SPAAM2 - Session Notes

Session 5 - Reuse the refuse: applying new analytical methods beyond current practices

Chairs Anna Fotaki & Alexander Hübner
Icebreaker Presenters Maxime Borry & Antonio Fernandez Guerra

Session Abstract

Throughout its history the field of ancient genomics has traditionally been playing catch-up by creatively adapting the advances achieved in modern genomics to their own types of datasets. Although this process has not been without setbacks, it steadily pushed the field forward and allowed it to apply the same methods as when using modern DNA.

The newest technological advancement that took over metagenomics is de-novo assembly, whereby short reads from an environmental sample are assembled into novel microbial genomes. Its application to sequencing data from present-day samples has led to a rush of thousands of new metagenomic-assembled genomes (MAGs) from uncultivated microorganisms and allowed scientists to investigate the “unknown” at large scales for the first time. However, its application to sequencing data from ancient samples has just started and still requires thorough evaluation and fine-tuning of its parameters to make it suitable for this special type of sequencing data.

In this session, we want to discuss the challenges that come along with applying de-novo assembly strategies to ancient metagenomic data and highlight potential pitfalls in the current workflows. Furthermore, we want to develop some guidelines on how to interpret this new type of results and establish which conclusions can be drawn from them, in order to avoid long-lasting mistakes that come along when adapting new methods to ancient genomics. Finally we would like to encourage novel ways of approaching datasets building upon previous sessions of SPAAM2 regarding public data sharing deposition

Session Introduction (Anna Fotakis & Alexander Hübner)

- Question from the chairs (hand raising).
    - Who knows what genome-resolved metagenomics is?: ~½ audience
    - Who is applying these concepts to their research?: 2-3 people
    - People who have no idea what we are talking about :2-3 people

Icebreaker 1 (Maxime Borry)

Presentation

- Lots of early skepticism of assembly of ancient microbial genomes from supervisor: ‘didn’t work before’, ‘it’s too difficult’, etc.

- Common belief was damage makes too many break-points in contigs
    - Simulations from Maxime and Alex Hübner shows, actually no, damage not really a problem.
- More important: coverage!
    - Counteracts any mistakes from damage
- If assembly does work: how to validate authenticity (assuming sufficient damage)?
    - Look through hundreds/thousands of damage profile figures? Not realistic to do.
    - Switch to statistical scoring!
    - mapDamage, is accurate but slow (robust bayesian model). HOPS is fast but very approximate (only looks at first 10 basis, no real single statistical output value(s) for cut-offs)
- What about trying other aDNA criteria to also model?

Q & A

- **Q**: how are contigs clustered for each genome? Unsupervised clustering is very difficult
    - **A**: There are lots of tools out there, with sufficient depth it's less problematic that data is ancient. Contigs are contig, damage/short read lengths obstacles already removed.

Icebreaker 2 (Antonio Fernandez Guerra)

Presentation

- Antonio is new to aDNA, actually modern metagenomicist by training.
- Asked by his boss to enter ancient DNA, to take on challenges of applying modern methods to ancient data.
- This is a challenge, lots of complexity: short reads, low coverage, and damage.
- Primarily interested in function reconstruction of ancient metagenomes:
    - Unresolved questions being: effect of damage on amino acid space, substitution matrices etc.
    - Skeptical it's possible - most modern tools don't work.
    - Therefore: Goes his 'own way'.
- Performed some simulations (with/without damage, different coverages)
- Applying concept of 'Super-Reads'
    - Careful extension of reads from both ends (expand to only 100 nucleotides first),
    - Gets a little more information about each read, and extends to length than can be usefully assembled with modern tools(?).
    - For functional inference from metagenomes: 100-200bp sufficient for amino acid sequence reconstruction!
    - Result: when you get super-reads, you then get to a point you can start doing functional analysis of reads, regardless of damage (10x coverage best!).
- How to perform assembly?
    - This remains an open question, but doing simulations with MEGAHIT and unpublished PenguiN (see ref list) provided some optimism.
    - Emphasises that this will NOT always be possible. Most ancient assemblies will be very fragmented - particularly with very low coverage, very short reads, and high damage.

- Coding space?
    - Looking into damage and effect on codons.
    - Interested into what to expect in terms of effects of codon reconstructions, i.e. effect of damage on functional reconstruction.
    - Initial results have found some possible patterns, with some codons changed more often than others due to damage, these are mostly non-synonymous changes!
- Important consideration of functional reconstruction analysis: choosing the right amino-acid substitution matrix
    - Some are affected more than others g. BLOSUM - most common default in tools, most badly affected, whereas PAM30 appears to be best?
    - Choosing PAM30 might be restricting yourself, you get much more reliable function hits.
- Word of caution when using modern methods if you use short reads:
    - Always double check!
    - Short reads results in extremely short amino acids (AA) sequences useless.
    - Short AA sequences do not include enough info for functional reconstruction (will fail!)

