Date : 22nd September 2020
Chair: Clio Der Sarkissian

SPAAM

Standards, Precautions & Advances in Ancient Metagenomics

Session 4: Recycling the Trash (Part 2)

Authentication, Standards, and Reproducibility in Ancient Metagenomics

Session Scope

- How to ensure responsible conduct in ancient metagenomics research?
  ETHICS + AUTHENTICATION + REPRODUCIBILITY
  - What ethical practices should be implemented and followed starting from study design to code and data sharing?
  - What is the minimum line of evidence for aDNA authentication that should be reported? How?
  - What are current obstacles to analytical/data reproducibility?
  - What solutions can we propose?
- How do we communicate on guidelines within the ancient metagenomic community?
- How do we communicate on guidelines outside the community?
- Icebreaker speakers:

  Miriam Bravo (International Laboratory for Human Genome Research, Mexico)
  Nicolás Rascovan (Pasteur Institute, France)
  James Fellows Yates (MPI for the Science of Human History, Germany)
  Antonio Fernandez-Guerra (GLOBE Institute, Denmark)

Definitions

- **Reproducibility**: ability to run another researcher's analyses and get the same results
  - E.g. specify versions of software, deposit code and *raw* data on established archives
- **Comparative datasets**: community-defined datasets researchers can use to compare their new data against
  - E.g. soil datasets to check level of environmental DNA in a calculus microbiome sample
- **Metadata**: contextual information relating to the samples
  - E.g. sampling date, geographic location, age, library treatment
- **Standards**: community-wide definition of analysis and terms
  - E.g. always report aDNA damage patterns; always provide a date for a sample in BP

SPAAM

Standards, Precautions & Advances in Ancient Metagenomics

Ethics in ancient pathogen genomics

Miriam Bravo

LIIGH-UNAM, Mexico

MiriamJBravo1

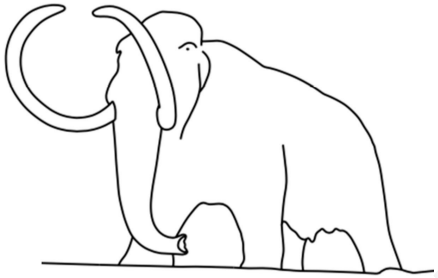
What is ethics?

“Branch of philosophy that involves systematizing, defending, and recommending concepts of right and wrong conduct”.

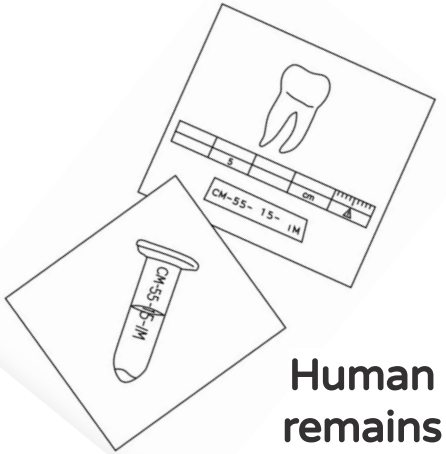
Internet Encyclopedia of Philosophy; Adventures in Archeological Science

Why we should care about it?

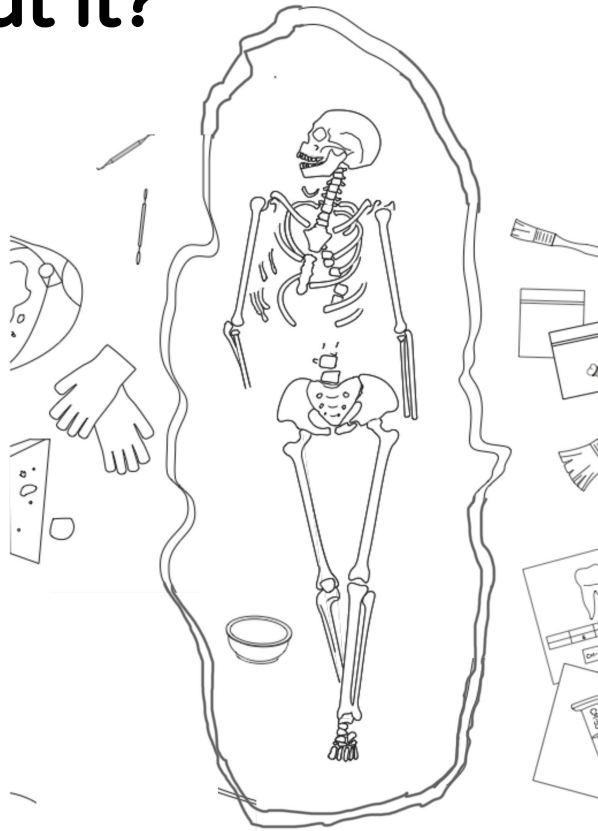
- **Ancient remains are not a limitless resource.**



