

# 1 Community-curated and standardised metadata of 2 published ancient metagenomic samples with 3 AncientMetagenomeDir

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

9 <sup>1</sup>Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, 07745, Jena,  
10 Germany

11 <sup>2</sup>Institut für Vor- und Frühgeschichtliche Archäologie und Provinzialrömische Archäologie, Ludwig Maximilian  
12 University, München, 80539, Germany

13 <sup>3</sup>Section for Evolutionary Genomics, GLOBE Institute, Faculty of Health and Medical Sciences, University of  
14 Copenhagen, Copenhagen, 1350, Denmark

15 <sup>4</sup>School of Life Sciences, University of Warwick, Coventry, CV4 7AL, United Kingdom

16 <sup>5</sup>International Laboratory for Human Genome Research, National Autonomous University of Mexico, Queretaro,  
17 76230, Mexico

18 <sup>6</sup>Section for GeoGenetics, GLOBE Institute, Faculty of Health and Medical Sciences, University of Copenhagen,  
19 Copenhagen, 1350, Denmark

20 <sup>7</sup>Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen,  
21 28359, Germany

22 <sup>8</sup>BioArCh, Department of Archaeology, University of York, York, YO10 5DD, United Kingdom

23 <sup>9</sup>Department of Earth Sciences, Natural History Museum, London, SW7 5BD, United Kingdom

24 <sup>10</sup>Human Genetics, University of Chicago, Chicago IL, 60637, USA

25 <sup>11</sup>The Arctic University Museum of Norway, UiT The Arctic University of Norway, Tromsø, 9037, Norway

26 <sup>12</sup>Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, 04103,  
27 Germany

28 <sup>13</sup>Department of Anthropology, Pennsylvania State University, Pennsylvania PA, 16802, USA

29 <sup>14</sup>Department of Anthropology, McMaster University, Hamilton, L8S4L9, Canada

30 <sup>15</sup>McMaster Ancient DNA Centre, McMaster University, Hamilton, L8S4L10, Canada

31 <sup>16</sup>Department of Anthropology, Harvard University, Cambridge MA, 02138, USA

32 \*corresponding author(s): James A. Fellows Yates ([fellows@shh.mpg.de](mailto:fellows@shh.mpg.de)) and Christina Warinner  
33 ([warinner@shh.mpg.de](mailto:warinner@shh.mpg.de))

## 34 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  
35 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.

## 36 Background & Summary

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

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

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

70 AncientMetagenomeDir is designed to track the development of ancient metagenomics through regular releases. As  
71 of release v20.09, this includes 87 publications since 2011, representing 443 ancient host-associated metagenome samples,  
72 269 ancient microbial genomes, and 312 sediment samples (Fig. 1), spanning 49 countries (Fig. 2). We expect Ancient-  
73 MetagenomeDir to deliver three key benefits. First, it will contribute to the longevity of important cultural heritage by guiding  
74 future sampling-strategies; reducing the risk of repeated or over-sampling of the same samples or regions. Second, it can form  
75 a starting point for the development of software to allow rapid aggregation and field-specific data processing. Finally, as a  
76 community-curated resource designed specifically for widespread participation, AncientMetagenomeDir will help the field to  
77 define a common standard of metadata reporting (such as with MIXs checklists<sup>9</sup>), facilitating the creation of further rich but  
78 consistent sample databases for future researchers.