Q & A

- **Q**: Why are you using amino acid information assemblies over DNA assembly for genes?
    - **A**: Might get more/maximise information *[Notetaker: missed second half of answer]*. Super-read approach helps a bit to get modern metagenomics assemblers running properly.

General Discussion

Authentication

- **Q**: What are your common strategies for those evaluating whether new methods could be applied e.g. *Yersinia pestis* vs *Yersinia pseudo*?
    - **A**: Requires per-organism specifications. Recombination differs per species, and the number of environmental species that may cross-map to the target of interest differs. General stuff they look out for: damage, read length, coverage,
- **Q**: Other labs? How are you doing authentication?
    - **A**: We use Kraken to identify possible pathogens (very clear differentiation between *Y. pestis* and *Y. pseudo*). Then damage, read length, coverage etc.. Sometimes we observe irregular peaks of coverage, but not sure what they are.
        - **A**: You should always expect some irregular coverages, due to conserved regions, so don’t worry too much. But if you’re looking at a well characterised genome or species, it can be useful to check if any crucial/characterising genes of that species are missing.

Getting into applying new techniques

- **Q**: Many people here might be out of their comfort zone. Often in studies we are using pipelines (because we likely will not have expertise in details behind each pipeline to run ‘manually’). Are there any questions from people who wanted to try a new type of analysis, and tried but failed? Or could you not even get into it ? Any feedback on what you could do to move forward?

    - **A**: Originally tried to replicate human techniques/databases on Reindeer/Bear calculus. There was *a lot* of skepticism from people working on humans, saying 'this isn't real'. But actually since, then they continued anyway and found the bear does look quite real! Tip is to just try anyway if it doesn't cost extra! Maybe good stuff comes out if you don't understand it all at first.
- **Q**: What can you do downstream with the Metagenome-Assembled-Genomes (MAGs)?
- **A**: Depends on your question? Only then decide what to do? E.g. Single copy genes? Bit difficult, field currently still trying to improve compatibility for QC stuff (e.g. for anvi'o). Issue is often how to compare between different assemblies?
- **Poll** based on previous answer - do your projects take a hypothesis-testing approach or an exploratory approach?
    - **Result**
        - Hypothesis-testing: 30% (9)
        - Exploratory: 55% (17)
        - Both: 15% (5)
    - **A:** Majority exploratory. Does grants with large funds lead to more exploratory projects? Generally, the poll results seem to suggest 'Yes'.
    - **A**: But big funding still not perfect, still waiting for a long time for samples to run e.g. MALT because not enough capacity.
    - **A**: *de novo* assembly seems only works if you have enough sequencing depth - but difficult for most people to understand (don't have enough experience). Many people have had previous difficulties convincing (modern) reviewers, although it looks like (from own experience) that we are appearing to start to convince them. Maybe exploratory approach is due to this: can we ever truly do aDNA assembly reconstruction? Too much time spent convincing reviewers which is why hypotheses not focused on in our projects.
- **Q**: As starting to be able to convince reviewers, we should start thinking about future directions. What about pan-genomics?
    - **A**: Once you do the assembly, you can already directly infer function (from genes on shorter contigs) before you have the whole genome. So almost easier to infer function first. This might provide an alternative concept/perspective for future projects: some groups already have projects to use gene-level assembly (don't care so much about genomes).

De novo Assembly

- **Q:** how can we design studies using these newer analytical techniques? Do people have questions on how to define these studies? What the pitfalls could we foresee in planning new projects with new techniques?
    - **A**: Primary concern remains about complexity of libraries and genome coverage. Most metagenomic samples remain comparatively very shallow due to prohibitive cost of deep sequencing.
    - **A**: One option to explore could be using 'debris' contigs of graphs to resolve strains (preprint from Chris Quince on this). Other stuff too.
    - **Q**: Have you tried these methods on real aDNA samples?
    - **A**: Not yet, only just started. We are still trying to understand the basic effects of aDNA

on methods using synthetic datasets.

- **Q**: What observations have people seen when applying de novo methods to aDNA data?
    - **A**: Samples dominated by caries-infected teeth appeared to have higher coverage and helped resolving MAGs?
    - **A**: On *Y. pestis* it's a bit special, as it is pretty slow evolving. But nothing special was seen, some rearrangements - however not much knowledge in modern *Y. pestis* on the effects of these rearrangements other than that it happens. It could be interesting to explore further, but few projects in the field on this yet (e.g. investigating these would be more interesting for more open genomes species)
    - **A**: Only been doing mapping based analysis at the moment, and would like explore more, but in our sample sets read lengths have been too short
    - **A**: My personal opinion: these methods are very complex and data rarely worked, and thus I am still a bit skeptical. Back in 2016 we tried assembly/binning but didn't really work with standard methods. We would like to try again with more recent tools - *if* these can account for damage.

