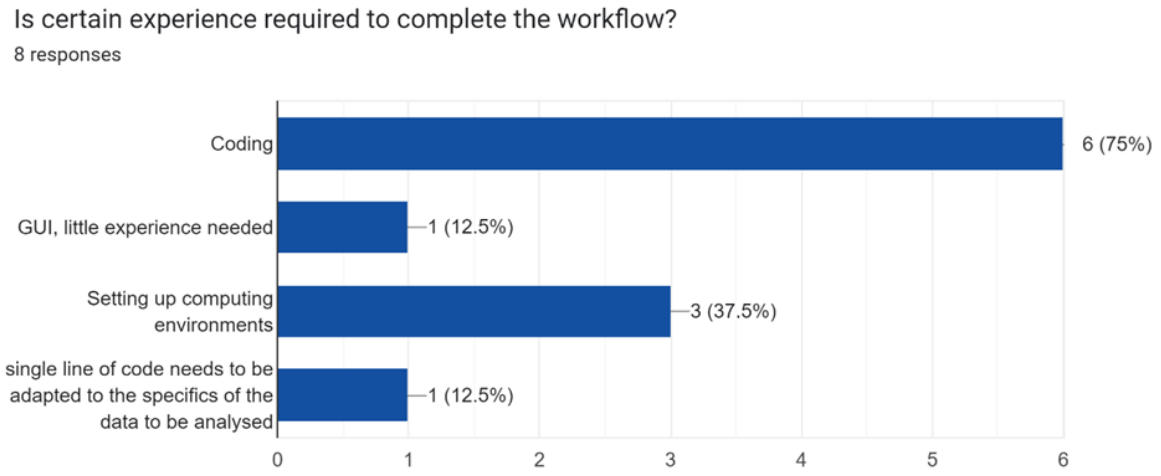
Which is your preferred pipeline at the moment and where do you run the bioinformatic analysis? Is it publicly available, is it dockerised, does it run in a cloud, etc.? More than one answer possible. Examples: PEMA for COI metabarcoding. Runs on our local pc. DADA2 for 18S metabarcoding. Runs on our local pc. Custom pipeline (link to publication or github repository). Runs on the national supercomputer. 8 responses

- We run most of the metabarcoding analyses on a custom pipeline in Galaxy. Some tools are pre-existing (e.g. FLASH for merging, Cutadapt for primer trimming), some tools are custom-made (e.g. our LCA taxon resolver). All the wrappers and tools can be found on our Github repository: https://github.com/naturalis. Galaxy now runs on the local cloud, but we are also working on being able to deploy on any instance (e.g. AWS) and local hardware (MaaS solutions).
- First we use a few bash scripts (written at ILVO) for quality control and trimmomatic, then we use DADA2 for metabarcoding, which runs on the genomic server of ILVO.
- QIIME2, VSEARCH, DADA2 plugin in QIIME2. Both VSEARCH and QIIME2 are publicly available. I run these tools in university administered bioinformatics server

- We use Decona, in-house developed pipeline for nanopore data analysis. https://github.com/Saskia-Oosterbroek/decona Also tested several others like MOTHUR. We run it mostly on our local computers or sometimes Wageningen University HPC
- Dada2 for COI, 12S and 18S based on illumina data; decona for Minion data; runs on our ILVO server;
- DADA2 enhanced with decontaom on 16S .. data / decona https://github.com/Saskia-Oosterbroek/decona run on the server at our company
- DADA2 for COI and 18S metabarcoding Run on our local supercomputer. Assemblies and gene annotations Run on the supercomputer from Senckenberg bioinformatic department (TBG)
- DADA2, own scripts local PC

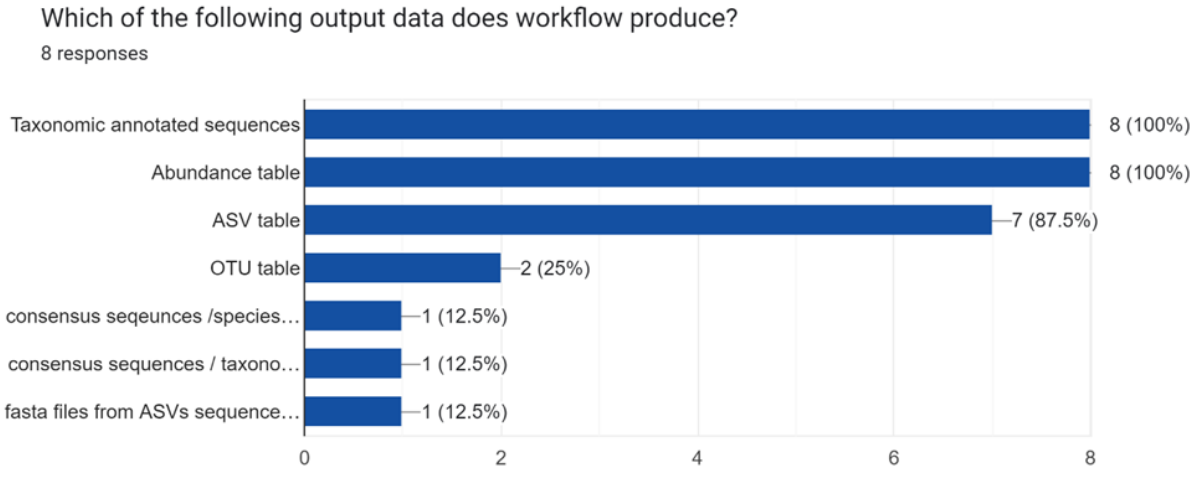
What is your main motivation for using your preferred workflow as opposed to others? 8 responses

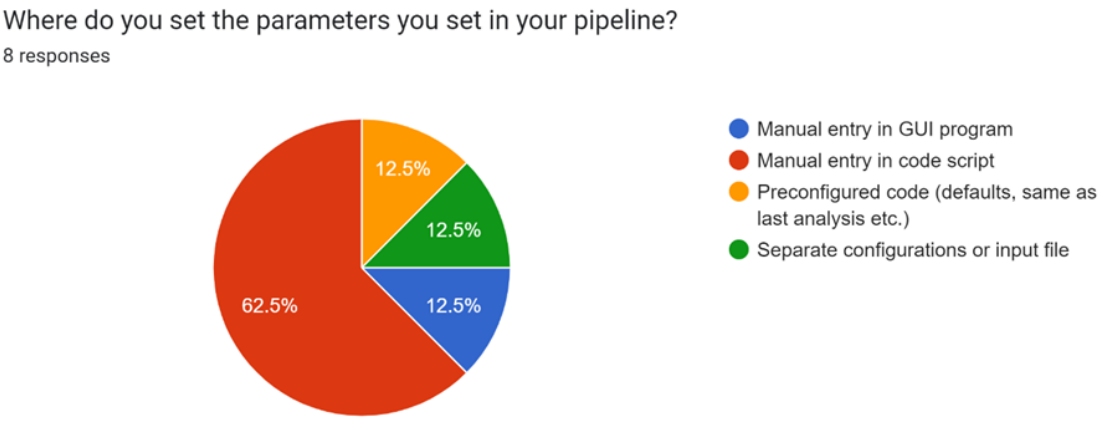
- User interface (no command line necessary).
- Dada2: 1) The learnErrors method learns this error model from the data, by alternating estimation of the error rates and inference of sample composition until they converge on a jointly consistent solution, 2) use a native implementation of the naive Bayesian classifier method for this taxonomic assignment.3) it retains a summary of the quality information associated with each unique sequence. The consensus quality profile of a unique sequence is the average of the positional qualities from the dereplicated reads. These quality profiles inform the error model of the subsequent sample inference step, significantly increasing DADA2's accuracy
- Open access, huge community to help with.
- fits the Nanopore data best. Still not completely satisfied and looking for alternatives / actively developing alternatives ourselves.
- illumina data: resolution to ASVs instead of OTU's, runs in R, very user friendly
- Easily runnable/understandable and written in known languages
- For metabarcoding using DADA2 to categorize the sequences into ASVs rather than implementing a certain similarity threshold for the entire community to obtain OTUs in other pipelines.
- ASVs have no similarity threshold like OTUs



