Community-curated and standardised metadata of published ancient metagenomic samples with AncientMetagenomeDir

James A. Fellows Yates[1,2,*], Aida Andrades Valtueña[1], ˚Ashild J. Vågene[3], Becky Cribdon[4], Irina M. Velsko[1], Maxime Borry[1], Miriam J. Bravo-López[5], Antonio Fernandez-Guerra[6,7], Eleanor J. Green[8,9], Shreya L. Ramachandran[10], Peter D. Heintzman[11], Maria A. Spyrou[1], Alexander Hübner[1,12], Abigail, S. Gancz[13], Jessica Hider[14,15], Aurora F. Allshouse[16], and Christina Warinner[1,16,*]

[1]Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, 07745, Jena, Germany
[2]Institut für Vor- und Frühgeschichtliche Archäologie und Provinzialrömische Archäologie, Ludwig Maximilian University, München, 80539, Germany
[3]Section for Evolutionary Genomics, GLOBE Institute, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, 1350, Denmark
[4]School of Life Sciences, University of Warwick, Coventry, CV4 7AL, United Kingdom
[5]International Laboratory for Human Genome Research, National Autonomous University of Mexico, Queretaro, 76230, Mexico
[6]Section for GeoGenetics, GLOBE Institute, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, 1350, Denmark
[7]Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen, 28359, Germany
[8]BioArCh, Department of Archaeology, University of York, York, YO10 5DD, United Kingdom
[9]Department of Earth Sciences, Natural History Museum, London, SW7 5BD, United Kingdom
[10]Human Genetics, University of Chicago, Chicago IL, 60637, USA
[11]The Arctic University Museum of Norway, UiT The Arctic University of Norway, Tromsø, 9037, Norway
[12]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, 04103, Germany
[13]Department of Anthropology, Pennsylvania State University, Pennsylvania PA, 16802, USA
[14]Department of Anthropology, McMaster University, Hamilton, L8S4L9, Canada
[15]McMaster Ancient DNA Centre, McMaster University, Hamilton, L8S4L10, Canada
[16]Department of Anthropology, Harvard University, Cambridge MA, 02138, USA
[*]corresponding author(s): James A. Fellows Yates (fellows@shh.mpg.de) and Christina Warinner (warinner@shh.mpg.de)

ABSTRACT

Ancient DNA and RNA are valuable data sources for a wide range of disciplines. Within the field of ancient metagenomics, the number of published genetic datasets has risen dramatically in recent years, and tracking this data for reuse is particularly important for large-scale ecological and evolutionary studies of individual microbial taxa, microbial communities, and metagenomic assemblages. AncientMetagenomeDir (archived at https://doi.org/10.5281/zenodo.3980833) is a collection of indices of published genetic data deriving from ancient microbial samples that provides basic, standardised metadata and accession numbers to allow rapid data retrieval from online repositories. These collections are community-curated and span multiple sub-disciplines in order to ensure adequate breadth and consensus in metadata definitions, as well as longevity of the database. Internal guidelines and automated checks to facilitate compatibility with established sequence-read archives and term-ontologies ensure consistency and interoperability for future meta-analyses. This collection will also assist in standardising metadata reporting for future ancient metagenomic studies.

Background & Summary

A crucial but often overlooked component of scientific reproducibility is the efficient retrieval of sample (meta)data. While the field of ancient DNA (aDNA) has been celebrated for its commitment to making sequencing data available through public archives[1], the retrieval of this data is not always trivial. The field of ancient metagenomics benefits from large sample sizes and the ability to reuse previously published datasets. However, the current absence of standards in basic metadata reporting can make data retrieval tedious and laborious, leading to analysis bottlenecks.

