Date : 21st September 2020
Chair: James Fellows Yates

SPAAM

Standards,
Precautions &
Advances in
Ancient
Metagenomics

Session 3: Recycling the Trash (Part 1)

Authentication, Standards, and Reproducibility, in Ancient Metagenomics

Session Scope

- What criteria should be part of a minimal authentication line of evidence?
  - Defining **recommended minimal** experimental and computational **criteria** (from Session 1, from Session 2 and more...)
  - Providing **guidelines for how to use authentication criteria** in publications

- Icebreaker speakers:
  Sterling Wright (Penn State University, USA)
  Nikolay Oskolkov (Lund University, Sweden)

- Tweeting ALLOWED/NOT ALLOWED

Definitions

- Authentication

Providing evidence or a line of evidence supporting the validity of:

- the identification of taxa or strains,
- reported sequences or genotype calls,
- the reconstruction of community compositions

SPAAM

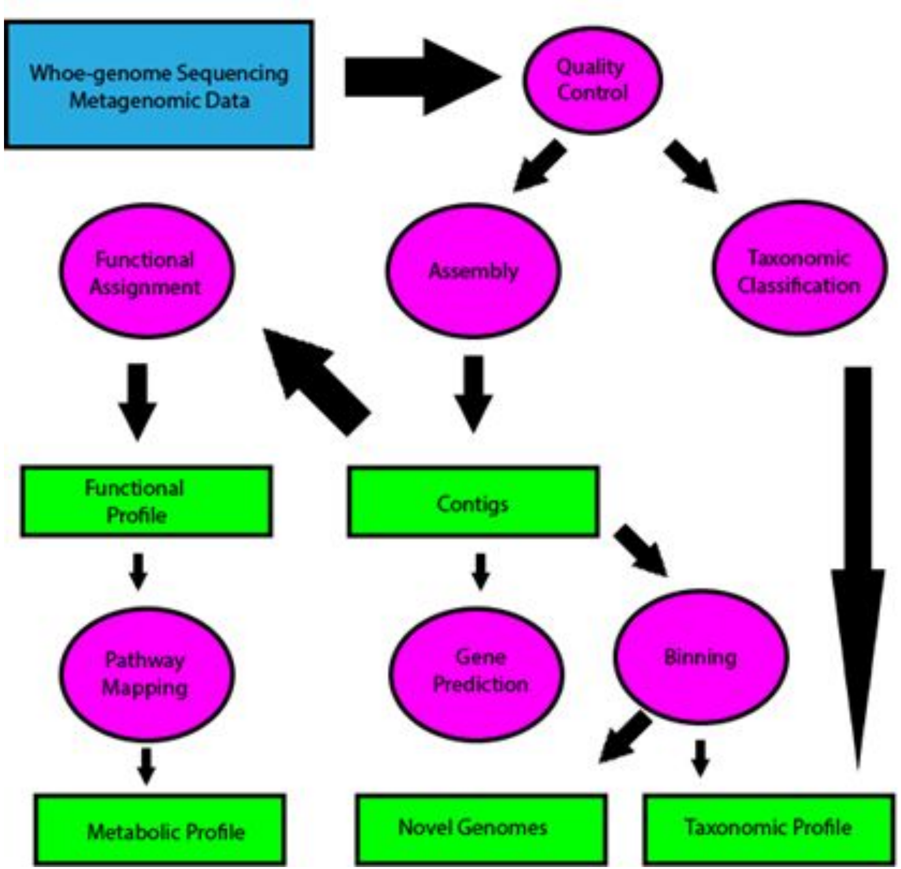
Standards,
Precautions &
Advances in
Ancient
Metagenomics

Improving current and future practices in ancient microbiome analyses: reference-based authentication

Sterling

Taxonomic Classification

- **Taxonomic Classification**: accurately assign query sequences to their respective group in the reference taxonomy
  - MALT, MetaPhlAn2, MIDAS, CLARK-S, and QIIME/UCLUST
  - Each program has tradeoffs that are context dependent (Velsko et al. 2018, Msystems)