Which coding languages do you need to master for the workflow? 8 responses

- For using the tools no coding is required, only for the "behind the scenes" work with development and update of tools (but we have a dedicated team for that).
- A few commands for the bash script (very little experience needed) and R for the Dada2 pipeline
- basic bash command should be enough
- Non, basic experience in bash and R needed
- R language; demultiplexing requires python scripting
- R/ bash/shell
- R and C++
- R, bash



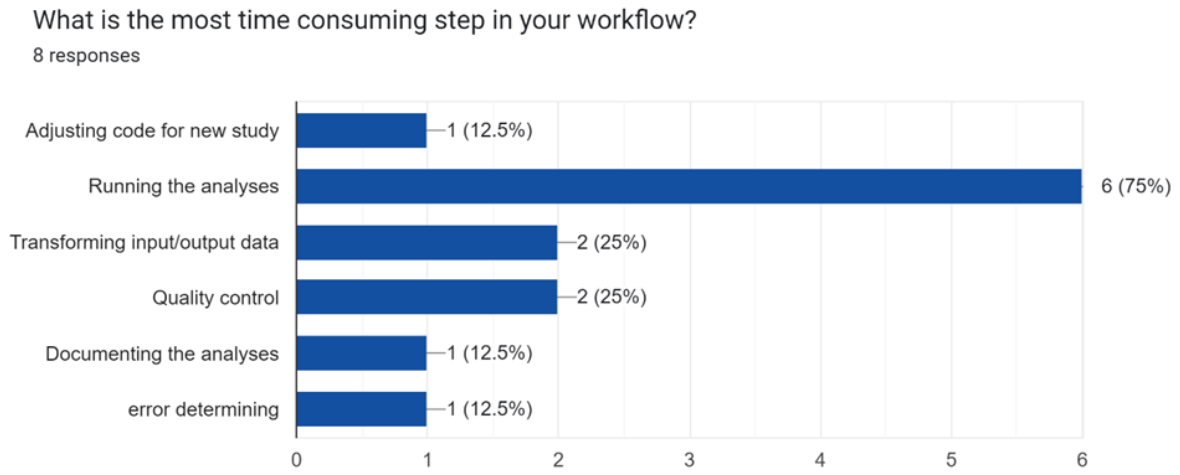


Does it do any provenance tracking (= documentation of where a piece of data comes from and the processes by which it was produced)? If so, describe how: 6 responses

- No
- Galaxy allows for the creation of workflows/histories.
- all actions are saved in a log-file
- not sure what is meant here, you have to define the datasets and the functions in R to generate the data, so everything is in the R script
- During the process each ASV gets a taxonomic assignment of the best hit from NCBI and/or our own reference barcode libraries (if available); the assignment gets annotated with accession numbers from NCBI to be able to track back the assignment. In addition, the percentage ID, query coverage and Evalue are also retrieved from NCBI through the pipeline.

How long does the workflow take you on average? 8 responses

- Hands-on time for an average MiSeq dataset would be roughly one hour, tasks/tools can be queued or performed as part of a pre-set workflow. Work is executed in the cloud, so no need to wait.
- Estimate: 2 hours hands-on time
- less than 24 hours
- depends on amount of data, between 30 minutes and 24h on local server
- depends on the size of the dataset, I would say two days
- depends on the amount of input data and the nature of the data
- depends on the number of reads, 1 to 4 hours
- 30 minutes



Where do you think you could save time? 7 responses

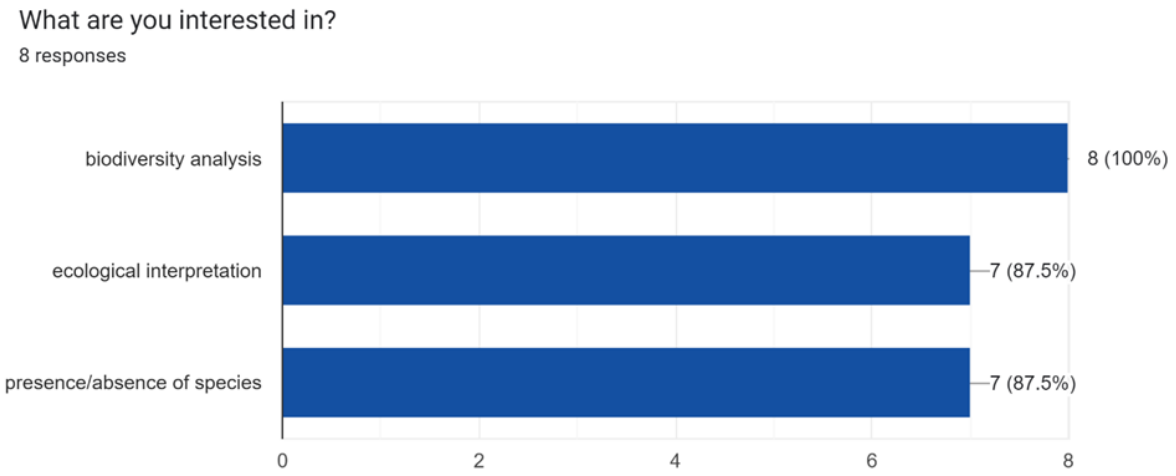
- Right now, correcting for controls and combining replicates etc. takes the most time (the bio-info pipeline "stops" at the MOTU/ASV table, the rest of the statistical analyses are done in R). Some of this work is easily done by predefined R scripts, some other steps need more individual attention in R.
- Now we run the separate chunks of the R script separately, but you still need to wait a few moments before you can go to the next chunk. This can be improved. Maybe just in general, I think there is room for improvement in multiple tasks around the data analysis (so not only the analyses itself, but also the documenting, ...) by making these tasks more automatic.
- Adjusting the code
- improved and more clear process of taxonomic assignment, to reduce quality control time needed.
- making the code more automated
- adjusting code for a new study
- no need to save time

