Date : 21.09.2020
Chairs: Åshild (Ash) Vågene

SPAAM

Standards,
Precautions &
Advances in
Ancient
Metagenomics

Session 2: Removing persistent trash

Challenges in genotyping and filtering out contaminant reads from microbial genome alignments

Session Scope

Accurate genotyping lays the foundation for downstream analyses that interrogate single genomes based on variant positions

- What are the challenges in genotyping ancient microbial single genomes derived from metagenomic contexts? (i.e. aDNA damage, cross-mapping reads, mis-mapping reads etc.)
  - What are the best approaches to:
    - evaluate the level of mis-mapping reads in an alignment?
    - mitigate the effects of mis-mapping/ contaminant reads on genotyping outcomes?
    - perform strain/genome separation (either for multiple strains of the same microbe or for different genetically close microbes?)

- Icebreaker speakers:
  - Susanna Sabin (Stone lab, Arizona State University, USA)
  - Kun Huang (Segata lab, CIBIO - University of Trento, Italy)

Definition of terms

- **Genotyping** = the process of determining the DNA sequence—the **genotype**—at specific positions across the genome of an single organism, usually done in comparison to a reference
- **SNP** = single nucleotide polymorphism
- **Indel** = Insertion or deletion of bases in the genome
- **UDG treatment** = enzymatic treatment with uracil DNA glycosylase (UDG) and endonuclease VIII to remove uracil residues from ancient DNA and repair resulting abasic sites
- **Position-specific damage probability:** The probability of observing a C->T (G->A) due to a post-mortem damage on a position of a read.
- **Damage probability cap:** A maximum threshold of position-specific damage probability for a base on the reads to be considered in building consensus alleles.

SPAAM

Standards,
Precautions &
Advances in
Ancient
Metagenomics

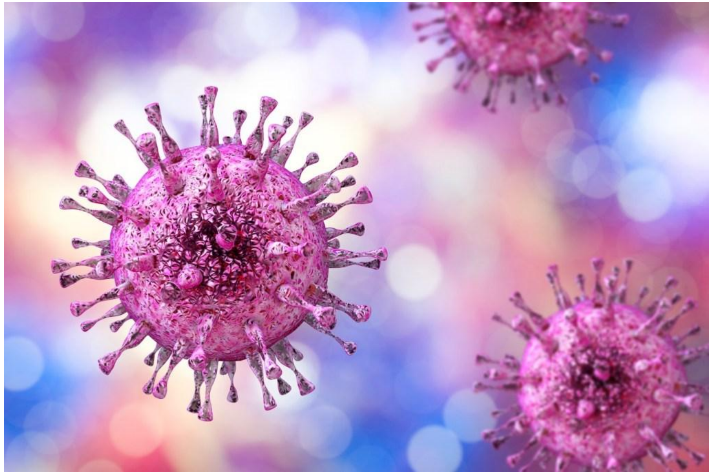
One man's trash…

The untapped potential of intra-host diversity in microbial ancient DNA

Susanna Sabin, PhD
Center for Evolution & Medicine, ASU

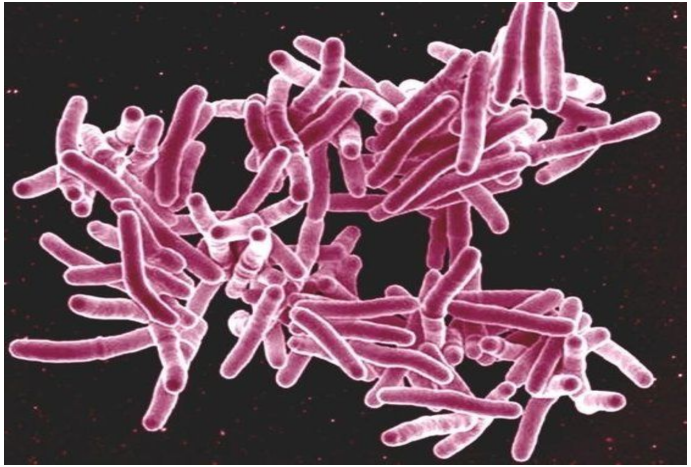
Intra-host diversity happens!

It happens in highly recombining and highly mutating organisms…



e.g. human cytomegalovirus

… and slowly mutating, non-recombining organisms.

