SPAAM3 - Session Notes

Session 3: Urgent challenges and future steps in ancient microbiome research

Chairs - Sterling Wright and Kelly Blevins

Presenters - Abigail Gancz, Nathan Martin, Allison Mann, Nikolay Oskolkov, Barbara Moguel

Session Abstract

Over the past decade, advances in DNA sequencing technologies have revolutionized the amount of data that can be recovered from archaeological materials. In particular, metagenomic techniques have been able to recover ancient microbial DNA from dental calculus, paleofeces, plants, soils, animal remains, and so forth. However, several related challenges remain, such as best practices for the collection of samples, decontamination, and authentication criteria. The lack of standardization makes it difficult to control for noise across datasets generated from different research groups. The goals of this SPAAM session are to (1) Address the current issues in ancient host-associated and environmental microbiome research and (2) Discuss possible ways in which the ancient metagenomic data is more reproducible and easier to compare across research teams.

Session Objectives

01 Share expert knowledge about the ancient microbiome field, present new research findings, and facilitate new collaborations.

02 Promote open and frank discussions about the current gaps in the field and how to optimize and advance them.

03 Outline a series of guidelines that can be recommended for future research design.

Part I: Systemizing Pre-Laboratory Set Up: Optimizing and standardizing sample collection for reproducibility in ancient metagenomics

Abigail Gancz - *The Importance of Documenting Oral Geography*

Presentation notes

Please contact Abby with any questions. Email for information: asg5573@psu.edu

POLL:
Q: How do you acquire your dental calculus samples?

- I always sample it myself - 5.6% (n=1)
- I sample it and collect it myself when I can - 61.1% (n=11)
- I have never sampled it/collected it (sent by collaborators) - (n=6)

Part II: From bag to tube: Identifying the challenges and opportunities to advance laboratory practices

General Discussion

- Q: If you work in an aDNA lab and you are the first person to do ancient microbiome research, what are some key things you need to think about and do?
  - A: (James Fellows Yates) - Ask SPAAM these things! Nico's (Rascovan) group is really great about that. Reach out to SPAAM community on slack and get good input.
- Q: Is there any value in studying the microbiome of one sample set from one population in one time period that isn't providing any sort of transition or time transect? Or do you really need a broader dataset?

- ○ A (Irina Velsko) - Valuable from a microbiological perspective and those organisms and how they are evolving in their interactions with each other and a host.

Making protocols public

(Irina Velsko) Find good protocols and take a look at protocols.io and everyone should continue to publish their own protocols.

(Aida Andrades Valtueña) Making protocols public will not only help with reproducibility but will also allow us to evaluate how our lab protocols are influencing our data so this availability of protocols is good for the field as a whole.

(James Fellows Yates) There is a system where you can directly publish your protocols in PLOS One so if you make public your protocol on protocols.io and it becomes very popular there is a direct integration where you just write an abstract and the protocol can be rendered/published in PLOS One.

Single Stranded Libraries - yay or nay?

- Q: How many labs are transitioning to single stranded libraries? And what kind of changes do we see in the end, i.e. improvement in microbiome studies?
    - A: (Nasreen) Grad student Josh in UCSC lab developed a protocol similar to SRSLY libraries - two hours and 10 dollars a sample! https://doi.org/10.1093/jhered/esab012
    - (James Fellows Yates) But has anyone done any comparisons? They generated single stranded libraries from calculus but when they compared they found it wasnt beneficial - they had overall lower coverage but slightly increased percentage coverage because smaller fragments were filling in gaps in the genomes. Did anyone find that the ss libraries help for microbiome or on calculus specifically?
    - A: (Aida) We see some improvement in pathogen recovery in highly damaged samples. Wonder how much UDG treatment will help because you might lose all these fragments they want to keep.
- Q: How many basepairs on average do we lose from UDG treatment?

  - A: It would depend on the damage and overhangs and how fragmented your DNA is. The damage profile should tell you about how much you are losing.

POLL:
What kind of library is constructed in your lab?