Other assembly approaches and experiences

- **Q**: Has anyone tried binning reads before assembly?
    - **A**: Most people not. One person tried, but it was very difficult because reads so short and not enough information.
    - **Q**: What about using e.g. kraken for binning?
    - **A**: We tried with e.g. BGC databases, i.e. with more targeted methods. Binning works better with longer sequences e.g. genes. General approach was to BLAST, then gene prediction with gene scan, but with modified substitution model (accounting for damage). *[Notetaker: long, detailed and technical back-and-forth]*.
- **Q**: What about sedaDNA people?
    - **A**: Only trying mapping based approaches so far, many more problems even with established approaches (e.g. lacking databases) so still trying to tackle those.
    - **A**: Assembly is an obvious next step, but not tried yet. Haven't got enough data in that area to even try that. Furthermore, in our research area we don't even have modern data to compare - first trying to just create genomes from modern data!
    - **A**: Our lab doesn't even care about microbes (no offense!), we are more interested in eukaryotic stuff (jealous as microbial genomes so small). Most work is just identifying what's there: there is still so much we can't identify with a mapping based approach. Also trying to take different approach damage.
- **Q:** For those working in very high-throughput modern labs, what are your experiences?
    - **A**: we do have that sort of pipeline, but only just started with routine assembly. Reviewers always ask how to check against artefacts in MAGs that may have occurred from damage. We don't have serious projects on ancient MAGs yet, but thinking about ideas. It's already very tricky to validate modern genomes/MAGs as no reference, so moving to aDNA even harder. We therefore still have a very exploratory approach - still trying to sift through all the modern stuff (even there - in trying to identify what is soil/environmental vs. host associated is hard for modern researchers).
    - **A**: Still experimenting with a variety of approaches but nothing useful yet.

- **Q**: Talking about established vs new approaches. What about variation graphs (from genotyping sessions) - these take a similar approach to de novo graphs. Aligners also do similar things!
- **Q**: There are other interesting approaches - e.g. when going to gene level. Mapping approaches are useful but this means we are only comparing reads to 'known' stuff. In contrast bioSynthetic-SPAdes does assembly, but optimises at smaller contig lengths (but more complete biosynthetic gene sequences). Human level - but again short reads do not result in long enough amino acid sequences. Can we even get to protein level?

Tricks for improving assemblies

- **Q**: What is the minimum length of sequences that you need to use for assembly? Are there any guidelines?
  - **A**: Not really - Shorter sequences mean shorter contigs. There isn't really a minimum value - as long as they overlap it's 'assembled' but this results in a gradient of usefulness.
  - **A**: people generally set the cut off of 25bp in mapping, but it depends as some people try to go lower (e.g. labs focusing on middle Palaeolithic material).
  - **A**: Main influence of read length is on k-mer size; smaller k-mers result in worse assembly results.
  - **A**: We had an average of 42 bp for *Mycobacterium leprae* and didn't work.
  - **Q:** What was your minimum in your experience?
  - **A**: Most samples didn't work at 49-60bp, but if you sequence deep enough you can pick up enough longer reads to get 5-10 most abundant species to 50% coverage.
- **Q**: What about co-assembly to help with short reads?
  - **A**: tried assembling birch-tar pitch samples (which are complex samples). This didn't work well because 1) only one sample, and 2) complex mixture of DNA so need more samples to get that but these are rare.
  - **Q**: Do other people have similar experiences? Do you really need multiple samples?
  - **A**: I seem to remember from the ancient *Prevotella copri* (Tett et al. 2019) paper, that they only needed at least 5 samples for co-assembly?
  - **A**: The cleaning procedure is key. Need to filter out all the crap parts of the assembly. But we are still exploring this.
- **Q**: What do you mean the *M. leprae* assembly didn't work?
  - **A**: We tried to assemble but couldn't get any contigs to leprosy.
  - **A**: One of the problems with assembly is how repetitieve a genome is. When many repetitive points exist in the genome, it will lead to very fragmented assemblies. This is further problematic with very short reads as they won't span these repetitive regions.
  - **A**: Super-read approaches might help alleviate these problems, when every base counts.
- **Q**: Have people tried removing short reads and assembling on only longer ones?
  - **A**: I don't think it really matters because the assembler doesn't care so much about the reads themselves; the k-mers that are inferred from the assembly are more important.
  - **A**: Get as much coverage as possible, that's more important. This corrects for damage.

    - **A**: The convention in our lab is we filter reads after trimming if less than 30bp. Not sure if it will improve assembly necessarily, but it at least improves memory/computer resource requirements (this works well).
- **Q**: Please remember to not be too negative about the potential of these methods! Only now getting enough sequencing depth to experiment. Maybe we could make a list of successful aDNA assembly papers? *[Note taker: see ref list].*
- **Poll:** would you be interested in a training workshop specifically on genome resolved metagenomics and applications for ancient datasets?
    - Yes: 100% (25)
    - No: 0% (0)