# SPAAM2 Session Notes

*Session 4 - Sorting and Recycling the Trash: authentication, standards, and reproducibility in ancient metagenomic research*

**Chairs:** Clio der Sarkissian

**Icebreaker Presenters**: Miriam Bravo, Nikolay Oskolkov, James Fellows Yates (+ Antonio FernandezGuerra)

## Session Abstract

Authenticating ancient metagenomic datasets has long been considered equally a nearly impossible and crucial task to demonstrate that 'true' ancient metagenomes can be characterised. Postmortem fragmentation (typically via depurination) and base modification (via cytosine deamination) make aDNA analysis very sensitive to exogenous 'contaminating' DNA that can confound downstream analysis. Validation of results is therefore imperative not only to ensure scientific rigour, trust in the methods by researchers from related fields (e.g., anthropologists, archaeologists), but can also have important ethical consequences. Defining a consensus-based legitimate set of minimum-authentication criteria for ancient metagenomics is therefore crucial for gaining this trust in the field and facilitating further exploration of this type of data.

Secondly, following standard guidelines by researchers can only be expected if these standards and guidelines follow FAIR principles: Finable, Accessible, Interoperable, and Reusable. Making sure authentication methods and reporting guidelines can be used and applied by any type of lab (big or small, specialist or generalist), can only come by making sure software and data is accessible, usable and most importantly understandable.

In part one of this two-part session we will: i) get an overview of the obstacles and possible solutions of current and potential future approaches for authenticating metagenomic datasets, ii) begin to define a minimal authentication line of evidence and iii) discuss how to communicate this to collaborators inside and outside the ancient metagenomic community. In part two we will discuss i) how to improve accessibility and usability of ancient metagenomic datasets and software such as through collaboratively generated benchmarking/comparative datasets as well as metadata reporting, and ii) how to ensure responsible and ethical research conduct.

## Session Introduction (Clio der Sarkissian)

- Legal and ethical responsibilities as well as technical responsibility with our data ;
- Emphasizing the need for transparency with our work
- How the general public can perceive our work is also important

- We are a young field of research
    - Ethical questions and assistance with that
    - Minimal line of authentication a necessity?
    - How to communicate guidelines and how to make them evolve
    - Communicate guidelines outside SPAAM and other people in the field.

- Definitions of the session presented

# Icebreaker 1 (Miriam Bravo)

## Presentation

- Ethics not yet extended to ancient pathogenomics/metagenomics
    - define ethics
    - we should care because ancient remains are limited, dignified treatment of human and non human remains,: language, storage, techniques,
    - often used objectifying terms, treated as just data
    - poor storage conditions
- Legacies of colonialism  exporting ancient remains to foreign labs because of funcindg, n o long-term collaborations or capacity building
- Strong bias on ancient european studies because of funding
- "Chasing the next Nature paper" common situation
- Competition with question-driving research
- Side-effects of unethical practices
- repeated or redundant sampling, lack of publicly available genetic data
- no transparent methods
- Journals are not accessible to countries where samples are from
- Avoid do first ask later, research budgets must include sample return and continued engagement
- Consult with local scientists and indigeneous groups
- Collaborative projects are needed to progress the field

## Q & A

- **Q**: Are there points that are specific to ancient pathogen work?
    - **A**: Everything is equally important
- **Q**: How applicable is ethics to microbial samples ? competitive sampling ; transparency of methods and exact pipelines that we use
    - **A**:Competitive sampling remains an issue even in microbial samples. There is a lot of relevant cultural and biological significant information in microbial samples and there remains a lack of transparency in methods and pipelines used.
- **Q:** Are there any Mexican regulations regarding who owns the data? Would the data be owned by the place that generated the data or the origin of the sample?
    - Currently there are Mexican regulations prohibiting the analysis of samples unless there are Mexican researchers actively involved in the study. At their group there is a strong collaborative nature with archaeologists, with regular meetings and sharing of information.
-