Ancient metagenomics can be broadly defined as the study of the *total* genetic content of temporally-degraded samples[2]. Areas of study that fall under ancient metagenomics include studies of host-associated microbial communities (e.g., ancient microbiome studies of dental calculus or paleofeces[3]), genome reconstruction and analysis of specific microbial taxa (e.g., ancient pathogens[4]), and environmental reconstructions using sedimentary ancient DNA (sedaDNA)[5]. Genetic material obtained from ancient samples has undergone a variety of degradation processes that can cause the original genetic signal to be overwhelmed by modern contamination, requiring large DNA sequencing efforts to detect, quantify, and authenticate the remaining truly aDNA[6,7]. These studies have only become feasible since the development of next-generation sequencing (NGS), which employs massively parallel sequencing to generate large amounts of data that are mostly uploaded to and stored on large generalised archives such as the European Bioinformatic Institute's (EBI) European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena/) or the US National Center for Biotechnology Information (NCBI)'s Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra). However, because these are generalised databases used for many kinds of genetic studies, searching for and identifying ancient metagenomic samples can be difficult and time consuming, partly because of the absence of standardised metadata reporting for ancient metagenomics data. Consequently, researchers must resort to repeated extensive literature searches of heterogeneously reported and inconsistently formatted publications to locate ancient metagenomics datasets. Overcoming the difficulty of finding previously published samples is particularly pertinent in studies of aDNA, as palaeontological and archaeological samples are by their nature limited and avoiding repeated or redundant sampling is a high priority[8].

To address these issues, we established AncientMetagenomeDir, a community-curated collection of annotated sample lists that aims to guide researchers to all published ancient metagenomics-related samples. AncientMetagenomeDir was conceived by members of a recently established international and open community of researchers working in ancient metagenomics (Standards, Precautions and Advances in Ancient Metagenomics, or 'SPAAM' - spaam-workshop.github.io), whose aim is to foster research collaboration and define standards in analysis and reporting. The collection aims to be comprehensive but lightweight, consisting of tab-separated value (TSV) tables for three major sub-disciplines of ancient metagenomics. These tables contain essential, sample-specific information in the field of aDNA studies, including: geographic coordinates, temporal data, sub-discipline-specific critical information, and accession codes of public archives that guides researchers to associated sequencing data (see Methods). Keeping these tables in a simple format, and together with our comprehensive contribution guides, encourages continuous contributions from the community and facilitates usage of the resource by researchers coming from non-computational backgrounds, something common in interdisciplinary fields such as archaeo- and palaeogenetics.

AncientMetagenomeDir is designed to track the development of ancient metagenomics through regular releases. As of release v20.09, this includes 87 publications since 2011, representing 443 ancient host-associated metagenome samples, 269 ancient microbial genomes, and 312 sediment samples (Fig. 1), spanning 49 countries (Fig. 2). We expect AncientMetagenomeDir to deliver three key benefits. First, it will contribute to the longevity of important cultural heritage by guiding future sampling-strategies; reducing the risk of repeated or over-sampling of the same samples or regions. Second, it can form a starting point for the development of software to allow rapid aggregation and field-specific data processing. Finally, as a community-curated resource designed specifically for widespread participation, AncientMetagenomeDir will help the field to define a common standard of metadata reporting (such as with MIXs checklists[9]), facilitating the creation of further rich but consistent sample databases for future researchers.

Methods

Repository Structure

AncientMetagenomeDir[10] is a community-curated set of tables maintained on GitHub, that contains metadata from published ancient metagenomic studies (https://github.com/SPAAM-workshop/AncientMetagenomeDir). While most submissions are made by SPAAM members, anyone with a GitHub account is welcome to propose and/or add publications for inclusion. Submitted publications must be published in a peer-reviewed journal; the purpose of AncientMetagenomeDir is not to act as a quality filter and it currently does not make assessments based on data quality. The tables are formatted as tab-separated value (TSV) files in order to maximize accessibility for all researchers and to allow portability between different data analysis software. Valid samples for inclusion currently fall under three categories: (1) host-associated metagenomes (i.e., host-associated or skeletal material microbiomes), (2) host-associated single genomes (i.e., pathogen or commensal microbial genomes), and

(3) environmental metagenomes (e.g., sedaDNA). In addition, a fourth category is currently planned: (4) anthropogenic metagenomes (e.g., dietary and microbial DNA within pottery crusts, or microbial DNA and handling debris on parchment). To be included, samples must have been sequenced using a shotgun metagenomic or genome-level enrichment approach, and sequence data must be publicly available on an established or stable archive. Publications included in the current release were selected for inclusion based on direct contributions by authors and literature reviews. Publications are initially added as a GitHub 'Issue'. Publications may belong to multiple categories, and the corresponding issue is tagged with relevant category 'labels' to assist with faster evaluation and task distribution.