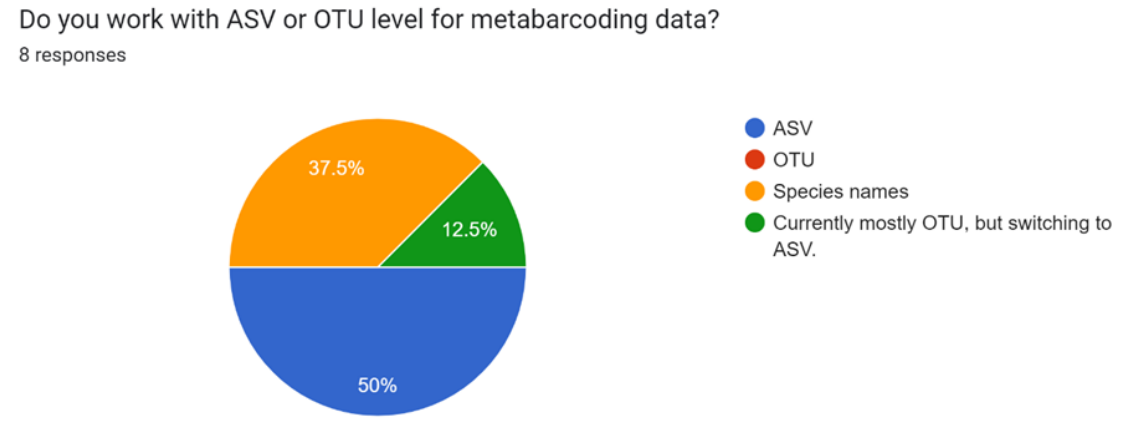
What are the issues or pitfalls you experience while using the pipeline? What would you like to see improved? 8 responses

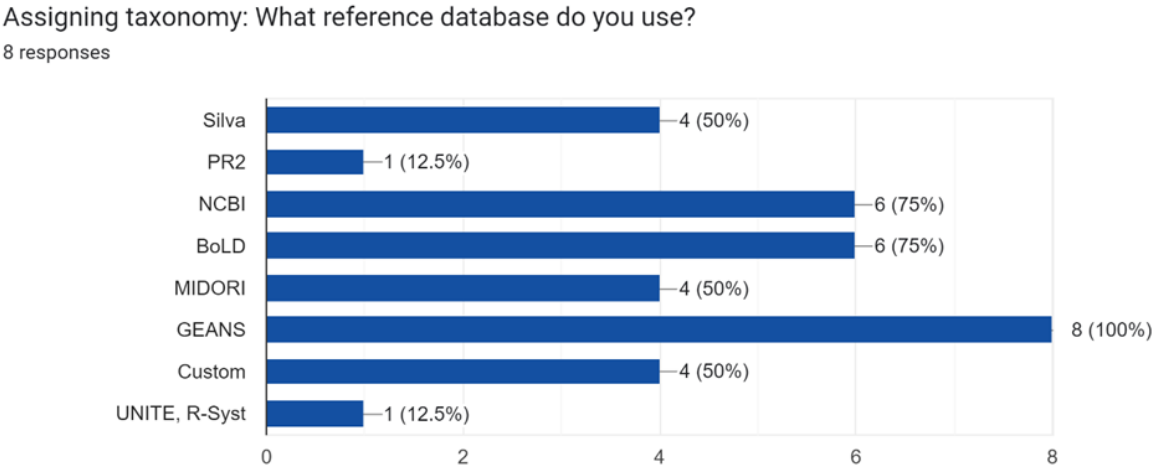
- We are still experiencing some issues to get DADA2 to work within the Galaxy environment, but that's currently being worked on. There's a few post-OTU table processing tools, we can probably add some more, esp. regarding the automated checking and correcting for contaminations based on control samples.
- see above
- Taxonomy assignment and reference database
- data output from Decona currently not well annotated taxonomically. Integration with R-workflow is currently being developed for a version 2

- manual setting of the parameters for ASV generation, depends on the primers used and of the quality of the reads
- reference database
- We use the WoRMS website to confirm the taxonomic assignments. During this process if a species name (retrieved from NCBI) has a code incorporated to the name, WoRMS would not recognize the species and will assign “no-match” as taxonomic assignment. It would be very useful to unify the NCBI records to only hold plain species name without any code or extra letter within species.
- happy with the pipeline

Section 2- Analysis

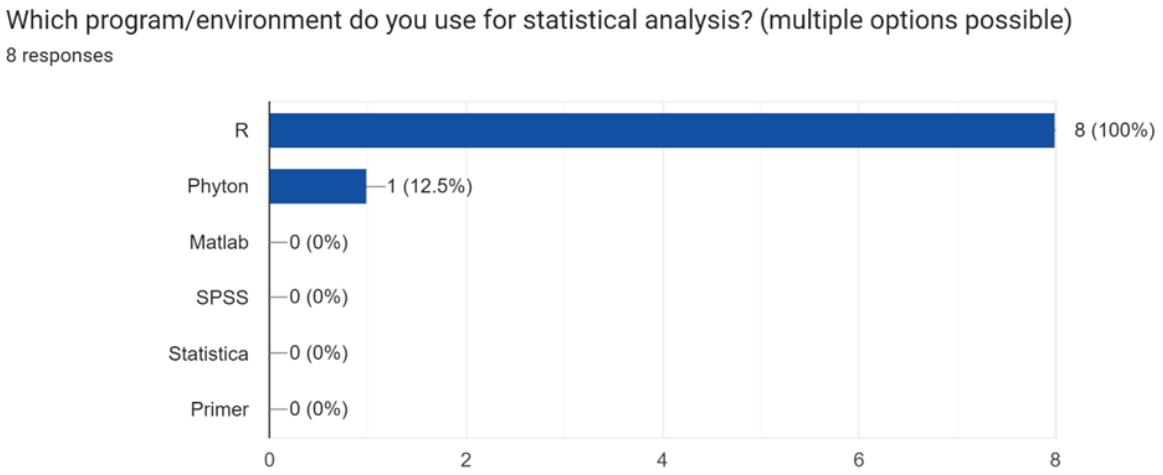






How do you assure the accuracy of the taxonomic assignment of your ASVs/OTUs? E.g. Similarity threshold (Which?), Phylogenetic methods (Which tool/method?), Baysian/maximum likelihood approach (which tool/method?). 8 responses

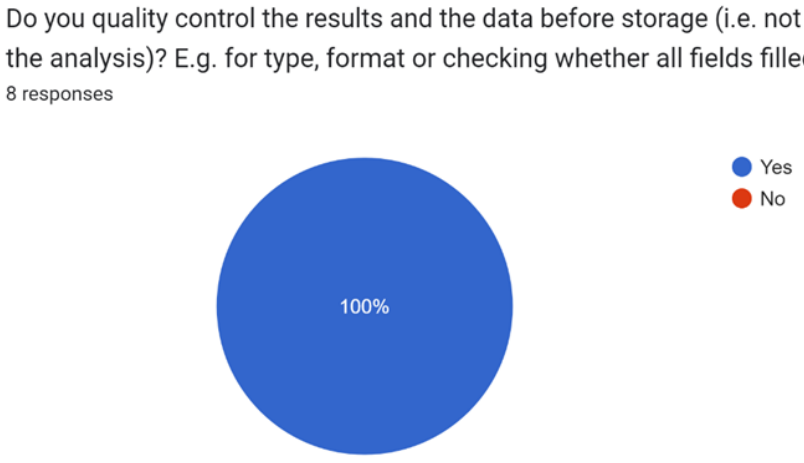
- Similarity threshold (ca. 97-98% depending on taxon/marker), higher taxonomic assignments in case of no species-level matches are done with a custom LCA tool based on MEGAN (https://github.com/naturalis/galaxy-tool-lca)
- The DADA2 package provides a native implementation of the naive Bayesian classifier method for this purpose. The assignTaxonomy function takes as input a set of sequences to be classified and a training set of reference sequences with known taxonomy, and outputs taxonomic assignments with at least minBoot bootstrap confidence.
- Phylogenetic methods using maximum likelihood method
- Similarity threshold at 98%, database with species occurrence, manual curation
- rdp with minboot set at 80
- 80%bootstrap / 95% accuracy / 70 % coverage
- 1.Similarity threshold for COI, 97% and for 18S 98 or 99%. 2.phylogenetic method using Bayesian approach and Yule tree prior using GMYC method
- Grade of percent identity and query coverage