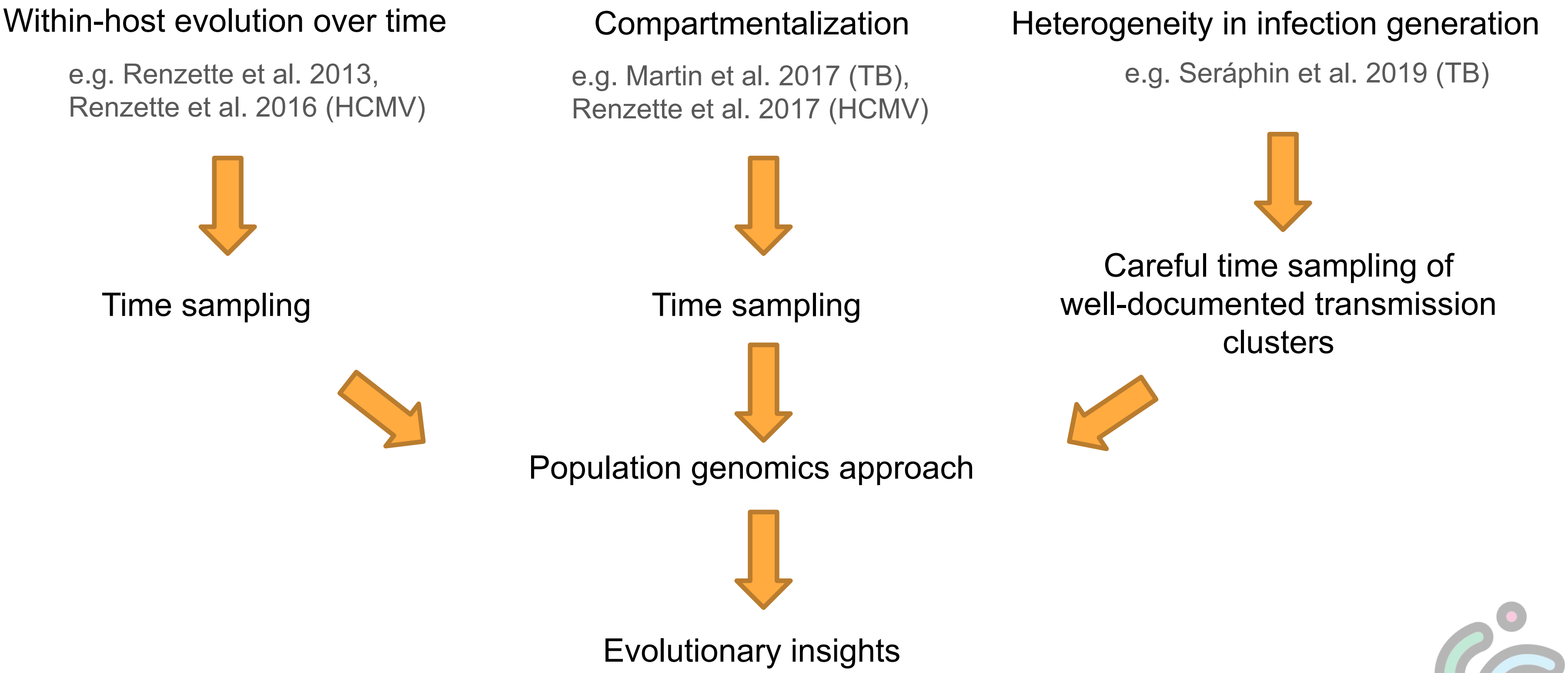


e.g. tuberculosis

Renzette et al. 2013 “Rapid intrahost evolution of Human Cytomegalovirus is Shaped By Demography and Positive Selection”

Trauner et al. 2017 “The within host population dynamics of Mycobacterium tuberculosis *vary with treatment efficacy”*

Clinical Realities



Within-host evolution over time

e.g. Renzette et al. 2013,
Renzette et al. 2016 (HCMV)

Time sampling

Compartmentalization

e.g. Martin et al. 2017 (TB),
Renzette et al. 2017 (HCMV)