## 79 Methods

### 80 Repository Structure

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

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

## 96 **Data Acquisition**

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

103 The metadata in each table covers five main categories: publication metadata (project key, year, and publication DOI),  
104 geographic metadata (site name, coordinates, and country), sample metadata (sample name, material type, and (meta)genome  
105 type) and sequencing archive information (archive, sample archive accession ID). Due to inconsistency in the ways metadata  
106 are reported in publications and archives, and to maintain concise records we have specified (standardised) approximations  
107 for the reporting of sample ages, geographic locations, and archive accessions, following where possible MIXS<sup>9</sup> categories.  
108 This approach allows researchers to access sufficiently approximate information during search queries to identify samples of  
109 interest (e.g., 3700 BP), which they can subsequently manually check to obtain the exact specifications reported in the original  
110 publication (e.g., Late Bronze Age, 3725 +/- 15 BP). Geographic coordinates are restricted to a maximum of three decimals,  
111 with fewer decimals indicating location uncertainty (e.g., if a publication only reports a province rather than a specific site).  
112 Dates are reported (where possible) as uncalibrated years Before Present (BP, i.e., from 1950), and rounded to the nearest  
113 100 years, due to the range of calculation and reporting methods (radiocarbon dating vs. historical records, calibrated vs.  
114 uncalibrated radiocarbon dates, etc.). For sequence accession codes, we opted for using *sample* accession codes rather than  
115 direct sequencing data IDs. This is due to the myriad of ways in which data are generated and uploaded to repositories (e.g.,  
116 one sample accession per sample vs. one sample accession per library; or uploading raw sequencing reads vs. only consensus  
117 sequences). We found that in most cases sample accession codes are the most straightforward starting points for data retrieval.  
118 However, we did observe errors in some data accessions uploaded to public repositories, such as multiple sample codes assigned  
119 to different libraries of the same sample, and insufficient metadata to link accessions to specific samples reported in a study.  
120 Overall, we found that heterogeneity in sample (meta)data uploading was a common problem, which highlights the need for  
121 improvements in both training and community-agreed standards for data sharing and metadata reporting in public repositories.  
122 In addition to metadata recorded across all sample types, we have added table-specific metadata fields to individual categories  
123 as required (e.g., species for single genomes and community type for microbiomes). Such fields can be further extended or  
124 modified with the agreement of the community.

## 125 **Data Validation**

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

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

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

## 144 Data Records

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

## 150 Technical Validation

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

## 157 Usage Notes

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

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

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

## 169 Code availability

170 R notebooks used for generating images can be found at [10.5281/zenodo.4011751](https://doi.org/10.5281/zenodo.4011751). Code for validation of the dataset (with ver-  
171 sion 1 used for the first release of AncientMetagenomeDir) can be found at <https://github.com/SPAAM-workshop/AncientMetagenomeDirCheck>  
172 and <https://doi.org/10.5281/zenodo.4003826>.

## 173 References

- 174 1. Anagnostou, P. *et al.* When data sharing gets close to 100%: what human paleogenetics can teach the open science  
175 movement. *PLoS one* **10**, e0121409, [10.1371/journal.pone.0121409](https://doi.org/10.1371/journal.pone.0121409) (2015).
- 176 2. Warinner, C. *et al.* A robust framework for microbial archaeology. *Annu. review genomics human genetics* **18**, 321–356,  
177 [10.1146/annurev-genom-091416-035526](https://doi.org/10.1146/annurev-genom-091416-035526) (2017).
- 178 3. Warinner, C., Speller, C., Collins, M. J. & Lewis, C. M., Jr. Ancient human microbiomes. *J. human evolution* **79**, 125–136,  
179 [10.1016/j.jhevol.2014.10.016](https://doi.org/10.1016/j.jhevol.2014.10.016) (2015).
- 180 4. Spyrou, M. A., Bos, K. I., Herbig, A. & Krause, J. Ancient pathogen genomics as an emerging tool for infectious disease  
181 research. *Nat. reviews. Genet.* **20**, 323–340, [10.1038/s41576-019-0119-1](https://doi.org/10.1038/s41576-019-0119-1) (2019).
- 182 5. Edwards, M. E. The maturing relationship between quaternary paleoecology and ancient sedimentary DNA. *Quat. Res.* **96**,  
183 39–47, [10.1017/qua.2020.52](https://doi.org/10.1017/qua.2020.52) (2020).
- 184 6. Dabney, J., Meyer, M. & Pääbo, S. Ancient DNA damage. *Cold Spring Harb. perspectives biology* **5**, [10.1101/cshperspect.  
185 a012567](https://doi.org/10.1101/cshperspect.a012567) (2013).

