SOFTWARE TOOL ARTICLE

Facilitating accessible, rapid, and appropriate processing of ancient metagenomic data with AMDirT [version 1; peer review: 1 approved, 1 approved with reservations]

Maxime Borry [ID][1,2], Adrian Forsythe[3], Aida Andrades Valtueña[2], Alexander Hübner[1,2,4], Anan Ibrahim [ID][5], Andrea Quagliariello[6], Anna E. White [ID][7,8], Arthur Kocher[2,9], Åshild J. Vågene[10], Bjørn Peare Bartholdy [ID][11], Diāna Spurīte[2,12], Gabriel Yaxal Ponce-Soto[13], Gunnar Neumann[2,14], I-Ting Huang[15], Ian Light[16], Irina M. Velsko[2], Iseult Jackson [ID][17,18], Jasmin Frangenberg [ID][5], Javier G. Serrano[19], Julien Fumey [ID][13,20], Kadir T. Özdoğan[21,22], Kelly E. Blevins[23,24], Kevin G. Daly[18], Maria Lopopolo [ID][13], Markella Moraitou[25], Megan Michel[2,14,26], Meriam van Os [ID][27], Miriam J. Bravo-Lopez[28,29], Mohamed S. Sarhan [ID][30,31], Nihan D. Dagtas[32], Nikolay Oskolkov [ID][33,34], Olivia S. Smith[35], Ophélie Lebrasseur[36,37], Piotr Rozwalak[38], Raphael Eisenhofer[39], Sally Wasef [ID][40], Shreya L. Ramachandran [ID][41], Valentina Vanghi[32], Christina Warinner [ID][2,4,42,43], James A. Fellows Yates [ID][2,4,5]

[1]Cluster of Excellence "Balance of the Microverse", Leibniz Institute for Natural Product Research and Infection Biology Hans Knöll Institute, Adolf-Reichwein-Straße 23, Jena, Thuringia, 07745, Germany
[2]Department of Archaeogenetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Pl. 6, Leipzig, Saxony, 04103, Germany
[3]Department of Animal Zoology, Uppsala Universitet, Norbyvägen 18D, Uppsala, 752 36, Sweden
[4]Associated Research Group of Archaeogenetics, Leibniz Institute for Natural Product Research and Infection Biology Hans Knöll Institute, Adolf-Reichwein-Straße 23, Jena, Thuringia, 07745, Germany
[5]Department of Paleobiotechnology, Leibniz Institute for Natural Product Research and Infection Biology Hans Knöll Institute, Adolf-Reichwein-Straße 23, Jena, Thuringia, 07745, Germany
[6]Department of Comparative Biomedicine and Food Science, Universita degli Studi di Padova, Viale dell'Università 16, Legnaro, Padova, 350250, Italy
[7]Section for Molecular Ecology and Evolution, Globe Institute, Faculty of Health and Medical Sciences, Københavns Universitet, Øster Farimagsgade 5, Copenhagen K, 1353, Denmark
[8]BioArCh, Department of Archaeology, University of York, York, England, YO10 5DD, UK
[9]Transmission, Infection, Diversification and Evolution Group, Max Planck Institute for Geoanthropology, Kahlaische Str. 10, Jena, Thuringia, 07745, Germany
[10]Section for Hologenomics, Globe Institute, Faculty of Health and Medical Sciences, Københavns Universitet, Oester Voldgade 44747, Copenhagen K, 1350, Denmark
[11]Department of Archaeological Sciences, Universiteit Leiden, Einsteinweg 2, Leiden, 2333 CC, The Netherlands
[12]Institute of Ecology and Evolution, Friedrich-Schiller-Universität Jena, Jena, Thuringia, 07743, Germany
[13]Microbial Paleogenomics Unit, Institut Pasteur, Universite Paris Cite, CNRS UMR 2000, Rue du Docteur Roux 25-28, Paris, Île-de-France, F-75015, France
[14]Max Planck-Harvard Research Center for the Archaeoscience of the Ancient Mediterranean (MHAAM), Max Planck Institute for Evolutionary Anthropology, Deutscher Pl. 6, Leipzig, Saxony, 04103, Germany

[15]Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford St., Cambridge, Massachusetts, 02138, USA

[16]Max Planck Institute for Infection Biology, Virchowweg 12, Berlin, Berlin, 10117, Germany

[17]SFI Centre for Research Training in Genomics Data Science, University of Galway, Galway, H91 TK33, Ireland

[18]Smurfit Institute of Genetics, The University of Dublin Trinity College, Dublin, Leinster, D02 VF25, Ireland

[19]Department of Biochemistry, Microbiology, Cell Biology and Genetics, Universidad de La Laguna, San Cristóbal de La Laguna, Santa Cruz de Tenerife, 38200, Spain

[20]Bioinformatics and Biostatistics Hub, Institut Pasteur, Université Paris Cité, Rue du Docteur Roux 25-28, Paris, Île-de-France, F-75015, France

[21]Animal Ecology, Wageningen Environmental Research, P.O box 47, Wageningen, Gelderland, 6700 AA, The Netherlands

[22]Department of History and Art History, Universiteit Utrecht, Drift 6, Utrecht, Utrecht, 3512 BS, The Netherlands

[23]Center for Bioarchaeological Research, Arizona State University, Candy Mall, Tempe, Arizona, 85281, USA

[24]Department of Archaeology, Durham University, South Road, Durham, County Durham, England, DH1 3LE, UK

[25]Institute of Ecology and Evolution, School of Biological Sciences, The University of Edinburgh, Charlotte Auerbach Road, Edinburgh, Scotland, EH9 3FL, UK

[26]Department of Human Evolutionary Biology, Harvard University, Divinity Avenue 11, Cambridge, Massachusetts, 02138, USA

[27]Department of Anatomy, University of Otago, 270 Great King St, Dunedin, Otago, 9016, New Zealand

[28]International Laboratory for Human Genome Research (LIIGH), Universidad Nacional Autonoma de Mexico, La Mesa 3001, Juriquilla, Queretaro, 76230, Mexico

[29]Center for Genomic Sciences (CCG), Universidad Nacional Autonoma de Mexico, Cuernavaca, Morelos, 62210, Mexico

[30]Institute for Mummy Studies, Eurac Research, Drususallee 1, Bolzano/Bozen, Autonome Provinz Bozen, 39100, Italy

[31]Centre for Integrative Biology (CIBIO), Universita degli Studi di Trento, Via Sommarive 9, Povo, Trentino, 38123, Italy

[32]Department of Anatomy and Anthropology and Department of Human Molecular Genetics and Biochemistry, Sackler Faculty of Medicine, Tel Aviv University, Ramat Aviv, Tel Aviv-Yafo, 69978, Israel

[33]National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Tomtebodavägen 23, Stockholm, 17165, Sweden

[34]Department of Biology, Lunds Universitet, Sölvegatan 35, Lund, 223 62, Sweden

[35]Department of Integrative Biology, The University of Texas at Austin, Speedway 2415, Austin, Texas, 78712, USA

[36]Instituto Nacional de Antropología y Pensamiento Latinoamericano, 3 de Febrero 1370 (1426), Ciudad Autónoma de Buenos Aires, C1426BJN CABA, Argentina

[37]Center for Anthropobiology and Genomics of Toulouse, CNRS/Universite Toulouse III Paul Sabatier, Allées Jules Guesde 37, Toulouse, Occitanie, 31000, France

[38]Department of Computational Biology, Uniwersytet im Adama Mickiewicza w Poznaniu, Uniwersytetu Poznanskiego 6, Poznań, Wielkopolska, 61-614, Poland

[39]Center for Evolutionary Hologenomics, Globe Institute, Københavns Universitet, ester Voldgade 44747, Copenhagen, Copenhagen K, 1350, Denmark

[40]Defence Genomics, Centre for Genomics and Personalised Health, Queensland University of Technology, Musk Ave 60, Kelvin Grove, Queensland, 4059, Australia

[41]Department of Human Genetics, The University of Chicago, E. 58th St. 920, Chicago, Illinois, 60637, USA

[42]Department of Anthropology, Harvard University, Divinity Avenue 11, Cambridge, Massachusetts, 02138, USA

[43]Faculty of Biological Sciences, Institute of Microbiology, Friedrich-Schiller-Universität Jena, Neugasse 25, Jena, Thuringia, 07743, Germany

Open Peer Review

Approval Status ✔ ❓

	1	2
version 2 (revision) 28 May 2024		

Abstract

Background: Access to sample-level metadata is important when selecting public metagenomic sequencing datasets for reuse in new biological analyses. The Standards, Precautions, and Advances in Ancient Metagenomics community (SPAAM, https://spaam-community.github.io) has previously published AncientMetagenomeDir, a collection of curated and standardised sample metadata tables for metagenomic and microbial genome

datasets generated from ancient samples. However, while sample-level information is useful for identifying relevant samples for inclusion in new projects, Next Generation Sequencing (NGS) library construction and sequencing metadata are also essential for appropriately reprocessing ancient metagenomic data. Currently, recovering information for downloading and preparing such data is difficult when laboratory and bioinformatic metadata is heterogeneously recorded in prose-based publications.

Methods: Through a series of community-based hackathon events, AncientMetagenomeDir was updated to provide standardised library-level metadata of existing and new ancient metagenomic samples. In tandem, the companion tool 'AMDirT' was developed to facilitate automated metadata curation and data validation, as well as rapid data filtering and downloading.