Time sampling

Heterogeneity in infection generation

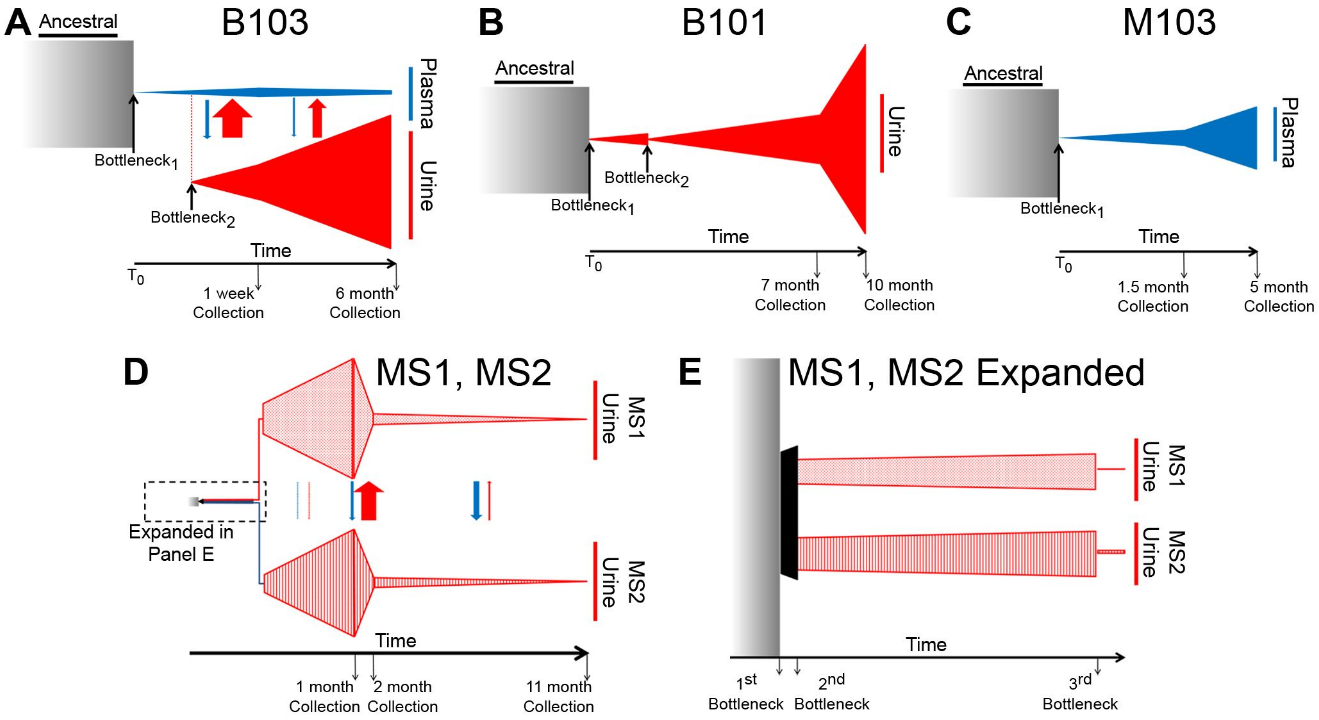
e.g. Seráphin et al. 2019 (TB)

Careful time sampling of well-documented transmission clusters

Population genomics approach

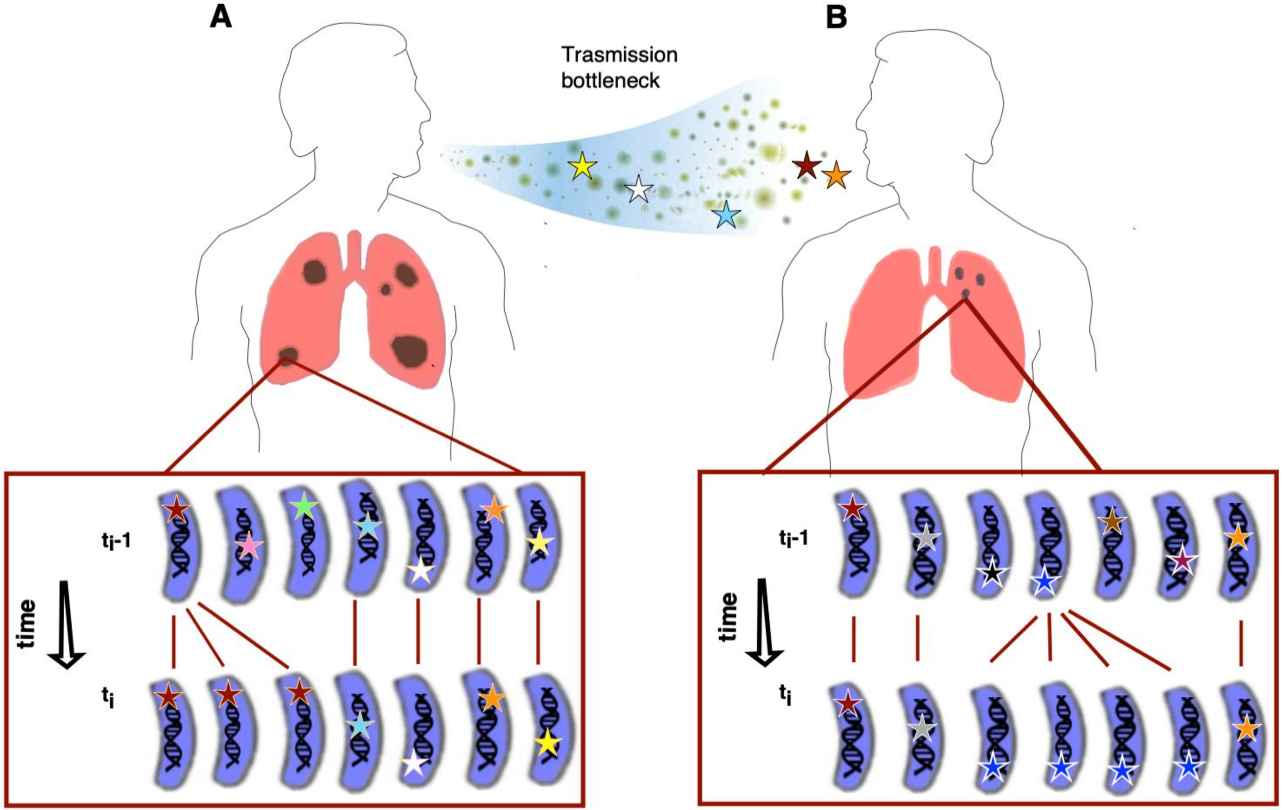
Evolutionary insights

e.g. Reconstruction of Detailed Infection and Population Dynamics (HCMV)