- 186 7. Peyrégne, S. & Prüfer, K. Present-Day DNA contamination in ancient DNA datasets. *BioEssays: news reviews molecular,*  
187 *cellular developmental biology* e2000081, [10.1002/bies.202000081](https://doi.org/10.1002/bies.202000081) (2020).
- 188 8. Prendergast, M. E. & Sawchuk, E. Boots on the ground in africa’s ancient DNA ‘revolution’: archaeological perspectives  
189 on ethics and best practices. *Antiquity* **92**, 803–815, [10.15184/aqy.2018.70](https://doi.org/10.15184/aqy.2018.70) (2018).
- 190 9. Yilmaz, P. *et al.* Minimum information about a marker gene sequence (MIMARKS) and minimum information about any  
191 (x) sequence (MIxS) specifications. *Nat. biotechnology* **29**, 415–420, [10.1038/nbt.1823](https://doi.org/10.1038/nbt.1823) (2011).
- 192 10. FellowsYates, J. A. *et al.* Spaam-workshop/ancientmetagenomedir: v20.09.1: Ancient ksour of ouadane. *Zenodo*  
193 <https://doi.org/10.5281/zenodo.4011751> (2020).
- 194 11. Borry, M. & Yates, J. A. F. Spaam-workshop/ancientmetagenomedircheck: Ancientmetagenomedircheck v1.0 (version  
195 1.0). *Zenodo* <https://doi.org/10.5281/zenodo.4003826> (2020).
- 196 12. Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy  
197 ontology. *Genome biology* **13**, R5, [10.1186/gb-2012-13-1-r5](https://doi.org/10.1186/gb-2012-13-1-r5) (2012).
- 198 13. Buttigieg, P. L. *et al.* The environment ontology: contextualising biological and biomedical entities. *J. biomedical*  
199 *semantics* **4**, 43, [10.1186/2041-1480-4-43](https://doi.org/10.1186/2041-1480-4-43) (2013).
- 200 14. Buttigieg, P. L. *et al.* The environment ontology in 2016: bridging domains with increased scope, semantic density, and  
201 interoperation. *J. biomedical semantics* **7**, 57, [10.1186/s13326-016-0097-6](https://doi.org/10.1186/s13326-016-0097-6) (2016).

## 202 Acknowledgements

203 We would like to thank the wider SPAAM community ([spaaam-workshop.github.io](https://spaaam-workshop.github.io)) for their input in developing the project.  
204 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  
205 supported by grant ERC-2015-StG 678901-FoodTransforms (to Philipp W. Stockhammer, Ludwig Maximilian University,  
206 Germany). B.C. is supported by grant ERC-2014-ADG 670518 (to V. Gaffney, University of Bradford, United Kingdom).  
207 A.J.V. is supported by Carlsbergfondet Semper Ardens grant CF18-1109 (to M. Thomas P. Gilbert, University of Copenhagen,  
208 Denmark). A.H. is partly supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under  
209 Germany’s Excellence Strategy—EXC 2051—Project-ID 390713860 (to C. Warinner, Friedrich Schiller University, Germany).  
210 E.J.G. is supported by Arts & Humanities Research Council (grant number AH/N005015/1) and Natural History Museum  
211 (London, United Kingdom). M.J.B.-L. is supported by grant Wellcome Trust Seed Award in Science 208934/Z/17/Z, and by  
212 project IA201219 PAPIIT-DGAPA- UNAM (to María Ávila Arcos, LIIGH, Mexico). M.S. is supported by grant ERC-CoG  
213 771234 PALEORIDER (to Wolfgang Haak, Max-Planck-Institute for the Science of Human History, Germany). A.S.G. is  
214 supported by NSF GRFP Grant No. DGE1255832 (any opinions, findings, and conclusions or recommendations expressed in  
215 this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation). S.L.R. is  
216 supported by NIH Genetics and Regulation Training Grant 5T32GM007197-46. J.H. is supported by the Social Sciences and  
217 Humanities Research Council (Canada). I.V., M.B., and C.W. are supported by Werner Siemens Stiftung (Paleobiochemistry)  
218 (to C. Warinner, Leibniz Institute for Natural Product Research and Infection Biology, Germany).