Results: AncientMetagenomeDir was extended to include standardised metadata of over 5000 ancient metagenomic libraries. The companion tool 'AMDirT' provides both graphical- and command-line interface based access to such metadata for users from a wide range of computational backgrounds. We also report on errors with metadata reporting that appear to commonly occur during data upload and provide suggestions on how to improve the quality of data sharing by the community.

Conclusions: Together, both standardised metadata and tooling will help towards easier incorporation and reuse of public ancient metagenomic datasets into future analyses.

Keywords
metagenomics, environmental, palaeogenomics, aDNA, microbiome, metadata, microbial, FAIR data

	1	2
version 1 02 Aug 2023	✔ view	❓ view

1. **Timothy Read**, Emory University, Atlanta, USA

   Robert Petit III, Wyoming Public Health Laboratory, Cheyenne, USA

2. **Jonas Coelho Kasmanas** (iD), University of São Paulo, São Paulo, Brazil

Any reports and responses or comments on the article can be found at the end of the article.

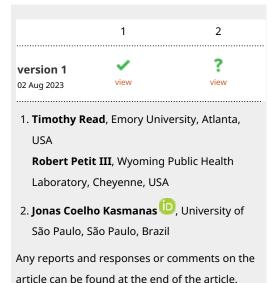This article is included in the Bioinformatics gateway.

This article is included in the Max Planck Society collection.

This article is included in the Evolutionary Genomics collection.

Corresponding authors: Maxime Borry (maxime_borry@eva.mpg.de), Christina Warinner (christina_warinner@eva.mpg.de), James A. Fellows Yates (james_fellows_yates@eva.mpg.de)

Author roles: Borry M: Conceptualization, Methodology, Software, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Forsythe A**: Data Curation, Writing – Review & Editing; **Andrades Valtueña A**: Data Curation, Writing – Review & Editing; **Hübner A**: Data Curation, Software, Writing – Review & Editing; **Ibrahim A**: Data Curation, Writing – Review & Editing; **Quagliariello A**: Data Curation, Writing – Review & Editing; **White AE**: Data Curation, Writing – Review & Editing; **Kocher A**: Data Curation, Writing – Review & Editing; **Vågene ÅJ**: Data Curation, Writing – Review & Editing; **Bartholdy BP**: Data Curation, Writing – Review & Editing; **Spurīte D**: Data Curation, Writing – Review & Editing; **Ponce-Soto GY**: Data Curation, Writing – Review & Editing; **Neumann G**: Data Curation, Writing – Review & Editing; **Huang IT**: Data Curation, Writing – Review & Editing; **Light I**: Data Curation, Software, Writing – Review & Editing; **Velsko IM**: Data Curation, Writing – Review & Editing; **Jackson I**: Data Curation, Writing – Review & Editing; **Frangenberg J**: Data Curation, Software, Writing – Review & Editing; **Serrano JG**: Data Curation, Writing – Review & Editing; **Fumey J**: Software, Writing – Review & Editing; **Özdoğan KT**: Data Curation, Writing – Review & Editing; **Blevins KE**: Data Curation, Writing – Review & Editing; **Daly KG**: Data Curation, Writing – Review & Editing; **Lopopolo M**: Data Curation, Writing – Review & Editing; **Moraitou M**: Data Curation, Writing – Review & Editing; **Michel M**: Data Curation, Writing – Review & Editing; **van Os M**: Data Curation, Writing – Review & Editing; **Bravo-Lopez MJ**: Data Curation, Writing – Review & Editing; **Sarhan MS**: Data Curation, Writing – Review & Editing; **Dagtas ND**: Data Curation, Writing – Review & Editing; **Oskolkov N**: Software, Writing – Review & Editing; **Smith OS**: Data Curation, Writing – Review & Editing; **Lebrasseur O**: Data Curation, Writing – Review & Editing; **Rozwalak P**: Software, Writing – Review & Editing; **Eisenhofer R**: Data Curation, Writing – Review & Editing; **Wasef S**: Data Curation, Writing – Review & Editing; **Ramachandran SL**: Data Curation, Writing – Review & Editing; **Vanghi V**: Data Curation, Writing – Review & Editing; **Warinner C**: Conceptualization, Funding Acquisition, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Fellows Yates JA**: Conceptualization, Data Curation, Investigation, Methodology, Project Administration, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Introduction

The field of palaeogenomics has been praised as a role model for scientific data reporting and data availability.[1] When compared against FAIR principles (Findability, Accessibility, Interoperability, and Reusability),[2] ancient DNA (aDNA) sequencing data have been consistently made available in standard data formats on public data repositories, satisfying the principles of *accessibility*, *interoperability* and, to a certain extent, *reusability*. However, the *findability* of the uploaded data still poses challenges, often due to the lack of inclusion of key metadata specific for aDNA in the standardised sample metadata fields used by public sequencing repositories such as the European Bioinformatic Institute's European Nucleotide Archive (EBI ENA), the US National Center for Biotechnology Information's Sequence Read Archive (NCBI SRA), and the Japanese National Institute of Genetics' DNA Data Bank of Japan (NIG DDBJ). To improve findability of ancient metagenomic samples in public data repositories, the SPAAM community (https://spaam-community.github.io) previously developed the AncientMetagenomeDir project, a set of curated standard sample metadata for ancient host-associated shotgun-sequenced metagenomes, ancient environmental metagenomes, and/or host-associated microbial genomes.[3] However, while sample-level metadata already help with the discovery of suitable comparative data, library-level metadata is also needed to further facilitate data reuse in dedicated aDNA analysis pipelines such as PALEOMIX,[4] nf-core/eager,[5] aMeta,[6] and nf-core/mag.[7] aDNA researchers often build many different types of NGS libraries[8] and may generate (meta)genomic data using multiple different sequencing platforms that require different bioinformatic pre-processing workflows. Furthermore, the library-level metadata currently available in public repositories often lack key information about aDNA library treatments and other laboratory information needed to reproducibly reanalyse palaeogenomic datasets obtained from different studies.

As the number of ancient metagenomics samples and shotgun sequenced library files steadily increases (currently >2300 host-associated metagenome, >2200 single-genome, and >600 environmental metagenome sequencing run accessions as of March 2023; Figure 1), the need to find ways to efficiently identify, curate, and download such data is becoming more pressing. Although the original AncientMetagenomeDir releases provided project- or sample-level accession numbers that point to data primarily hosted by the ENA, SRA, and DDBJ, the metadata tables did not provide direct links to the data themselves. This meant that researchers still needed to manually search for each project or sample accession number in public data repositories and then manually identify and download the relevant associated files. Researchers were then required to parse and evaluate each sequencing file for inclusion in their study by consulting the original scientific publications for laboratory, library, and sequencing metadata. As with sample metadata, the reporting of this information within publications can be heterogeneous, may appear in the main text or supplement, and may take the form of prose text, tables, supplementary spreadsheets, or citations to other publications or protocols. While other tools for exploring public data repositories exist, such as NCBImeta,[9] SRA-Explorer,[10] and ffq,[11] they are generally limited to a restricted set of metadata available for inspection, or require the use of command-line filtering tools, an interface not always accessible to all palaeogenomics researchers, who often have varying levels of computational experience.

Figure 1. Growth of studies curated in the AncientMetagenomeDir as of v23.03. (a) Number of ancient metagenomic publications published per year with open sequencing data and included in AncientMetagenomeDir. The original AncientMetagenomeDir publication was in 2020. (b) Cumulative sum of the number of published samples with publicly accessible sequencing data. (c) Cumulative sum of the number of ancient metagenomic sequencing data accessions of the samples in panel b. Data from Fellows Yates et al.[12]

Here we present AMDirT (**A**ncient**M**etagenome**Dir T**oolkit), a tool designed to assist researchers in using a new extension of AncientMetagenomeDir that now includes aDNA library- and sequencing-level metadata. AMDirT is designed to provide a solution to three different challenges. First, it helps researchers curate newly published aDNA sequencing data in AncientMetagenomeDir by automatically retrieving relevant library-level metadata available from sequencing archives (`autofill` command), and preparing semi-filled data entry tables for submission. Second, the tool can perform a variety of data validation tasks on completed entry tables to ensure consistency (`validate` command, an improved version of the ancientMetagenomeDirCheck tool from Ref. 3). Third, AMDirT provides a web browser-based, user-friendly tabular graphical-user interface (GUI) that allows users to explore relevant ancient metagenomics-related sequencing datasets in AncientMetagenomeDir tables (`viewer` command) and to export scripts to download the filtered data. As an additional functionality, AMDirT can provide template input configuration files for a suite of standard aDNA metagenomics-related pipelines in order to further automate and accelerate the processing of such aDNA data. Finally, most of the AMDirT `viewer` functionalities are additionally mirrored in a complementary command line interface (CLI) for more experienced bioinformaticians (`convert` command). AMDirT is available for installation via PyPI (https://pypi.org/) or Bioconda,[13] with source code on GitHub under https://github.com/SPAAM-community/AMDirT.

Methods
Implementation
AMDirT tool implementation. Members of the SPAAM community (https://spaam-community.github.io) developed AMDirT through a series of code sprints and hackathons, using Python v3.9 (https://www.python.org/; RRID: SCR_008394). It is accessible via a command-line-interface written using Click or via a python API (https://click.palletsprojects.com/). The `autofill` command that automates metadata retrieval from sequencing archives to help with data curation is using the ENA portal API (https://www.ebi.ac.uk/ena/portal/api/)[14] to query and return metadata associated with the sequencing library level. Data validation in the `validate` subcommand is performed using the jsonschema python library (https://python-jsonschema.readthedocs.io/) and it uses a variety of checks written using Pandas[15] to avoid data duplication and ensure consistency of new entries.