Archaeofaunal remains



Human remains



- **Dignified treatment of human and non-human remains.**

  Language, storage and treatment

Pálsdóttir *et al.*, 2019; Wagner *et al.*, 2020; Adventures in Archeological Science

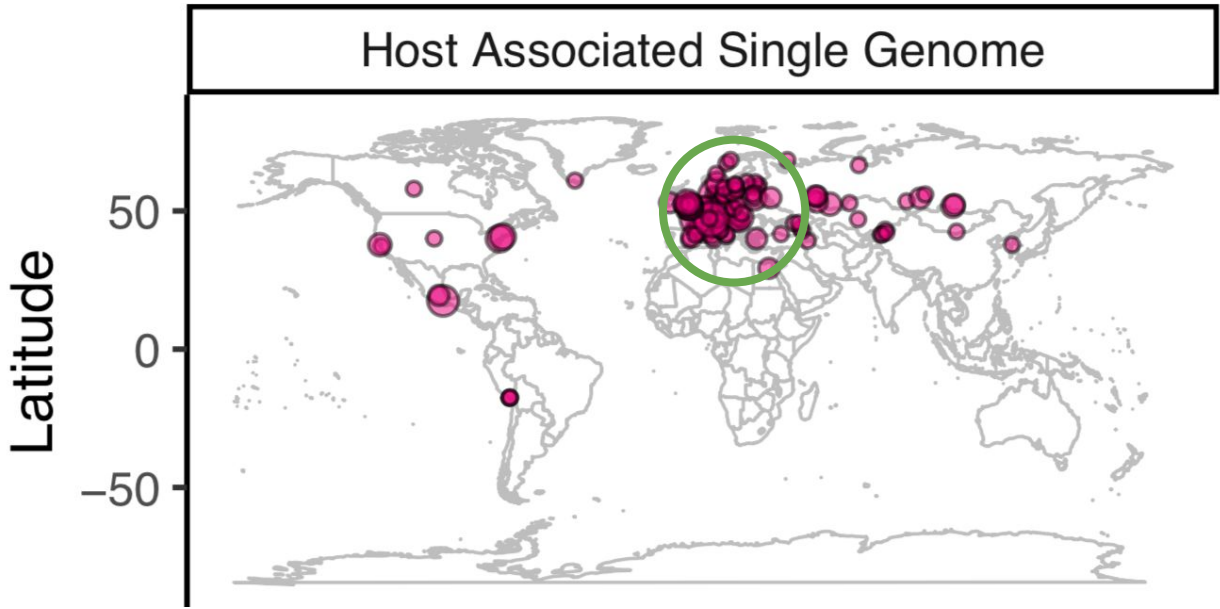
Unethical practices in aDNA research

- **Legacies of colonialism.**

★ Exporting ancient remains to a foreign laboratory.

★ Lack of long-term collaborations and local capacity-building.

- **Competition without question-driven research.**



Fellows Yates *et al.*, 2020, biorxiv doi: 10.1101/2020.09.02.279570

Prendergast & Sawchuk, 2018

Side effects of unethical practices on ancient metagenomics

- Repeated or redundant sampling.
- Not publicly available genetic data.
- Lack of transparency in methods.
- Publishing results in journals not accessible in the countries from which the remains were taken.

Towards responsible aDNA research: considerations for practice

- **It is imperative to avoid a 'sample first, ask questions later' approach.**
- **Research budgets must account not only for sampling, but also for sample return and continued engagement with collaborating institutions.**
- **Building up local research capacity to improve science.**

Hudson *et al.*, 2020

Towards responsible aDNA research: considerations for practice

- Consultation with (1) with local scientists and (2) with the indigenous groups who may be invested in the treatment and legacy of the human remains.
- Collaborative projects are required to progress the field.