# Icebreaker 2 (Nikolay Oskolkov, James Fellows Yates (+ Antonio FernandezGuerra))

## Presentation

- Summary from yesterday: Who's there, where do they come from, source-tracked data ancient or modern, am I doing this right?
- How far can we go with cleaning?
- Supervised database guided vs unsupervised intrinsic read composition
- Tax level genus, species, strains, whole genomes, etc?
- Correct naming? Phylogenetic techniques, kmers with training sets?
- Is what you're comparing the same as what you're comparing it to?
- Worth discussing
- Reproducibility, parity guidelines (differences in access to computer resources, software should be designed so most people can use it), metagenomic analysis reducing artefacts, improving quantifications (comparability within and between labs, standardizing genotypes that are shared)
- Define projects beforehand
- Centralized place for discussing
- Secure continuity regular spaam meetings
- Reducing entropy working committees
- **Poll**: what are topics of most interest (part 1)?
    - Data clean-up: How far can we go?: 18% (3)
    - Taxonomic levels: What's the right size for the lens?: 0% (0)
    - Naming the things right: the wheat and the chaff: 0% (0)
    - SOPs, good practices and reproducibility: Is that method good/better?: 82% (14)
    - Total Votes: 17

- **Poll**: what are topics of most interest (part 1)?
    - Reproducibility and continuity: 41% (7)
      Parity guidelines (computing resources): 0% (0)
    - Metagenomic analyses: Reducing artifacts, improving quantifications: 59% (10)
    - Total Votes: 17

## Standards in metadata reporting

- aDNA quite good at sharing data publicly
- But how do we get metadata after because finding that can be quite convoluted
- Differences in where to find it and how it's reported

## AncientMetagenomeDir

- tables with uniformized metadata for samples
- tsv files easy to access
- everyone can contribute and download
- regular releases, with Zenodo id that an be cited
    - There's a LOT of variation in metadata reporting that makes extraction difficult
- some general problems: cal/uncal C14, differed BC/BP/AD
- location not there or different language
- sample names aren't consistent between pub and ENA/SRA
- run/sample/library codes are not consistently used
- Upload TSV or excel tables and not TXT embedded tables that can't be parsed

- How to ensure consistency in reporting?
    - What is the crucial information to report?
    - Where are samples/individuals stored?
    - How to maintain such standards?
    - Follow a method that fits INSDC?
    - Min info about any sequences: MIxS as a framework for our metadata reporting
    - MInAS: Min Info about an Ancient Sequence  a group to set these standards (if you want to join let them know)
- Proposal for 3 components: human popgen, pathogen, environmental, each with set levels defined by people in the field (so they need people to join and help out)

## Q & A

- *[Note taker: no immediate questions]*

**Polls**:

- How many people in your lab are also working in metagenomics?
    - I'm all alone: 26% (7)
    - 2-3: 44% (12)
    - 4-6: 11% (3)
    - more than 6: 19% (5)
    - Total Votes: 27
        - **A**: It's two in my lab, but we're both starting basically from scratch

# General Discussion

- SOPs, good practices and reproducibility: Is that method good/better?
- What specifically?
- How to select a tool for analysis
- **Q**: difficult to start analysis because very few papers published to use as guides and LOTS of software coming out  how to know which was appropriate to use or to try?
- **Q**: for someone with no experience in benchmarking, what metrics to use to select the best tool for the data?
- Dealt with this issue by just using what made most sense from reading papers (modern?)
    - i.e. Gloor paper and rarefaction papers
    - https://www.frontiersin.org/articles/10.3389/fmicb.2017.02224/full
    - https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003531
- (The community) need to read or do more benchmarking specifically for aDNA
- Ancient standard for testing so that tools are tested equally are needed (such as with CAMI)
- Explanation of the CAMI challenge because only about 4 people raised in in knowing what it's about
- Has anyone used reference frames
- **Q**: How can we screen all the new methods that come out?
    - **A:** list of what they tried and they found and what they chose and why would be helpful
- Reasons people don't take up testing: Funding issues for testing, PIs not interested in testing, people aren't comfortable in taking this up for those reasons; time and cost benefit not in favour of testing
- The purpose of SPAAM: use this community for support in this