Databases

- Too big to fit in many cases (e.g. NT with MALT ← aDNA optimised blast)

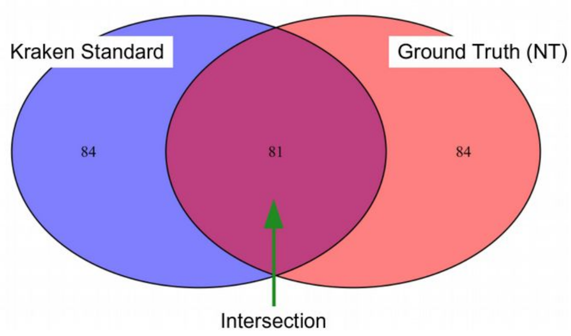
  Different approaches: can we use stepwise/iterative analyses? Representive analyses (e.g. SPARSE/GTDB)? Chunking databases (e.g. minimap2)

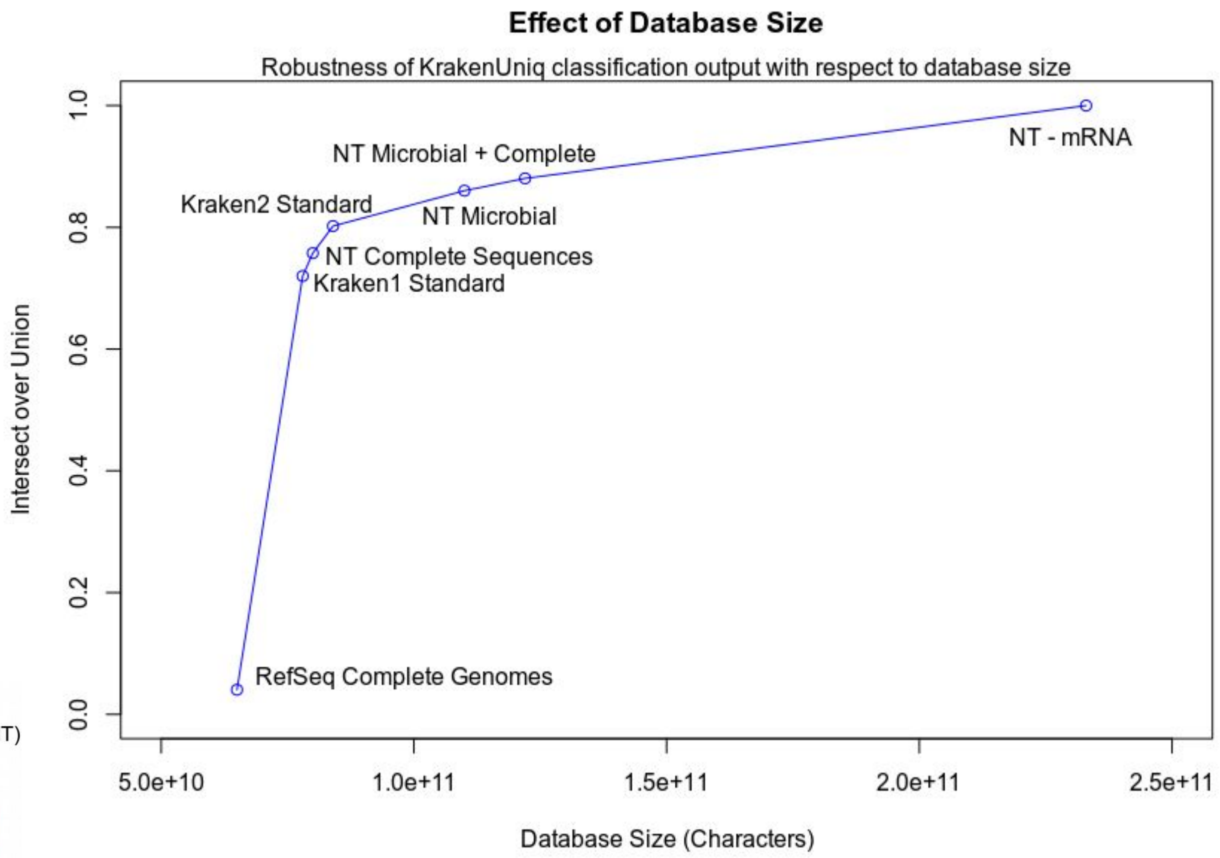
  Project idea: what are the different sizes of different database sizes of taxonomic profiler e.g. compare kraken vs MALT vs SPARSE (etc.) at different database sizes refseq, GTDB, NT etc.

