SPAAM3 - Session Notes

Session 2: Current challenges and biases in ancient patho(meta)genomics

Chairs - Miriam Bravo
Presenters - Richell Maribet Ramirez Molina, Meriam Guellil, Ian Light, Kelly Blevins

Session Abstract

Ancient remains are limited and are an invaluable part of our cultural patrimony. Thus, maximization of scientific output should be a priority in the ancient DNA (aDNA) field, given the destructive nature of the sampling procedure required by most of the studies. DNA extracted from ancient remains contains information beyond the host's genetic component; it can also harbour pathogen DNA present at the time of the host's death. The recovery of ancient DNA from pathogens, along with historical and archeological evidence, offers new insights into the origins, etiology, nature, and evolutionary history of ancient infectious diseases. As the number of recovered ancient pathogen genomes grows, so have concerns about the wet and dry lab strategies employed. This SPAAM session will be focused on discussing challenges and biases in the retrieval and analysis of ancient pathogen genomes. Furthermore, we would like to discuss new analytical avenues, such as epidemiological modeling or functional evolution, for ancient pathogenomics, which could provide additional valuable information for the modern microbial community.

Session Objectives

01 **Hear the current challenges and biases in lab and bioinformatic techniques in ancient patho(meta)genomics.**
02 **Exchange views on the key steps during the detection and reconstruction of ancient pathogen genomes.**
03 **Discuss the importance of paleopathological evidence in ancient patho(meta)genomics.**

Part I: Looking for the invisible: lab and bioinformatic techniques to detect pathogens

Richell Maribet Ramirez Molina - *Methodology for rapid scanning of aDNA extracts based on real-time PCR melting curves*

Presentation Notes

Methodology for rapid scanning of aDNA extracts based on melting curves generated in real-time PCR. Analyzing pathological specimens with possible tuberculosis and treponemal cases - working to molecularly confirm these pathogens in recovered individuals from two Archaeological sites: Pajones, Zacatecas and Tamtoc, San luis Potosí.

- 20 individuals, 10 from each site
- These individuals show skeletal evidence of TB and/or treponemal infection.
- 20 individuals (17 teeth and 16 bones - some individuals had both and others not) with a total of 140 aDNA extracts.
- Use real time PCR with sequence specific assays to detect the pathogens
- Amplification of DNA in real time PCR is not specific but shows all DNA (evn. = environmental)
- Melting Curve
    - HRMA - High Resolution Melting Analysis used to show difference between species and subspecies
    - The confidence of this analysis allows to identify putative positives - offering a more financially feasible method
    - They were able to identify the specific markers for TB and syphilis

Meriam Guellil - *Few and far between? Detection and reconstruction of ancient pathogen genomes*

Presentation Notes

Detection and reconstruction of ancient pathogen genomes using shotgun screening. Will be focusing on metagenomic screening and hit validation (if time allows).

- Basic screening workflow includes
    - 1) Data Preparation (Deduplication)
    - 2) Metagenomic Screening (Software? Database?)
    - 3) Data Aggregation and Support Calculations
    - 4) Filtering for Pathogenic Hits
    - 5) Mapping and Validation for hits.
- You have to make a choice of what software you will use and what database you will build.
- Focusing on Kmer based approaches - a lot of false positives and false negatives (low signals can remains undetected within the noise)
- Noise or true hit? This can be influenced by:
    - Sequencing depth?
    - Genome size?
    - Species?
    - Reference Genome Quality?
- KrakenUniq - counts the number of unique k-mers for each taxon. General idea - k-mer count should always be several times higher than the read count. It also provides statistics on duplication rates and coverage of available unique k-mers per taxon.
    - You can use multiple databases hierarchically
    - Taxonomy can be extended to include nodes for strains and plasmids
    - NCBI viral genome resource data is available for database building
    - Output -> read count, unique k-mer count, duplication rate, coverage
- She is working with current db: bacteria, viruses, Archaea and Protozoa (Complete/Chromosome level and Dusted), Human Genome, NCBI UniVec and NCBI Viral Neighbor Genomes Database (~210 GB)
- Statistics from KrakenUniq - Difference between false and positive is based on the way this is calculated - E-value: (Kmer count/Read count) x Coverage
    - KrakenUnique heatmap is based on this E-value (cutoffs between 0.001-0.1)
    - Output is aggregated and plotted across multiple plot types
    - Filtered for organisms of interest
    - KrakenUniq - effectively minimizes noise and false positives.
- Final validation via mapping:
    - Mapping statistic
    - Distribution of coverage across the sequence
    - Presence/absence of plasmids/chromosomes
    - Presence/absence of specific regions