SPAAM
Standards, Precautions & Advances in Ancient Metagenomics

Acknowledgements

María C. Ávila Arcos

Advisor

LIIGH-UNAM

Kelly Elaine Blevins

Arizona State University

James

Irina

Ash

Clio

Anna

Alex

Short Questions (2 mins!)

SPAAM

Standards, Precautions & Advances in Ancient Metagenomics

Improving current and future practices in ancient microbiome analyses

Nico Rascovan

Tuesday 22nd 13h10, Session 4

Ask questions!

- Vote on main questions with emoji reactions!
  (posted in #spaam2-open slack channel as a poll)
- Upvote questions you also want to ask.
- We will go through them at the end of this presentation!

Brief summary of Part 1 *(yesterday session)*

- Who is there? (but really)
  Databases to be used for taxonomic classifications
- Where are they coming from?
  Comparison datasets to be used for decontamination
- Is my source-tracked data ancient or modern?
  Considering aDNA signatures in the analyses
- Am I doing things right?
  Authentication criteria, standardized procedures and FIRE principles

Improving practices: *possible paths*

- Data clean-up: *How far can we go?*
  Supervised (i.e., using database-guided annotation)
  Unsupervised (i.e., using intrinsic read composition)
- Taxonomic levels: *What's the right size for the lens?*
  Genera, species, strain signatures, WG from curated strains
- Naming the things right: *the wheat and the chaff*
  Phylogenetic techniques (supervised)
  K-mers (unsupervised) with training datasets
- SOPs, good practices and reproducibility: *Is that method good/better?*
  Designing benchmark strategies, checklists, writing readable codes, software maintenance

Specific points that may be worth discussing

- Reproducibility and continuity

  Good coding (written to be assimilated and maintained by others). Guidelines?
  Ensure continuity of tools (not rely on a single person)
  Helping tools to get assimilated in the field
  Checklist to be followed when reviewing articles or colleagues' tools, etc.

- Parity guidelines

  Computing power requirements
  Database fitting

- Metagenomic analyses: Reducing artifacts, improving quantifications

  Eliminating low-complexity, noisy and confounding sequences
  Data normalization and transformation
  Sample comparison (within/between projects)
  Standardized results readily usable by others (like genotypes)

Competition vs Cooperation
Can we work together?

- Define projects beforehand to get feedback
  Avoiding 'scooping' risks
  Accounting for everyone needs: PhDs/postdocs ≠ Permanent bioinformaticians
- Creating a centralized place for discussing
  Which platform? Memberships, codes of honour?
  Open forums, crowdsourcing, help others/get help, reward systems?
  Bulletin board for projects (beta testing, pre-registrations, etc.)
- Securing continuity: *Regular SPAAM meetings?*
- Reducing entropy: *Working committees?*

Short Questions (2 mins!)

SPAAM

Standards, Precautions & Advances in Ancient Metagenomics

Towards standards in metadata reporting in ancient metagenomic studies: experiences from AncientMetagenomeDir

James Fellows Yates (w/ Antonio Fernandez-Guerra)

Data Retrieval and Metadata Reporting?

- Big challenge in any project: how to generate comparative datasets but also *metadata*
    - How to get the types of samples and data I want?
    - Have I missed something?
    - How to get library-level information?
    - Differently reported in every paper?
    - Misleading information (raw FASTQs! Only of mapped reads…)

Example github.com/SPAAM-community/AncientMetagenomeDir

- One attempt:

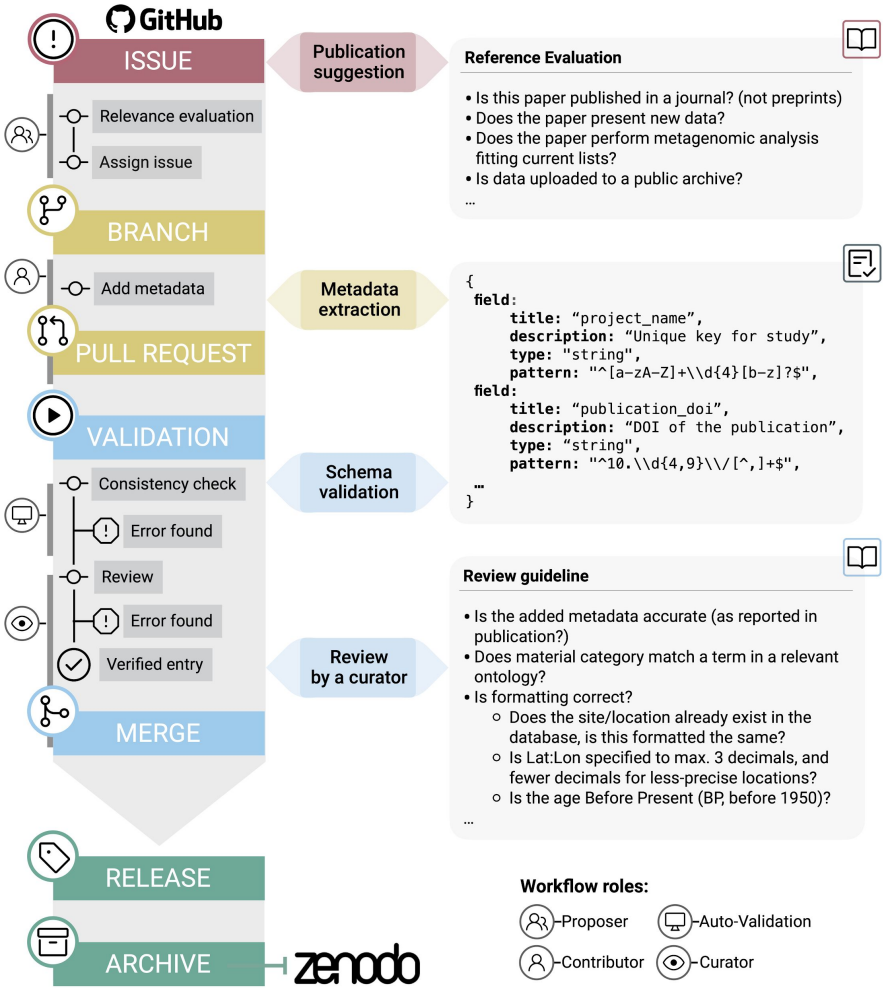
Ancient Metagenome Dir

- [No Table emoji, wtf?] Tables containing published ancient metagenomic samples
- 👭 Community curated - crucial!
- 🔓 Lightweight but 📚 comprehensive
- 📏 Standardised metadata

SPAAM

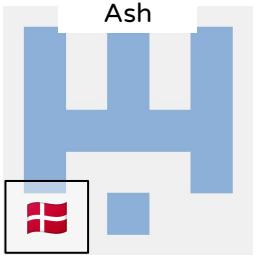
- One list for Ancient Metagenomic-related sub-fields
  - 💩 Host-Associated Metagenomes (e.g. microbiomes, mummy samples)
  - 🦠 Host-Associated Single Genomes (e.g. microbial pathogens)
  - 🏔️ Environmental (e.g. sedaDNA)
  - 🏺 *Proposed:* Anthropogenic (e.g. pottery crusts, parchment debris)
- Minimum metadata
  - 📰 DOIs
  - 🗺️ Spatial information
  - ⏳ Temporal information
  - 💻 Data location
- Additional crucial sub-field metadata
  - 🦌 e.g. host species, sample material

project_name	publication_year	publication_doi	site_name
Warinner2014	2014	10.1038/ng.2906	Dalheim
Warinner2014	2014	10.1038/ng.2906	Dalheim
Weyrich2017	2017	10.1038/nature21674	Gola Forest
Weyrich2017	2017	10.1038/nature21674	El Sidrón Cave
Weyrich2017	2017	10.1038/nature21674	El Sidrón Cave
Weyrich2017	2017	10.1038/nature21674	Spy Cave
Weyrich2017	2017	10.1038/nature21674	Spy Cave
Weyrich2017	2017	10.1038/nature21674	Dudka
Weyrich2017	2017	10.1038/nature21674	Dudka

