SPAAM2 - Session Notes

Session 2 - Removing persistent trash – challenges in genotyping and filtering out contaminant reads from genome alignments

Chairs Åshild Vågene
Icebreaker Presenters Susanna Sabin, Kun Huang

Session Abstract

Genotyping is the characterization of variants (SNPs, Indels) that occur at the single genome level. Accurate genotyping of ancient microbial genomes is crucial for downstream analyses, such as phylogenetic placement, SNP effect analysis and molecular dating. Genotyping is also important for taxonomic profiling when trying to perform strain separation. Thus, inaccuracies in genotyping/variant calling can have an enormous impact on the interpretation of the results.

Confounding factors related to these genomes being a) ancient, b) from metagenomic backgrounds, can negatively impact genotyping efforts by causing multiallelic variant calls that infringe on thresholds for calling homozygous variants. These factors include ancient DNA damage, short reads prone to mis-mapping in regions of genomic rearrangement and cross-mapping of contaminating sequences derived from genetically similar organisms from the specimens' burial/storage environment. Approaches such as edit distance and analysing the allele frequency of multiallelic positions are commonly used to give an indication of the level of 'contamination' in an alignment. Both molecular and computational approaches exist to mitigate the effects of contaminant and mis-mapping reads in genotype calling.

In this session we will discuss issues and problems relating to genotyping of ancient microbial single-genomes, with emphasis on (but not restricted to) haploid bacterial and viral genomes. The aim of the session is not only to make others aware of specific issues/scenarios that may arise from working with different types of microbes, but also to have a general discussion of the best approaches to tackle such problems.

Session Introduction (Åshild Vågene)

- Brief introduction of general terminology

Icebreaker 1 (Susanna Sabin)

Presentation

- Intra-host diversity happens for both fast and slow evolving human pathogens
- Reasons: within-host evolution over time; compartmentalization heterogeneity in infection generation, this requires sampling over time
- Time sampling allows to study population dynamics
- Genotyping is complicated by aDNA damage and contamination; however, in the end it's about deciding on a threshold for the consensus allele

Icebreaker 2 (Kun Huang)

Presentation

- Build computational framework for targeting and removing damaged bases prior to building the consensus allele; three-step framework
    - 1. Aligning reads to one single reference
    - 2. Calculate position-specific probability of observing aDNA damage
    - 3. Reconstruction consensus genome sequence
- Tested against set of ancient calculus samples and focus on *Methanobrevibacter oralis*
- Filtering of sites with high probability of being affected by aDNA damage reduces the number of single-nucleotide variants → higher accordance with published mutation rates based on modern samples (https://github.com/SegataLab/cmseq.git: consensus_aDNA.py)

General Discussion

- **Q**: Chair runs poll - Who is working on single genomes?
    - I am
        - I am: 67% (6)
        - I am not: 33% (3)

Validation of genetic variation from contamination

- **Q**: What are prerequisites for discriminating real intra-host variation from contamination?
    - **A**: So far, most samples were excellently preserved making it feasible to study intra-host variation; mainly, it requires very deeply sequenced samples that are preferentially well preserved
    - **A**: New concepts like sequence variation graphs that allow simultaneous alignment against multiple genomes, e.g. published by Martiniano *et al.* (Genome Biology, 2020)
    - **A**: Great idea to align samples against multiple-reference genomes at the same time that makes it easier to identify ancient DNA damage, however, still hard to use in every-day set-up
- **Q**: How big is the reference bias of the tool introduced by Kun during his icebreaker talk?
    - **A**: It requires high-quality reference genomes and it is suggested to align against multiple reference genomes if there is a large variation among the different species, e.g. *Prevotella copri*. The tool is still immature, but hopefully in the future there are even reference-free approaches to distinguish them
    - **A**: Suggestion to move away from the concept of a reference genome because a reference genome is just a single haplotype from and not a standardised unit. Therefore, future tools should include the diversity of the reference genome in their genotype model.
    - **Q:** Questions if we should move to genotype-probability based tools like ANGSD for low-coverage samples rather than calling a strict genotype
    - **A**: The suite of genotype-probability based tools (e.g. ANGSD) are currently hard to use outside of their respective tool; not an easy solution for phylogenomics
- **Q**: Is capture-enrichment the solution for studying whole-genome diversity when only low-coverage genomes are available?
    - **A**: It's challenging for microbial genomes that are genetically related to contaminants, e.g. *Mycobacterium*; it will work better for animal genomics
- **Q**: Are there tools available that consider both ancient DNA damage and variation known from