Section 3 – Archiving

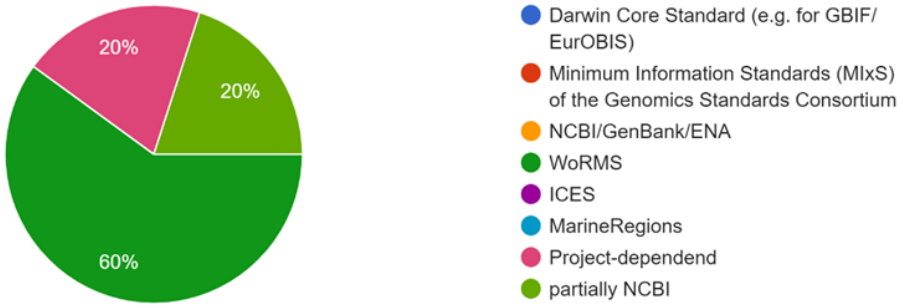
What long-term repositories do you use to store your data? Public (which?) or internal? 8 responses

- Published data will usually go to the NCBI SRA, long term (non-public) data storage is currently being overhauled, but will most likely go to SURF, but can potentially be anywhere via iRODS.
- Internal: on our server at ILVO and public: ncbi BIOproject, MDA (will be public after publication)
- University server and NCBI SRA
- both internal, ENA and for GEANS data MDA/IMIS
- Genbank, ILVO server (internal)
- internal
- internal, BoLD and NCBI
- still no choice



Do the output fields (e.g. table headers) in the results files follow a standard nomenclature?

5 responses

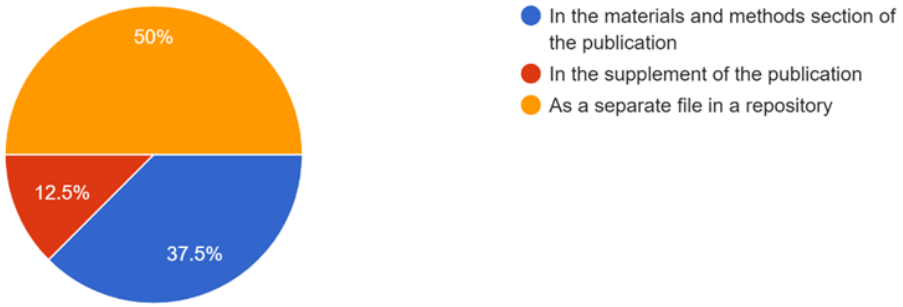


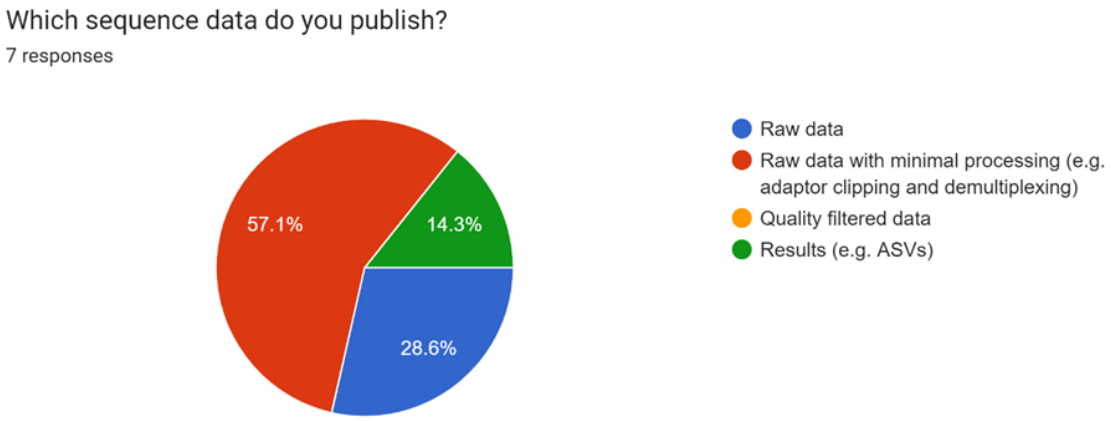
Do you put extra effort into filling in optional fields?

8 responses

How do you publish metadata?

8 responses





Where do you publish your processed ASVs: 5 responses

- Supplement in publication, but we are working on ways to publish these through our own platform.
- Supplementary tables
- n/a
- suppl info of paper
- NCBI, Dryad

Where do you publish your processed OTUs: 5 responses

- Supplement in publication, but we are working on ways to publish these through our own platform.
- Supplementary tables
- with raw data in same repository
- NA
- NCBI, Dryad

Where do you publish your code and scripts: 7 responses

- GitHub for tools, scripts (R) as supplements when required for manuscript.
- on gitlab
- Supplementary file with the MS
- github and/or in publication on the topic
- gitlab
- github and supplementary of the paper
- github

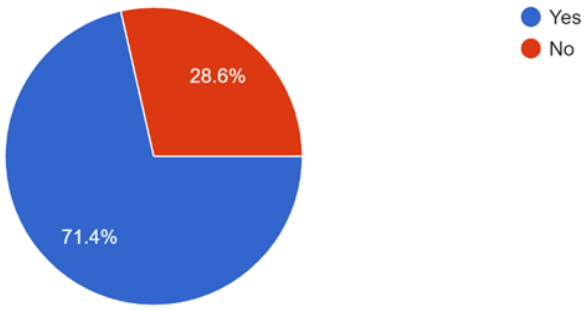
Where and how do you publish your Version numbers, parameter settings, etc.? 5 responses

- Some details only in internal reports, versions of tools in M&M section of paper.
- in the paper where the data is used
- Materials and methods section or Supplementary files
- local R script
- within the scripts

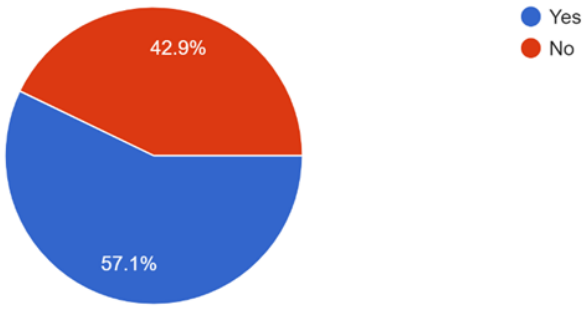
Under which license do you publish your data and code and why? 6 responses

- Published data are open access, tools are also mostly open access (MIT license)
- Mostly open access
- depends on project,
- No idea
- Nagoya
- GPL

