SPAAM3 - Session Notes

Session 4: Tool up or die - gaps and solutions in ancient metagenomics analysis

Chairs - Aida Andrades Valtueña

Presenters - Irina Velsko, Maxime Borry, James Fellows Yates, Frederic Lemoine

Session Abstract

Workflow and tool development in the field of ancient DNA (aDNA) has been largely focused in the study of ancient human genomics. A combination of High Throughput Sequencing and increased availability of High Performance Computing has opened the possibility of analysis of whole metagenomes. These advancements have led to broadening the types of specimens analysed, new avenues of research, and a myriad of data types being generated. Therefore only recently has ancient metagenomics come to the forefront of palaeogenomics and is primed for the development of tools to tackle the challenges of ancient metagenomics, such as, but not limited to incomplete and biased databases, validation of taxonomic profiling, authenticating datasets, contamination with modern or ancient microbes, etc. In this session we would like to (1) discuss new tools/workflows being developed for the analysis of ancient metagenomic datasets, as well as (2) review commonly used workflows and tools in the field of ancient metagenomics and how to improve them.

Session Objectives

01 Hear about tools developed by SPAAM3 members

02 Discuss which tools we would like to have to work in ancient metagenomics

03 Create a trusting environment where we can discuss new tool ideas (and maybe establish collaborations)

Part I: Let's start from the beginning: Metagenomic classification

Irina Velsko - *Which taxonomic classifier should I use?*

Presentation Notes

Important preface: There is no one-fits-all classifier for ancient metagenomics, because your question will determine which is the most appropriate to use.

Note of encouragement: You don't need a computer-related background! Irina started her postdoc in ancient metagenomics without having worked on the command line at all. And crash-course googling is a standard way to learn things!

Metagnomic profilers create taxonomies to approximate the original microbial community composition

Are modern metagenomic profilers biased when used for aDNA?
Be mindful that aDNA is usually degraded into short, damaged reads, which affect specificity and accuracy of profilers, respectively.

Keep in mind: all programs were developed for a specific purpose! Some prioritize speed, accuracy, database flexibility, or strain differentiation…how to find out? Read the original publication.

Tested five programs using different assignment methods (Velsko et al., 2018)

- **16S rRNA**
  - QIIMEI/UCLUST/GreenGenes
  - DADA2(QIIME2)/GreenGenes - wasn't appropriate for these data
- **Single copy markers**
  - MetaPhlAn2
  - Midas
- **K-mer matching**
  - CLARK-S (similar to Kraken)
  - MALT

Simulated data sets of modern and ancient data with gargammel.

Observed biases:

- **Alpha diversity metrics (within community/sample)**
    - **Most programs overestimate the # of species, even when setting a threshold for minimum abundance**
        - MetaPhlAn2 seems to be the only program that doesn't overestimate species diversity
    - **Phylogenetic diversity is consistently underestimated by most programs**
    - Biases will be specific to the diversity metric you choose
- Beta diversity metrics (between communities/samples)
    - Weighted unifrac t**o determine "accuracy"**
        - MIDAS performed best
    - Unweighted unifrac distance
        - MetaPhlAn2 performed best
- **Species abundance**
    - Biased more by **database** than by assignment method (e.g., does your database include viruses?)
    - Know if your profiler is calculating abundance based on read count of cell count (i.e. is it normalizing based on genome size?), for example, MetaPhlAn2.

Important aspects to consider:

- Damage patterns differ
    - By species
    - By GC content
    - By age

Irina's guide for choosing a classifier:

1. Know your question
2. Know the biases of the programs you're considering? (see the original publication)
3. Know the biases of the databases

   a. Which species are present/absent
   b. Are some species over-represented?
4. Know what your program reports
   a. Read counts?
   b. Relative abundance?
      i. Read counts or cells?
5. Which program has biases that are least likely to affect your question?

Maxime Borry - *Computing lowest common ancestors on SAM files with sam2lca*

Presentation Notes

- **Problem**: currently forced into using tools that combine aligning and LCA algorithm without preserving mapping quality (MAPQ), such as MALT, kraken, etc.
- **Solution**: split aligning and LCA assignment using your favorite mapper and Sam2lca - computing lowest common ancestor on SAM files

Fragmented, short sequences are sources of ambiguity in taxonomic assignment. When a short sequence is compared to a sequence database, it may match to multiple references/organisms. So how do we deal with assigning a read to the correct species?