Data Acquisition

After an Issue (i.e., publication) is suggested, any member of the open SPAAM community can assign themselves to the Issue. The member then creates a git branch off of the main repository, manually extracts the relevant metadata from the given publication, and adds it to the corresponding table (e.g. host associated metagenome, or environmental metagenome). Extensive documentation is available to assist contributors to ensure correct entry of metadata, with one README file per table that contains column definitions and guidelines on how to interpret and record metadata. Extensive documentation on submissions, including instructions on using GitHub, are available via tutorial documents and the associated repository wiki.

The metadata in each table covers five main categories: publication metadata (project key, year, and publication DOI), geographic metadata (site name, coordinates, and country), sample metadata (sample name, material type, and (meta)genome type) and sequencing archive information (archive, sample archive accession ID). Due to inconsistency in the ways metadata are reported in publications and archives, and to maintain concise records we have specified (standardised) approximations for the reporting of sample ages, geographic locations, and archive accessions, following where possible MIxS[9] categories. This approach allows researchers to access sufficiently approximate information during search queries to identify samples of interest (e.g., 3700 BP), which they can subsequently manually check to obtain the exact specifications reported in the original publication (e.g., Late Bronze Age, 3725 +/- 15 BP). Geographic coordinates are restricted to a maximum of three decimals, with fewer decimals indicating location uncertainty (e.g., if a publication only reports a province rather than a specific site). Dates are reported (where possible) as uncalibrated years Before Present (BP, i.e., from 1950), and rounded to the nearest 100 years, due to the range of calculation and reporting methods (radiocarbon dating vs. historical records, calibrated vs. uncalibrated radiocarbon dates, etc.). For sequence accession codes, we opted for using *sample* accession codes rather than direct sequencing data IDs. This is due to the myriad of ways in which data are generated and uploaded to repositories (e.g., one sample accession per sample vs. one sample accession per library; or uploading raw sequencing reads vs. only consensus sequences). We found that in most cases sample accession codes are the most straightforward starting points for data retrieval. However, we did observe errors in some data accessions uploaded to public repositories, such as multiple sample codes assigned to different libraries of the same sample, and insufficient metadata to link accessions to specific samples reported in a study. Overall, we found that heterogeneity in sample (meta)data uploading was a common problem, which highlights the need for improvements in both training and community-agreed standards for data sharing and metadata reporting in public repositories. In addition to metadata recorded across all sample types, we have added table-specific metadata fields to individual categories as required (e.g., species for single genomes and community type for microbiomes). Such fields can be further extended or modified with the agreement of the community.

Data Validation

After all metadata has been added, a contributor makes a Pull Request (PR) into the master branch. Every PR undergoes an automated continuous-integration check via the open-source companion tool AncientMetagenomeDirCheck[11] (https://github.com/SPAAM-workshop/AncientMetagenomeDirCheck, License: GNU GPLv3). In order to ensure consistency within and between metadata fields, this tool checks that the entries for each column match a given regex or category defined in controlled JSON 'enum' lists (stored in an 'assets' directory in the repository). For example, valid country codes are guided by the International Nucleotide Sequence Database Collaboration (INSDC) controlled vocabulary (http://www.insdc.org/country.html), host and microbial species names are defined by the NCBI's Taxonomy database (https://www.ncbi.nlm.nih.gov/taxonomy), and material types are defined by the ontologies listed on the EBI's Ontology Look Up service (https://www.ebi.ac.uk/ols/index) - particularly the Uberon[12] and Envo ontologies[13,14]. These controlled vocabularies, alongside stable linking (via DOIs), ensures reliable querying of the dataset, and allows future expansion to include richer metadata by linking to other databases. Descriptions for the minimum required fields for an AncientMetagenomeDir table are provided in Table 1.

Once automated checks are cleared, a contributor then requests a minimum of one peer-review performed by another member of the SPAAM community. This review involves checking the entered data for consistency against the table's README file and also for accuracy against the original publication. Once automated checks and the peer-review are both passed, the publication's metadata are then added to the master branch and the corresponding Issue is closed. For each added publication, a CHANGELOG is maintained to track the papers included in each release and to record any corrections that may have been

made (e.g., if new radiocarbon dates are published for previously entered samples). The CHANGELOG or Issues pages on GitHub can be consulted to check whether a given publication has already been added (or excluded) from a table.