Do you reuse data from other researchers?
7 responses

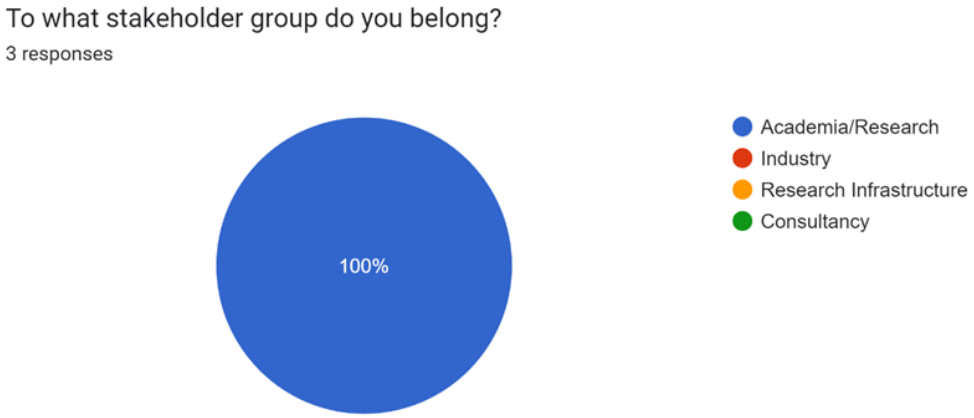


Has your data been reused by other researchers?
7 responses

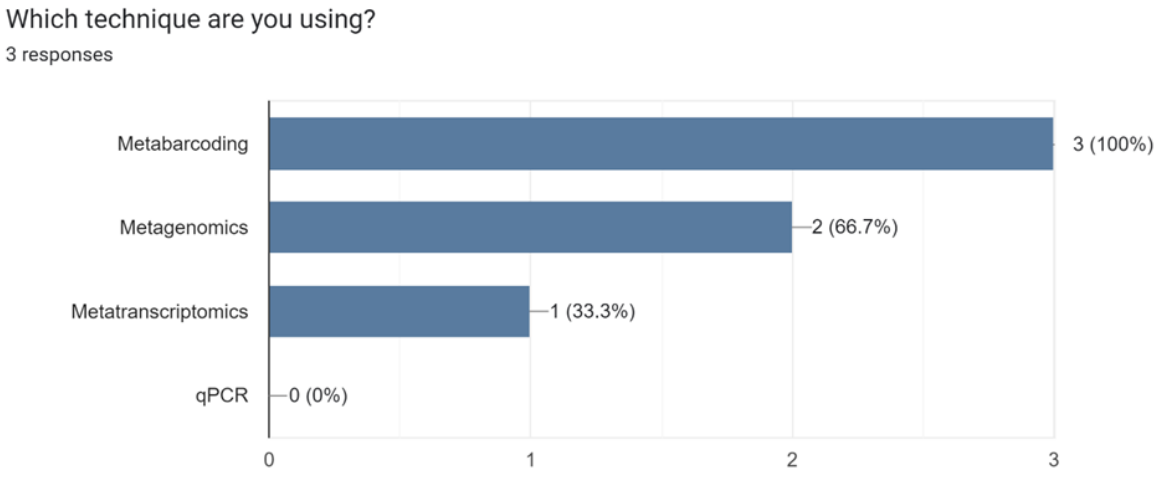


- WP4 - BIOINFORMATICS SURVEY OUTSIDE GEANS CONSORTIUM

As part of the GEANS project (Genetic Tools for Ecosystem Health Assessment in the North Sea region), we are working on harmonization and standardization of (genetic) protocols and tools in Europe. To get a view of bioinformatic pipelines used, we would like you all to fill in the following questionnaire on bioinformatics and biodiversity informatics. This information will help us to further consolidate protocols, standardize workflows and help us to identify steps in workflows where even experts are struggling with, loose time or which are prone to introduce errors. This questionnaire concerns only metabarcoding and metagenomics (i.e. only techniques that produce nucleotide sequences).

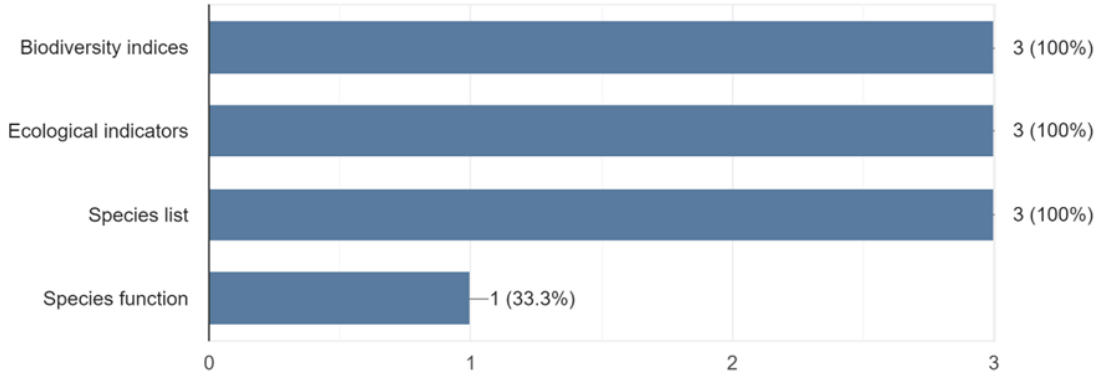


Section 1 - Processing



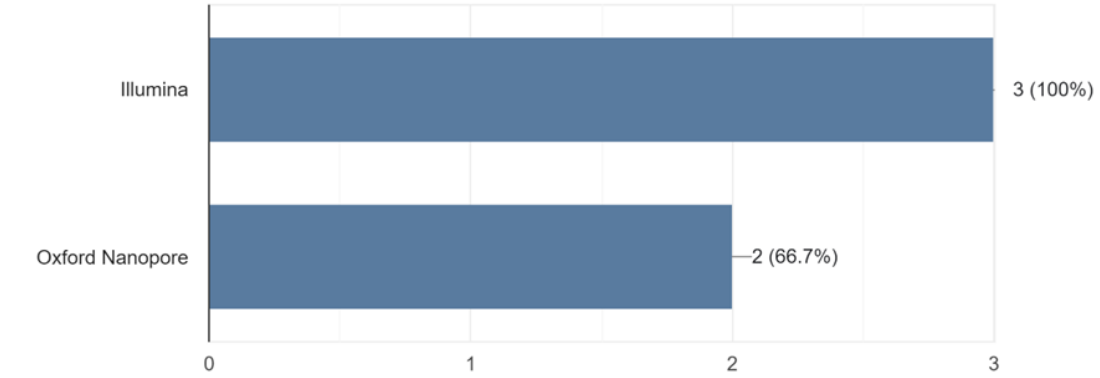
What type of result do you aim for?

3 responses



Which sequencing platform are you using?