The GUI data exploration interface of the `viewer` command was developed using Streamlit (https://streamlit.io/), and the streamlit-aggrid library[16] is utilised to allow the end user to interactively filter and prepare configuration files to process ancient (meta)genomic data in bioinformatics pipelines. Finally, AMDirT is packaged thanks to setuptools,[17] and is distributed on PyPi and Bioconda.[13] The source code is available on GitHub (github.com/SPAAM-community/AMDirT), and associated documentation is provided online (amdirt.readthedocs.io).

AncientMetagenomeDir library metadata aggregation. To extend the original AncientMetagenomeDir[3] repository to include library metadata, we created new tab-separated value (TSV) tables and their associated validation checks in the form of JSON schema files, following the original AncientMetagenomeDir structure.

We retained the TSV format for maximum software compatibility, as originally described in Ref. 3. Fields included in the new library-level schema were selected after consultation with ancient metagenomics researchers via the SPAAM community (an international and open community of over 300 researchers), and, where relevant and possible, by mirroring existing metadata fields and controlled vocabulary from the ENA repository. Newly added library information columns include the library name (how data are typically reported in original publications), the aDNA library generation method (e.g., double-stranded or single-stranded libraries), the library indexing polymerase (e.g., proof-reading or non-proofreading), and the library pretreatment method (e.g., non-Uracil-DNA Glycosylase (UDG), full-UDG, or half-UDG treatments). The latter three fields represent information about the sequencing library construction that influence the presence of aDNA damage, a factor that is critical for the processing of aDNA NGS data.[8,18] Sequencing metadata columns include instrument model, library layout (single- or paired-end), library strategy (whole genome sequencing, targeted capture, etc.), and read count. This metadata is also critical for correct processing of aDNA data. For example, whether an instrument uses 2- or 4-colour sequencing chemistry determines if poly-G tail trimming is required to remove sequence artefacts that arise in aDNA reads that are normally shorter than the number of sequencing cycles. Library layout is also necessary to indicate whether read-pair merging needs to be applied prior to mapping, or whether unmerged read pairs are available for *de novo* assembly. The remaining columns provide information about storage and file retrieval of sequencing data: direct URLs to FASTQ files, 'md5 checksum' strings (for post-download integrity verification), and download sizes (for storage space usage estimation). Tables may also be extended to contain field-specific metadata useful for data processing under specific conditions. For new non-ENA/SRA supported fields, such as library polymerase or library treatment, we defined fixed lists via new JSON-based 'enum' files stored in the AncientMetagenomeDir repository, as with the sample-level metadata.

For compiling the initial set of 'raw' library metadata, we took the sample-level publication data from each of the three sample-level metadata tables and their respective ENA and/or SRA sample-level accession numbers. We then used the

ENA's REST API to pull relevant metadata for the AncientMetagenomeDir library tables for each of the sample-level accession IDs. This was performed using R scripts (stored in the GitHub repository under 'assets/utility/') using the following libraries: tidyverse,[19] jsonlite,[20] httr,[21] and xml2.[22]

Via a series of community events, we then performed hand-curation of this initial set of metadata, and also completed data entry for columns of metadata not stored in the ENA by comparing the ENA stored metadata with the methods descriptions in original publications. This procedure identified multiple instances of inconsistencies between the two sources, as well as incorrectly uploaded data and metadata in previously published articles. We describe some of the common issues we encountered in the Discussion below. In cases of conflict between the publication and the ENA metadata, we endeavoured to contact the original authors of the publication for confirmation. When this was not possible, we used 'unknown' or another missing-data value to indicate uncertainty. Each library-level metadata addition under-went automated validation and peer-review following the same procedure described in Ref. 3. Since the community events, the AMDirT autofill command has been developed to improve the submission experience by community members contributing new metadata. The autofill sub-command automates the pulling of ENA metadata into a 'draft' library-table format during the continuous integration tests (CI) of a GitHub pull request of sample-level metadata, replacing and improving upon the R scripts used in the initial pull-down for the community events. Submitters to AncientMetagenomeDir can then copy over much of the metadata and fill in the remaining missing metadata not covered by the existing ENA metadata fields.

Operation

AMDirT requires a UNIX-based terminal (e.g., Linux, OSX, Windows Subshell for Linux) for both installation and initial usage; however, the toolkit is written in Python and can therefore be used on a wide range of platforms and operating systems.

To install, users are recommended to use the pip or conda package managers. Users who wish to use the GUI based table viewer and downloader will also require any modern web browser supported by Streamlit (https://streamlit.io/).

For example, to install and load the help message:

```
$ pip install amdirt
```

or via conda

```
$ conda install -c bioconda amdirt
```

The general help of AMDirT is available from the CLI:

```
$ AMDirT --help
Usage: AMDirT [OPTIONS] COMMAND [ARGS] …

  AMDirT: Performs validity check of AncientMetagenomeDir datasets
  Authors: AMDirT development team and the SPAAM community
  Homepage & Documentation: https://github.com/SPAAM-community/AMDirT

Options:
  --version  Show the version and exit.
  --verbose  Verbose mode
  --help     Show this message and exit.

Commands:
  autofill  Autofills library and/or sample table(s) using ENA API and…
  convert   Converts filtered samples and libraries tables to eager,…
  validate  Run validity check of ancientMetagenomeDir datasets…
  viewer    Launch interactive filtering tool
```

Most tools follow a standard CLI based interface. For example, converting a user-filtered ancient metagenome host-associated AncientMetagenomeDir table (e.g. in R) to a curl download script can be performed as follows:

```
$ AMDirT convert --curl <filtered_table>.tsv ancientmetagenome-hostassociated -o ./
```

In the case of 'convert', the input metadata tables must be supplied by the user (i.e., downloaded from AncientMetagenomeDir manually).

The resulting file `AncientMetagenomeDir_curl_download_script.sh` will be present in the directory specified in the command. The user can then simply run the bash script to download all libraries of the samples present in the input table.

```
$ bash AncientMetagenomeDir_curl_download_script.sh
```

For the template pipeline input sheets, these can be supplied to the pipelines themselves, after checking for accuracy.

The other AMDirT tools follow a similar scheme, with help messages and documentation on the AMDirT website providing more how-to information (https://amdirt.readthedocs.io/).

For the GUI-based 'viewer' tool, a user simply enters the following command in their terminal, after which their web browser will automatically load. Alternatively, the reported local or network address can be manually entered into the user's web browser. In comparison to the 'convert' subcommand, the input tables are automatically pulled from the AncientMetagenomeDir for the user, without requiring any manual input.

```
$ AMDirT viewer
AMDirT [INFO] :
[AMDirT] To close app, press on your keyboard: ctrl+c

  You can now view your Streamlit app in your browser.

  Local URL: http://localhost:8501
  Network URL: http://172.16.9.75:8501
```

Once completed, the user can close the tab and cancel the command in their terminal (e.g., with `ctrl + c`).

Alternatively, for individuals who wish to use the `viewer` but do not wish to deal with software installation and/or are not comfortable with command line interfaces, a hosted webserver version of the AMDirT `viewer` is available at https://spaam-community.github.io/AMDirT/.

Use cases
Here we will describe a common use case for when users may wish to use the AMDirT package, namely filtering for a particular subset of metagenomic aDNA data, downloading the resulting data, generating a corresponding semi-prepared input sheet for nf-core/eager, and creating a citations file. Full tutorials in text and video format for the following and other AMDirT commands can be found on the AMDirT website (https://amdirt.readthedocs.io/).

This example scenario demonstrates how to download all publicly available ancient host-associated metagenomes published since 2020 from samples originating from Spain. In this hypothetical example, a user may wish to compare the microbial taxonomic profiles of archaeological dental calculus and other skeletal elements in Spain at different time points. In order to distinguish modern and aDNA, the user will likely want to examine the DNA for evidence of degradation, which can be used to authenticate aDNA. To do this, the user will already have selected their preferred dedicated aDNA analysis workflow, such as nf-core/eager,[5] that integrates DNA damage analysis into its pipeline. nf-core/eager requires an input 'sample sheet' that describes whether a particular sample has been sequenced over multiple lanes or libraries, and whether aDNA damage has been already removed during laboratory processing. We will show how AMDirT can assist with the creation of this sample sheet.

This example assumes that the user has already installed AMDirT and nf-core/eager, and has downloaded a *Homo sapiens sapiens* reference genome.

Figure 2. Example workflow of using AMDirT viewer. (a) The viewer opens in a user's web browser, where the desired AncientMetagenomeDir version and table is selected. (b) Interaction with columns follows standard operations common to most spreadsheet software. Samples for download are selected using check boxes. (c) After pressing 'Validate selection', buttons appear for downloading various download scripts, reference, and pipeline input sheets.

To load the GUI based viewer and downloading tool, a user enters the following command into their terminal:

```
$ AMDirT viewer
```

As shown in Figure 2a, the viewer is loaded into the user's web browser. Using the left sidebar, the user can navigate drop-down menus to select the release of AncientMetagenomeDir to use (for reproducibility purposes), and the desired AncientMetagenomeDir table to explore (i.e., ancient environmental metagenomes, ancient host-associated metagenomes, or ancient microbial single-genomes). The user can also specify the number of rows for the table to display and which tool to use for the download scripts generated later in the example.