## 219 Author contributions statement

220 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.  
221 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.,  
222 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.

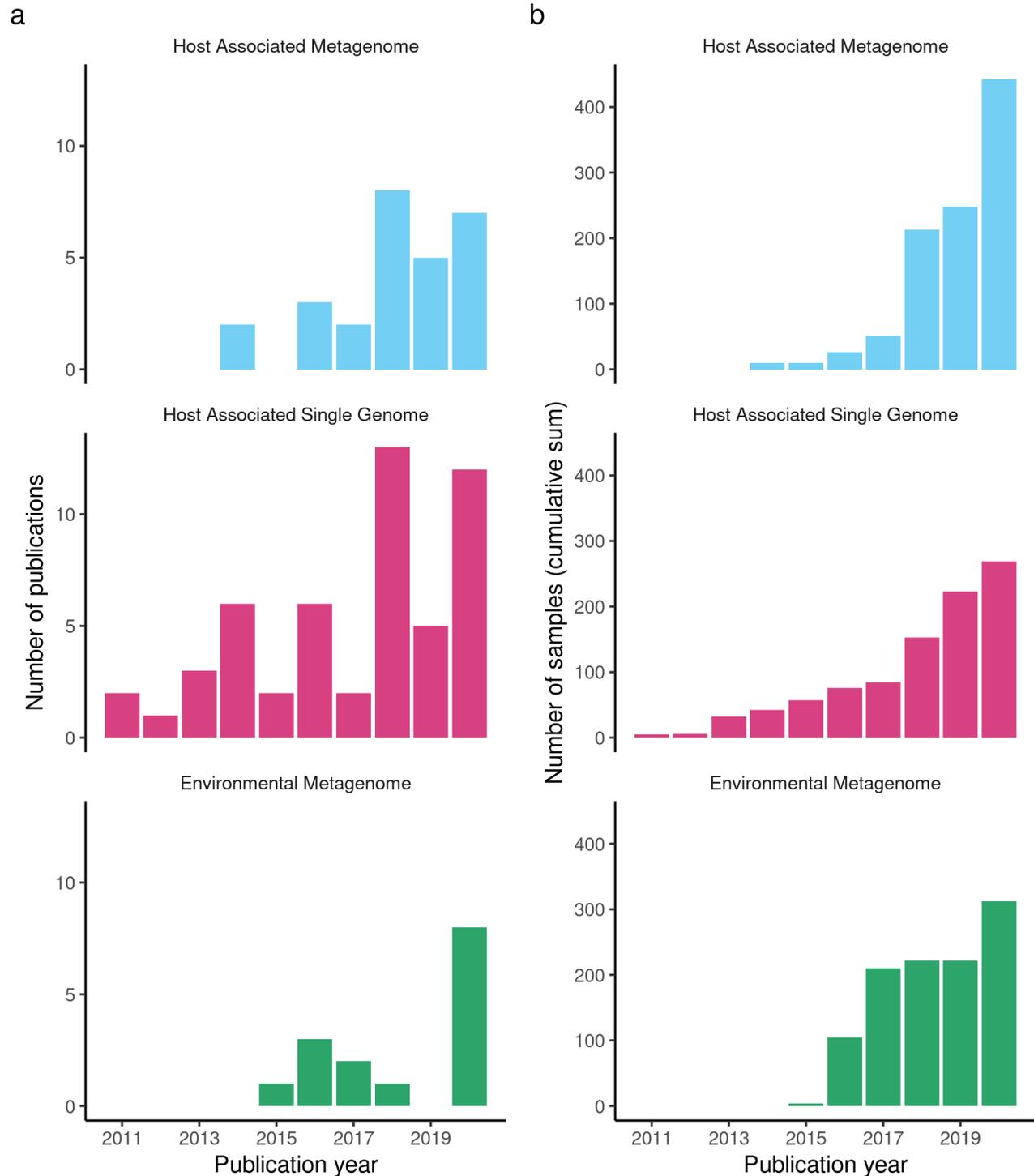
