Improving access to ancient metagenomic data with the community-curated metadata repository

Ancient Metagenome Dir

James A. Fellows Yates,
Aida Andrades Valtueña,
Åshild J. Vågene, Becky Cribdon, Irina M. Velsko,
Maxime Borry, Miriam J. Bravo-López, Antonio Fernandez-Guerra,
Eleanor J. Green, Shreya L. Ramachandran, Peter D. Heintzman, Maria A. Spyrou,
Alexander Hübner, Abigail S. Gancz, Jessica Hider, Aurora F. Allshouse, Christina Warinner.

Finding relevant sequencing data is not trivial

Public data is useful for:	But can be hard to find
- comparative data generation - meta-analysis variables - public data reanalysis - field development tracking	- incomplete metadata (missing fields) - inconsistently reported metadata (lat:lon vs. site name; cal vs uncal ^{14}C) - data in different archives (ENA; Institute FTP) - stored in papers, not with data itself

Simplify with community curated sample lists

AncientMetagenomeDir is:

Ancient metagenomic sample lists (TSV file) of *basic* but *essential* and *consistent* metadata

Citation information: *DOIs*

Geographic location: *Controlled country list, Lat:Lon*

project_name	publication_year	publication_doi	site_name	latitude	longitude	geo_loc_name	sample_name	sample_host
Warinner2014	2014	10.1038/ng.2906	Dalheim	51.565	8.84	Germany	B61	Homo sapiens
Warinner2014	2014	10.1038/ng.2906	Dalheim	51.565	8.84	Germany	G12	Homo sapiens
Weyrich2017	2017	10.1038/nature21674	Gola Forest	7.657	-10.841	Sierra Leone	Chimp	Pan troglodytes
Weyrich2017	2017	10.1038/nature21674	El Sidrón Cave	43.386	-5.328	Spain	ElSidron1	Homo sapiens nean
Weyrich2017	2017	10.1038/nature21674	El Sidrón Cave	43.386	-5.328	Spain	ElSidron2	Homo sapiens nean
Weyrich2017	2017	10.1038/nature21674	Spy Cave	50.48	4.674	Belgium	Spy1	Homo sapiens nean
Weyrich2017	2017	10.1038/nature21674	Spy Cave	50.48	4.674	Belgium	Spy2	Homo sapiens nean

| sample_age	sample_age_doi	community_type	material	collection	archive	archive_project	archive_accession
900	10.1038/ng.2906	oral	dental calculus	NA	SRA	PRJNA216965	SRS473742,SRS473743,SRS47374
900	10.1038/ng.2906	oral	dental calculus	NA	SRA	PRJNA216965	SRS473747,SRS473746,SRS47374
100	10.1038/nature21	oral	dental calculus	NA	SRA	PRJNA485265	SRS7890499
49000	10.1038/nature21	oral	dental calculus	NA	SRA	PRJNA485265	SRS7890498
49000	10.1038/nature21	oral	dental calculus	NA	SRA	PRJNA485265	SRS7890496
35800	10.1038/nature21	oral	dental calculus	NA	SRA	PRJNA485265	SRS7890491
36000	10.1038/nature21	oral	dental calculus	NA	SRA	PRJNA485265	SRS7890487

Sample ages: *rounded to last 100y*

Sample type: *term ontologies*

Links to data: *archives, accession IDs*

Release v20.09 stats:

- Publications: 87
- Total Samples: 1024
- Ancient host microbiomes: 443
- Ancient microbial genomes: 269
- Ancient sediment metagenome: 312
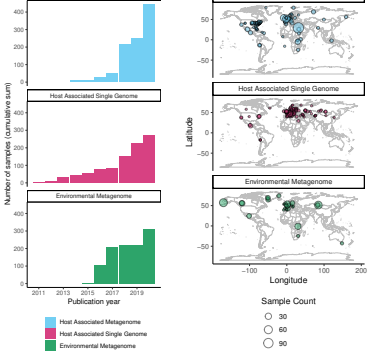
Benefits for your research

- Metadata easy to access and to update (simple spreadsheet format)
- Standardised fields: rapid filtering to samples that you actually want
- Takes you to the raw data (or back to original publication)
- Comprehensive: includes most of published samples of field
- Regularly updated: quarterly releases

Ensuring quality metadata: template for other fields

Suggestions for other fields - use:

Community
- Field-wide consensus of definitions
- Peer-review to check for accuracy
- Wider 'net' for literature screening

Collaboration
- Use collaborative tools (github, slack)
- Online platforms allows communication across time-zones (inclusivity)

Automation
- Metadata field specification checks
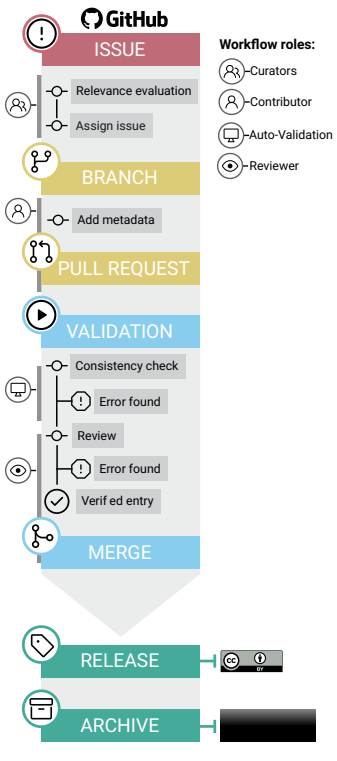- Long-term release archiving

Accessibility
- Write documentation for new contributors
- Offer training workshops for new concepts and skills (e.g. github)

Integration
- Utilise existing resources for linked data (term ontologies, data archives)

Workflow roles:
- Curators
- Contributor
- Auto-Validation
- Reviewer

GitHub
- ISSUE
  - Relevance evaluation
  - Assign issue
- BRANCH
  - Add metadata
- PULL REQUEST
- VALIDATION
  - Consistency check
    - Error found
  - Review
    - Error found
    - Verified entry
- MERGE
- RELEASE
- ARCHIVE

aDNA metadata reporting needs improving

Future directions

- Expand to library-level metadata
- Make project infrastructure template for use by other fields
- Develop standardised aDNA-wide metadata reporting sheet?

**Want to get involved?
Join the SPAAM community!**
https://spaam-community.github.io/

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

Affiliations

Archaeogenetics, Max Planck Institute for the Science of Human History
Institut für Vor- und Frühgeschichtliche Archäologie und Provinzialrömische Archäologie, LMU München
Evolutionary Genomics, GLOBE Institute, University of Copenhagen
School of Life Sciences, University of Warwick
LIIGHM, National Autonomous University of Mexico
GeoGenetics, GLOBE Institute, University of Copenhagen
BioArCh, Department of Archaeology, University of York
Human Genetics, University of Chicago
The Arctic University Museum of Norway, UiT The Arctic University of Norway
Department of Anthropology, Pennsylvania State University
Department of Anthropology, McMaster University
Department of Anthropology, Harvard University
Department of Biology, Università degli Studi di Firenze