Once the AncientMetagenomeDir table is selected, the main window will load the corresponding table. The user can manipulate the columns as customary in most spreadsheet software, such as re-sizing the columns by dragging the bars between each column, dragging column names to reorder them, etc. Users can use keyboard arrows or scroll bars to navigate further along the columns. To filter the columns, the user can press the 'hamburger' menu of each column, which reveals a range of column operation options. In the example in Figure 2b, the 'geo_loc_name' column has already been filtered to display only samples with the value 'Spain', and by pressing the funnel on an integer column, such as 'publication_year', the user can specify to only display rows 'greater than or equals' 2020. Additional filter specifications can be added using the AND/OR operators in each filter menu. The user can then select which samples to be exported from AMDirT. They can either use the 'Select All' checkbox in the top left of the table, or select each sample using the check boxes at the beginning of each row on a sample-by-sample basis.

Once the user is satisfied with their selection, they can press 'Validate Selection' to display a range of option buttons (Figure 2c). Here, we recommend that users always download the corresponding AncientMetagenomeDir libraries table, as well as the BibTeX citation file. In this example, the user would also download the 'curl download script' and the 'nf-core/eager' input TSV using the corresponding buttons. Hovering over the download script button also provides an estimate for the user of how much hard-drive space the download of all selected data will use.

To close the GUI viewer, the user can close the tab in their browser, and then in their terminal press ctrl + c on their keyboard to stop the Streamlit server.

Once downloaded, the user should check the AncientMetagenomeDir libraries table to ensure that all desired libraries are present. If there are extra libraries with specifications that the user does not want, they should remove those entries from the curl download script and nf-core/eager input TSV files. The user should also review the nf-core/eager pipeline input sheet to check for accuracy in regards to the pipeline's specifications.

After filtering of the scripts and pipeline input TSV sheets, the user can then use their terminal to navigate to a directory, move the curl script into it and begin the download. Due to the large sizes of sequencing data, in most cases we recommend that a user do this in a 'screen' or 'tmux' session (or similar) to ensure that the downloading can continue in the background:

```
$ bash AncientMetagenomeDir_curl_download_script.sh
```

Once the sequencing data are downloaded, the user can provide their AMDirT generated nf-core/eager input sheet to the following Nextflow[23] command (the nf-core/eager input sheet assumes that the command is being run in the same directory as the downloaded data):

```
$ nextflow run nf-core/eager -r 2.4.6 -profile conda \
> --input ancientMetagenomeDir_nf_core_eager_input.tsv \
> --fasta hg19.fasta --outdir ./results
```

Discussion
Results of library level metadata aggregation
Since the original publication of AncientMetagenomeDir[3] and the release of version v20.09, the SPAAM community has nearly doubled the number of manually curated publications in the AncientMetagenomeDir from 87 to 161 studies as of version v23.03. The number of samples has increased from 443 to 1358 for ancient host-associated metagenome samples, 269 to 583 for ancient microbial genome level sequences, and 312 to 563 for sediment samples (Figure 1).

During the series of 'hackathon' events carried out by the community to scrape library metadata from previous publications and subsequent submissions of new studies, a total of 2332 ancient host-associated metagenome libraries, 2278 ancient microbial genomes libraries, and 627 ancient environmental metagenome libraries have been curated and included.

Common issues in ancient metagenomic library metadata

During the aggregation and clean up of the library metadata by the SPAAM community, a range of problems were repeatedly encountered across multiple studies that made data entry and the determination of appropriate preprocessing procedures difficult. Here, we describe the most common issues encountered, as well as possible solutions, listed from most to least severe. We hope by raising these common mistakes and problems, that the field can help improve data uploads that will both benefit the contributing to AncientMetagenomeDir but also the field as a whole.

Inconsistent sample and library naming. *Problem:* A common problem encountered when cross-referencing ENA or SRA metadata with information provided in original publications was inconsistencies in sample, library, and/or sequencing file names. This often made it difficult for the community member to correctly infer which library was associated with which sample, or even which sequencing file went with which library. *Example:* In studies where two sets of libraries were generated (for example, one with UDG-based DNA damage removal and one without), this palaeogenomic-specific information was often not indicated in the library or file names. Given that this information is not supported in the ENA/SRA metadata schema, this is the only location where such aDNA-specific information could be reasonably recorded. In such cases, we found that while library pretreatment procedures were documented in the original publication, the uploaded metadata and sequencing files generally lacked this information and in some instances used internal laboratory IDs instead of the sample or library codes recorded in the publication. Without a key linking the published IDs with the internal laboratory IDs, other researchers cannot know which files to use for their particular analyses or how to process the data appropriately, and this can lead to downstream problems. For example, if a user does not know that a sequencing file was generated from damage-removed libraries, they may inappropriately apply additional *in silico* trimming steps to remove DNA damage, and thus unnecessarily truncate the sequences. *Solutions:* We suggest two solutions: first, ancient metagenomic researchers should ensure that library and sample names are descriptive (i.e., in a structured system in which a certain level of information can be inferred just by the name) and that sequencing metadata uploads match those reported in the publication; second, where this is not possible (e.g., if upload is carried out by a third party), then researchers should at a minimum include a key in their supplementary files. This could be in the form of a table that includes all ID codes for each sample, library, and sequencing batch, including internal laboratory codes, other-analysis codes, and external sequencing archive accession codes.

Metadata discrepancies about sequencing methods. *Problem:* Another relatively common issue was the discrepancies between the metadata reported in the sequence archive and in the original publication. It was generally difficult to resolve such discrepancies without contacting the authors. Discrepancies occurred most frequently in the reporting of the sequencing platform. *Example:* In several cases, the particular sequencing platform recorded in the sequence archive metadata, such as 'Illumina HiSeq 4000', did not match that reported in the publication, e.g. 'NextSeq 500'. *Solutions:* Researchers should be sure to cross-reference their metadata upload sheets with their manuscripts prior to upload. In cases where Illumina sequencing was carried out externally (and where limited information may have been provided by the sequencing centre), researchers can generally inspect the headers of the FASTQ file to determine which platform was used, as in example provided here.

Methods description in secondary or tertiary citations. *Problem:* For journals with strict word or character count limits, it was qualitatively observed that there was an increased tendency in these publications to rely on secondary or tertiary non-protocol specific citations for describing laboratory methods used for DNA library construction and sequencing. This practise is problematic as secondary or tertiary citations may describe multiple protocols, and it was not always possible to determine which protocol was actually used in the study. *Example:* In one case, a publication reporting an ancient microbial genome reconstruction referred to library protocols used in an earlier related publication that described data generation for a human population genetics study. However, upon closer inspection, this cited study itself referred to an even earlier publication that included extensive protocol experimentation and development. Neither the primary nor secondary publication indicated which of the experimental protocols from the original methods study was actually used. *Solutions:* Ancient metagenomic researchers should make an effort to more clearly describe their protocols and explicitly indicate which library protocol is linked to each sequencing file. At a minimum, the information provided should include critical metadata for downstream analysis, such as library treatment protocols that affect DNA damage. This can be accomplished by providing expanded, plain-language descriptions of laboratory methods in article supplementary information files (rather than simply citing and reciting) and providing a supplementary table that lists each library name and their corresponding treatments. For improved compliance with FAIR principles, researchers are

encouraged to further provide or cite a protocol written up in a citable protocol format and/or on open platforms. For example, platforms such as protocols.io[24] allow critical protocol information to be communicated consistently and unambiguously, by providing a persistent identifier (DOI) that points to a specific version of a given protocol.

Uploading of mapped BAM files or merged FASTQ files rather than raw metagenomic data. *Problem:* Occasionally, we found that in some cases ancient metagenomic researchers uploaded mapped BAM files or merged FASTQ files rather than 'raw' FASTQ files (i.e., against ENA/SRA specifications). Both formats present obstacles for downstream analysis. For example, mapped BAM files include only reads mapped to a particular reference genome and thus do not represent a full metagenomic dataset. For BAM files containing reads mapped to the human genome, microbial DNA will be absent, including ancient pathogen DNA that could be highly relevant for an archaeological study. While an 'unmapped' BAM (uBAM) format exists and FASTQ files can be partly reconstructed from such data, raw FASTQ files are the preferred format for data reporting. Unmapped BAM files indicate that a certain level of data preprocessing has already occurred, and such files often combine multiple libraries into a single BAM file in order to achieve sufficient genomic coverage for analysis. If the process of generating the BAM file is not sufficiently described, it can be difficult for other researchers to disentangle the data originating from different libraries or sequencing batches prior to reanalysis. Providing FASTQ files containing merged paired-end reads also limits data reuse. Although read merging is a common first step in some ancient bioinformatics pipelines, it is incompatible with others. Base quality scores are often altered during the read merging process, which can interfere with tools reliant on such scores, and most *de novo* sequence assemblers either require or perform better on unmerged reads. Furthermore, many metagenomic tools, including taxonomic classifiers, do not accept merged paired-end FASTQ files or BAM files as an input format. *Example:* An ancient metagenomic author generates both damage and damaged-removed libraries, but merges them together in BAM format and uploads to a sequencing archive. However another researcher wishes to analyse only the data deriving from the damage-removed libraries. *Solution*: Ideally researchers should upload FASTQ files that match the 'raw' output from sequencing, i.e., demultiplexed datasets separated per library, applying only the preprocessing steps recommended by the sequence repository (e.g., for the ENA/SRA, adapter removal but not read merging). If this is not possible, authors should, at a minimum, describe exactly how the merging steps were performed so that other researchers can manually separate merged sequences (e.g., using sequencing read headers) when required for downstream analysis.