- Databases construction often not reported sufficiently, or not reproducible
  - E.g. RefSeq Bacteria genomes on X day [no track of historical dates]

NT Microbial:
archaea, bacteria, fungi, parasitic worms, protozoa, viruses

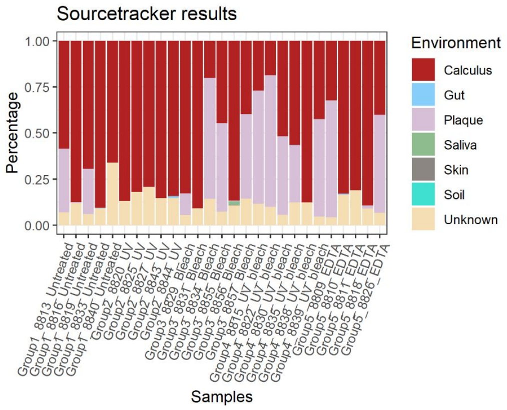
Complete:
Human, invertebrate, plants, vertebrate, mammalian, vertebrate other

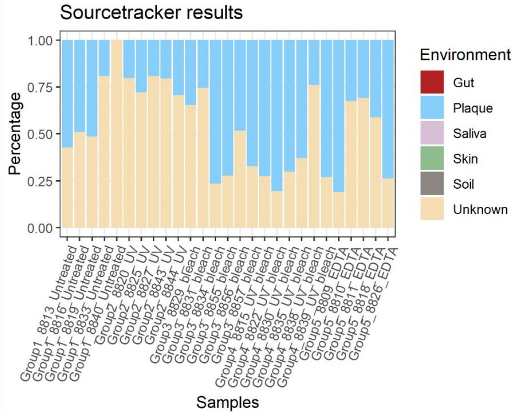




SourceTracker: A widely-used authentication tool

- A Bayesian source-prediction tool to estimate the proportion of contamination
- Requires DNA sequences from ‘sources’ (e.g. modern dental calculus, oral plaque, soil, and skin)
- Many studies include SourceTracker but use different datasets and for different purposes
  - Hagan et al. (2020)
  - Ottoni et al. (2019)
  - Mann et al. (2018)





Contamination Estimation datasets?

- Standardised sequencing data/collections of samples for use to estimation contamination
- Considerations:
    - Should a standard soil dataset be used, or should researchers aim to collect soil samples surrounding a burial?
    - Should we use extraction blanks as a source in the analysis or use them to filter out sequences?
    - Should we make custom datasets or standardize a benchmarked dataset?
    - Should we have a standardized screening tool?

Short Questions (2 mins!)