Data Records

AncientMetagenomeDir (https://github.com/SPAAM-workshop/AncientMetagenomeDir) and AncientMetagenomeDirCheck (https://github.com/SPAAM-workshop/AncientMetagenomeDirCheck) are both maintained on GitHub. AncientMetagenomeDir has regular periodic releases, each of which has a release-specific DOI assigned via the Zenodo long-term data repository. Both the collection and tools are archived in the Zenodo repository with generalised DOIs: https://doi.org/10.5281/zenodo.3980833 and https://doi.org/10.5281/zenodo.4003826 respectively. The full workflow can be seen in Figure 3.

Technical Validation

All data entries to AncientMetagenomeDir undergo automated continuous-integration validation prior to submission into the protected main branch. These tests must pass before being additionally peer-reviewed by other member(s) of the community. Validation tests consist of regex patterns to control formatting of specified fields (e.g. DOIs, project IDs, date formats), and cross-checking of entries against controlled vocabularies defined in centralised JSON-format enum lists. Entries must also have valid sample accession IDs corresponding to shotgun metagenomic or genome-enriched sequence data uploaded to established and stable public archives.

Usage Notes

Usage of the resource typically consists of copying or downloading the TSV file of interest for subsequent analysis using software such as Microsoft Excel, LibreOffice Calc, or R. The data table can be subsequently sorted or queried to identify datasets of interest. It should be noted that certain metadata fields (e.g., sample_age, latitude, and longitude) are approximate and do not provide *exact* values; rather, if exact values for these fields are required, they must be retrieved from the original publication. All selected data retrieved using AncientMetagenomeDir and used in subsequent studies should be cited using the original publication citation as well as AncientMetagenomeDir.

Retrieval of sequencing data using sample accession codes can be achieved manually via a given archive's website, or via archive-supplied tools (e.g., Entrez Programming Utilties for NCBI's SRA (https://github.com/enasequence/enaBrowserTools), or enaBrowserTools for EBI ENA (https://github.com/enasequence/enaBrowserTools).

Contributions to the tables are also facilitated by extensive step-by-step documentation on how to use GitHub and AncientMetagenomeDir, the locations of which are listed on the main README of the repository and the associated wiki page.

Code availability

R notebooks used for generating images can be found at 10.5281/zenodo.4011751. Code for validation of the dataset (with version 1 used for the first release of AncientMetagenomeDir) can be found at https://github.com/SPAAM-workshop/AncientMetagenomeDirChecl and https://doi.org/10.5281/zenodo.4003826.

References

1. Anagnostou, P. *et al.* When data sharing gets close to 100%: what human paleogenetics can teach the open science movement. *PloS one* **10**, e0121409, 10.1371/journal.pone.0121409 (2015).

2. Warinner, C. *et al.* A robust framework for microbial archaeology. *Annu. review genomics human genetics* **18**, 321–356, 10.1146/annurev-genom-091416-035526 (2017).

3. Warinner, C., Speller, C., Collins, M. J. & Lewis, C. M., Jr. Ancient human microbiomes. *J. human evolution* **79**, 125–136, 10.1016/j.jhevol.2014.10.016 (2015).

4. Spyrou, M. A., Bos, K. I., Herbig, A. & Krause, J. Ancient pathogen genomics as an emerging tool for infectious disease research. *Nat. reviews. Genet.* **20**, 323–340, 10.1038/s41576-019-0119-1 (2019).

5. Edwards, M. E. The maturing relationship between quaternary paleoecology and ancient sedimentary DNA. *Quat. Res.* **96**, 39–47, 10.1017/qua.2020.52 (2020).

6. Dabney, J., Meyer, M. & Pääbo, S. Ancient DNA damage. *Cold Spring Harb. perspectives biology* **5**, 10.1101/cshperspect.a012567 (2013).

7. Peyrégne, S. & Prüfer, K. Present-Day DNA contamination in ancient DNA datasets. *BioEssays: news reviews molecular, cellular developmental biology* e2000081, 10.1002/bies.202000081 (2020).

8. Prendergast, M. E. & Sawchuk, E. Boots on the ground in africa's ancient DNA 'revolution': archaeological perspectives on ethics and best practices. *Antiquity* **92**, 803–815, 10.15184/aqy.2018.70 (2018).