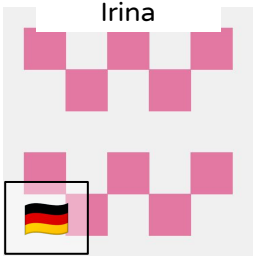


Acknowledgments


Aida

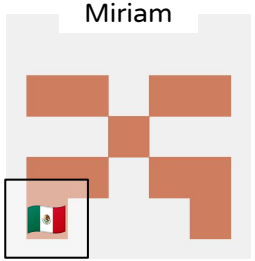

Ash


Becky


Irina


Maxime


Tina

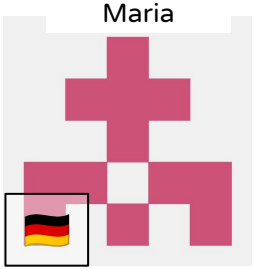

Miriam


Antonio


Ele

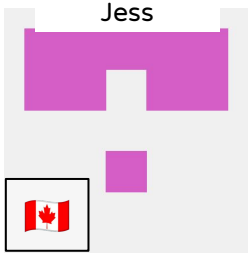

Shreya


Pete

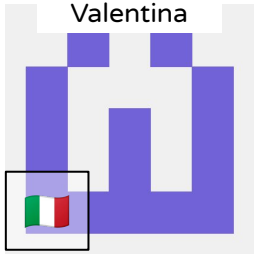

Maria


Alex


Abby


Jess


Aurora


Valentina

Tina

🏳️‍🌈: Lab Locations

Fellows Yates *et al.* 2020 bioRxiv 10.1101/2020.09.02.279570

How to standardise sample metadata reporting?

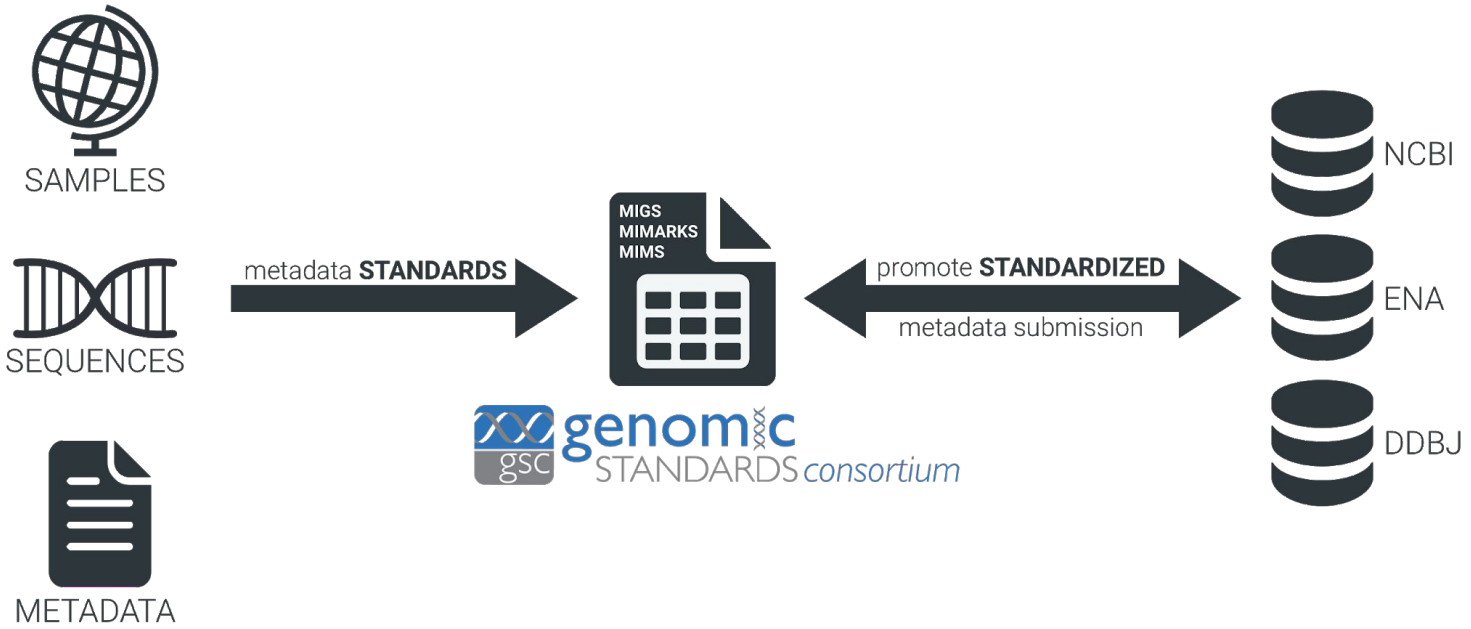
Not always a good time...