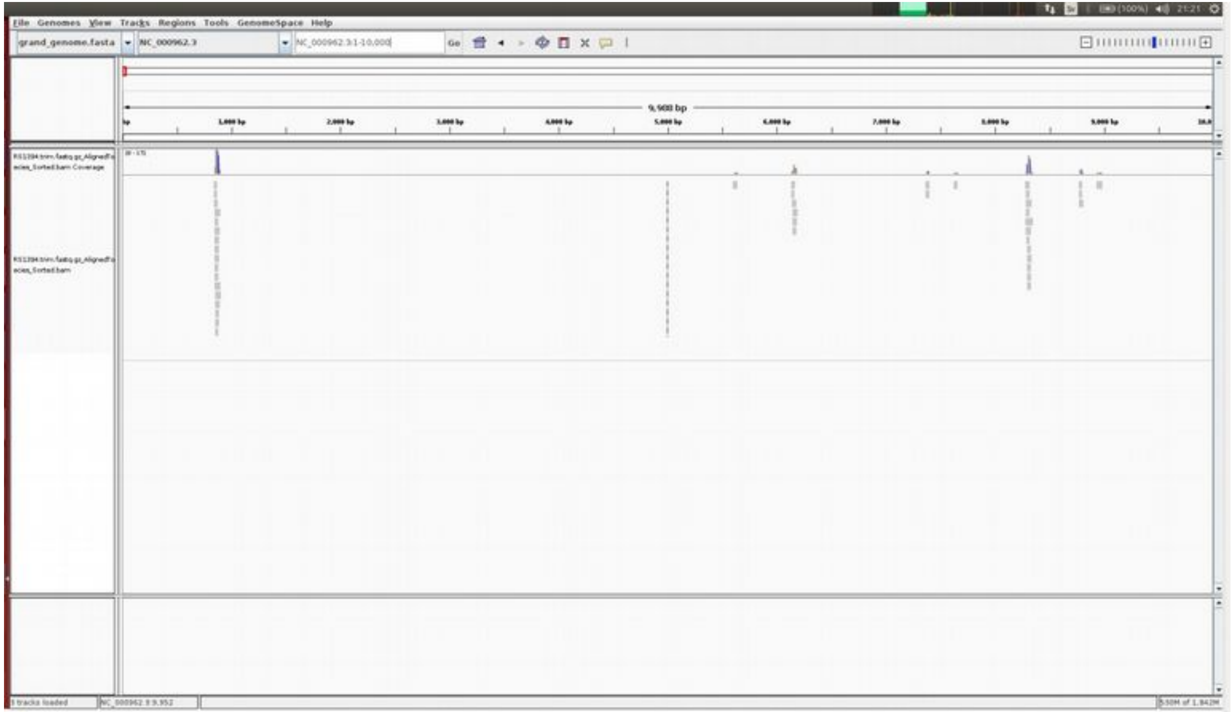
SPAAM

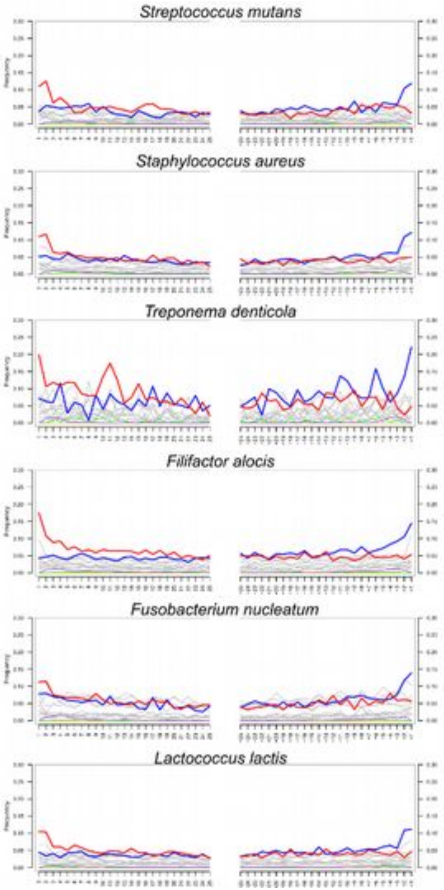
Standards,
Precautions &
Advances in
Ancient
Metagenomics

Improving authentication tools for ancient microbiome analysis

Nikolay Oskolkov

Ancient Microbes (Especially Viruses)