1) Only double stranded - 36.9% (n=7)
2) Only single stranded - 10.5% (n=2)
3) Both and will continue both - 52.6% (n=10)
4) Both but moving to single stranded - 0% (n=0)

- Q: Is it important to record osteological information?
  - A: Osteological information we need to have special and even formal training. Therefore without that formal training we should include
- Q: How to get people on board, especially archeologists, in metadata collection.
  - A: Create documentation with detailed information and pictures

POLL:
Q: Do you have experience in osteological analysis?

- Yes - 44% (n=16)
- No - 56% (n=20)

Part III: What is minimal and what is better? Best practices for authenticating and analyzing ancient microbiome datasets

Nathan Martin - *Impacts of agro-pastoral societies on biodiversity: palaeogenomic, palaeoecological and archaeological approaches*

Presentation Notes

Researching the impacts of agro-pastoral societies on biodiversity primarily through retrieving DNA from sedimentary DNA.

- What was the best method to retrieve sedaDNA? Depending on the environment, the methodology strategy varies. Different chemical compositions or climatic/environmental conditions will influence the preservation of aDNA.
- Many changes exist in evaluating and analyzing sedaDNA data. There are no current standards for the bioinformatic analysis of sedaDNA. Taxonomic assignment based on mapping and eventually sometimes BLAST+LCA.
- Here he presents a strategy to enhance biomolecular analysis of sedaDNA of aDNA studies.
- Metabarcoding and shotgun sequencing presented multiple challenges: Metabarcoding it's hard to authenticate and with shotgun sequencing the endogenous DNA is overwhelmed with unwanted modern DNA.
- A targeted capture sequencing was used, and the baits were designed based on archeological evidence. Designed four databases (matK, rbcL, cytb and complete mitochondria).
- We observed that complementing the kraken and mapping information could increase the genomic information of the taxa observed in sedaDNA, for future analysis.

Allison Mann - *Archaeogenetic detection of eukaryotes in host-associated microbial communities: challenges and considerations*

Presentation Notes

Dental calculus is the best place to start to reconstruct diet and allows the preservation of microbial communities. We don’t really know how dietary DNA is preserved in calculus.

- There have been few studies which have studied the diet in ancient populations (Warinner et al., 2014; Weyrich, 2017). However, some food sources can be misidentified.
- One of the most common problems - most of your hits will be eukaryotes and they are not specific enough to give you any insight into diet and you can’t rule it out from the environmental contamination.
- Second (and possibly more serious) problem - really specific hits to organisms and not being able to tell if they are true hits. Example - Ottoni study (cucumbers and watermelons and mummified baboons study) - cucumbers were not introduced yet though watermelons may be possible.

The authors stress that “extreme caution is needed when inferred food sources from shotgun reads, in particular without authentication strategies or other sources of evidence”.

- Why are putative dietary organisms difficult to detect in metagenomic datasets generated from dental calculus, which otherwise serves as a rich source of aDNA?
    - Fragmented and damaged, low number of dietary molecules compared to bacteria, differences in genomic architecture of the species, reference database errors. (Mann et al., 2020).
- Designed an experiment to test how reliable these analyses and taxonomic profilers and identifiers are:
    - Representative genomes of 10 various species - generated DNA fragments and removed any suspected microbial contamination, added DNA damage
    - We randomly sampled 50, 500 and 5k reads from the synthetic datasets
    - Spiked them into oral microbiome and ended up with datasets for 3 organisms totalling 5 million reads each
    - Results: proportion of reads that could be appropriately assigned pretty much remained the same. Tomatoes were either assigned correctly to species level or left unassigned, red deer were never correctly assigned or unassigned. Most alarming was the number of reads that were incorrectly assigned at the species level.
    - Analyzing different previously published dental calculus - ended up finding many species that were identified as fish with the common carp being detected in high abundance - this could be because the common carp genome is notoriously contaminated with other sequencing reads. Reads detected of species that may make sense but comparable numbers of reads detected from species that are very unreasonable, i.e. wine grapes and tasmania devil.
- Things to consider:
    - Potential sources of contamination?
    - Choosing the right database? - if you use a custom database there should be a plan in place to disseminate that database
    - Using other proxy evidence to back up your interpretation
    - Verification/confirmation criteria
    - Report the weirdness!!! - as a community we should be better at communicating weird results in publications and workshops.