literature?
  - **A**: At the moment, there are no tools that combine all these; ANGSD is hard to combine with other tools; snpAD considers aDNA damage estimated from the data but doesn't consider known reference variation
  - **A**: However, isn't the reference information more informative?
  - **A**: For pathogens, some labs use the known variation e.g. from *Mycobacterium tuberculosis*

SNPs versus genes - what is a strain?

- **Q**: However, one can also be interested in gene diversity rather than SNP diversity over time.
  - **A**: Pan-genome analysis is able to identify gene diversity over time; however, strain inference from ancient DNA samples would improve pan-genome analysis very much, but require a lot of mathematical development
  - **A**: In modern samples at the Segata lab, for strain transmission or separation analysis only the major strain is usually used and minor strains are ignored; however, it is already very challenging for modern samples with much larger sequencing depth
  - **A**: However, strains can be identified either by haplotype or by gene content; it might be more feasible to do pan-genome analysis instead if trying to infer correct SNPs → allows to estimate diversity using coalescence-based approaches
- **Q**: Are there wet-lab tools that can help bioinformatic tools?
  - **A**: Soil samples from the same excavation site could be off big help because they would allow to filter likely contaminants prior to do the strain analysis
- **Q**: What do we understand under the term "strain"? Set of genes or set of SNPs?
  - **A**: For very clonal microorganisms, there is not much chance to distinguish strains based on gene content; for these, SNPs are more helpful
  - **A**: Gene loss appears in some species that are living in biological niches so this might be still informative to look at
  - **A**: Gene content would be especially informative for diverse species when sticking to a single reference genome might lead to underestimation of diversity
- **Q**: How does one deal with pathogens/microorganisms with only a few published strains for comparison but that shows diversity among their samples?
  - **A**: Here, gene content analysis might be helpful or use marker-gene based approach
  - **A**: For oral microbiome, there has been a lot of gene-based studies on particular virulence genes that have been well studied; suggests to do literature research to identify known SNPs in these virulence genes

Tools used for genotype analysis

- **Q**: Which aligner (BowTie2/BWA) is better for working in ancient DNA work?
  - **A**: In a paper by the Orlando lab (Poullet and Orlanda, 2020: https://doi.org/10.3389/fevo.2020.00105), it was evaluated for ancient human samples; however, only marginal gains in certain settings were observed
  - **A**: One shouldn’t run BowTie2 in local mode or BWA mem because soft-clipping might otherwise remove the signal of aDNA damage that is used for validation
  - **A**: Systematically test with different aligners with respect to microorganisms might be worth as a future community project