    - Deamination
    - If needed: comparative mapping
- Get to know the genomes of species of interest!!! - close commensal/environmental relatives? Genetic diversity within species? Mappability issues? Choosing the right reference sequence will help you maximise what you get out of your data.
- Most cost-effective way is a multi-organism panel but it is better suited for small genomes.
- Costs can be reduced by playing with hybridization temperature and bait mismatch rates.
- Most ancient samples can be captured efficiently using half - or less - of a bait aliquot.
- Kraken2 vs KrakenUniq - KrakenUniq will clean up spurious reads

Advice:

- When working with pathogens it is important to ask: Does it make sense to detect this organism in the sampled tissue?
- Get to know the genomes of species of interest - this way you can detect mappability issues
- Choosing the right reference is important to the quality and confidence of your work
- Capture or Shotgun Sequencing? Most effective way is a multi-organism panel but this is better suited for small genomes

Ian Light - *Quantitative Construction of Metagenomic Screening Databases*

Presentation Notes

Quantitative construction of metagenomic databases (Db) is work he has done in the early stages of his PhD. Tackling the problems of large reference databases. Nucleotide Db are becoming too large for efficient metagenomic screening.

- Large databases
    - Inefficient to query
    - Resources intensive
    - Unreliable for some sequences
- Random subset - randomly subsampled
    - Possible over representation of some species

    - Lack of reproducibility due to constant new releases
    - Possible loss of complexity/diversity

Quantitative approach to Database Reduction:

Started out with:

- Source data (RefSeq db - archaea, bacteria)
- Issues: biased representation, species definitions, mislabeled genomes
- Strategies: ANI > Clustering > BLAST
- Rosselló-Móra & Amann 2015 - shows the utility of ANI (Average Nucleotide Identity)

Overview of Strategy:

- All genomes in Taxid -> compute ANI and find meioids (by clustering) - Pairwise ANI Matrix -find medioids (by clustering) > confirm medioids in BLAST to confirm

Benefits

- Information about db
- Limit loss of complexity
- Identify mislabeled genomes
- Reduce size
- Higher reproducibility

Potential drawbacks

- Possible removal of species or taxids
- High use of computational resources
- Unexpected behavior of algorithms

Finding representative sequences by clustering

- Find the point in the cluster that most represent the diversity of the group
- Considerations:
    - Behavior run time outputs
    - Required inputs
- Ideal algorithm: (no algorithm contains all of these so you will have to compromise)
    - No prior needed
    - Clusters rep sections of diversity
    - Identify outliers

Currently using a “one size fits all approach” so getting some unexpected clustering.

25430 -> 11626

- Kicked out poor data and previously misnamed data
- Mainly driven by being largely over represented by taxa IDs

Further Directions: Working to optimize current algorithm, investigating other algorithms for clustering, expand use cases (where it can be applied)