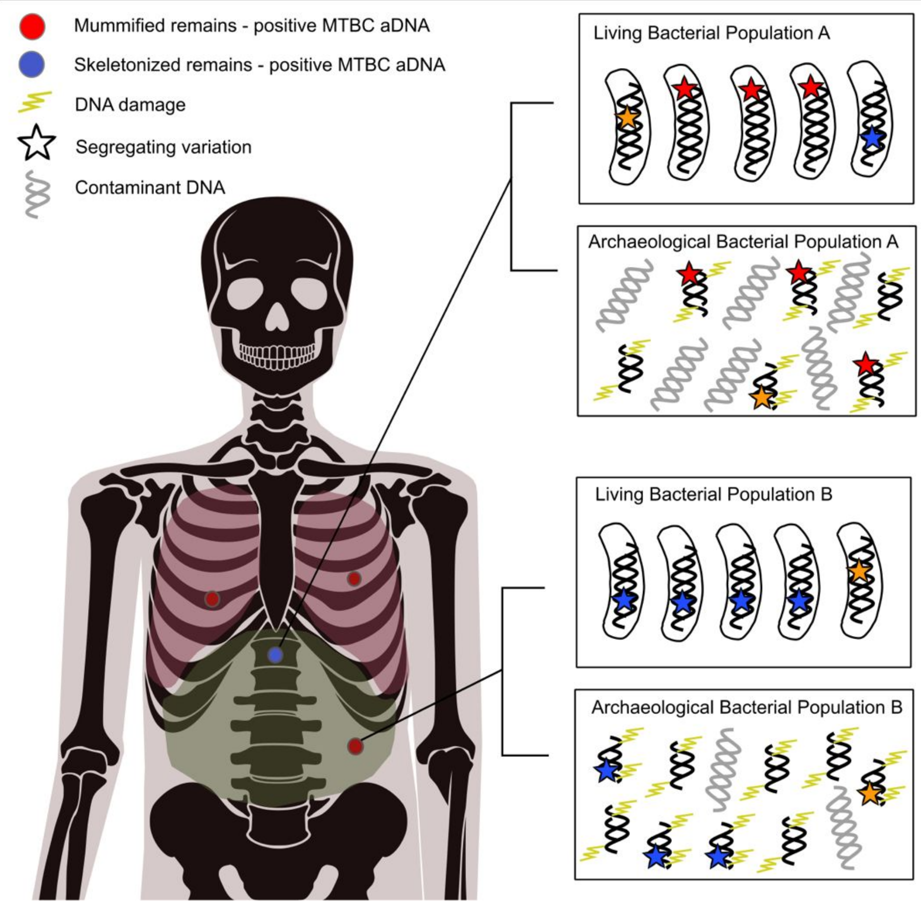
Renzette et al. 2013

e.g. Reconstruction of Detailed Infection and Population Dynamics (TB)



Morales-Arce et al. *in review*

What are we missing in aDNA land?



● Mummified remains - positive MTBC aDNA
● Skeletonized remains - positive MTBC aDNA
DNA damage
☆ Segregating variation
Contaminant DNA
Living Bacterial Population A
Archaeological Bacterial Population A
Living Bacterial Population B
Archaeological Bacterial Population B