Nikolay Oskolkov - *The problem of false discoveries in ancient microbiome analysis*

Presentation Notes

The Problem of False Discoveries in Ancient Microbiome Analysis. This is pretty common in ancient microbial analysis. Screening using Kraken2 and rank them by abundance and you might see bacteria but you might see eagles, wolves, etc. - this is in a blank! So unlikely to exist in a blank but if this was sediment how would you deal with this problem?

- Common carp genome was assembled on illumina adapters - this genome is full of adapters and shows up in taxonomic identifiers if your adapters are not sufficiently removed.
- Example: in a mammoth tusk - many other mammals are detected. Intuition: all the mammals have reference genomes of better quality than the elephant reference. There are missing pieces in the elephant reference.
- BSA reagent - introduces bovine DNA to samples - Leonard et al., 2007.
- If you know who the host is you are very lucky BUT sediments are a mix of ancient and modern reads, it is unclear who is the host. AN example: Cattle (modern) genome and ancient microbes having damage existing in soil that bioinformatically can become mixed up and look genomically like ancient cattle. Check out Vernot et al., 2021 supplementary where they discuss this phenomenon. They identified misidentified reads by evaluating regions of excess reads.
- Depth of coverage may be misleading:
    - Reads can be “forced” to map to a reference genome and stack in the alignment making deceiving metrics of 5000 reads mapping and 1.2x genome coverage but not being actually detected.
    - A good metric is distribution of coverage.
    - With competitive mapping you can identify confidently reads actually mapping to a pathogen. Example of Y. pestis shows that you can detect this pathogen in a random sample when mapping with only the reference genome - as we increase the database size (to 10k bacteria) the signal is eventually gone.
    - Databases enriched for pathogens can bring false positive pathogenic hits. The larger the database is the fewer false positives

we get. By increasing our database size our analysis is less and less biased.

- Ancient Microbiome Workflow Overview:
    - KrakenUniq: screening and filtering and use MALT as a validation step.
    - MALT is not easily used as a screening tool because most people cannot afford the space for a large database. KrakenUniq is much easier to eliminate false positives and then you can build a database for MALT depending on the KrakenUniq results to validate your KrakenUniq results.
    - MALT/HOPS used for validation authentication
    - KrakenUniq - delivers coverage information and you can immediately narrow down your list and remove likely false positives.
    - MALT is very specific but not sensitive and KrakenUniq is very sensitive but not specific. Together you strengthen your analysis by combining them.
    - He wants to maximise specificity - prefers to find nothing rather than detect false positives.

Barbara Moguel - *Study of biodiversity and the past environments by DNA in sediment records from Chalco Lake, Mexico*

Presentation Notes

Biodiversity and the past environmentant by DNA in sediment records from Lake Chalco, Mexico. What were the changes in biotic diversity related to climate environment, and human impact, corresponding to the Holocene period. Examined this by recovering 200+ sediment core samples of the dried lake. This is the first study of aDNA of lake sediments to reconstruct this environment.

- Samples underwent multiple analyses - sediment smear slides for fossil diatom analysis, x-ray diffraction for geochemical analysis and DNA extraction for metagenomic analysis of shotgun sequencing data.
- They used MG-RAST for metagenomic analysis: RefSeq, LSU, SSU, ITS databases with Subsystem for functional genes. Decided to work with just LSU database - RefSeq gave a lot of false information, ITS didn’t give much information and there was not much difference between ITS and LSU .