- **Q**: Which tools are currently used to clean inferred SNP datasets? Just MultiVCFAnalyzer?
    - **A**: Some use freeBayes instead for genotyping instead of GATK, and then use vcftools/bcftools instead for filtering; allows to filter for known sites
    - **A**: MultiVCFAnalyzer is a bit out-dated and doesn't work with many newer VCF tools formats, so it might be worth switching forward to new tools
- **Q**: Chair runs poll- What genotyper do you use?
    - gatk(ug): 59% (10)
    - gatk(hc): 0% (0)
    - freebayes: 24% (4)
    - other : 18% (3)
    - Further comments:
        - htsbox majority consensus caller by Heng Li so far, but haven't done a careful evaluation
        - Angsd, but I've only used it for single genes so far.
- **Q**: Did anyone observe differences between GATK UnifiedGenotyper and freeBayes?
    - **A**: FreeBayes has returned slightly better results than GATK (lower number of obscure SNPs), but not systematically tested; however, in freeBayes one can easily fix ploidy to haploid and pool samples together
    - **A**: GATK has bias towards reference allele calling in default model because of purpose to identify medical SNPs in human or other model organism populations; freeBayes doesn't have this bias
    - **A**: MultiVCFAnalyzer does SNP calling just based on allele frequencies and ignores genotype probabilities of GATK; however, most information are available also in other programs
    - **A**: wouldn't suggest to use GATK for genotype calling for non-model organisms because of the genotype calling algorithm
- **Q**: How does one separate good from bad SNPs?
    - **A**: One could try to infer this competitive mapping
- **Q**: Should one remove homoplasy or singletons for species with known modern phylogenies?
    - **A**: It’s hard to decide but modern genomes might help to identify if this a real signal
    - **A**: However, be suspicious of large clusters of ancient strains that were dated to a much younger age than is suggested by the molecular date inferred from the alignment

- **Poll**: what tools do you use to filter genotype data?
    - MultiVCFAnalyzer: 13% (1)
    - Other: 0% (0)
    - VCFtools: 88% (7)
- Total Votes: 8

Strain analysis on highly similar genomes

- **Q**: How does one disentangle different strains from highly similar species in the oral cavity?
    - **A**: It is very tricky and we haven’t tried it on a larger scale. For their latest study, we also did no pan-genomic approach because this would require higher amount of

coverage
  - **A**: In most cases, highly similar genomes could be treated as a population and work on averages similarly like human population genetics works with multiple individuals from the same population
- **Q**: Are their tools available that can disentangle these signals in separate strains? How much data does one need for this?
  - **A**: There are multiple tools available, for e.g. StrainPhlAn, but the amount of data varies largely across species. Conserved genes might have a much higher coverage than other regions, which further complicates it.
  - **A**: More discussion on this will take place during the session on the de-novo assembly. However, there are a number of tools, e.g. StrainPhlAn and DESMAN, available but more and more labs switch to working on assembly graphs directly instead.
  - **A**: When working with cultural heritage samples be aware of the importance of these samples and give them their best shot. Such datasets might not be as great as what is expected by standards of the modern metagenomics fields but still worth trying
- **Q**: Is it possible to identify recombination events in ancient samples?
  - **A**: It is very hard to study due to the short read lengths.
  - **A**: Recombination is typically only a problem if you want to infer the phylogeny across the whole genome, therefore, it highly depends on research questions
  - **A**: For highly recombinant species like *Helicobacter pylori*, phylogenetic analysis might be difficult but researchers typically adapt whole-genome analysis tools from human population genetics, e.g. Structure and ChromoPainter

Community standards and standing of ancient DNA research

- **Q**: What could sensible community standards be? Or don't we need one yet?
  - **A**: The field is still in an early stage, but people should be open about their strategies
  - **A**: Community could push for reproducibility by reporting more specifically on which tools were used
  - **A**: While people are verbose when describing their own data, they usually do not describe how they treat published data → community should push to encourage people to describe this process, too.
- **Q**: What are typical criticisms from reviewers that the network has experienced?
  - **A**: Many reviewers do not understand that many of the analyses are still very experimental because of the lack of experience; often directly dismissed by reviewers from modern fields; suggestion: get in touch with experts of the field, e.g. species expert for pathogen studies, to collaborate
  - **A**: Quite commonly reviewers do not understand the scope of the studies; Segata lab usually includes experienced researchers to counter these criticisms; the SPAAM support network might be helpful in the future
  - **A**: Next to including archaeologists as experts, it is particular important to also focus on including experts from the modern fields