Unique sample accessions applied to multiple libraries of the same sample. *Problem:* Another common error was found to occur when researchers mistakenly uploaded each library or sequencing dataset with a unique sample accession code. While often not a critical error, because in these cases the correct sample could usually be inferred from the file name, this nevertheless makes automated data processing more difficult and requires manual intervention. To reuse such data, a researcher must manually reassociate the library datasets with the correct sample based on the file names, rather than relying on the sequence archive sample accession ID, as expected by metadata schemas of the ENA/SRA data repositories. *Solution:* Researchers should review sequencing archive documentation to ensure they correctly construct upload sheets at both the library and sample levels (e.g. https://ena-docs.readthedocs.io/en/latest/submit/general-guide/metadata.html). Furthermore, researchers should ensure that library names have a consistent pattern such that other researchers can unambiguously associate each library with the correct sample.

Note on AMDirT generated pipeline scripts
It is important to note that the aim of the pipeline TSV sheets generated by AMDirT is to provide a *template* for data input to the pipelines. Due to the high heterogeneity in the way that sequence (meta) data are uploaded, some information in AncientMetagenomeDir may be missing or erroneous, despite the 'best efforts' of the SPAAM community to standardise the information, resolve ambiguities, and correct errors. However, we hope that by providing this functionality, it reduces the time it takes to create such input sheets from scratch.

Future development
We envision that the future development of the AncientMetagenomeDir project will be to further extend and also standardise the types of metadata currently recorded. For example, when recording the age of samples, AncientMetagenomeDir currently only records a single value of an extremely rough, rounded-up, and approximate date. This poses challenges for analyses requiring exact dates and probability intervals such as tip dating for phylogenetic trees and other analyses of evolutionary divergence. At present, however, heterogeneity in the reporting of radiocarbon dates (the most common dating method in archaeology and palaeogenomics) and associated modelling information currently limits our ability to add such dating information to AncientMetagenomeDir, and to consistently apply calibration and reservoir effect correction across studies. This is despite the fact that there is already standard reporting guidance.[25] However, we also call on ancient metagenomics researchers to report both uncalibrated *and* calibrated dates and associated metadata (radiocarbon lab code, calibration curve, software, etc.), and to not rely solely on secondary citations to facilitate adding such data to repositories such as AncientMetagenomeDir, as well as refinement of chronological modelling in the future.

In the same vein, we also aim to synchronise AncientMetagenomeDir with upstream standardised sequencing data metadata schemas and repositories such as MIxS checklists[26] via another SPAAM project, MInAS (https://mixs-minas.org/), to further ensure common standards across both modern and ancient sequencing data.

Given the functionalities of AMDirT provide simple data exploration without requiring advanced computational knowledge, but also offers semi-prepared templates for aDNA and metagenomics, the dataset and tooling are ideal for further generation of community resources. Following other projects that have been developed for modern microbiome data,[27] the ancient metagenomics community could also consider providing standardised and pre-made taxonomic profiles (e.g., for microbiome or environmental samples) or VCF files (for single genomes) that could allow integration into current analysis workflows to assist users in more rapidly integrating public data into their analyses from a single source. This could be particularly useful for screening ancient microbiome samples for preservation (e.g., by comparing a newly sequenced sample to all previously published ancient metagenomes), in order to assess whether a sample falls within the variation of known well-preserved or environmentally degraded samples.

Conclusions

By extending AncientMetagenomeDir to include library-level metadata, not only do we make ancient metagenomics data more findable, but we also make them more accessible by providing improved transparency of the diverse library and sequencing treatments performed in the field of palaeogenomics. Furthermore, AMDirT has been designed to improve the experience of researchers in the downloading and processing of previously published ancient metagenomics data. By providing both a graphical user interface and a command-line interface to filter and generate relevant download scripts and input sample sheets for aDNA analyses, we provide more flexibility and choice for the wide range of computational backgrounds that ancient metagenomic researchers can have. Finally, we hope that by informing researchers about inconsistencies in past data and metadata uploads, and by providing templates of standardised metadata for future publications, we will contribute to improving aDNA data reporting and FAIR data sharing.

Data availability

Source data
The source data for the sample-level metadata used by AMDirT is from the AncientMetagenomeDir project originally published in Ref. 3 under a CC-BY 4.0 license.

The existing sample-level and new library-level data is stored on GitHub:

https://github.com/SPAAM-community/AncientMetagenomeDir

Each release is archived on Zenodo: https://doi.org/10.5281/zenodo.3980833.

Underlying data
New sequencing library-level metadata is also stored in the AncientMetagenomeDir project from version v22.09 (Pyu Ancient Cities) onwards.[12] The version of the data set used for the demonstration of AMDirT, statistics, and figures in this manuscript is v23.03 (Rocky Necropolis of Pantalica).[28]

Zenodo: SPAAM-community/AncientMetagenomeDir: v22.12: Joya de Cerén. https://doi.org/10.5281/zenodo.7470037.

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Software availability
- Software available from:

    - PyPi: https://pypi.org/project/AMDirT/

    - Bioconda: https://bioconda.github.io/recipes/amdirt/README.html

    - Hosted web version of AMDirT `viewer`: https://spaam-community.github.io/AMDirT

- Source code available from: https://github.com/SPAAM-community/AMDirT/

- Archived source code at time of publication: https://doi.org/10.5281/zenodo.7891473

- License: GNU General Public License v3.0.

Acknowledgements

References

1. Anagnostou P, Capocasa M, Milia N, *et al*. **When data sharing gets close to 100%: what human paleogenetics can teach the open science movement.** *PLoS One.* March 2015; **10**(3): e0121409. 1932-6203.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

2. Wilkinson MD, Dumontier M, Aalbersberg IJJ, *et al*. **The FAIR guiding principles for scientific data management and stewardship.** *Sci. Data.* March 2016; **3**: 160018. 2052-4463.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

3. Fellows Yates JA, Andrades Valtueña Å, Vågene ÅJ, *et al*. **Community-curated and standardised metadata of published ancient metagenomic samples with AncientMetagenomeDir.** *Sci. Data.* January 2021; **8**(1): 31. 2052-4463.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

4. Schubert M, Ermini L, Der Sarkissian C, *et al*.: **Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX.** *Nat. Protoc.* May 2014; **9**(5): 1056–1082. 1754-2189, 1750-2799.
PubMed Abstract | **Publisher Full Text**

5. Fellows Yates JA, Lamnidis TC, *et al*.: **Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager.** *PeerJ.* March 2021; **9**: e10947. 2167-8359.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

6. Pochon Z, Bergfeldt N, Kırdök E, *et al*.: **aMeta: an accurate and memory-efficient ancient Metagenomic profiling workflow.** *bioRxiv.* October 2022; page 2022.
Publisher Full Text

7. Krakau S, Straub D, Gourlé H, *et al*.: **nf-core/mag: a best-practice pipeline for metagenome hybrid assembly and binning.** *NAR Genom. Bioinform.* January 2022; **4**(1).
PubMed Abstract | **Publisher Full Text** | **Free Full Text** | **Reference Source**

8. Orlando L, Allaby R, Skoglund P, *et al*.: **Ancient DNA analysis.** *Nat. Rev. Methods Primers.* February 2021; **1**(1): 1–26.
Publisher Full Text

9. Eaton K **NCBImeta: efficient and comprehensive metadata retrieval from NCBI databases.** *J. Open Source Softw.* February 2020; **5**(46): 1990. 2475-9066.
Publisher Full Text

10. Ewels P, Duncan A, Fellows Yates JA: **ewels/sra-explorer: Version 1.0.** March 2023.
Reference Source

11. Gálvez-Merchán Á, Min KHJ, Pachter L, *et al*.: **ffq: A tool to find sequencing data and metadata from public databases.** 2022.
Reference Source

12. Fellows Yates JA, Andrades Valtueña A, Vågene ÅJ, *et al*.: **SPAAM-community/AncientMetagenomeDir: v22.09.2.** August 2022.
Reference Source

13. Grüning B, Dale R, Sjödin A, *et al*.: **Bioconda: sustainable and comprehensive software distribution for the life sciences.** *Nat. Methods.* July 2018; **15**(7): 475–476.
PubMed Abstract | **Publisher Full Text**

14. Harrison PW, Ahamed A, Aslam R, *et al*.: **The european nucleotide archive in 2020.** *Nucleic Acids Res.* January 2021; **49**(D1): D82–D85.
PubMed Abstract | **Publisher Full Text** | **Free Full Text** | **Reference Source**

15. The pandas development team: **pandas-dev/pandas: Pandas.** 2020.
Publisher Full Text

16. Fonseca P: **streamlit-aggrid: Implementation of Ag-Grid component for streamlit.** 2023.
Reference Source

17. Python Packaging Authority: **setuptools: Official project repository for the setuptools build system.** 2023.
Reference Source

18. Dabney J, Meyer M, Pääbo S: **Ancient DNA damage.** *Cold Spring Harb. Perspect. Biol.* July 2013; **5**(7). 1943-0264.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

19. Wickham H, Averick M, Bryan J, *et al*.: **Welcome to the tidyverse.** *J. Open Source Softw.* November 2019; **4**(43): 1686. 2475-9066.
Publisher Full Text

20. Ooms J: **The jsonlite package: A practical and consistent mapping between JSON data and R objects.** *arXiv.* March 2014.
Reference Source

21. Wickham H: **httr: Tools for working with urls and http.** 2021.
Reference Source

22. Wickham H, Hester J, Ooms J: **xml2: Parse xml.** 2021.
Reference Source

23. Di Tommaso P, Chatzou M, Floden EW, *et al*.: **Nextflow enables reproducible computational workflows.** *Nat. Biotechnol.* April 2017; **35**(4): 316–319.
PubMed Abstract | **Publisher Full Text**