Morales-Arce et al. *in review*

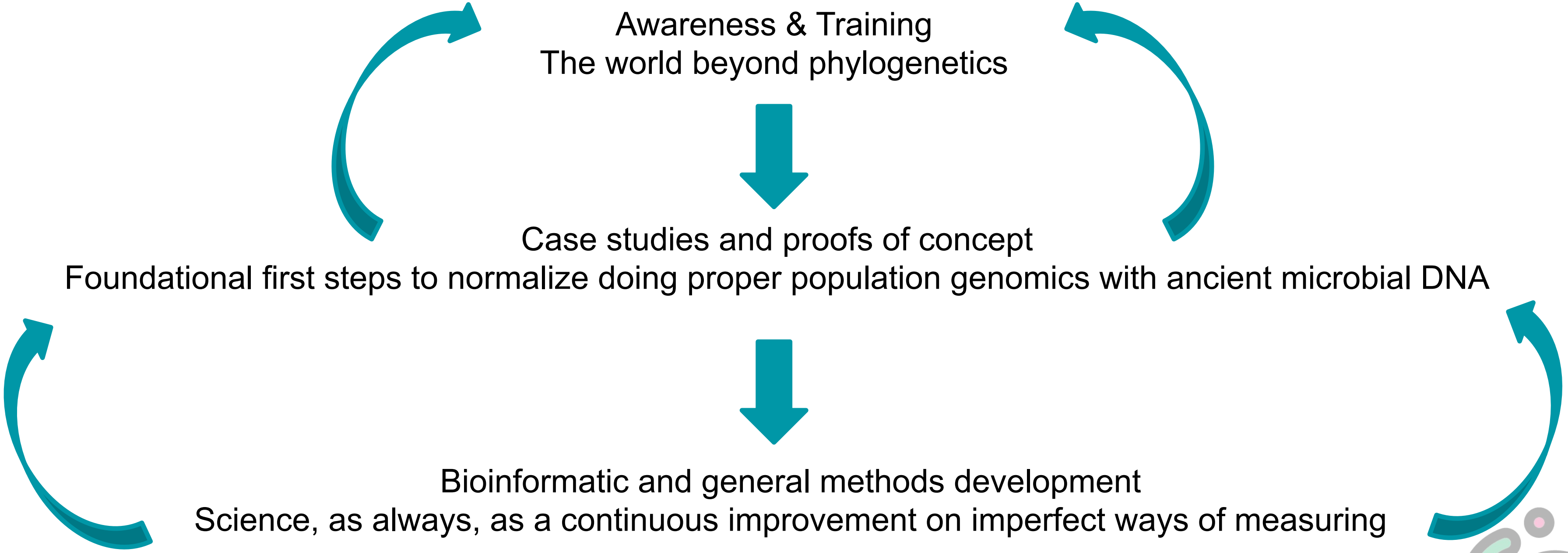
Accessing within-host, intra-species variation is necessary to gain evolutionary insights.
This requires either persistent serial sampling, or the quantification of minority alleles.

For genetic studies of modern samples, this means thoughtful experimental design and deep sequencing.

For the ancient DNA community, this means a massive headache.

BUT, it also expands the horizons of what ancient microbial DNA can teach us.

To Do List for ancient microbial metagenomics to open the door population genomics



Awareness & Training
The world beyond phylogenetics

Case studies and proofs of concept
Foundational first steps to normalize doing proper population genomics with ancient microbial DNA

Bioinformatic and general methods development
Science, as always, as a continuous improvement on imperfect ways of measuring

SPAAM

Standards,
Precautions &
Advances in
Ancient
Metagenomics

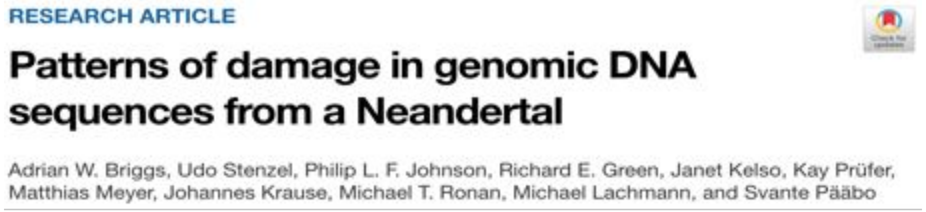
@Kun_L_H
@cibiocm

Building consensus alleles with awareness of position-specific damage probability

Speaker: Kun Huang

COMPUTATIONAL METAGENOMICS
SEGATA LAB

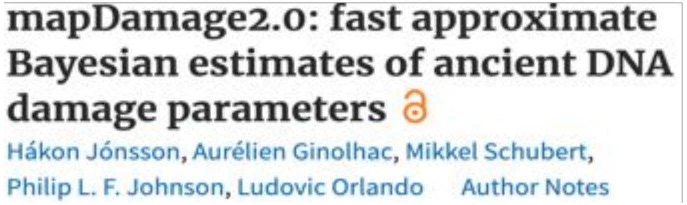
Cytosine deamination of ancient DNA: instrumental, and problematic as well



RESEARCH ARTICLE

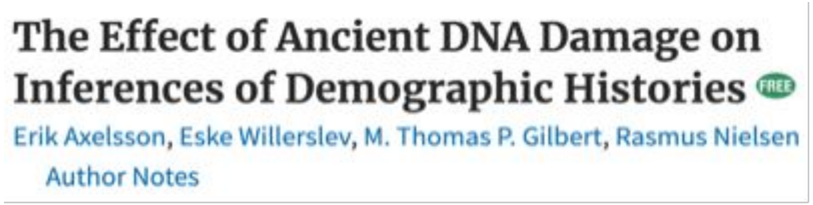
Patterns of damage in genomic DNA sequences from a Neandertal

Adrian W. Briggs, Udo Stenzel, Philip L. F. Johnson, Richard E. Green, Janet Kelso, Kay Prüfer, Matthias Meyer, Johannes Krause, Michael T. Ronan, Michael Lachmann, and Svante Pääbo