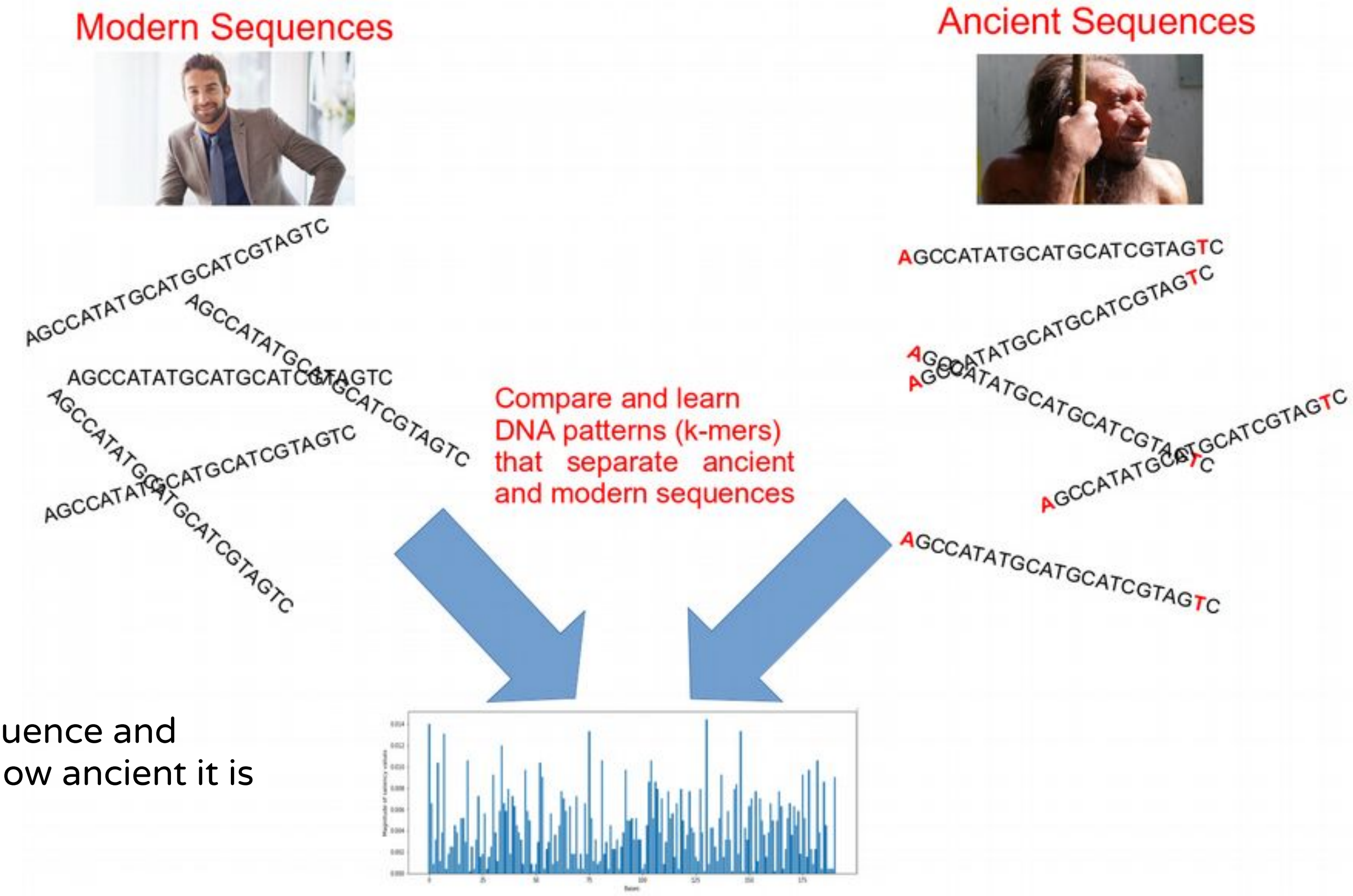

File Genomes View Tracks Regions Tools GenomeSpace Help
grand_genome.fasta
NC_000962.3
NC_000962.3:1-10,000
Go
9,900 bp
1,000 bp
2,000 bp
3,000 bp
4,000 bp
5,000 bp
6,000 bp
7,000 bp
8,000 bp
9,000 bp
RS1394.trim.fastq.gz_AlignedReads_Sorted.bam Coverage
RS1394.trim.fastq.gz_AlignedReads_Sorted.bam
3 tracks loaded



- Too low coverage to infer deamination pattern
- Needs reference genome and alignment step

Herbig et al.,
https://www.biorxiv.org/content/10.1101/050559v1

How can Machine Learning help to solve the problem?