9. Yilmaz, P. *et al.* Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. biotechnology* **29**, 415–420, 10.1038/nbt.1823 (2011).

10. FellowsYates, J. A. *et al.* Spaam-workshop/ancientmetagenomedir: v20.09.1: Ancient ksour of ouadane. *Zenodo* https://doi.org/10.5281/zenodo.4011751 (2020).

11. Borry, M. & Yates, J. A. F. Spaam-workshop/ancientmetagenomedircheck: Ancientmetagenomedircheck v1.0 (version 1.0). *Zenodo* https://doi.org/10.5281/zenodo.4003826 (2020).

12. Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome biology* **13**, R5, 10.1186/gb-2012-13-1-r5 (2012).

13. Buttigieg, P. L. *et al.* The environment ontology: contextualising biological and biomedical entities. *J. biomedical semantics* **4**, 43, 10.1186/2041-1480-4-43 (2013).

14. Buttigieg, P. L. *et al.* The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *J. biomedical semantics* **7**, 57, 10.1186/s13326-016-0097-6 (2016).

Acknowledgements

We would like to thank the wider SPAAM community (spaam-workshop.github.io) for their input in developing the project. J.A.F.Y., A.A.V., I.V., M.B. A.H. and C.W. acknowledge the Max Planck Society for financial support. J.A.F.Y. is partly supported by grant ERC-2015-StG 678901-FoodTransforms (to Philipp W. Stockhammer, Ludwig Maximilian University, Germany). B.C. is supported by grant ERC-2014-ADG 670518 (to V. Gaffney, University of Bradford, United Kingdom). A.J.V. is supported by Carlsbergfondet Semper Ardens grant CF18-1109 (to M. Thomas P. Gilbert, University of Copenhagen, Denmark). A.H. is partly supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2051—Project-ID 390713860 (to C. Warinner, Friedrich Schiller University, Germany). E.J.G is supported by Arts & Humanities Research Council (grant number AH/N005015/1) and Natural History Museum (London, United Kingdom). M.J.B.-L. is supported by grant Wellcome Trust Seed Award in Science 208934/Z/17/Z, and by project IA201219 PAPIIT-DGAPA- UNAM (to María Ávila Arcos, LIIGH, Mexico). M.S. is supported by grant ERC-CoG 771234 PALEoRIDER (to Wolfgang Haak, Max-Planck-Institute for the Science of Human Hisory, Germany). A.S.G is supported by NSF GRFP Grant No. DGE1255832 (any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation). S.L.R. is supported by NIH Genetics and Regulation Training Grant 5T32GM007197-46. J.H. is supported by the Social Sciences and Humanities Research Council (Canada). I.V., M.B., and C.W. are supported by Werner Siemens Stiftung (Paleobiochemistry) (to C. Warinner, Leibnitz Institute for Natural Product Research and Infection Biology, Germany).

Author contributions statement

J.A.F.Y and C.W. conceptualised the project. J.A.F.Y designed the project and infrastructure with input from all co-authors. M.B. developed software. J.A.F.Y., A.A.V., I.M.V., B.C., A.J.V., M.J.B-L., A.F.-G., E.J.G., S.L.R., P.D.H., M.A.S., A.H., A.S.G., J.H., A.F.A., and C.W. acquired data. J.A.F.Y drafted the manuscript with input from all co-authors.

Competing interests

The authors declare no competing interests.

Figures & Tables

Field	Description	Field Type	Field Format
project_name	Unique AncientMetagenomeDir key for study	String	FirstAuthorYYYY
publication_year	Publication year of study	Integer	YYYY
publication_doi	Publication DOI (or library permalink)	String	Regex
site_name	Specific locality name where sample taken from	String	Free text
latitude	Latitude in decimal coordinate (WGS84 projection)	Number	Max. 3 decimals
longitude	Longitude in decimal coordinate (WGS84 projection)	Number	Max. 3 decimals
geo_loc_name	Present-day country name that locality resides in	String	Restricted enum
sample_name	Name of sample as reported in publication or archive	String	Free text
sample_age	Approximate date (before 1950, rounded to last 100 years)	Integer	YYYYY
sample_age_doi	DOI of source of date. Can be more recent publication.	String	Regex
collection_date	Date sample was taken for genetic analysis	Integer	YYYY
archive	Name of established data repository	String	Restricted enum
archive_accession	Sample-level accession code in data repository	String	Free text

Table 1. Core fields that are required for all AncientMetagenomeDir sub-discipline tables, including field type and standardised formatting description. Field formats are defined in a JSON schema, against which each new study submission is cross-checked. Further sub-discipline specific fields are included in the corresponding table, as required by the community.