MAPQ - the probability of uniqueness of alignment segment, which is normally used to filter reads that map in more than one location.

MAPQ

- Computation differs between aligners, but generally you get a score up to 42.
- You use this measure as a threshold, everything below the threshold is discarded
- Exclude reads mapped multiple times.

Why discard reads that map to multiple locations/references? Aligners can keep a record of multiple alignments, for example, Bowtie2.

Bowtie2
Search for multiple alignments, report best one
-k mode search for one or more alignments
-a mode: search for and report all alignments

Taxonomy and LCA to the rescue!
Instead of discarding the reads, drop them off at a less specific taxonomic level

SAM files - tracks the multiple alignments
Sam2lca (https://github.com/maxibor/sam2lca) uses this info from SAM files and the lowest common ancestors (LCA) algorithm to compute lowest common ancestors of multiply-aligned reads.
(sam files from MALT lose their LCA and MAPQ assignments)
It can be used in combination with your favourite read aligner.

General Discussion

Slack question

From Nikolay to Irina

- **Q: Super interesting work, thank you! When you say "detected" or "not detected", this would depend on your filtering strategy, right? E.g. Methaphan gives you normalized abundance values from 0 to 1, how do you decide if a species is present?**
  - **A:** Initial abundance is affected by threshold parameters, especially for low abundance taxa
  - Setting an abundance threshold may not be the best way to filter data; a factor analysis is another way to perform filtering
  - Also depends on your focus: are you interested in specific taxa or community abundance, because low abundance taxa will always be affected by filtering method
- **Q: Nikolay to Maxime: what does the alignment.bam look like? I mean the one you get out of "sam2lca analyze input.bam"? This is important because it seems like with Bowtie and sam2lca you could use a larger database more efficiently than MALT?**
  - A: Most/all aligners set MAPQ to 255 (unavailable) for secondary+ alignments; this is a limitation of aligners themselves

  - **LCA information is not included in the .bam file currently, but the sam/bam format is flexible and allows for the addition of other tags, so it could be added**
- **Q: Nathan Martin to Maxime: Have you noticed a better taxa detection with LCA directly computed on the sam file than with LCA done from a blast of reads presenting multihits after a mapping step (e.g. process done by Megan)?**
  - **A: By default sam2lca computes the LCA on all mapped segments, so all reads that had multiple mappings; there are some filtering options already implemented, such as a 95% sequence identity.**
- **Q: Claudio Ottoni to Irina: Thanks @irinavelsko for the talk! Question open to all those using databases of complete genomes (and doing microbiomes): how do you deal with genome lengths? Any normalizations?**
  - A: Irina hasn't done this and doesn't know of any tools that take count of reads and normalize by genome length - may not be substantial for a lot of taxa, but we may be overestimating abundance for smaller-genome organisms.
- Q: Nasreen to Irina: I have a not-so-fully-formed question/thing I need clarification on: So some of the species on your graph (notably Bordetella pertussis) are listed as "under-detected" are species I tend to routinely see in my outputs (which I highly suspect are false positives based on soil samples and context). Does this just mean I'm using a different database? Part of the problem is that I'm not the person who set up the programs for our lab so the database stuff seems like a bit of a black box for me.
  - A: It largely is determined by input dataset and abundance. Irina and collaborators have noticed that there is a set of taxa that is unusually abundant in many samples. They seem to be real genomes, so possibly they are abundant in older samples but poorly studied in modern samples.
- Q: Nico to Maxime: Have you tested how much disk space it takes to make a bowtie database of the NCBI nt database, and how much RAM it takes to load it for the analyses? How many