24. Teytelman L, Stoliartchouk A, Kindler L, *et al*.: **Protocols.io: Virtual communities for protocol development and discussion.** *PLoS Biol.* August 2016; **14**(8): e1002538.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

25. Millard AR **Conventions for reporting radiocarbon determinations.** *Radiocarbon.* 2014; **56**(2): 555–559.
Publisher Full Text | **Reference Source**

26. Yilmaz P, Kottmann R, Field D, *et al*.: **Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications.** *Nat. Biotechnol.* May 2011; **29**(5): 415–420.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

27. Pasolli E, Schiffer L, Manghi P, *et al*. **Accessible, curated metagenomic data through ExperimentHub.** *Nat. Methods.* October 2017; **14**(11): 1023–1024.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

28. Fellows Yates JA, Andrades Valtueña A, Vågene ÅJ, *et al*.: **SPAAM-community/AncientMetagenomeDir: v23.03.0: Rocky necropolis of pantalica.** March 2023.
Reference Source

Open Peer Review

Current Peer Review Status: ✔ ❓

Version 1

Reviewer Report 18 October 2023

https://doi.org/10.5256/f1000research.147881.r210768

❓ **Jonas Coelho Kasmanas** iD

University of São Paulo, São Paulo, Brazil

The submitted manuscript sheds light on an imperative yet often overlooked aspect of current research, specifically metagenomic research: the accessibility and standardization of sample-level metadata, particularly concerning ancient samples. The topic is relevant to the current scientific landscape, especially as the volume of metagenomic sequencing datasets continues to burgeon. With the surge in data, the need for standardized, accessible, and reusable metadata becomes paramount. This manuscript aptly addresses this gap, making it an essential read for researchers in metagenomics, palaeogenomics, and related fields.

The authors have channeled the collective effort of the research community through hackathon events to update and enhance the 'AncientMetagenomeDir.' Even contacting researchers and actively encouraging them to correct their metadata annotation. Therefore, I understand that this manuscript evolves the AncientMetagenomeDir and provides an additional resource to further increase its accessibility and usability: the AMDirT.

The introduction of 'AMDirT' as the companion tool to 'AncientMetagenomeDir' is a significant stride forward. By offering functionalities ranging from metadata template generation, and guidance to data validation, the tool exhibits promise in addressing the prevalent challenges in metadata management. Additionally, it helps researchers with less computational background to download samples systematically and gives them access to a GUI. My review goes toward the message delivery, especially during the initial sections of the manuscript.

During the introduction section, I missed a clear definition of ancient metagenome. This should include a description of what requires the ancient samples to receive special treatment regarding metadata annotation. Naturally, this leads to comparing your effort to past works like the GOLD (https://doi.org/10.1093/nar/gky977), the MetaSRA (https://doi.org/10.1093/bioinformatics/btx334), the gcMeta (https://doi.org/10.1093/nar/gky1008), or the HumanMetagenomeDB (https://doi.org/10.1093/nar/gkaa1031).

I appreciate the extensive AMDirT documentation, including tutorials and videos. However, the

manuscript lacks in the description of the tool's workflow and usability. The "Use cases" section of the manuscript makes the tool much clearer. At the same time, it gives the impression that AMDirT is limited to the GUI exploration and template generation guidance. Specifically, in my opinion, the manuscript poorly describes and explores the "Validate" and "Autofill" commands.

For example, the Validate command does "a variety of checks." It would help if you made it more transparent. Do you have a document detailing the checks and standards? If so, it should be placed here; if not, I would like to see it in the manual. For the Autofill command, I missed a more detailed description of the "improvements" done using R scripts after pulling the metadata from ENA. Specifically, since those commands could be used outside the realm of ancient metagenomic samples. I believe a figure containing the complete schematic workflow from AMDirT would help.

Regarding the software operation, I missed a storage requirement specification. The paper mentions that the installation requires Python 3.9, but nothing is specified in the GitHub. The installation via pip and conda should include guidance on accessing those tools and creating a specific environment for the AMDirT. Finally, as a minor opinion, have you considered using specialized sample accessing software? For instance, the SRAtoolkit can download samples from the SRA more efficiently than curl.
If possible, I would like to see a slight reformat in the discussion "Problem/Solutions" section to make it more visible. A simple extra paragraph or bullet points that clearly separate problems and solutions should solve the problem.

Running and installing the pipeline.

I had a relatively smooth experience. However, I kept receiving the following warning upon any usage:
"WARNING streamlit.runtime.caching.cache_data_api: No runtime found, using MemoryCacheStorageManager"
Additionally, the convert command did not work. After successfully installing the tool and downloading the AncientMetagenomeDir. I received the following:
$ AMDirT convert --curl SPAAM-community-AncientMetagenomeDir-60f8a00/ancientmetagenome-hostassociated/libraries/ancientmetagenome-hostassociated_libraries.tsv ancientmetagenome-hostassociated -o ./

2023-10-17 19:50:32.289 No runtime found, using MemoryCacheStorageManager
Traceback (most recent call last):
  File "/mnt/tools/miniconda3/envs/amdirt/lib/python3.9/site-packages/streamlit/runtime/caching/storage/in_memory_cache_storage_wrapper.py", line 87, in get
    entry_bytes = self._read_from_mem_cache(key)
  File "/mnt/tools/miniconda3/envs/amdirt/lib/python3.9/site-packages/streamlit/runtime/caching/storage/in_memory_cache_storage_wrapper.py", line 137, in _read_from_mem_cache
    raise CacheStorageKeyNotFoundError("Key not found in mem cache")
streamlit.runtime.caching.storage.cache_storage_protocol.CacheStorageKeyNotFoundError: Key not found in mem cache

Traceback (most recent call last):

```
  File "/mnt/tools/miniconda3/envs/amdirt/lib/python3.9/site-
packages/pandas/core/indexes/base.py", line 3790, in get_loc
    return self._engine.get_loc(casted_key)
  File "index.pyx", line 152, in pandas._libs.index.IndexEngine.get_loc
  File "index.pyx", line 181, in pandas._libs.index.IndexEngine.get_loc
  File "pandas/_libs/hashtable_class_helper.pxi", line 7080, in
pandas._libs.hashtable.PyObjectHashTable.get_item
  File "pandas/_libs/hashtable_class_helper.pxi", line 7088, in
pandas._libs.hashtable.PyObjectHashTable.get_item
KeyError: 'archive_accession'

  File "/mnt/tools/miniconda3/envs/amdirt/lib/python3.9/site-
packages/streamlit/runtime/caching/storage/in_memory_cache_storage_wrapper.py", line 137, in
_read_from_mem_cache
    raise CacheStorageKeyNotFoundError("Key not found in mem cache")
streamlit.runtime.caching.storage.cache_storage_protocol.CacheStorageKeyNotFoundError: Key
not found in mem cache
```

References
1. Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, et al.: Genomes OnLine database (GOLD) v.7: updates and new features.*Nucleic Acids Res*. 2019; **47** (D1): D649-D659 PubMed Abstract | Publisher Full Text
2. Bernstein MN, Doan A, Dewey CN: MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive.*Bioinformatics*. 2017; **33** (18): 2914-2923 PubMed Abstract | Publisher Full Text
3. Shi W, Qi H, Sun Q, Fan G, et al.: gcMeta: a Global Catalogue of Metagenomics platform to support the archiving, standardization and analysis of microbiome data.*Nucleic Acids Res*. 2019; **47** (D1): D637-D648 PubMed Abstract | Publisher Full Text
4. Kasmanas JC, Bartholomäus A, Corrêa FB, Tal T, et al.: HumanMetagenomeDB: a public repository of curated and standardized metadata for human metagenomes.*Nucleic Acids Res*. 2021; **49** (D1): D743-D750 PubMed Abstract | Publisher Full Text

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the

findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, metagenome, metadata standardization, machine learning

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 06 May 2024

James Fellows Yates

Thank you for your generally positive review. We apologise for the long time between your review and our revision, this was due to parental leave and extensive travelling of the corresponding authors.

Our responses to your specific comments are below. Line numbers in our revised manuscript should correspond to the submitted revised word document (required by F1000), once track changes are accepted. Please note depending on which software we opened it, the line numbers were not indicating the same content. The line numbers should correspond to what is opened in Microsoft Office for Mac Word v16.78.3.

Manuscript Comments

- *During the introduction section, I missed a clear definition of ancient metagenome. This should include a description of what requires the ancient samples to receive special treatment regarding metadata annotation. Naturally, this leads to comparing your effort to past works like the GOLD (https://doi.org/10.1093/nar/gky977), the MetaSRA (https://doi.org/10.1093/bioinformatics/btx334), the gcMeta (https://doi.org/10.1093/nar/gky1008), or the HumanMetagenomeDB (https://doi.org/10.1093/nar/gkaa1031). I appreciate the extensive AMDirT documentation, including tutorials and videos. However, the manuscript lacks in the description of the tool's workflow and usability. The "Use cases" section of the manuscript makes the tool much clearer. At the same time, it gives the impression that AMDirT is limited to the GUI exploration and template generation guidance. Specifically, in my opinion, the manuscript poorly describes and explores the "Validate" and "Autofill" commands. For example, the Validate command does "a variety of checks." It would help if you made it more transparent. Do you have a document detailing the checks and standards? If so, it should be placed here; if not, I would like to see it in the manual.*

Response: For the ancient metagenome definition: this was originally defined in the AncientMetagenomeDir publication but we've added a short sentence rephrasing and citing that definition on line 36-37.