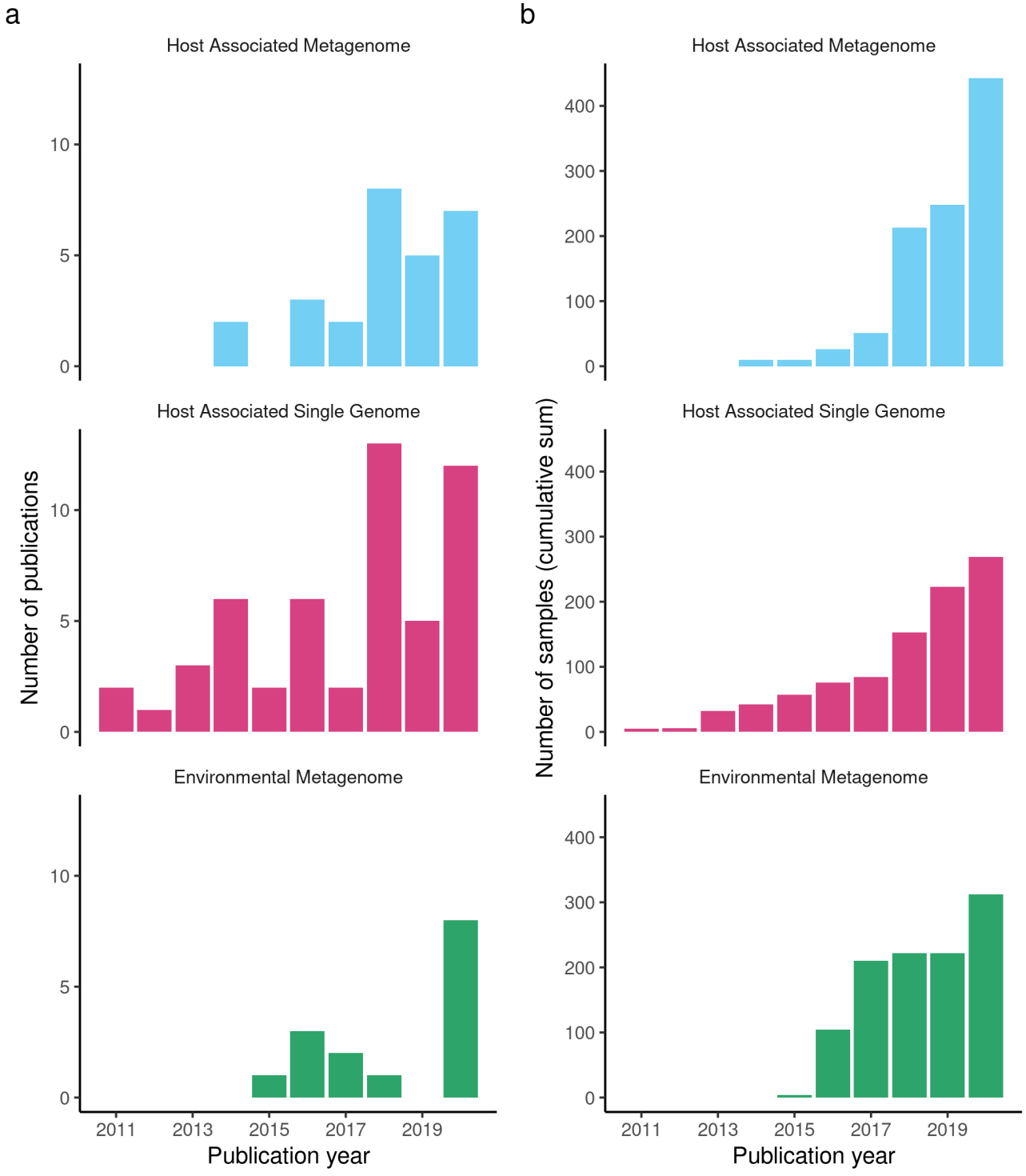

Figure 1. Timelines depicting the development of sub-fields of ancient metagenomics as recorded in AncientMetagenomedir as per release v20.09. (**a**) Number of ancient metagenomic publications per year. (**b**) Cumulative sum of published samples with genetic sequencing data or sequences in publicly accessible archives.