3 responses



Which is your preferred pipeline at the moment and where do you run the bioinformatic analysis? Is it publicly available, is it dockerised, does it run in a cloud, etc.? More than one answer is possible. Examples: PEMA for COI metabarcoding. Runs on our local pc. DADA2 for 18S metabarcoding. Runs on our local pc. Custom pipeline (link to publication or GitHub repository). Runs on the national supercomputer. 3responses

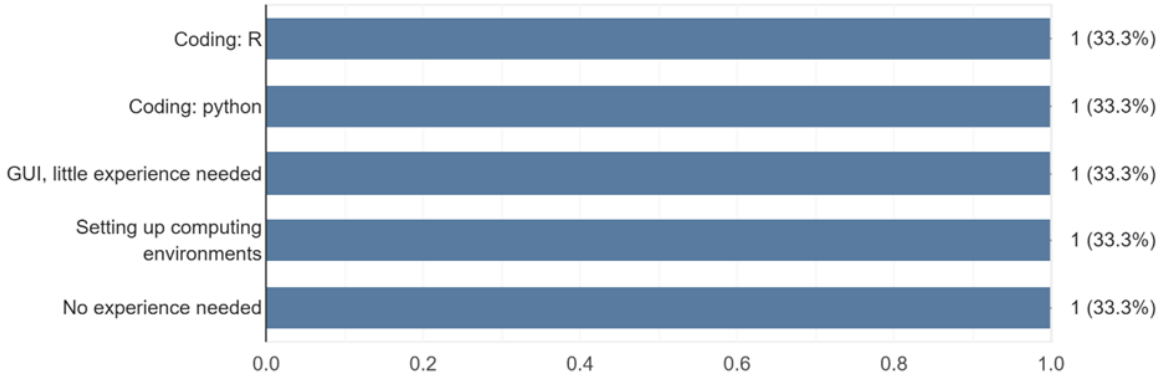
- Anvio for metagenomics (local PC and HPC server), ORP for metatranscriptomics (HPC server), Galaxy pipeline with USEARCH (cloud on local server)
- DADA2 for amplicon data, custom local pipeline for metagenomes
- Cascabel - a snakemake pipeline built in-house by NIOZ but publicly available. Runs with pear, Qiime and dada2

What is your main motivation for using your preferred workflow as opposed to others? 3responses

- Reliable and tested, easy to use for students
- We have complete control over it
- Availability and support at our own workplace

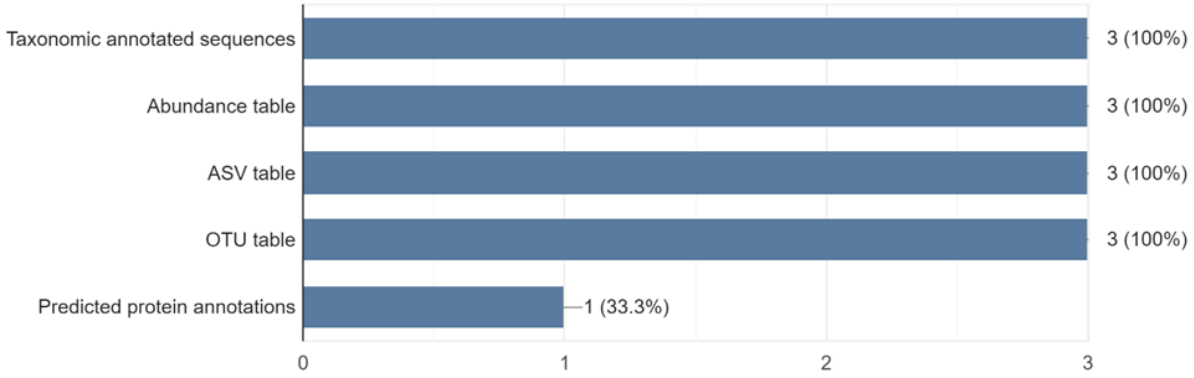
Is certain experience needed to master for the workflow?

3 responses



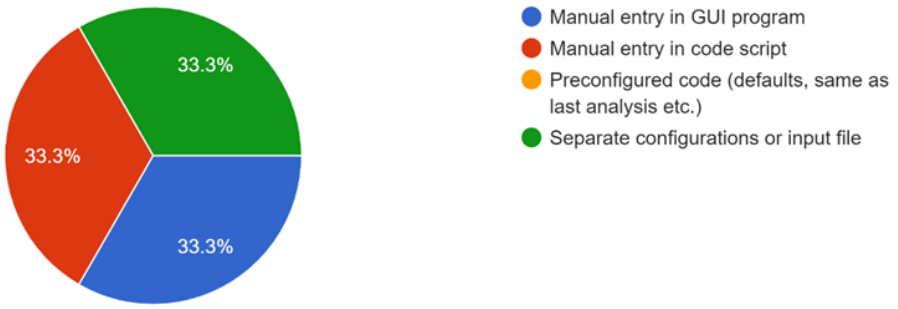
Which of the following output data does the workflow produce?

3 responses



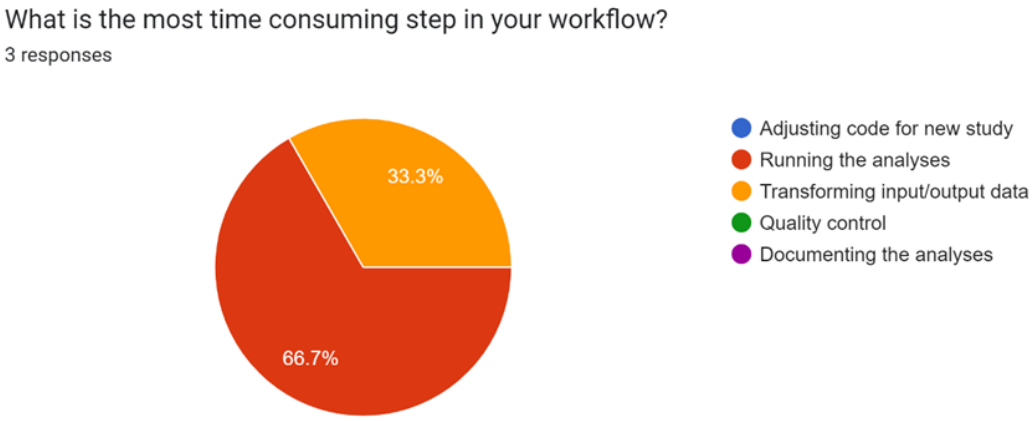
Where do you set the parameters you set in your pipeline?

3 responses



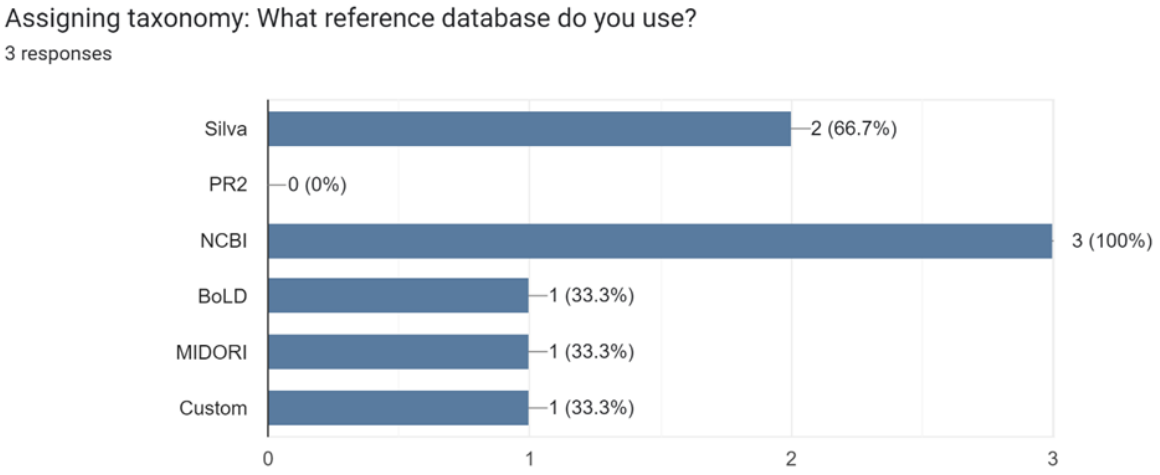
Does it do any provenance tracking (= documentation of where a piece of data comes from and the processes by which it was produced)? If so, describe how: 3 responses

- No, not automatically. Depends on the metadata provided and taken along during the workflow.
- Yes
- Yes, full workflow report with used programs, versions etc.