- Less interest in methods/software testing from PIs either from being shut down or just they aren't interested (because it's not publishable (in high-impact journals))
- Why don't people think about this? Need the right dataset and how does one get this to be able to start? More interesting to do data analysis and get a story than to do the testing
- Being the only one in a group doing this work and having no one in the group to go to for help is very hard; not a lot of encouragement to get outside help from PIs also
- lots of stuff to get through and no clear place to start, too many options
- lack of technical support and pressure from PIs to get projects analyzed and done
- need time and resources that aren't always there
- **Q**: what problems do people have? -people are hesitant to share (no volunteers)

## Data upload/sharing

- Think about the metadata being organized and recorded as for the Dir at the time of data upload to encourage full reporting and consistency
- **Q**: Are there templates that can be used? For example we use different methods to upload blanks - how can we do this consistently?
- **Q**: how did people deal with issues that came up in uploading? Look for instructions, google the issue, email someone?
- Consider metadata from the beginning of the project (fits in with Miriams' presentation about ethical setup)
- Handbook for project design and development like used for Earth Microbiome Project outlines the data that needs to be collected
- There are people who will help with project design and management and make a data management plan (within Germany but not sure if it's available outside Germany) https://www.gfbio.org/
- CASCADE (Toulouse's LIMS) metadata pilot test went well suitable for any sample; excavation site, contact info for people who gave samples, protocols, and more that's user-defined (free to Germans or people in Germany)
- Who has a handbook or directions/training for data upload? Only 2 people
- First time upload was the "training" for 1  learnging by doing; done by self
- MPI-SHH has a wiki page with directions for ENA upload
- Toulouse's LIMS has a function that makes the tables for upload
- Differences in what popgen and microbe people upload  bam vs raw fastq
- ENA separates out the reeds that map to separate chromosomes(?) and hides this in a separate column
- Uploading only the mapped bam files is increasingly done for pathogens only maybe because The human DNA isn't yet published this is problematic
- A pledge to upload the full raw data? If only mapped bam is uploaded because of using the data in different publications (human, pathogen1, pathogen2, etc)
- OSD handbook: http://www.assembleplus.eu/sites/assembleplus.eu/files/public/manual/OSD_HandbooK_2016.pdf
- And the standard: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4511511/

## Rights to uploaded data

- A discussion of who has rights to the data that's uploaded?
- People can use your data and you don't have a control over it
- PIs always add archaeologists who gave data on papers after the initial publication even when the data is publicly available, is it the same for people who upload the data? Should they be approached and invited to be put on papers by other groups that use their publicly available data?
- One strong no: if it's publicly available, it's available for everyone
- From Antonio: https://researchparasite.com/, https://academic.oup.com/gigascience/article/9/1/giz148/5691298, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5710834/
- Some people do approach the original people who generated the data because they feel it's the right thing to do, or to keep potential collaborators happy, or to respect their "right" to it as the generators of it

## Data upload timing

- Also a problem of waiting to publish data too long, it could be lost like when the copenhagen servers got flooded
- Some kind of pledge to provide raw data on a limited time period
- Antino works with a system that automatically uploads data to ENA and it's available for everyone
- This comes from taxpayer money and it's not the researchers to own by that funders require that data be made publicly available
- In Mexico different situation no money from government to do research upload only bam files for a recent paper; publications are required for degrees and so they need the data for that to guarantee getting the degree
- Example guidelines of authorship roles:
  - http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html
  - https://casrai.org/credit/
- When no authorization to work on human part, upload bam files to avoid uploading the human data
- **Poll**: Is it a requirement from funding bodies to publish ALL generated data with the grants money ?
  - Yes: 20% (1)
  - No: 80% (4)
  - Total Votes: 5
    - **A**: For European funding normally is
    - **A**: For EU projects it 100% is

## Working with ethical approval boards

- Experience with getting ethical approval for research: working with collaborators who know the ethics committee made it easier
- **Q:** Who has experience with ethical considerations?

**A:** Two participants only - and the ethical committees were quite helpful in assistance. The forms required a description of what was to be done with the samples from the lab work till the final analysis.