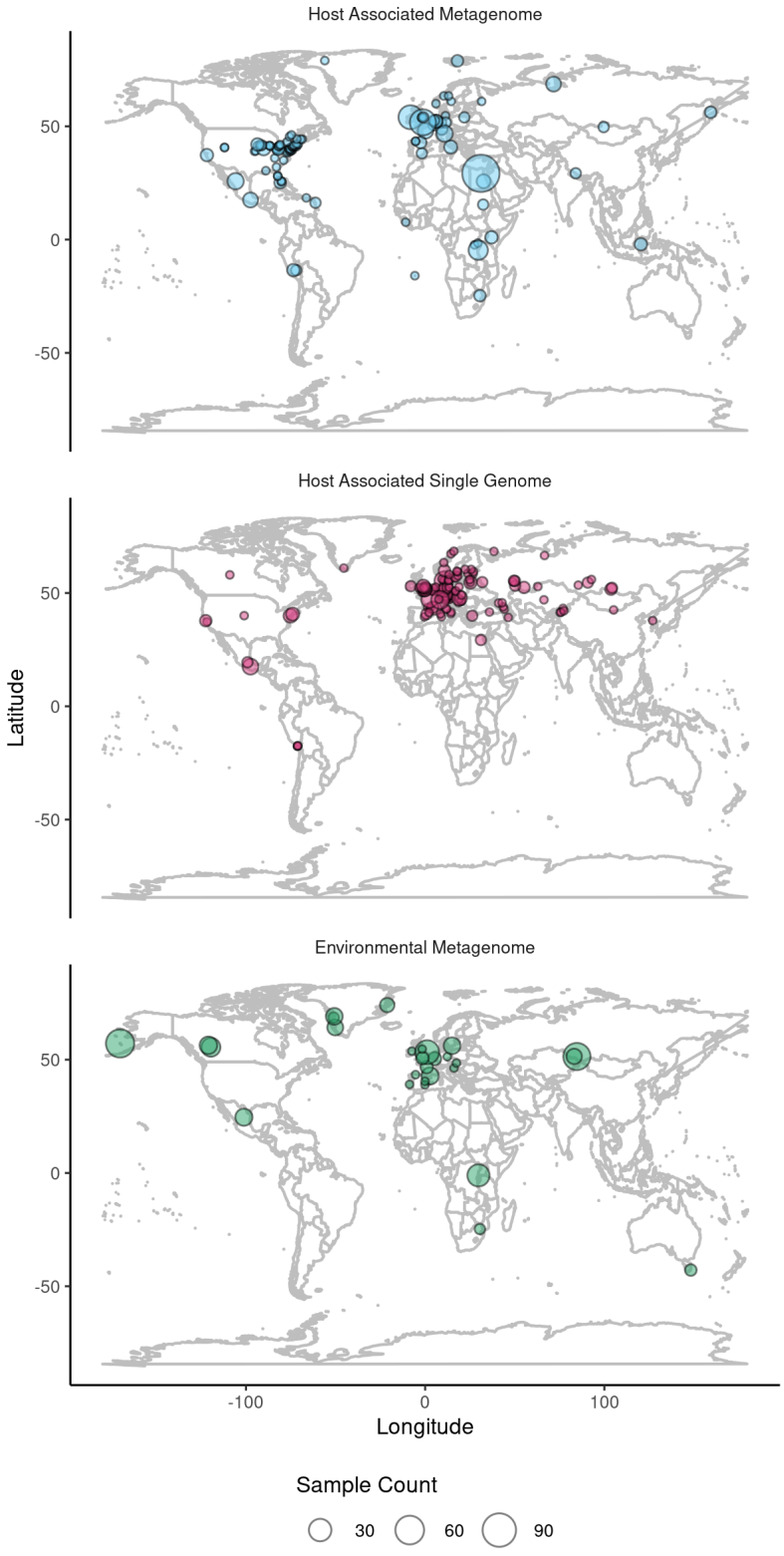

Figure 2. Maps depicting the geographic spread of samples with latitude and longitude information, between sub-fields of ancient metagenomics as recorded in AncientMetagenomeDir as per release v20.09.