General Discussion

- Q: Nikolay Oskolkov to Meriam Guellil: Very interesting, I confirm that we also use KrakenUniq and to our experience is the most robust tool. I think it is under used for Kraken2 even though it minimizes false positives. But recently they did implement "minimizer" in Kraken2. Did you test or compare KrakenUniq vs Kraken2 with the new minimizer feature of Kraken2 (that should be equivalent to unique kmer count in KU)?
    - A: Has not directly compared them but they are not directly comparable. KrakenUnique is only part of the code of Kraken2 - you only get some output, less output without the coverage information. But did notice a massive slowing down. I don't think the loss of information is worth the few minutes you will save. Also the database building and loading between the two is very different where KrakenUniq allows for a very customizable database.
- Q: Allie Mann to Ian Light: I may have missed this during your talk, but in your database reduction steps - where are the genomes from? Are you only using full genome sequences or contigs? If not, does that result in some taxonomic drop out?
    - A: We only took chromosomes and complete genomes. We had species of interest and if those didn't meet those criteria we may have included them (at the scaffold level).
- Q: Allie Mann to Ian Light: What about those species/strains that might only have genomes at the scaffold/contig/etc level of assembly?
    - A: Only took chromosomes and complete genomes in refseq. But also added a selection of genomes that if they didn't fall in that category were added anyway.
- James Fellows Yates to Ian Light:
    - Q: Have you compared your workflow against the database construction part of the SPARSE pipeline (Mark Achtmans group)? It also uses ANI (but difficult to use outside of SPARSE itself)
        - A: Have not directly compared it to sparse but main motivation is to have a database that we built that we understood how it

was going through and could potentially use it for downstream analysis.
    - Q: Do you have an approach to check for 'mislabelled' genomes that may accidentally be picked as a representative from a cluster (if I understood correctly)?
        - A: All the genomes that were picked as representatives ended up going through a BLAST search to increase confidence that they are what they are.
- Q: Nikolay Oskolkov to Ian Light: Cool work @Ian Light, how do you compute distance between your sequences and wouldn't you do multiple sequence alignment of all the sequences for all the ~25 000 ref genomes?
    - A: Average identity compilations and then building clusters from that. Once we iron out the approach we might do a reduction on strain or genus level. It might be better to go to a higher level. They will test that.
    - Tool listed by Felix Key: we use fastANI for average nucl. identity calc. it is kmer based and impressively fast. https://www.nature.com/articles/s41467-018-07641-9
    - Tool posted by Nikolay that helps reduce redundancy in a database: https://pubmed.ncbi.nlm.nih.gov/28158639/
- Q: Kelly Blevins to Richell Ramírez: Awesome project. I can't wait to see what you recover! Do you know how the samples were selected? And how did you subsample the elements? Did you prioritize the lesions?
    - A: Selected samples based on lesions and arch report. They identified TB using rtPCR - next step is to capture DNA to build libraries.
- Q: Miriam to Richell: How was the dialog between archaeologist and paleopathologist? Was it a close relationship, were they interested in your results?
    - A: Archaeologists support and they want to share more. They have a good relationship.

POLL:

Which taxonomic identifier do you use:

- MALT - 37% (n=7)
- Kraken - 63% (n=12)
- BLAST - 0
- Comments:
    - Nico Rascovan: I also map on a custom database we built. BLAST on whole NR is a good validation tool tough, for whatever read found with the other less accurate methods

    - Claudio Ottoni: I also used METAPHLAN
    - Meriam Guellil: I have used Metaphlan, Blast, Diamond, Kraken, Kraken2 and KrakenUniq

- Q: Miriam to Meriam: What criteria do you use to include for capture?
    - A: Based on the number of hits/reads since it can be difficult to reconstruct from so few reads. There are some thresholds based on whether it is a virus or bacteria.

- Lesley Sitter to Ian: Hey @Ian Light, have you also taken a look into how much unique genetic diversity is lost by your reduction method?
    - Not yet, the main thing we are concerned about is over or under fitting of our clustering algorithm. Not quantified yet.
- Q: Felix asking community: How does KrakenUniq compare to MALT in terms of specificity and sensitivity? My (Felix) previous experience with Kraken is that the false positive rate just goes through the roof given the k-mer based approach. Would the representation of evenness of clusters help to increase specificity and sensitivity?
    - A: Meriam: I haven't made any formal comparison. But has found differences in results - this could be database dependent.
    - Nikolay: Both deliver breadth and eveness of coverage. KrakenUniq is not as fast but it does have robustness. With KrakenUniq you can use any size database (unlimted) so we can assume it would be more sensitive than MALT. I haven't seen a formal comparison but KrakenUniq allows basically an unlimited database size while MALT does not so I would expect KrakenUniq would be at least more sensitive, not sure about specificity though
    - Meriam: Find the speed is similar so she thinks it depends on the service setup.
    - Felix: Both are faster than MALT because MALT is extremely slow?
    - Meriam: Found KrakenUniq isable to find very low hits, especially in the viral department that are recognized.
    - Nikolay: I did not compare MALT vs. KrakenUniq on synthetic data but ran MALT and KrakenUniq (using database built on same reference sequences) on a few benchmark samples where we are pretty sure we know the microbes. The results delivered by both tools were mostly in agreement, so no major discrepancy.