- Know your part very well and don't make statements you can't support
- Find out if your university /institute has someone who can help with ethics committee prep
- Norway has a special ethical committee that needs to be applied to, and also need to apply to the institute you're working with
- Be very open about your project design, outcomes, potential ethical issues (not destroy material, extent of material being intact, can't sample indigeneous Sami individuals in Norway without permission from Sami government, if find have Sami in human popgen data analysis must be stopped)
- From Maxime: https://peerj.com/about/policies-and-procedures/#authorship-criteria
- Taking modern samples for comparison there's even more you need to make sure is covered in your applications
- How to deal with human data in samples (include in the application that it will be screened but not fully analyzed)
- Can remove the human DNA from the raw files and upload that from raw data (but it will vary by the mapping parameters)

## Large-scale internal data screening

- Screening internal data  how much screening data is shared and uploaded?
- many labs are screening for pathogens or other signatures of DNA
- Then we do captures based on screening
- Screen many many more samples than are captured later and even mentioned in the data
- should the screening data be uploaded? Can be thousands for the maybe 10 that are eventually analyzed
- One lab can't do everything, so why hide the data from others?
- Not everyone considers it "hiding"  they are scared of scooping and have multiple projects they'd like to flush out
- Interest in an open letter about data sharing? We need to come to a consensus about it
- There are lot of disparate opinions about this so we need to come to some kind of consensus within the community
- **Poll**: Who would feel comfortable e.g. writing an open letter to request improved data sharing (even of 'failed' samples)
    - Yes: 100% (16)
    - No: 0% (0)
    - Total Votes: 16

## Sharing scripts:

- Are people comfortable sharing code? Some are, and soso
- Why not? We aren't bioinformaticians and can't help anyone with questions about it
- Needs to be fully independent and that takes a lot of work; code is clear, optimized, readable takes time and effort
- Tools need documented for use in publications to be useful
- lack of training in bioinformatics to do this well and we don't know how to document properly because it's assembled like a random jigsaw puzzle

- Maybe need a course on how to document your tools
- Guidelines on what to do report/how to document
- Reproducible just means it works and will run for everyone, and there does need to be training for reproducibility
- AncientMetagenomeDir also meant as a gentle introduction to github
- Just try and don't be shy and don't hide
- Issues using conda environments on cluster (takes lots of memory) and docker isn't supported on clusters because need sudo powers, singularity is but still tricky to use
- Issues with memory sharing across nodes, etc getting help really depends on your IT
- Snakemake is recommended
- Request that code is made available during the review process (some journals already request it but as a reviewer you can do that too)
- Need to get editors to understand the importance of having code available for review and to get authors to include it in initial submission
- Reproducibility of lab protocols too: protocols.io
- Methods have too much "cite a paper that cites a paper…" be thorough

## Communication with people outside the field

- Misconceptions about what can/can't be done with ancient DNA: how do you deal with this?
- Regular seminar with archaeologists where results are presented and archaeologists/anthropologist give feedback sometimes they weren't interested or they ignored, but no specific questions that are commonly asked
- Always ask about how much sample you need
- Having archaeology courses as part of the curriculum for scientists can be very beneficial
- More trust with older more established methods (C14 dating vs DNA extraction) - need to improve communication about new DNA techniques to improve trust?
- What are the backgrounds of the attendees? How many groups have "in house" archaeologists? Not many
- **Poll**: What is your background
    - Archaeology/Anth: 30% (9)
    - Biology: 67% (20)
    - Computer Science: 0% (0)
    - Other: 3% (1)
    - Total Votes: 30
        - **A**: I mean, my undergrad was economics but I jumped into "Biological Anthropology" as a masters/PhD but I guess I'm more biology
- **Poll:** Do you have an "in house" archaeologist?
    - Yes: 67% (10)
    - No: 33% (5)
    - Total Votes: 15
        - **A**: Does now-ancient-DNA from an archaeology background count?
        - **A**: Substitute 'archaeologist' with 'palaeontologist' or 'palaeoecologist', then yes.