- Found taxonomic diversity was with bacteria as the most abundant (81%), archaea (15%) and Eukaryota (3%).
- Also analysed the Alpha diversity, biodiversity - shannon index, fossil diatomes, characterized three different environmental zones (subsaline, hyposaline, freshwater.
- Beta diversity with cluster analysis of the three proposed zones - Moguel B et al., 2021
- Cyanobacteria was found in high abundance in the hyposaline environment 6,000 years before present which corresponds with the first human settlements in that region (based on archaeological records).
- Their results correlate with fossil diatom and geochemistry zones proposed and show changes in past environments at Lake Chalco with human occupation.
- Identified a transition stage between the hyposaline and subsaline zones mainly due to change in vegetation, mosquitos, higher precipitation, pathogenic fungi and crops.
- Hypothesis: the first human activities had a large and rapid impact on the ecosystems in this area.
- They could not separate between ancient and modern microorganisms.
- Shared the paper she presented: https://doi.org/10.1038/s41598-021-92981-8

General Discussion

- **Q: Nikolay to Nathan Martin: thank you very much for your very interesting talk! Did the bovine reads you detected from the sedDNA samples have a damage pattern (I didn't quite understand it)? And how would you explain the presence of elephant reads in your sedDNA samples?**
    - **A: Nathan to Nicolay: Yes we have been able to generate nice damage patterns for bovine reads. For the elephant, in Morocco it was an hypothesis that Loxodonta could have been present in Morocco and we detect it in 2 samples. So a bit surprising but not that much to retrieve some**
- **Q: Katerina to Allie: When using environmental controls, would you suggest removing any eukaryotic species also detected in the environment?**
    - **A: It would probably depend on the context. In general, if you find eukaryotes in your controls (blanks) then remove that.**

  - A: James: That's true for dietary but be careful for microbiome stuff. Quite common that people just take it out but that doesn't apply to all analyses. Seems like a simple and conservative method but not best for everyone.
  - A: Barbara: We use aContam software to help avoid this. COuld be a good approach to see the contamination in these samples.
- Q: Ian Light to Allie Mann: Any thoughts on using a tool like KrakenUniq to look at whether the # reads classified is just a result of very low complexity/highly conserved regions (few unique k-mers)?
  - A: I actually was curious on how different our results would have been with krakenUniq — I started downloading the database last night but unfortunately didn't get the analysis done before today.
- Q: Arumi to Allie: Awesome talk! what would be the considerations when choosing the right database?
  - A: Nikolay responds - I always try to use the full NCBI NT, basically all organisms ever sequenced by human being
- Q: Sterling to Allie: I was wondering if you were aware of any studies that include a larger dataset of modern calculus samples that also look at diet.
  - A: Not aware of any oral microbiome or dietary studies that provide this. Currently I have a dataset where they have this knowledge to establish some baseline but this is challenged by practices like brushing teeth and more modern diets.
- Q: Shreya to Nikolay: How do you sort through the hits from KrakenUniq to decide what to put into the MALT custom database for each project? Do you check for a list of "usual suspects" pathogens?
  - A: We apply depth and breadth of coverage thresholds reported by KrakenUniq for eliminating obvious false-positive organisms and use all organisms reliably detected in at least one sample for building a project-specific MALT database. Ideally those thresholds should be sample-specific, we are working on it, there are some ideas but so far hard thresholds.
- Q: Leslie to Nikolay: Do you have your workflow published or deposited somewhere for reference? We have done a similar approach but with 2 malt runs, one with a tiny database with our species of interest to filter out anything that is definitely not what we are looking for. Then extract all the aligned reads, and run them through a big Malt database with all bacteria… but as you said, the memory usage is a big problem. Your way of flipping the approach around is much more HPC friendly.

    - A: It is going to be published soon. I think during this conference I heard a similar line of thinking a few times already, i.e. using some way of pre-selecting reliable candidates and building MALT database on them. Glad to hear that other people also thought about it.
- Q: Nico to Nikolay: What is the best tool in your hands that grabs all possible reads? In my experience KrakenUniq misses a lot of things. We want to capture as many interesting reads as possible. What can do this? What tools and databases
    - A: If you want to be as sensitive as possible than you should include the full NT NCBI database. I would start on KrakenUniq with the largest database (everyone can build the largest database). Then to be more specific you need to do an alignment like MALT. The Kmers of Kraken allow it to be more sensitive and then MALT/HOPS (with a custom database based on your Kraken results) will perform an alignment for specificity.