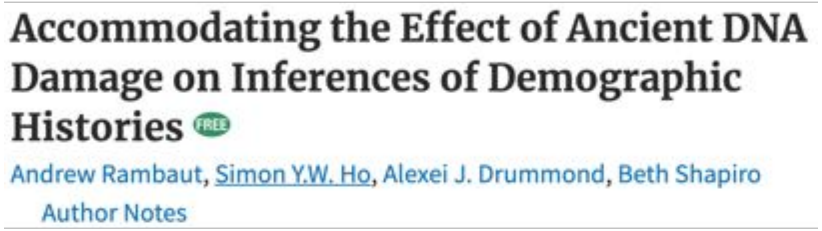
mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters

Hákon Jónsson, Aurélien Ginolhac, Mikkel Schubert, Philip L. F. Johnson, Ludovic Orlando Author Notes



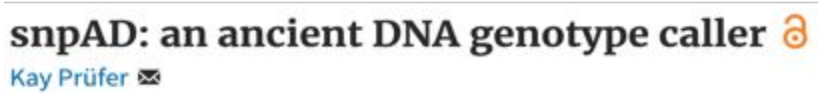
The Effect of Ancient DNA Damage on Inferences of Demographic Histories FREE

Erik Axelsson, Eske Willerslev, M. Thomas P. Gilbert, Rasmus Nielsen
Author Notes



Accommodating the Effect of Ancient DNA Damage on Inferences of Demographic Histories FREE

Andrew Rambaut, Simon Y.W. Ho, Alexei J. Drummond, Beth Shapiro
Author Notes



snpAD: an ancient DNA genotype caller

Kay Prüfer ✉

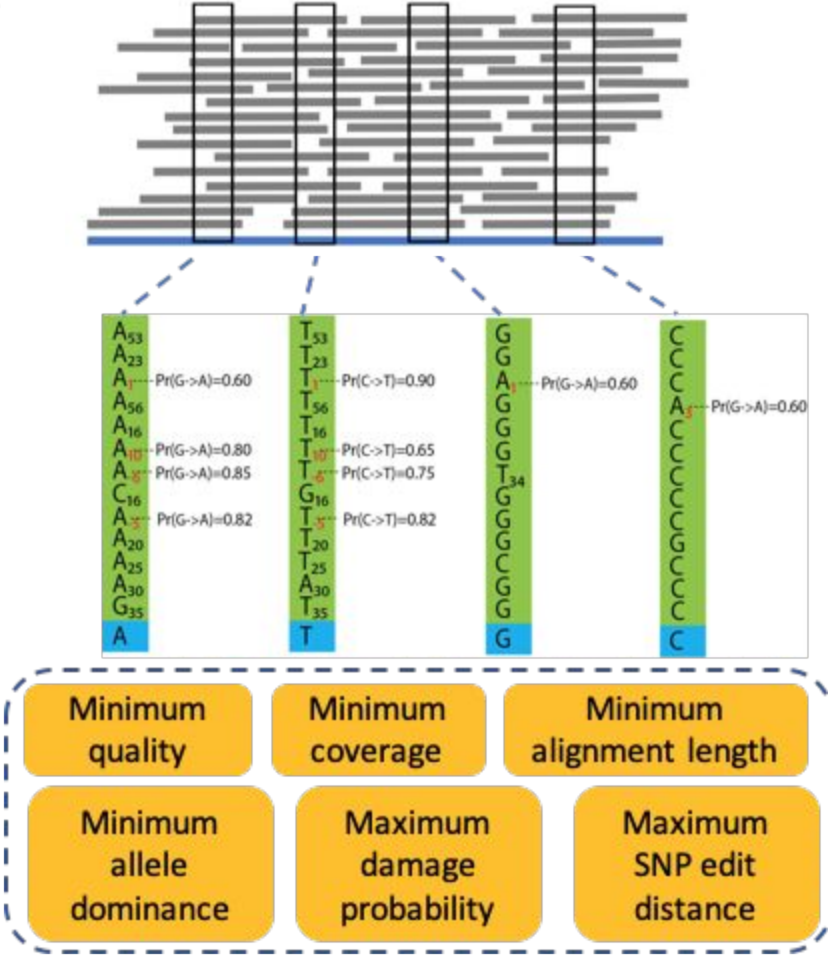
SPAAM

A computational framework for targeting and removing damaged bases prior to building consensus allele

Step 1. Aligning reads to one single reference

Step 2. Calculating position-specific probability of observing C->T (or G->A) due to a postmortem damage. (**mapDamage2**)

Step 3. Reconstructing consensus genome sequence


A_{53} A_{23} A_{1} A_{56} A_{16} A_{10} A_{6} C_{16} A_{5} A_{20} A_{25} A_{30} G_{35}
Pr(G->A)=0.60
Pr(G->A)=0.80
Pr(G->A)=0.85
Pr(G->A)=0.82
A
T_{53} T_{23} T_{1} T_{56} T_{16} T_{10} T_{6} G_{16} T_{5} T_{20} T_{25} A_{30} T_{35}
Pr(C->T)=0.90
Pr(C->T)=0.65
Pr(C->T)=0.75
Pr(C->T)=0.82
T
G G A_{1} G G G T_{34} G G G C G G
Pr(G->A)=0.60
G
C C C A_{5} C C C C C G C C C
Pr(G->A)=0.60
C