    - **Felix Key - we will explore it and come back with information for SPAAM4!**

- **Q: Marcel Keller to Ian Light but also the whole audience: with more sophisticated approaches to build custom databases, have you thought about how to publish your methods guaranteeing full reproducibility?**
    - **A: Felix: hi marcel we have not decided on that. I guess a snakemake would be an option (and easy to rerun by other people) but not sure if that is what is going to happen at the end.**
    - **Marcel: Thanks, but do you think it would also make sense to publish the output (e.g. a file with all IDs of included refseqs)? Otherwise I would be afraid that screening pipelines of different labs are turning into 'black boxes' hampering comparability and also troubleshooting.**
    - **Felix: sure databases grow constantly and interfere with explicit reproducibility. The approach that Ian presented is mainly a reproducible (non-random) method of how genomes are selected and it should lead to a reasonable output whatever version of, lets say refseq, is used. Of course for mirroring an exact database someone can always share the IDs used.**

Part II: How to present and discuss your findings

Kelly E. Blevins - *Paleopathology-informed sampling strategies for Mycobacterium tuberculosis complex aDNA recovery*

Presentation Notes

Paleopathology informed sampling strategies for *Mycobacterium tuberculosis* complex (MTBC) aDNA recovery. Thinking about the low recovery rate of aDNA from skeletal samples (Duchene et al., 2020). Sampling site location is very important to consider particularly when we think about pathogens that are not systemic. We need to develop more mindful sampling strategies.

- Recovery of MTBC ancient DNA
    - Infection of tuberculosis (TB) in some patients can leave the lungs and travel to the skeleton - most commonly the vertebrae leaving destructive lesions.

- Categorized bone:
    - No obvious lesion and no DNA
    - TB in element but not where they anticipated/sampled so it has been missed
    - Swing and a hit: they sampled vertebrae and got MTBC DNA
- Justifying a large scale sampling project
    - Contextual evidence
        - Is the pathogen of interest likely to be found here given characteristics of the lab?
    - Is there paleopathological/skeletal evidence?
    - Successful pilot study
- Site: Tenochtitlan-Tlatelolco (1300-1521 CE)
    - Characteristics suggest TB would have thrived
        - Densely populated area with a lot of people coming from many places
    - Examined and selected for bone and joint TB
    - High tuberculosis based on skeletal signs of evidence
- Pilot study:
    - Successful recovery

Main project:

- Sampled 50 vertebrae, 4 ribs, and 2 pelvic elements representing 48 individuals.
    - 160 libraries - sampling at different locations
    - Shotgun sequenced to 8 million reads
    - First map every sample to MTBC reference genome
    - Every sample has MTBC reads mapping
        - You can be pretty confident with samples with over 500 reads of TB are true
- Identifying true MTBC over environmental is important. Shotgun data workflow:
    - Map to MTBC ancestral genome > bam to fastq > husonlab/malt (custom database of 108 mycobacteria and 111 closely related genera reference sequences) > comparison of mycobacteria and MTBC summed up reads
        - If it summed at MTBC node = positive
        - If summed at Mycobacterium node = negative
        - Parameters: MTBC summed/Mycobacteria summed and the number of MTBC assigned reads -> because these are all from