folds bigger get the bam files of the results (when using the -a parameter), compared with the initial size of the fastq of a metagenome?
  - A: samtools command merge to merge sam files; don't build a monolithic database (hard to maintain, rebuild after every modification), rather index each genome independently because it is relatively easy to parallelize mapping; so the memory issue wouldn't be much of a problem
  - Q: Nico - but wouldn't that have some problems if you align to a single genome? Would it not be competitive mapping in that case?
  - A: Maxime: You map samples independently to references; THEN merge the sam file using samtools merge; THEN apply LCA algorithm through sam2lca
  - A: James: The sam2cla concept is the same as malt: competitive mapping! Unlike malt, however, sam2lca splits up the steps: first align then merge and apply LCA algorithm
  - Nikolay: are you splitting up eukaryotes and microbes? Won't that bias the results?
  - Maxime: no, not really. It's up to you to select your organisms for your database.
- Q. Nico to Maxime Have you benchmarked the results of sam2lca using NCBI nt DB compared to KrakeUniq on the same Database?
  - A: No, it is far down on the road map. This presentation is more of a call for ideas and collabs. You can use this tool already, but it is far from finished. Basically a teaser to get you to contribute haha!
- Q. Irina to Maxime where would sam2lca go in the eager pipeline?
  - A: It depends on what exactly you want to use it for - it's basically just the lca aspect of the metagenomic profiler.
- Q. Nico to Irina: what are you using now?
  - A: Irina used MetaPhlAn2 to avoid over-abundance estimates and she wasn't interested in low-abundance taxa. Now Irina uses MALT as her lab's SOP, but they are running into issues with improving and expanding their database - TOO BIG! Too many genomes.

Part II: Pipelines to analyse your data

James Fellows Yates - *High-throughput and scalable ancient metagenomic analysis using nf-core/eager*

Presentation Notes

What is nf-core/eager (https://github.com/nf-core/eager) and why should you use it?

- Nextflow pipeline designed for aDNA that has been extended for use for metagenomics analysis.
- Has command lines and GUI interface
- Extensive usage and output documentation that specify what you should be expected to look for in your ancient DNA data
- Open access and can use plots in your own teaching!
- Highly parallelized with asynchronous submission
- Auto-retry and auto-resubmit jobs
- Portable: local machine, schedulers, and cloud functionality!
- HIGHLY reproducible:
  - Containers
  - Profiles
  - Strict versioning
- Functionality for ancient metagenomics:
  - Complexity filtering using bbduk
    - Remove the reads that have very low sequence complexity (like AAAAAA or GGGGG - reads that do not contain any taxonomic information)
    - Decreases memory costs and run-time of taxonomic classifers
  - Aligner using MALT
    - Blast-like alignment
    - But takes a lot of memory
  - Profiling using kraken2
    - K-mers - no alignments
    - Read counts
  - Metagenomic aDNA authentication using hops

        - Damage and edit distance for reads by pathogen species
    - Strip FASTQ
        - Remove endogenous human DNA
        - Protects sensitive human DNA
    - Functional characterizing using bedtools coverage
    - Genotyping
        - GATKunified genotyper
        - Freebayes
    - Cross mapping and SNP alignments
        - MultiVCFAnalyzer

Frederic Lemoine - *Paleogenomics of ancient and modern pathogens with AMPHY*

Presentation Notes

AMPHY (Ancient and Modern Phylogenomics)

Ancient DNA from pathogens - is there anything better??!

- Persistence over time
- Ancient disease and past epidemiological contexts and societal responses
- Pathogen evolution from past to presence

Very challenging to get these data though:

- Short reads
- Damage
- Environmental contamination

Reconstruct phylogenetic trees from whole gnomes and must be

- Reproducible
- Generic
- Easy to use
- Executor independent
- Flexible

Start from raw data ->through to-> phylogeny

Useful for modern and ancient samples

AND public and experimental datasets

Nextflow workflow using a singularity container and git versioning

First step: sequence analysis
The pipeline will process your samples until obtaining a Multiple Sequence Alignment (MSA). This processing includes quality control, adapter trimming, mapping, removal of duplicates, aDNA damage correction, genotyping.

Important step prior to MSA
-select genomes and samples based on input parameters of sample coverage and gene coverage; these samples and genes then go into the MSA

Second step: phylogenetic analysis
This will include selection of sites for tree inference, tree building and bootstrapping. As outputs there will be the final tree plus a visualization of the final tree.