Minimum quality

Minimum coverage

Minimum alignment length

Minimum allele dominance

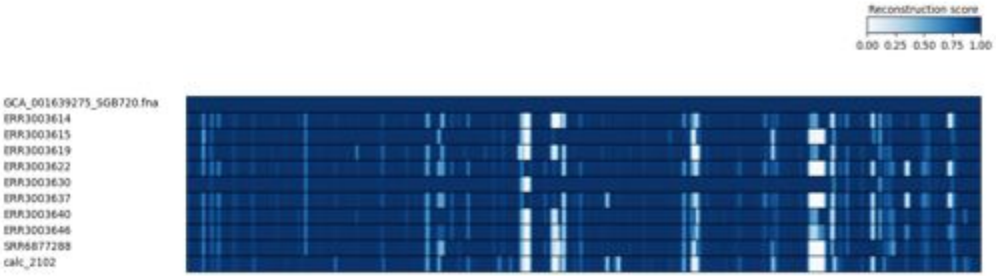
Maximum damage probability

Maximum SNP edit distance

SPAAM

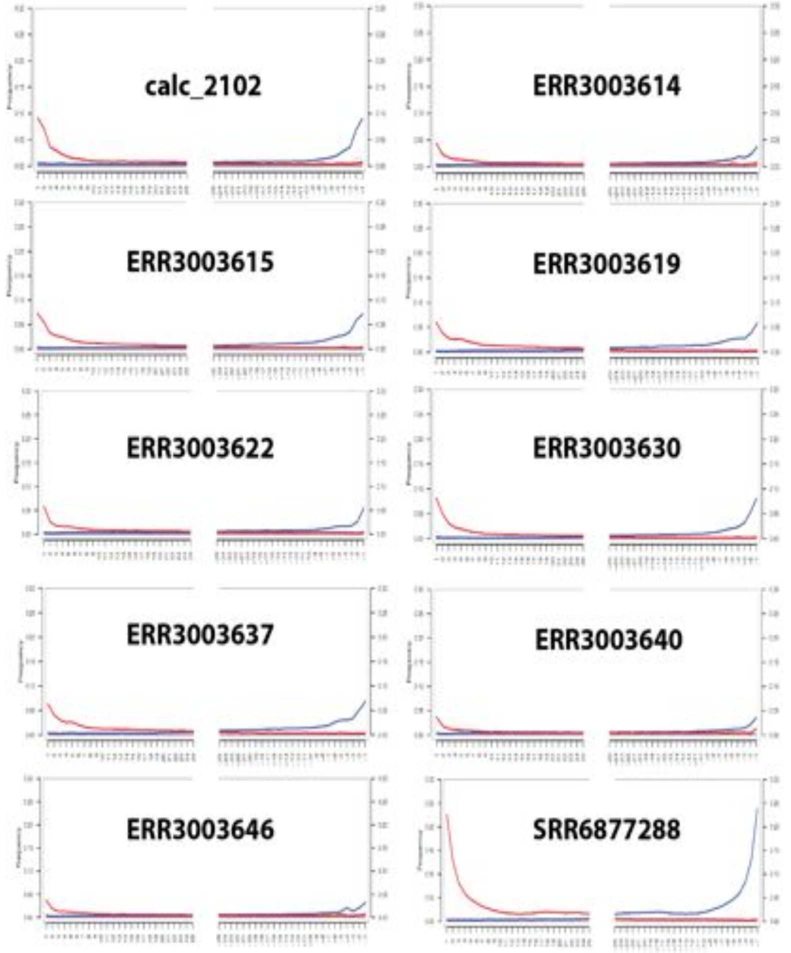
Benchmarking based on 10 ancient calculus metagenomes with focusing on *Methanobrevibacter oralis*

N=8 (Velsko et al. 2018)
N=1 (Mann et al. 2018)
N=1 (in this study)