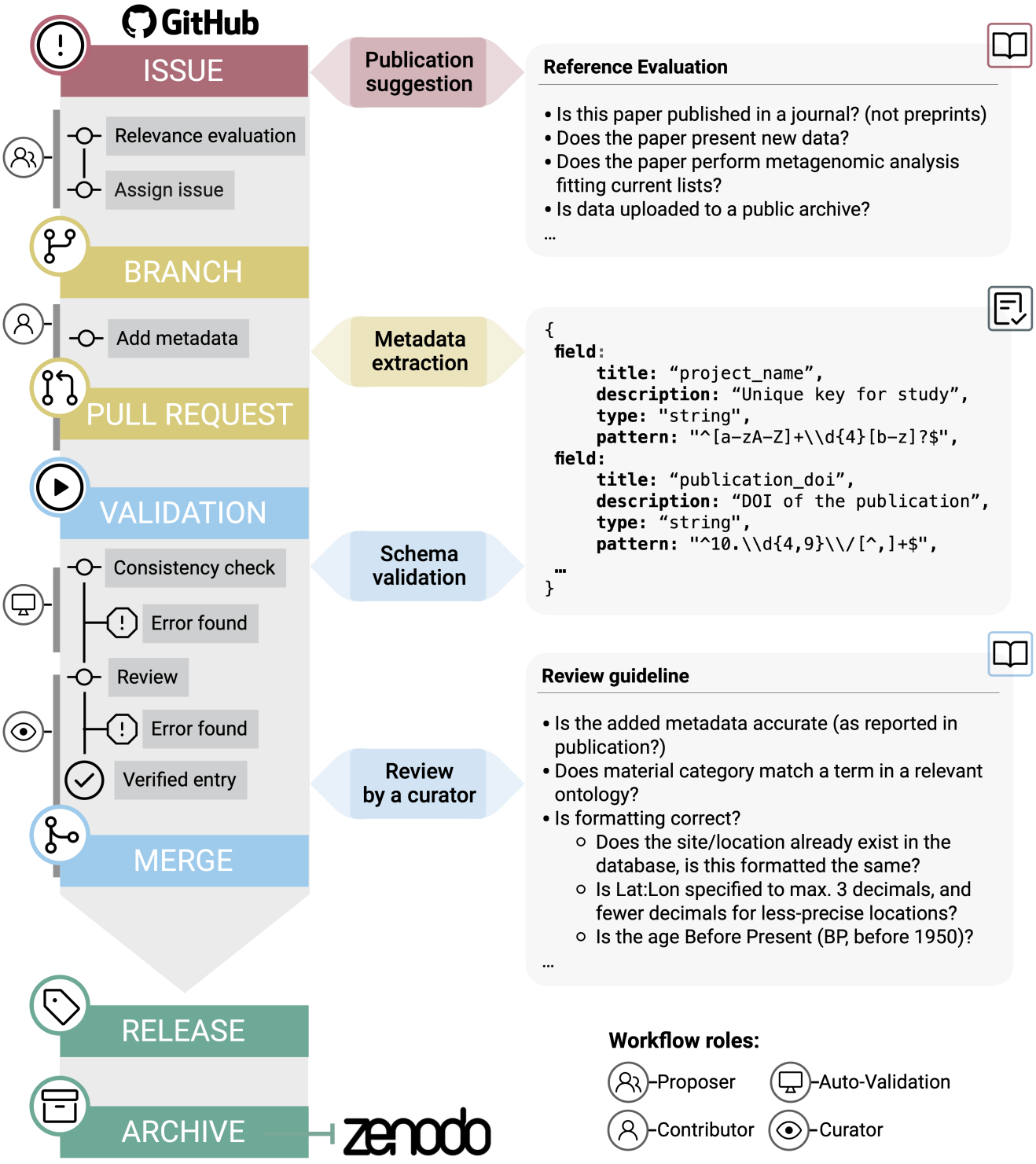

Figure 3. AncientMetagenomeDir submission and update workflow. The submission workflow is carried out on GitHub, and final releases archived at Zenodo. Submissions go through both automated computational validation and also human peer-review for consistency and accuracy.