For validate: we do not go into much detail on this, as it again was originally described in the first AncientMetagenomeDir paper (as cited in the main text) and as such does not

necessarily represent 'new' functionality other than a renaming. Furthermore, as it is not a user facing component of the tool kit, we feel it is less relevant for readers (who we presume many more will be potential users rather than developers wishing to adopt a similar scheme). However we have rephrased the text slightly to indicate which person each command is most relevant to (e.g. contributor vs maintainer vs user etc.). The same goes for autofill, in that it is not a user facing tool - as it typically will be just within GitHub actions. We have however updated the manuscript to provide a little more detail on the validate command (line 105-140) and for autofill (which calls the ENA API for information) in lines 100-103. We have otherwise also expanded the use case to also describe in more detail the 'convert' CLI command, which is also a primarily user-facing subcommand.

- *For the Autofill command, I missed a more detailed description of the "improvements" done using R scripts after pulling the metadata from ENA. Specifically, since those commands could be used outside the realm of ancient metagenomic samples. I believe a figure containing the complete schematic workflow from AMDirT would help.*

Response: We have removed references to the R scripts used in the very initial pulling of data in preparation for the first library-level hackathon. On reflection, indeed this could be confusing to the reader, particularly as these are no longer used - and were only used for a one off event. The same concepts have been added to autofill (which came later). In autofill, the ENA portal API is directly queried using Python. We have added more details on the retrieved metadata fields (line 100-103).
We have further now reduced the necessity of a user to run this command at all (see above about not being a user-facing tool), instead having a GitHub actions bot to run this command for a user on the AncientMetagenomeDir repository, during a pull request.
We have included a new figure (Figure 2) that shows the updated AncientMetagenomeDir workflow modified after the original AncientMetagenomeDir publication workflow diagram.

- *Regarding the software operation, I missed a storage requirement specification. The paper mentions that the installation requires Python 3.9, but nothing is specified in the GitHub. The installation via pip and conda should include guidance on accessing those tools and creating a specific environment for the AMDirT.*

Response: We unfortunately don't fully understand what the reviewer means exactly by 'storage requirement specification' here. However, based on our interpretation: all downloadable AMDirT files are small bash scripts, TSV, or bibtext files, all of which take up (bioinformatically) negligible amounts of space - single digit megabytes or less. Thus we do not feel it is necessary to mention this. However if the storage requirements refers to FASTQ files etc, as a convenience, AMDirT offers an estimate of the potential total size of the files if you were to run the download script. This is displayed by hovering over the download script download button - however AMDirT does not download FASTQ files itself. This functionality was already described on lines 452.

The AMDirT package zip archive itself is also less than a megabyte so we feel is not relevant here as it will not make an impact on storage requirements in the vast majority of cases.
For installation requirements: we are not entirely sure what the reviewer means by 'accessing these tools' - whether this is the downstream dependencies, or the package managers pip/conda themselves. If the former, the version of dependencies management is automatically handled by pip/conda, and is completely transparent for the user as this is

reported by each package manager when installing AMDirT. For advanced bioinformaticians, they are likely familiar with python packages setup.py and conda environment files, so should be able to find these themselves (as an everyday user does not need this). If the reviewer is referring to installation instructions of pip and conda, this is a fair point and we have added in the installation instructions URLs to the installation documentation pages of the respective package managers, and we have updated the README to provide updated guidelines to install AMDirT in a dedicated conda environment.

- *Finally, as a minor opinion, have you considered using specialized sample accessing software? For instance, the SRAtoolkit can download samples from the SRA more efficiently than curl.*

Response: As an alternative to curl, AMDirT already provides download scripts using other tools such as aspera and nf-core/fetchngs, which itself, already, provides the option to download with SRA-tools. We picked these tools based on the interest of AncientMetagenomeDir contributors - of which SRAToolkit was not one at the time, however we have added this option in AMDirT v1.6 at the reviewers request.

- *If possible, I would like to see a slight reformat in the discussion "Problem/Solutions" section to make it more visible. A simple extra paragraph or bullet points that clearly separate problems and solutions should solve the problem.*

Response: We are unfortunately somewhat beholden to the formatting of the publisher during typesetting, but we have attempted to reformat the paragraph according to the reviewer's wish with paragraphs (as each text is too long for a bullet point)

Running and installing the pipeline.
- *I had a relatively smooth experience. However, I kept receiving the following warning upon any usage:*
"WARNING streamlit.runtime.caching.cache_data_api: No runtime found, using MemoryCacheStorageManager"
This is unfortunately a known issue resulting from a design decision of one of AMDirT dependencies: the streamlit library and its caching system. As a stop gap solution, we now have 'monkeypatched' the streamlit function triggering this warning to silence it (https://github.com/SPAAM-community/AMDirT/pull/132).
  - *Additionally, the convert command did not work. After successfully installing the tool and downloading the AncientMetagenomeDir. I received the following:*

    $ AMDirT convert --curl SPAAM-community-AncientMetagenomeDir-60f8a00/ancientmetagenome-hostassociated/libraries/ancientmetagenome-hostassociated_libraries.tsv ancientmetagenome-hostassociated -o ./2023-10-17 19:50:32.289 No runtime found, using MemoryCacheStorageManager

Response: The convert command by default takes as input a *sample* table, not a library table, which is already reflected in the documentation of the convert command (amdirt.readthedocs.io/en/latest/how_to/convert) and the convert command CLI help.

```
$ AMDirT convert --help
Usage: AMDirT convert [OPTIONS] SAMPLES TABLE_NAME
  Converts filtered samples and libraries tables to eager, ameta, taxprofiler, and fetchNGS
input tables
  Note: When supplying a pre-filtered libraries table with `--libraries`, the
  corresponding sample table is still required!
  SAMPLES: path to filtered AncientMetagenomeDir samples tsv file
  TABLE_NAME: name of table to convert
Options:
 -t, --tables PATH   (Optional) JSON file listing AncientMetagenomeDir
                     tables
 --libraries FILE    (Optional) Path to a pre-filtered libraries table
                     NOTE: This argument is mutually exclusive with
                     arguments: [librarymetadata].
 --librarymetadata   Generate AncientMetagenomeDir libraries table of all
                     samples in input table NOTE: This argument is
                     mutually exclusive with  arguments: [libraries].
 -o, --output DIRECTORY  conversion output directory  [default: .]
 --bibliography      Generate BibTeX file of all publications in input
                     table
 --curl              Generate bash script with curl-based download
                     commands for all libraries of samples in input table
 --aspera            Generate bash script with Aspera-based download
                     commands for all libraries of samples in input table
 --fetchngs          Convert filtered samples and libraries tables to nf-
                     core/fetchngs input tables
 --sratoolkit        Generate bash script with SRA Toolkit fasterq-dump
                     based download commands for all libraries of samples
                     in input table
 --eager             Convert filtered samples and libraries tables to
                     eager input tables
 --ameta             Convert filtered samples and libraries tables to
                     aMeta input tables
 --mag               Convert filtered samples and libraries tables to nf-
                     core/mag input tables
 --taxprofiler       Convert filtered samples and libraries tables to nf-
                     core/taxprofiler input tables
 --help              Show this message and exit.
```