## 223 Competing interests

224 The authors declare no competing interests.

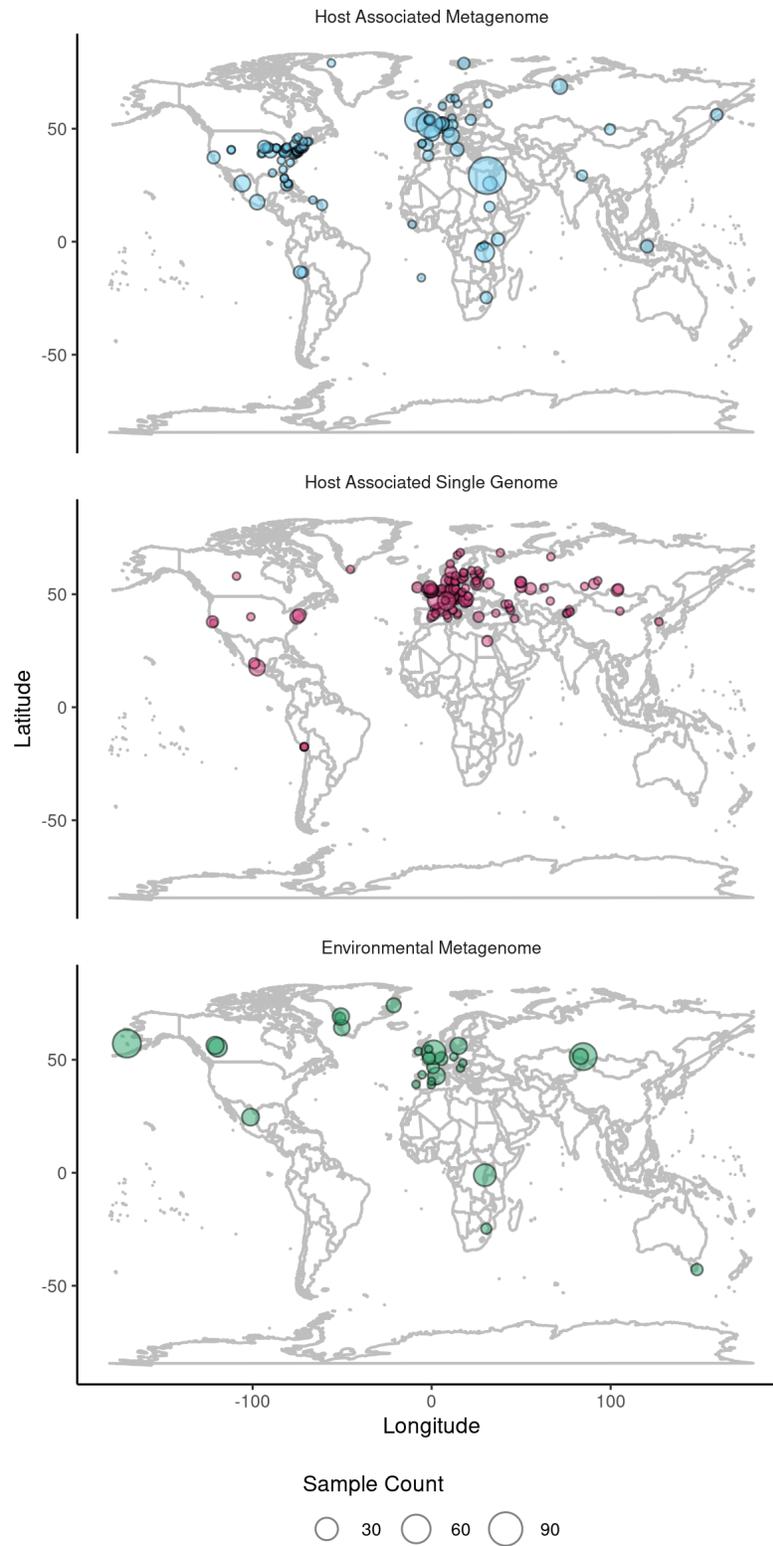
## 225 Figures & Tables

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

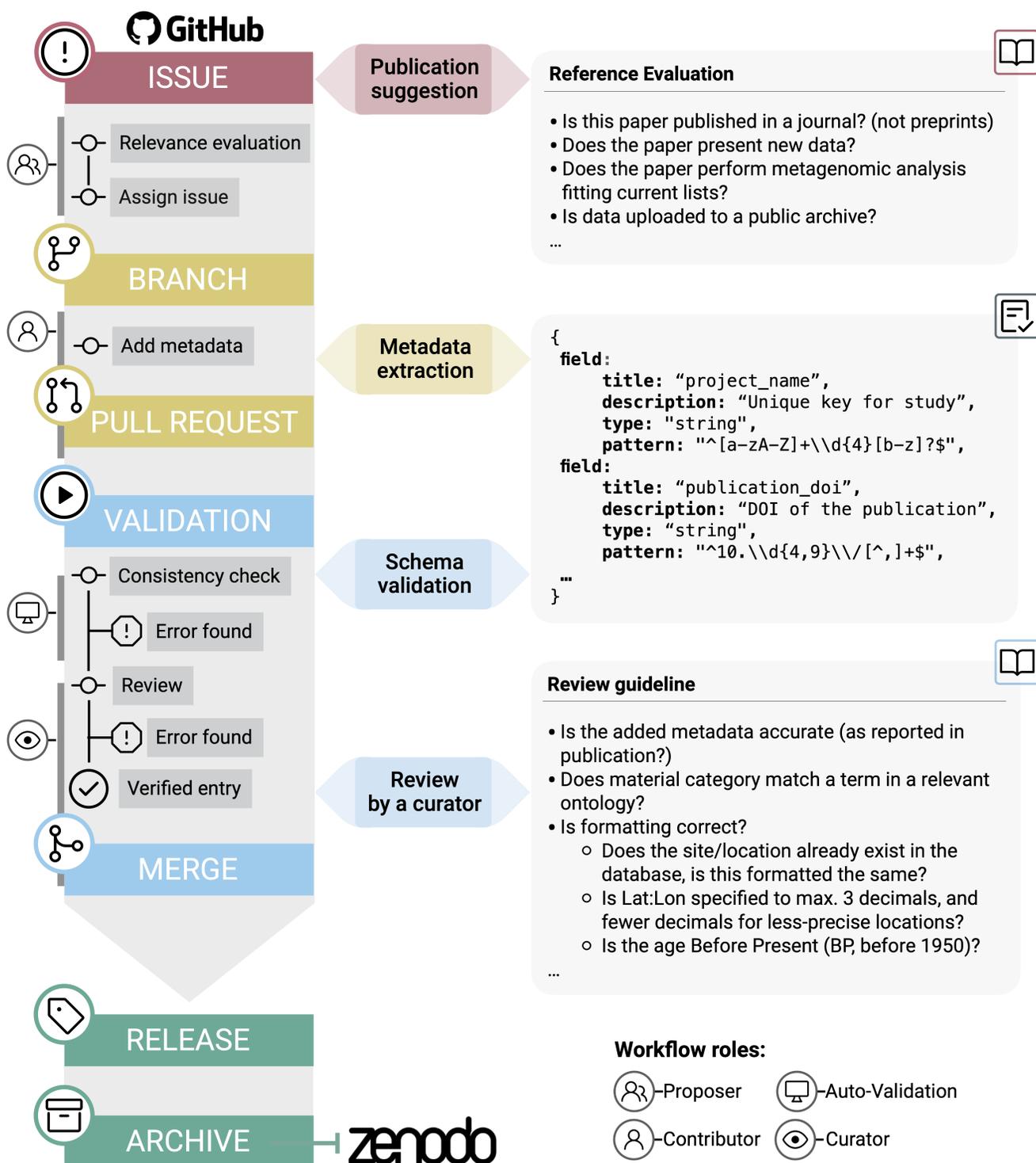
**Table 1.** Core fields that are required for all AncientMetagenomeDir sub-discipline tables, including field type and standardised formatting description. Field formats are defined in a JSON schema, against which each new study submission is cross-checked. Further sub-discipline specific fields are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