- Date reporting all over the place (calibrated/uncalibrated; AD/BC/BP; no 14C lab code; ranges vs. mid-point)
- Location reporting: varied
  - English vs original; Ethical challenges?
- Sample names inconsistent with data repository (just project code)
- Mis-assigned SRS codes
  - Same sample, different SRS codes (e.g. one per library, individual vs sample)
  - Reporting raw reads when only consensus or mapped-only!
- Data retrieval *still* not trivial

Areas to work on

- File formats!
  - For the love of all things please not tables embedded as images
- How to ensure consistency in field (rapid retrieval)
- How to get consensus?
  - Dating; locations; material types; data locations
- Up/Down-stream data?
  - Museum accessions?
  - Libraries; treatments; extractions?
- How to maintain and 'enforce' in field?
- Where to host?

Modeling Genomic and Environmental context



MIGS: The minimum information about a genome sequence | **MIMARKS:** Minimum information about a marker gene sequence | **MIMS:** Minimum information about a metagenomic sequence

Modeling Genomic and Environmental context

Minimum information about any (x) sequence

Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications

Pelin Yilmaz, Renzo Kottmann […] Frank Oliver Glöckner

Nature Biotechnology **29**, 415–420 (2011) | Download Citation

Specification projects	MIGS				MIMS	MIMARKS		New checklists
Checklists	EU	BA	PL	VI	ORG	metagenomes	survey	specimen	e.g., pan-genomes
Shared descriptors	collection date, environmental package, environment (biome), environment (feature), environment (material), geographic location (country and/or sea, region), geographic location (latitude and longitude), investigation type, project name, sequencing method, submitted to INSDC							
Checklist-specific descriptors	assembly, estimated size, finishing strategy, isolation and growth condition, number of replicons, ploidy, propagation, reference for biomaterial						target gene	
Applicable environmental packages (measurements and observations)	Air Host-associated Human-associated Human-oral Human-gut Human-skin Human-vaginal					Microbial mat/biofilm Miscellaneous natural or artificial environment Plant-associated Sediment Soil Wastewater/sludge Water		

MIxS: Yilmaz et al. 2011

Modeling Genomic and Environmental context

nature biotechnology

The minimum information about a genome sequence (MIGS) specification

Dawn Field, George Garrity […] Anil Wipat

nature biotechnology

Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea

Robert M Bowers, Nikos C Kyrpides […] Tanja Woyke

nature biotechnology

Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications

Pelin Yilmaz, Renzo Kottmann […] Frank Oliver Glöckner

nature biotechnology

Minimum Information about an Uncultivated Virus Genome (MIUViG)

Simon Roux, Evelien M Adriaenssens […] Emiley A Eloe-Fadrosh

nature chemical biology

Minimum Information about a Biosynthetic Gene cluster

Marnix H Medema, Renzo Kottmann […] Frank Oliver Glöckner

MIGS: Field et al. 2008 | **MIxS:** Yilmaz et al. 2011 | **MIBiG**: Medema et al. 2017 | **MISAG/MIMAG:** Bowers et al. 2017 | **MIUViG:** Roux et al. 2018