Still a work in progress.
Improvements to come:

- Automatic selection of the best ref genome
- Automatic selection of phylogenetic parameters
- Placement of low-coverage strains on a previous tree
- Phylogenetic-based pathogen-detection in raw seq data

General Discussion

- **Q: James to Frédéric Lemoine does the pipeline offer multiple insertion points? Could you (e.g. run eager to) generate the VCF files, and then start AMPHY from the MSA step?**
    - **A: Yes, that's something we want to incorporate**
    - **A: Nico comment: this is the workflow we use in our lab, so it was built to meet our specific needs.**
- Q: Betsy to Ferderic: Great talk! You can construct both full alignment and variant based alignment? Is there an option for partial deletion of sites to deal with the ambiguous sites we often have when working with ancient genomes where there is no call in the ancient genome but its covered in modern representatives?

    - A: If you have ambiguous snp calls, that would be taken into account in the tree. It should be built into the phylogenetic models
    - A: Aida followup: can you exclude regions like homoplasies from variant calling?
    - A: Frederic: yes, you can include and exclude certain regions
    - A: Nico: You need the phylogeny to detect homoplasies though
    - A: Aida: We know from modern studies of organisms that there are homoplasies and where they occur, but if you could include a test for this in the pipeline, it would serve as a good validation step and then would be useful to detect homoplasies in ancient, uncharacterized pathogens. I have been using HomoplasyFinder.
- Q:James to Frederic: Is it freely available yet?
    - A: No, not yet. But if anyone wants to try it out, let Frederic know.
- Q: Aida to Frederic: You are using mapdamage for your damage correction; have you done any testing to detect how mapdamage is introducing reference bias, since it relies on a reference genome?
    - A: Frederic: No we have not done these tests, but it is a good idea to do.
    - A: Nico: It may be possible to avoid this by mapping to the pan genome
- Q: Kelly to Frederic. Why did you decide to go with raxml as your ML tree tool? @everyone are people mostly using raxml? Does anyone have any experience with IQtree
    - A: Frederic: RAxML-ng has the same functionality as IQ-tree and is quite good. It has ways to correct for the biases introduced by variant-only alignments. The tools are considered to be equivalent in performance.
    - A: Nico: the goal is to have the pipeline run the model finder for you, so you don't have to choose the parameters
- Q: Aida to Frederic: How do you deal with different types of libraries?
    - A: Frederic: Seq technologies and library treatments: we need to think more deeply about pre-treatments and how to treat samples. We have mostly been using homogenous Illumina libraries.
    - Aida: We use many different types of libraries, single stranded, UDG-treated, non-UDG treated, etc, and we use different mappings parameters for them accordingly. It is important to map using less strict parameters when using non-UDG treated sequencing aDNA data.
    - James: AncientMetaGenomeDir plug; imagine a world where you can download all of the X genomes and the associated metadata - we need volunteers!
- Q: Aida to Frederic: How do you do variant calling with different types of libraries being included?

    - A: Frederic: Currently using GATK HaplotypeCaller open to suggestions

- Q: Kelly to all: Is UnifiedGenotyper going to become unusable because it is deprecated?
    - A: Aida: This is a good point. Someone should do a test of different variant calling tools. I would be happy to collaborate on such a project. Another good project idea is a VCF format translator!
- Q: Kelly to Frederic: How are you making your multiple sequence alignments?
    - A: Frederic: We want to try several options. One option is to take the GFF file of the reference annotations and concatenate all of the genes from that file to make a fake whole genome alignment. We also want to use the whole genome as well.
    - Nico: We want to divide the genome up into partitions to optimize the tree: protein-coding genes, tRNAs, non-coding sites, etc
    - Aida: This is interesting because we don’t use any RNAs in our trees because they are prone to contamination because they are conserved and they have a higher mutation rate and can throw off the phylogenetic signal of the tree
- Q: Aida to Frederic: When you are filtering the sites, would you remove heterozygous sites?
    - Frederic: It is not added yet, but it is something we should think more deeply about.
    - Meriam: It has to be a high-coverage strain before you could differentiate mixed strains from contamination