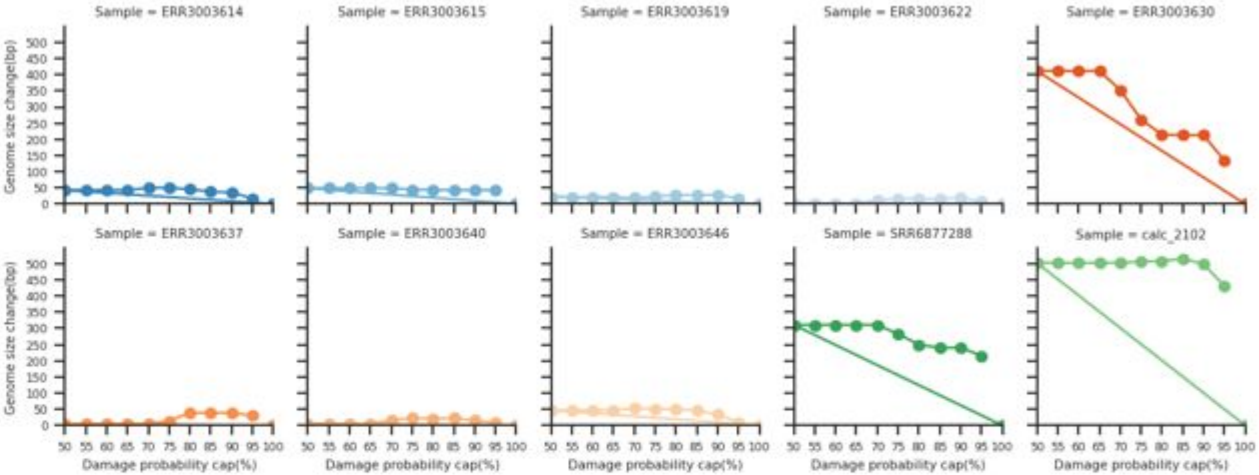


90% *M. oralis* reference genome is covered by sequencing reads at >3X

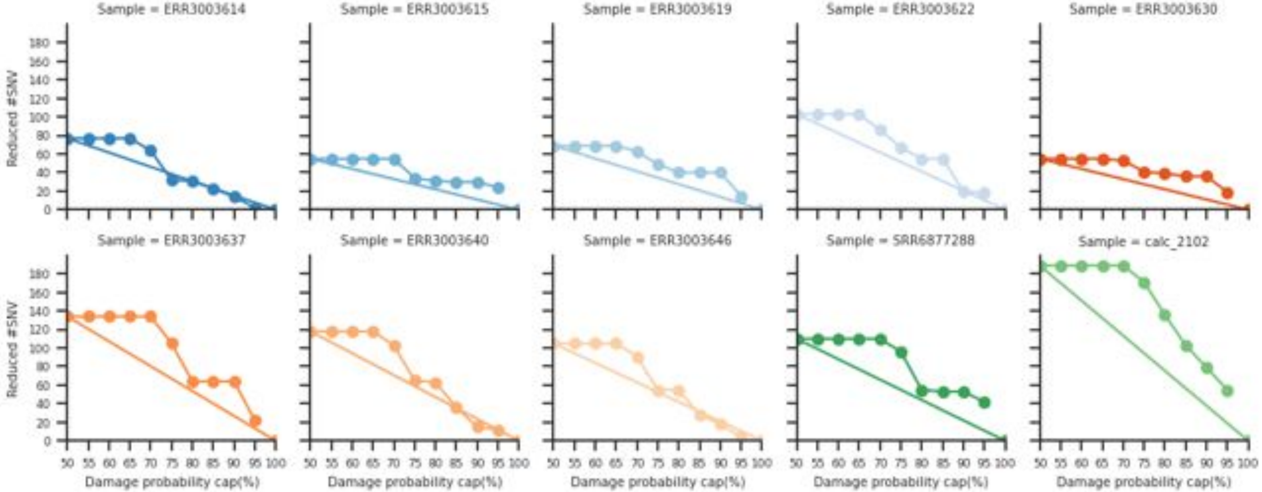
Reads of samples show various extents of damage pattern



Alignment quality = 30
Minimum coverage = 3X
Minimum allele dominance = 80%
Maximum SNP edit distance = 0.03



Damage probability cap: a maximum threshold of position-specific probability of observing C->T (or G->A) due to post-mortem damage for a nucleotide base on the reads to be considered in building consensus allele.



The effect of ancient DNA damage is observed on both reconstructed genome size and #SNV

Conclusion

1. Removing bases with high damage probability results in an extended reconstructed genome sequence.
2. With lowering damage probability cap for building consensus alleles, a clear reduction of number of single nucleotide variation (in comparison with RefSeq selected) was observed.
3. The effect varies according to the damage extent of ancient DNA.

Precaution

Damaged sites of ancient DNA should certainly be aware of when building consensus alleles for analyses which are sensitive to SNV-calling accuracy, such as demography inference, molecular dating and strain-level phylogeny.

GitHub https://github.com/SegataLab/cmseq.git

```
consensus_aDNA.py --mincov 5 -r reference.fna --pos_specific_prob_tab\
 Stats_out_MCMC_correct_prob.csv --pos_damage_prob_thrsh 0.95 mybam.sorted.bam
```

Questions?

Discussion Points

- Genotype callers: which tools are preferred? why?
- SNP filtering techniques, people's experiences. Benefits, drawbacks?
  - removal of specific classes of SNPs (singletons, homoplasies etc.)
  - manual inspection and curation
  - evaluating regions around SNP
- How to deal with:
  - low coverage data
  - aDNA damage, non-UDG treated data
  - multiple strains of same microbial species/subspecies
- Methods for removing and/or preventing contaminant/mis-mapping reads from being included in alignment? Evaluate extent of mis-mapping reads and effects on variant calling ?
- Is there a need for standardization and guidelines?