What are the issues or pitfalls you experience while using the pipeline? What would you like to see improved? 3 responses

- Metatranscriptomic and metagenomic analyses are still not as easy to use as established metabarcoding pipelines. It should also be easier to share and compare resulting datasets.
- RAM amounts
- None atm

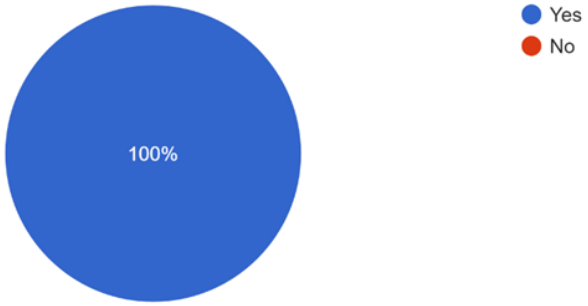


How do you assure the accuracy of the taxonomic assignment of your ASVs/OTUs? E.g. Similarity threshold (Which?), Phylogenetic methods (Which tool/method?), Baysian/maximum likelihood approach (which tool/method?). 2 responses

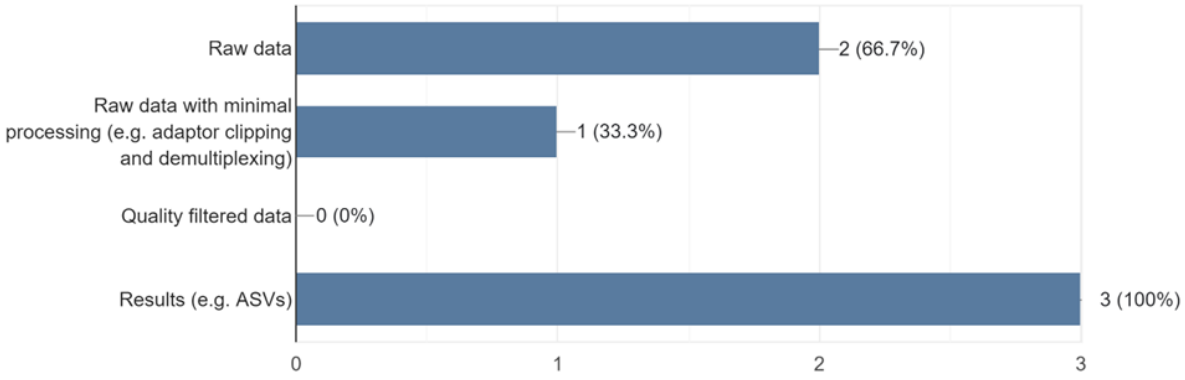
- Similarity depending on taxonomic group (commonly 98%id for species level), phylogenetic placement, annotation and LCA with various software packages implemented in Anvi'o, Diamond workflows, Megan, PhyloTree, BLASTn
- Rdp classifier, thresholds

Section 2 - Archiving

Do you quality control the results and the data before storage (i.e. not the quality control you do for the analysis)? E.g. for type, format or checking whether all fields filled consistently?
3 responses

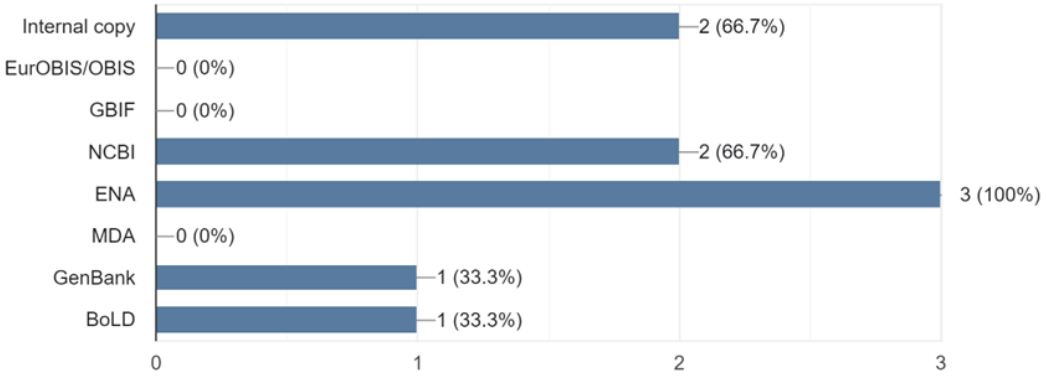


Which sequence data do you publish?
3 responses



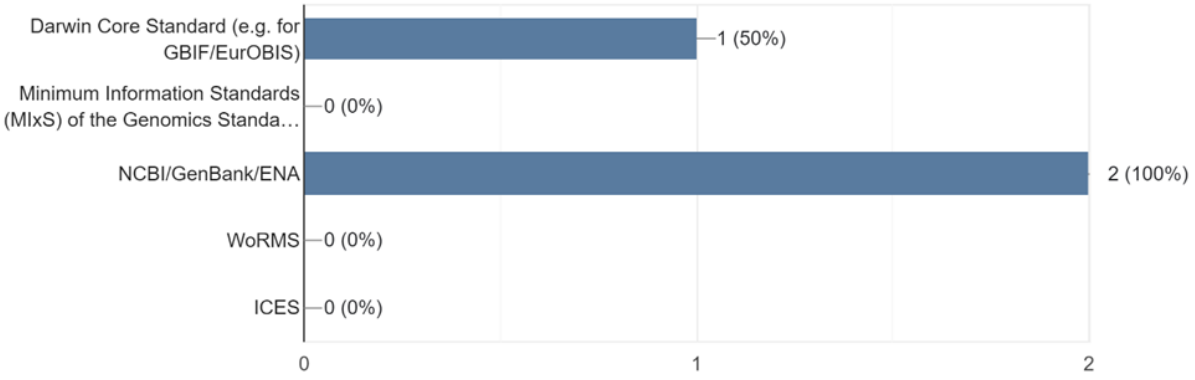
What long-term repositories do you use to store your sequence data? Public (which?) or internal?

3 responses



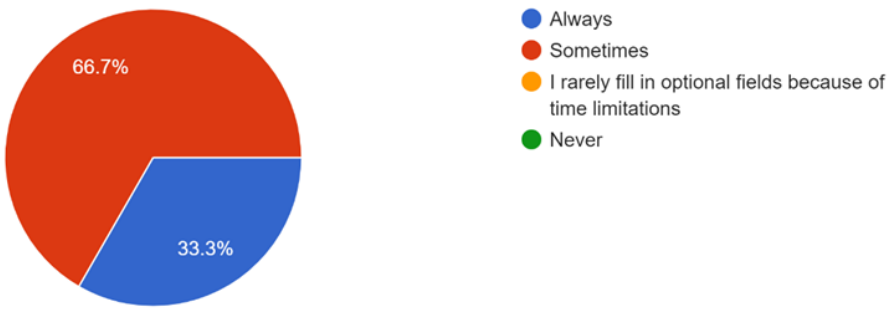
Do the output fields (e.g. table headers) in the results files follow a standard nomenclature?