Thus, for example, running the convert command with a sample table will generate this output:

```
$ AMDirT download -t ancientmetagenome-hostassociated -y samples -r v23.12.0
```

```
$ head -n 10 ancientmetagenome-hostassociated_samples_v23.12.0.tsv > test.tsv
$ AMDirT convert --curl test.tsv ancientmetagenome-hostassociated
AMDirT [WARNING]: We provide no warranty to the accuracy of the generated input sheets.
AMDirT [INFO]: Writing curl download script
$ head AncientMetagenomeDir_curl_download_script.sh
#!/usr/bin/env bash
curl -L ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR132/022/SRR13263122/SRR13263122_2.fastq.gz -
o SRR13263122_2.fastq.gz
curl -L ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR132/024/SRR13263124/SRR13263124_2.fastq.gz -
o SRR13263124_2.fastq.gz
curl -L ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR132/031/SRR13263131/SRR13263131_2.fastq.gz -
o SRR13263131_2.fastq.gz
curl -L ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR132/030/SRR13263130/SRR13263130_2.fastq.gz -
o SRR13263130_2.fastq.gz
curl -L ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR957/SRR957740/SRR957740.fastq.gz -o
SRR957740.fastq.gz
curl -L ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR957/SRR957744/SRR957744.fastq.gz -o
SRR957744.fastq.gz
curl -L ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR957/SRR957741/SRR957741.fastq.gz -o
SRR957741.fastq.gz
curl -L ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR957/SRR957738/SRR957738.fastq.gz -o
SRR957738.fastq.gz
curl -L ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR132/028/SRR13263128/SRR13263128_2.fastq.gz -
o SRR13263128_2.fastq.gz
```
```

However we've added an additional validation check to ensure that the input SAMPLE object is in the valid format expected by the tool, and if not asks the user to check if it is a sample table (https://github.com/SPAAM-community/AMDirT/pull/140, commit 7bb6b55)
We have also added support for supplying *library* level metadata tables (supplied alongside a corresponding samples table) to the convert command with the --libraries parameter. This, alongside the new command download (added since the submission of the original manuscript), allows for an almost end to end command line based interaction with AMDirT: download → (filtering with bash tools) → convert, as reflected in the updated use case section of the manuscript.

***Competing Interests:*** No competing interests were disclosed.

Reviewer Report 09 October 2023

✔ **Timothy Read**

Emory University, Atlanta, Georgia, USA

**Robert Petit III**

Wyoming Public Health Laboratory, Cheyenne, Wyoming, USA

This paper describes the 'AMDirT' open source software for interaction with ancient metagenome sequence sample metadata tables. The tables have been produced through a massive communal curation effort by the international SPAAM (The Standards, Precautions, and Advances in Ancient Metagenomics community) group to systematically improve on what is available from the public repositories of archived short read sequence data. AMDirT allows viewing and searching of the tables using the Python streamlit library and facilitates download of data and integration with open source nextflow analysis pipelines. AMDirT also helps with initial creation and validation of new metadata tables.

I successfully installed AMDirT version 1.4.6 using conda on my Intel MacBook following instructions in the manuscript. I was able to replicate the commands described. There is also a public server https://www.spaam-community.org/AMDirT/ for the AMDirT viewer. The documentation for the software is very good and includes video tutorials.

The manuscript is well-written and clearly outlines the functions of the software. The development of AMDirT represents a significant effort, not least because of the community-wide consultation. The need for community-based metadata addresses a well-known problem with the SRA/ENA/DDBJ databases and the discussion provides some nice examples of the type of issues faced when downloading. AMDirT is a valuable tool for both the ancient metagenomics research community and those outside the community interested in browsing and accessing the data.

**Specific Points**
- The streamlit-based viewer is pretty slick but I feel the importance of command line interface (CLI) is a little underplayed here. The viewer can be used by people without CLI proficiency but to actually take actions like making new sample tables, or download and process data, CLI is essential.

- The instructions in the text and on github for conda install should guide users to install into a fresh conda environment rather than into the base.

- There should be a list of computational environments that the software has been tested on (e.g do M1-3 macs work?)

- I did not see a mention in the text that there is actually a public facing server https://www.spaam-community.org/AMDirT/

- "Newly added library information columns include the library name (how data are typically reported in original publications), the aDNA library generation method (e.g., double-stranded or single-stranded libraries), the library indexing polymerase (e.g., proof-reading or non-proofreading), and the library pretreatment method (e.g., non-Uracil-DNA Glycosylase (UDG), full-UDG, or half-UDG treatments). The latter three fields represent information about the sequencing library construction that influence the presence of aDNA

damage, a factor that is critical for the processing of aDNA NGS data.8,18 Sequencing metadata columns include instrument model, library layout (single- or paired-end), library strategy (whole genome sequencing, targeted capture, etc.), and read count. " I could not work out how to search these fields through the viewer.  These are accessible as downloads post validation but it seems that users would want to search through samples based on these fields?

**Further questions**

Finally, I have some three questions about the project that it would be great to get comments on.
1. How sustainable is the community effort to maintain these databases moving forward into the future with the probability of the number of samples increasing each year?

2. Have you tried sending the improved metadata back to SRA/ENA?

3. How difficult would it be to adapt the software for a different community of researchers that wanted to improve annotation but use fields that would be different from the ancient metagenomes community?

**Is the rationale for developing the new software tool clearly explained?**
Yes

**Is the description of the software tool technically sound?**
Yes

**Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?**
Yes

**Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?**
Yes

**Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?**
Yes

*Competing Interests:* No competing interests were disclosed.

*Reviewer Expertise:* Bacterial genomics, metagenomics, bacterial genetics, antibiotic resistance

**We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Author Response 06 May 2024
**James Fellows Yates**

Thank you for your positive review. We apologies for the long time between your review and our revision, this was due to parental leave and extensive traveling of the corresponding authors.

Our responses to your specific points and further questions are as follows. Line numbers should correspond to the submitted revised word document (required by F1000), once track changes are accepted. Please note depending on which software we opened it, the line numbers were not indicating the same content. The line numbers should correspond to what is opened in Microsoft Office for Mac Word v16.78.3.

***Specific Points***
- *The streamlit-based viewer is pretty slick but I feel the importance of command line interface (CLI) is a little underplayed here. The viewer can be used by people without CLI proficiency but to actually take actions like making new sample tables, or download and process data, CLI is essential.*

**Response:** We have made more prominent references to the command line interface functionalities in the manuscript by further specifying in the introduction on lines 75-88, 101-140, and 151-155.
We also extended the 'Use case' section of the manuscript to further guide readers on how they could use AMDirT command line tools to perform the same workflow in the GUI as shown in Figure 3. Note that during the improvement of this use case, we also added an additional command called 'download' to further make it easier for CLI power-users to retrieve the tables for a more complete CLI based workflow.
- *The instructions in the text and on github for conda install should guide users to install into a fresh conda environment rather than into the base.*

**Response:** We have updated the documentation to reflect the new instructions in the code repository README, documentation pages (both can be seen in https://amdirt.readthedocs.io/en/master/README.html), and the manuscript line 263
- *There should be a list of computational environments that the software has been tested on (e.g do M1-3 macs work?)*

**Response:** We have added a brief section about hardware and OSs that have been tested in the github repository README installation page of the documentation ( https://amdirt.readthedocs.io/en/master/README.html).
- *I did not see a mention in the text that there is actually a public facing server https://www.spaam-community.org/AMDirT/*

**Response:** There was already a mention of the viewer interface at the line 369, but we have added more details lines 76, and 759

- *"Newly added library information columns include the library name (how data are typically reported in original publications), the aDNA library generation method (e.g., double-stranded or single-stranded libraries), the library indexing polymerase (e.g., proof-reading or non-proofreading), and the library pretreatment method (e.g., non-Uracil-DNA Glycosylase (UDG), full-UDG, or half-UDG treatments). The latter three fields represent information about the sequencing library construction that influence the presence of aDNA damage, a factor that is critical for the processing of aDNA NGS data.8,18 Sequencing metadata columns include instrument model, library layout (single- or paired-end), library strategy (whole genome sequencing, targeted capture, etc.), and read count. " I could not*

*work out how to search these fields through the viewer. These are accessible as downloads post validation but it seems that users would want to search through samples based on these fields?*

**Response:** We have now extended the AMDirT viewer and convert commands to support library level filtering. In the AMDirT viewer this is supported via a secondary table that is loaded after the sample selection has been made by the user. We've updated the use case example and Figure 2 to represent these changes.

### *Further questions*

*Finally, I have some three questions about the project that it would be great to get comments on.*
  - *How sustainable is the community effort to maintain these databases moving forward into the future with the probability of the number of samples increasing each year?*

**Response:** Ancient metagenomics as a field is still very young and therefore has indeed been very manageable in terms of numbers. But of course as it grows it could become more difficult.

In passive terms - our feeling so far is as we have started as the field is 'young' has benefited from the fact we will have good visibility and awareness of the resource within the community, something that would continue to spread as the research area expands through knowledge transfer. With this, we believe that we will have both a) more potential and willing contributors to cope with the increased number of samples and b) as they are already aware, will maybe already format and prepare their own sample and library metadata in a form easy to add to AncientMetagenomeDir.

More proactively, we will continue to survey the community to find out what are good motivating factors to continue contributing. We further plan to continue to refine the tooling and automation to make it as easy as possible for contributors to add and review new contributors. The SPAAM community has also recently become affiliated with the ISBA society ([https://isbarch.org](https://isbarch.org)), which can provide financial support for the previously 'volunteer' held hackathons. We also have various ideas for new publications that are typically the biggest motivating factor for researchers (e.g. extending the database to include more precise radiocarbon dates which are critical e.g. for phylogenetic dating analyses). Finally we have also expanded our aim of better metadata reporting of ancient DNA samples and sequencing data to a larger project covering the whole of palaeogenomics ([https://www.mixs-minas.org/](https://www.mixs-minas.org/)) that will also help further formalise such thinking in the community, and feed-back into the way AncientMetagenomeDir functions, as well as push such standardisation into the INDSC consortium databases (who use MIxS checklists themselves).

  - *Have you tried sending the improved metadata back to SRA/ENA?*

**Response:** We have not (yet) attempted this. We have not actively recorded all the corrections between the reported information in the publications and as changed for the AncientMetagenomeDir repository. These corrections have been also derived from a mixture of 'obvious' mistakes in the publication (e.g. single-end sequencing reported, but paired FASTQ files uploaded to the ENA or SRA), to personal communications with the authors of the publication. In some cases the original authors have themselves made the

corrections on ENA/SRA themselves. In any case, requesting changes on SRA/ENA would require permission from the original authors of the publications, which would require a large and extensive effort from the volunteers of SPAAM, with likely highly mixed chances of success (students move on, labs have changed/retired etc). In this vein we feel it would be more beneficial to invest time and effort in the MInAS project to correct future submissions at the source rather than historical data.

- ○ *How difficult would it be to adapt the software for a different community of researchers that wanted to improve annotation but use fields that would be different from the ancient metagenomes community?*

**Response:** AMDirT is mostly based around the JSON schemas that define and enforce the structure of the underlying data. Furthermore, because the validate subpackage is designed using a domain-driven design software architecture, the base class holding the different methods responsible for testing the data are generic and can be applied to any dataframe object, as long as they come with their associated JSON schema. In summary, adapting it to a different community would require some modifications (as some part of the code is specific to ancient metagenomics data), but most of AMDirT core functionalities could be adapted without too much effort.

***Competing Interests:*** We declare no competing interests.

---

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias

- You can publish traditional articles, null/negative results, case reports, data notes and more

- The peer review process is transparent and collaborative

- Your article is indexed in PubMed after passing peer review

- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000 Research