MInAS: **M**inimum **In**formation about an **A**ncient **S**equence

GSC21 - Vienna, April 2019

genomic STANDARDS consortium

Lynn Schriml
University of Maryland
Genomic Standards Consortium

Ramona Walls
University of Arizona
MIxS working group Leader

Guy Cochrane
EMBL-EBI
European Nucleotide Archive

Antonio Fernandez-Guerra
GLOBE Institute

Hannes Schroeder
GLOBE Institute

Fernando Racimo
GLOBE Institute

Mikkel Winther Pedersen
GLOBE Institute

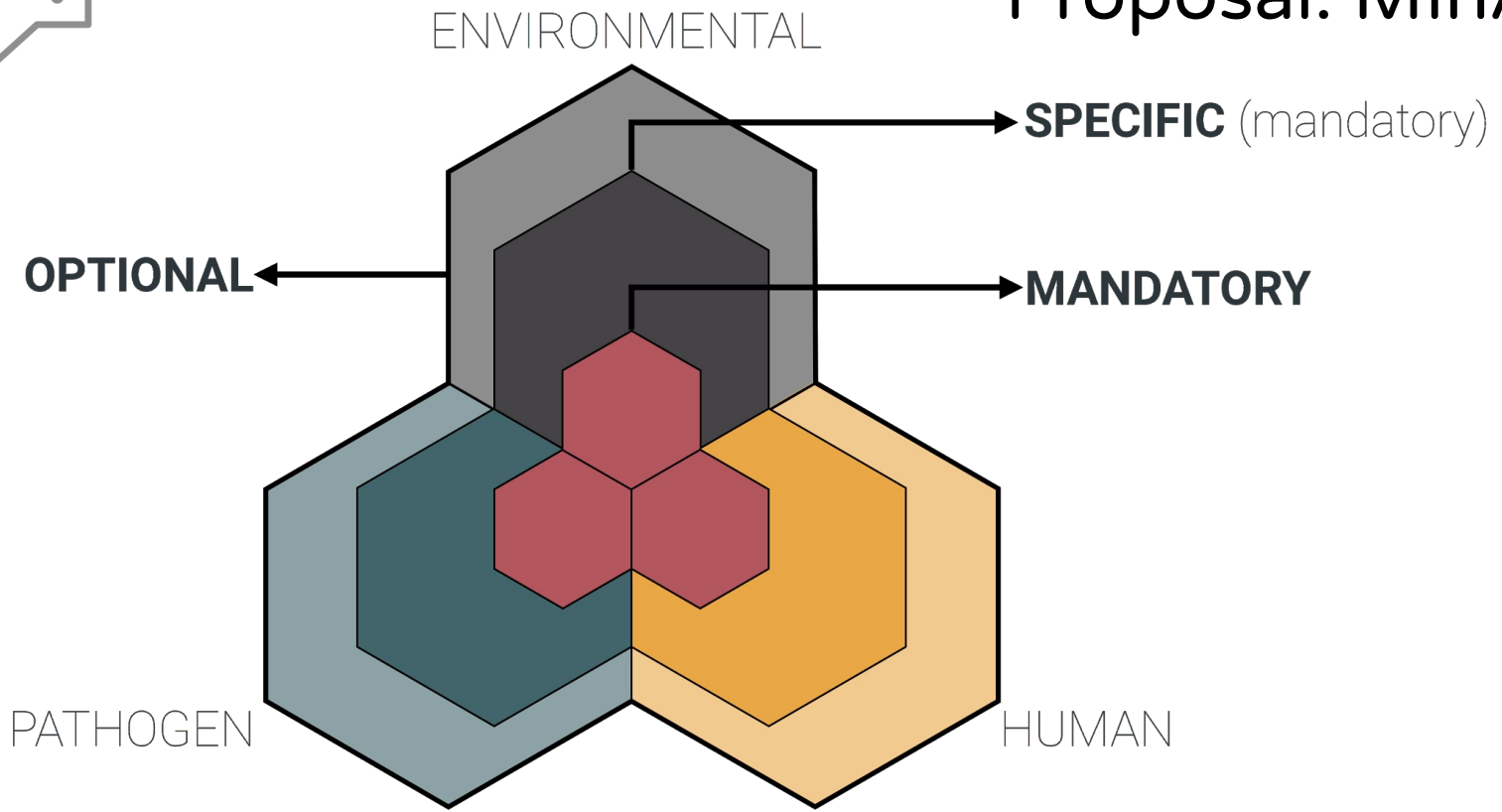
James Fellows Yates
MPI-SHH

Peter D. Heintzman
UiT

YOU?

- Identify descriptors that can be re-used for the ancient DNA samples from the MIxS core and packages
- Gather specific metadata types that are considered vital and important, but not required to collect samples.

Proposal: MInAS



MIxS core and package descriptors + aDNA new descriptors

Short Questions (2 mins!)

Discussion Points:

- **Do we want/need/have to formalize standards for responsible conduct in ancient metagenomics research?**
  - **Ethical requirements**: Consultation/collaboration with local stakeholders. Legal/ethical clearance. Appropriate and scientifically sound study design.
  - **Authentication guidelines**: Common comparative or benchmarking datasets? Mandatory types of analyses? Required minimal line of evidence? Reporting protocol?
  - **Reproducibility standards:** How to ensure data/code sharing consistency? What metadata must be reported and how to define?
- **TO KEEP IN MIND FOR SESSION 6:**

Regular SPAAM meetings? Working group and sub-groups for responsible conduct in ancient metagenomics research?

- defining/updating/disseminating standards **for the community**
- defining/updating/disseminating standards **outside the community**