2 responses



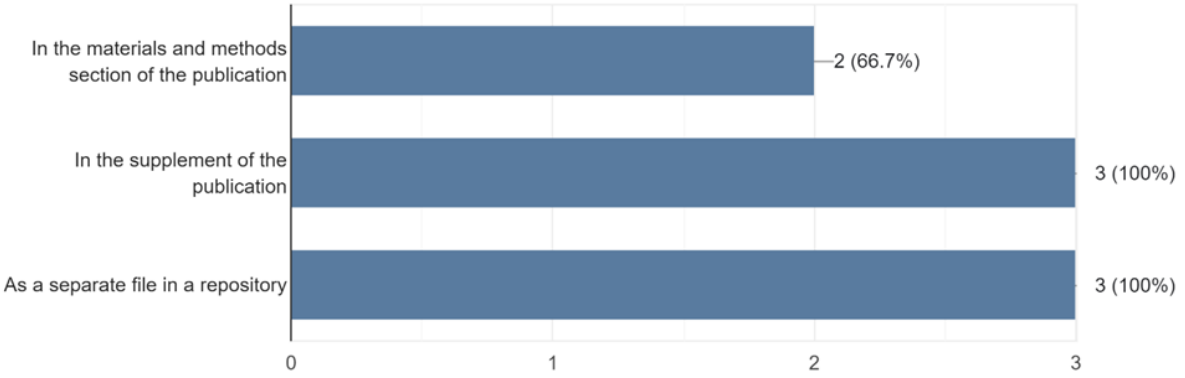
Do you put extra effort into filling in optional fields?

3 responses



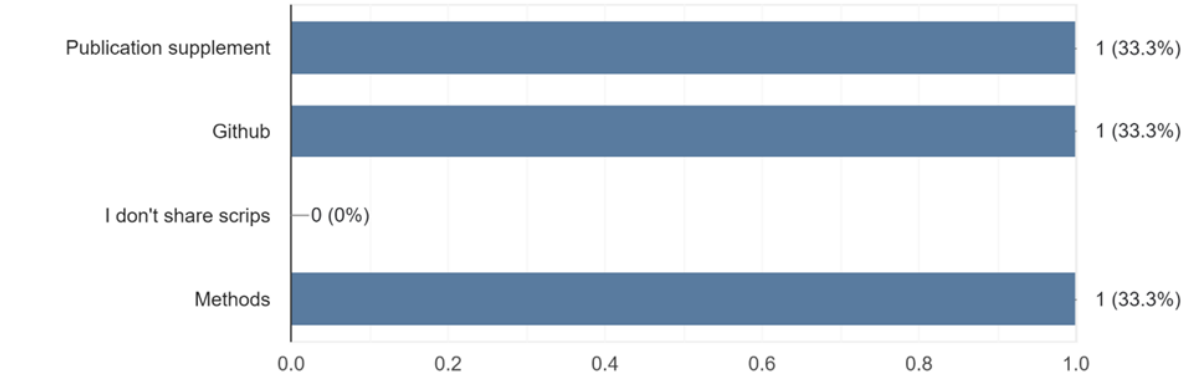
How do you publish metadata?

3 responses



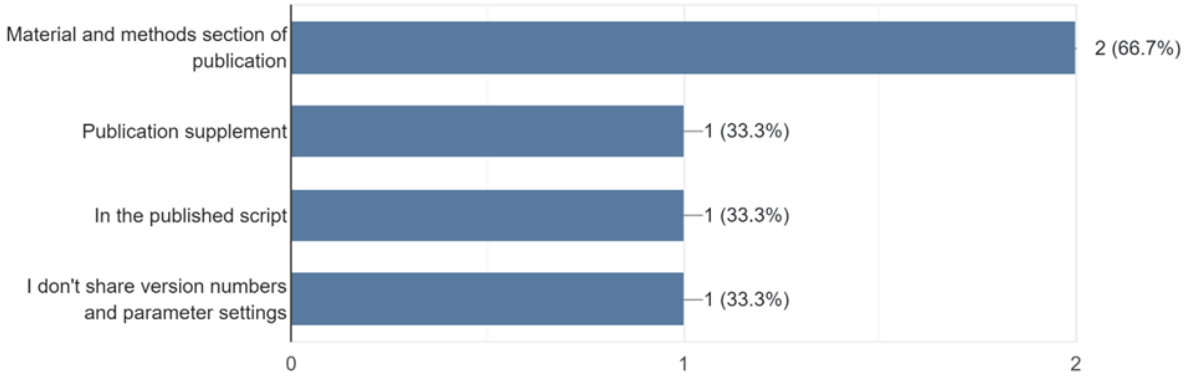
Where do you publish your code and scripts:

3 responses



Where and how do you publish your version numbers, parameter settings, etc.?

3 responses



Under which license do you publish your data and code and why? 3 responses

- Depends on the publisher. As free as possible
- MIT
- Open

- References

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu YX, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS 2nd, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019 Aug;37(8):852-857. doi: 10.1038/s41587-019-0209-9. Erratum in: Nat Biotechnol. 2019 Sep;37(9):1091. PMID: 31341288; PMCID: PMC7015180

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., 577 Holmes, S.P., 2016. DADA2: High-resolution sample inference from Illumina 578 amplicon data. Nat. Methods 13, 581–583.

Edgar,RC (2010) Search and clustering orders of magnitude faster than BLAST, Bioinformatics 26(19), 2460-2461. doi: 10.1093/bioinformatics/btq461

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C. et al. 2013. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Res. 41:D597–604.

GBIF: The Global Biodiversity Information Facility (year) What is GBIF?. Available from https://www.gbif.org/what-is-gbif [13 January 2020].

Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. 2007 Mar;17(3):377-86. doi: 10.1101/gr.5969107. Epub 2007 Jan 25. PMID: 17255551; PMCID: PMC1800929.

Machida, R., Leray, M., Ho, SL. et al. Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. Sci Data 4, 170027 (2017). https://doi.org/10.1038/sdata.2017.27

Marine Data Archive; editing status 2019-12-18; re3data.org - Registry of Research Data Repositories. http://doi.org/10.17616/R31NJMO4 last accessed: 2023-06-09

National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2023 06 09]. Available from: https://www.ncbi.nlm.nih.gov/

Nilsson RH, Larsson K-H, Taylor AFS, Bengtsson-Palme J, Jeppesen TS, Schigel D, Kennedy P, Picard K, Glöckner FO, Tedersoo L, Saar I, Kõljalg U, Abarenkov K. 2018. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. Nucleic Acids Research, DOI: 10.1093/nar/gky1022

OBIS (2023) Ocean Biodiversity Information System. Intergovernmental Oceanographic Commission of UNESCO. www.obis.org.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584. doi: 10.7717/peerj.2584

The Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update, Nucleic Acids Research, Volume 50, Issue W1, 5 July 2022, Pages W345–W351, doi:10.1093/nar/gkac247

Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO (2014) The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucl. Acids Res. 42:D643-D648

WoRMS Editorial Board (2023). World Register of Marine Species. Available from https://www.marinespecies.org at VLIZ. Accessed 2023-02-06. doi:10.14284/170

Interreg
North Sea Region
GEANS
European Regional Development Fund

EUROPEAN UNION

Genetic tools for Ecosystem health Assessment in the North Sea region

Interreg North Sea Region GEANS

European Regional Development Fund

EUROPEAN UNION

Genetic tools for Ecosystem health Assessment in the North Sea region