Modern Sequences

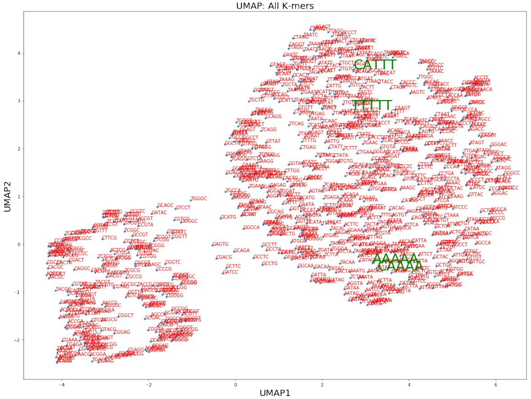
Ancient Sequences

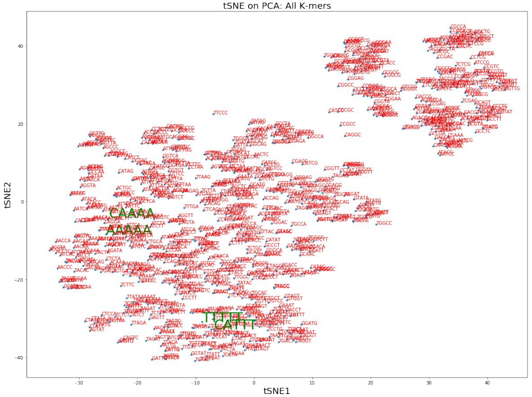
Compare and learn DNA patterns (k-mers) that separate ancient and modern sequences

Give me a sequence and
I will tell you how ancient it is

AI for Ancient Genomics: Natural Language Processing (NLP)