the same geographic area an increase in MTBC mapped reads should correlate to an increase in true MTBC reads in the samples.
    - In positive samples you should observe a positive relationship.
- 36 samples found to be positive
    - All sub-samples positive within the same element were positive
    - The sample set is small so its difficult to detect trends
    - MTBC positive category was primarily from juveniles/adolescents who died from 5-10 years
    - MTBC negative was also primarily populated by juveniles/adolescents
    - All ribs were negative and positives were lumbar and thoracic vertebrae
- Pathological elements should be prioritized for sampling, particularly lumbar verts. Look for classic TB lesions and lytic lesions.
- Subsamples from non-pathological areas on pathological bones can produce similar results to directly sampling lesions.
- Ratio of MTBC to mycobacteria assigned reads was predictive of genome recovery.

General Discussion Part II

- **Q: Nikolay Oskolkov to Kelly Blevins: super interesting, wouldn't competitive mapping against all bacteria help fishing out truly MTBC reads? To my understanding, you use competitive mapping on a later stage (after you have mapped to MTBC ref genome alone) to discriminate between MTBC and other mycobacteria, but wouldn't it be informative to use competitive mapping as well at the beginning against the full database of all possible bacteria?**
    - **A: I didn't use it from the beginning for time and to save on computational resources. I used a smaller database because we didn't have resources for a larger database. SO I took a subset and competitively mapped that.**
- **Q: Richell Ramirez to Kelly:** Good job! With regard to the samples that do not have lesions, did you study the teeth of the individuals? The interesting thing about this is that we study teeth because there are reports of treponemal DNA in teeth, so we analyzed the teeth to detect mycobacterium and we found the positive signal there.

    - A: I didn't examine teeth because we have very little evidence that MTBC DNA is preserved in the dental pulp cavity. TB is a more localized infection.
- Q: Aida to Kelly: Why not screen teeth to look for co-infections because there are some associations with pathogenic microbes causing comorbidity?
    - A: Yes! That's something she wants to do but it depends on research design.
- Q: Nasreen Broomand to Kelly: did you test the soil from the burial areas? With relatively low coverage I always worry that the mycobacteria I've detected are actually very closely related to Mycobacterium species--if you didn't test soil samples, do you have a way of verifying the hits are tuberculosis? Thanks!
    - A: No I didn't sample the soil because these were excavated in the 60's so there were not samples of soil. Second part of that question
- Q: Shreeya to Kelly: I find pathogenic bacteria but then I also find closely related species that have a damage pattern to provide authentication. SO I am wondering if I am actually finding two different taxa or if they are just closely related environmental bacteria and mapping. How do I know if I am finding the pathogenic bacteria or not? Having reads assigned to two microbes (pathogens or not). How do you distinguish these?
    - A: Aida answers: For certain bacterium if you find this pathogen it is likely it, it might be contaminated but it is there. For others there may be contamination and that it is not doing anything. Distinguishing these is difficult when you have an ancient damage pattern.
    - A: Meriam answers: Often times you have an environmental contaminant that is present but it could be that you have both. I think competitive mapping would be the way to go to figure out if you have one or two of them.
    - A: Nikolay answers: Aida you mentioned pestis, I often detect *Y. pestis* and *Y pseudotuberculosis* and *Y. enterocolitica* and these guys almost always pop up and they have fantastic damage profiles. And they have plenty of reads so even if you do competitive mapping you will still get tons of reads mapping to all of them.
    - A: Aida answers: Look at the plasmid. Many yersinia are in the soil so it is not uncommon to find several. Look for genes specific to the pathogene and they should be there.
    - A Meriam: If reads are well covered in the chromosomes for this that could help.

    - **A Aida: For example TB like Kelly presented is notoriously difficult and complicated to authenticate.**

Database bias

- This can sway your results when you don't have plasmids or specific signs
- We interpret our results of ancient pathogens through a modern lens
- Felix: Ultimately the proof of any species, you need to plug it into a phylogeny.
- Aida: What do we do with highly recombinant organisms?

Q: Aida: What sort of authentication criteria do you want to see in publications?

A: Phylogeny needs variant calling so we have to have the right reference to construct phylogeny. If we don't have the right reference we map to then it calls the wrong variants.