SPAAM2 - Session Notes

Session 3 - Sorting and Recycling the trash: Defining authentication guidelines for the research community

Chairs Clio Der Sarkissian & James Fellows Yates
Icebreaker Presenters Nikolay Oskolkov & Sterling Wright

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. Post-mortem fragmentation (typically via depurination) and damage 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 (James Fellows Yates)

- **Q**: What are criteria for minimal authentication for determining if aDNA is real?
    - This is the main focus of the session

Icebreaker 1 (Nikolay Oskolkov & Sterling Wright)

- **Topic**: Databases and taxonomic assignment tools (Nikolay)

- Concern about database sizes when screening for ancient microbes.The final conclusions drawn are based, to a large extent, on the type of database used and the taxa included/excluded in that database.
- Taxonomic assignment tools employ one of two types of read assignment:
    - Standard alignment: bowtie/malt/bwa
    - Pseudo-alignment: Kraken2
        - Pseudo-alignment uses certain sequences to quickly assign microbial sequences
- There is a need to find a balance between assignment of false positives/negatives and other trade-offs such as memory requirements/speed
    - By using small databases to offset e.g. memory requirements you not only miss microbial species that are present, but you bias/overestimate abundances for the ones you can detect.
    - A possible solution could be to experiment with krakenUnique + breadth of coverage information (provided by KrakenUnique) and different database sizes, see which are the best parameters.
    - Benefit of KrakenUnique/Kraken1-2 is that it is still possible to have very large databases.
    - MALT is prohibitive due to big memory requirements, many people do not have access to server nodes with 1-4TB of memory, which are required.
- **Topic**: Use of SourceTracker to estimate contamination in metagenomic datasets (Sterling)
- **Q**: Whether or not data is well preserved and if the database and references are valid.
- **Q**: Should sourcetracker be used as a screening tool? What do other people use?
- **Q**: Should we have a standardised sequencing dataset OR collections of samples for use to estimate contamination and to compare performance of various taxonomic assignment tools?

Icebreaker 2 (Nikolay Oskolkov)

- **Q**: For ancient microbes, especially viruses, What to do when the coverage is too low to infer a deamination pattern?This also requires a reference genome and alignment step.
- **Q**: What can we do about it?
    - **A**: Use machine learning to understand systematic motifs that are different between modern and ancient reads, where DNA is a text, k-mers are your words, the sequence is a sentence.
    - **A**: Use Natural Language Processing (NLP) to ask what k-mers are enriched in modern sequences compared to ancient sequences? how do they cluster and what do these clusters mean?
    - Neural Networks

- Salient maps - rank nucleotides by position in the read....are they at ends of reads?

- **Q**: What is this method doing?
    - **A**: There are two classes of data points and you are trying to separate them, you have observations and features. A typical format is a matrix, a feature can be a k-mer.
    - This is a way to find features that separate groups of samples.

General Discussion

- Machine Learning is non-linear way of looking at data, but it is non-trivial
- If you have batches you can investigate the differences, batch effects etc.

Discussion of microbial aDNA damage:

- See a lot of variation in the levels of aDNA damage for ancient pathogens from same time periods, it depends on the pathogen.
- GC content affects how the DNA molecule physically fragments
- Spore forming bacteria and eukaryotes: you would think they would be good to preserve and that it's DNA was not damaged
- Water is necessary to accumulate DNA damage (Alex Hübner on Slack)
- **Q**: Do we need to look into this more - maybe a group project?
- Archaea bias, same sort of pattern where methanobrevibacter has 60% abundance where it overgrew OR is better preserved due to better cell structure
- In thermophilic or acidophilic species I don't see any DNA damage (in ancient soil metagenomes?)
- Also depends on the source, *T. forsythia* has more damage in calculus compared to dentine, see Mann et al. 2018 *Scientific Reports*
- *P. propionicum* doesn't sporulate
    - **Q**: why do we see this taxa dominating? Is it database bias? Does it appear over-abundant because there are no other reference genomes to attract reads?
- **Poll:** Have you already heard the saying - Never map alone - for metagenomics ?
    - Yes: 44% (7)
    - No: 56% (9)

Building and curating databases for taxonomic assignment:

- **Q**: How to curate taxonomic databases when you have limited computational resources/funding/ lack of single nodes with enough memory to accommodate tools that only run on one node (i.e. MALT which usually needs 1-4TB)
    - **A**: Try SPARSE, using ANI to remove highly similar genomes
- **Poll:** How much RAM can you use for your databases?
    - Top 3 Choices of 6
    - 2 TB: 50% (7)

    - 1 TB: 29% (4)
    - >2 TB: 21% (3)
- **Q**: Should we build a common database for reproducibility?
    - **A**: It's a matter of continuous curation, which is time consuming. If everyone makes their own databases then it is not reproducible.
    - **A**: Software should have almost no differences if you run on a single genus.
    - **A**: Conterminator by Martin Steinegger, a suggestion. Trying to solve the issue of cross-contamination between genomes and sequences that shouldn't be in different genomes.
    - **A**: Segata lab curated metagenomics data, where they curated all metagenomic samples available and ran them through MetaPhlAn, when looking through human sources and metadata.However, MetaPhlAn1/2 is an outdated tool and will be phased out. Version 3 and 4 are coming out. Previously curated metagenomic samples/dataset will be run through MetaPhlAn3 to create the profiles for common use.
    - **A**: Building a universal database may not be so useful, it's a huge project.

- **Q**: Screen all samples with Metaphaln 2 quickly and then build a small scale reference database?
    - **A**: We need to find another solution that is stable, we have a lot of data that we need to put into categories and make successively smaller categories of databases to use.
- Databases are a challenge a lot of people have and feel uncomfortable with
- **Q**: What are the minimal authentication requirements that people use?
    - **A**: Damage and edit distance, even distribution across reference are common methods to authenticate
    - **A**: Damage and edit distance may not be so relevant for viruses if they are very divergent
    - **A**: CpG platypus instead of mapdamage. See one one end of read but not the other.
- Sourcetracker for authentication for microbiome identification
    - Keep an eye out for known human skin contaminants (*P. acnes*) and known lab contaminants

Statement: We need some PCA/MDS plot standard for microbiomes/source identification

- **Q**: With novel sample sources, how would you put this into a PCoA.
    - **A**: You have to look at what is in the blanks. You NEED the blanks for novel sample sources where there are no comparative datasets.

    - **A:** It is a problem in modern microbiome studies that they don't use blanks, often 16S and 18S.
    - **A:** Different extraction protocols can give different extraction outcomes
- **Q:** How well do we know the diversity of soil that we can use as a source? How different is it between sites? Is it enough to have a few soil samples, or do we need one from each site?
- **Q:** How do we best sample soil from an archaeological site?
    - **A:** Take the soil sample and put it on liquid nitrogen ASAP and keep it frozen. If you don't have access to liquid nitrogen then keep it cold/freeze it as soon as possible to stop the metabolic activity of the pathogens.
    - **A**: If the aim is to determine contamination of a bone by soil, then sampling the soil right next to/on the bone is probably best, if it's about characterising paleo-microbes at the site then a core from somewhere nearby might best.