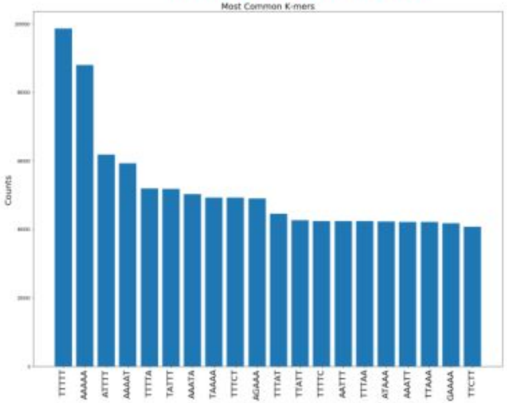
DNA is a Text

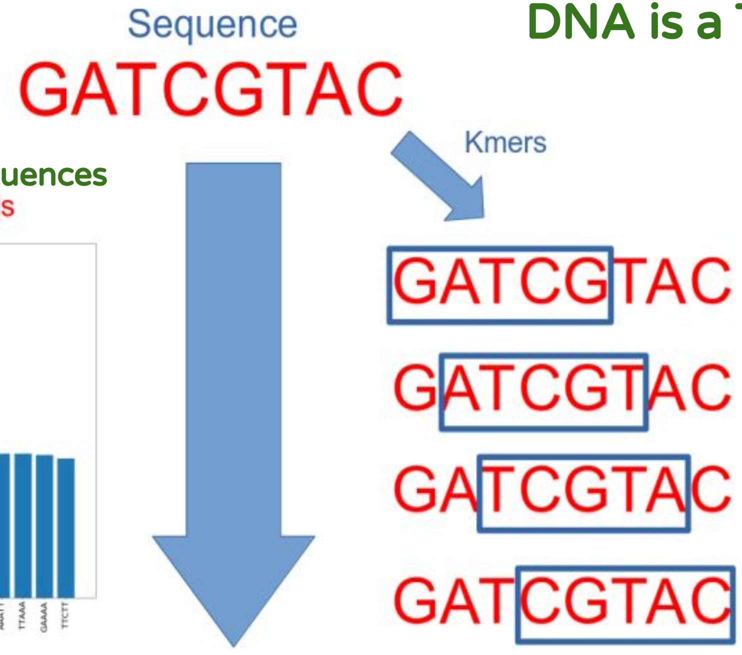

UMAP: All K-mers


tSNE on PCA: All K-mers

Ancient vs. Modern Sequences

NLP: Bag of Words


Most Common K-mers


Sequence

GATCGTAC

Kmers

GATCGTAC

GATCGTAC

GATCGTAC

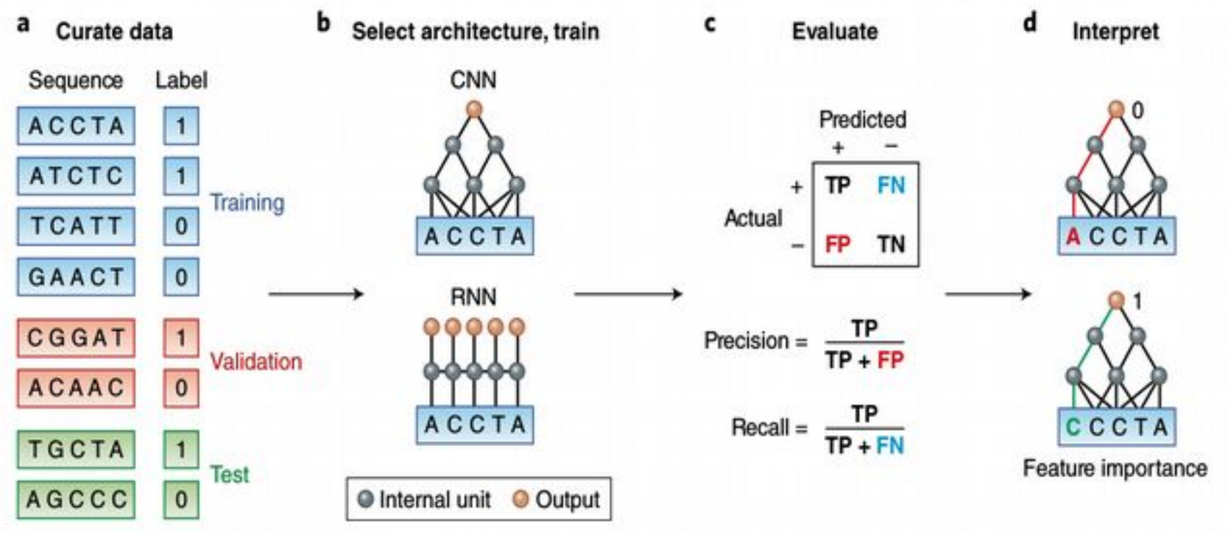
GATCGTAC

Sentence / Text

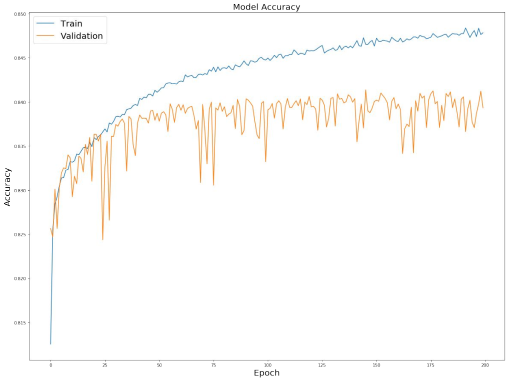
GATCG ATCGT TCGTA CGTAC

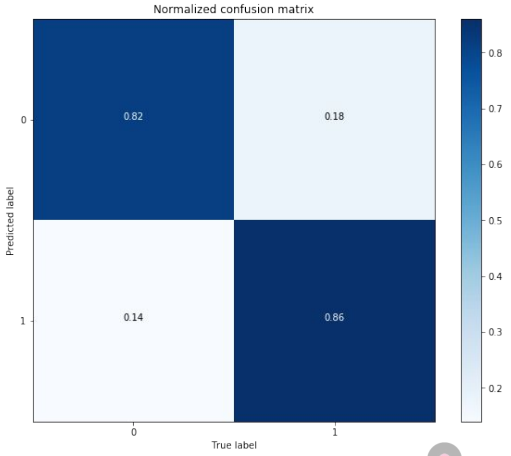
word word word word

AI for Ancient Genomics: Neural Network



Zou et al. *Nature Genetics* **51**, p. 12–18 (2019)

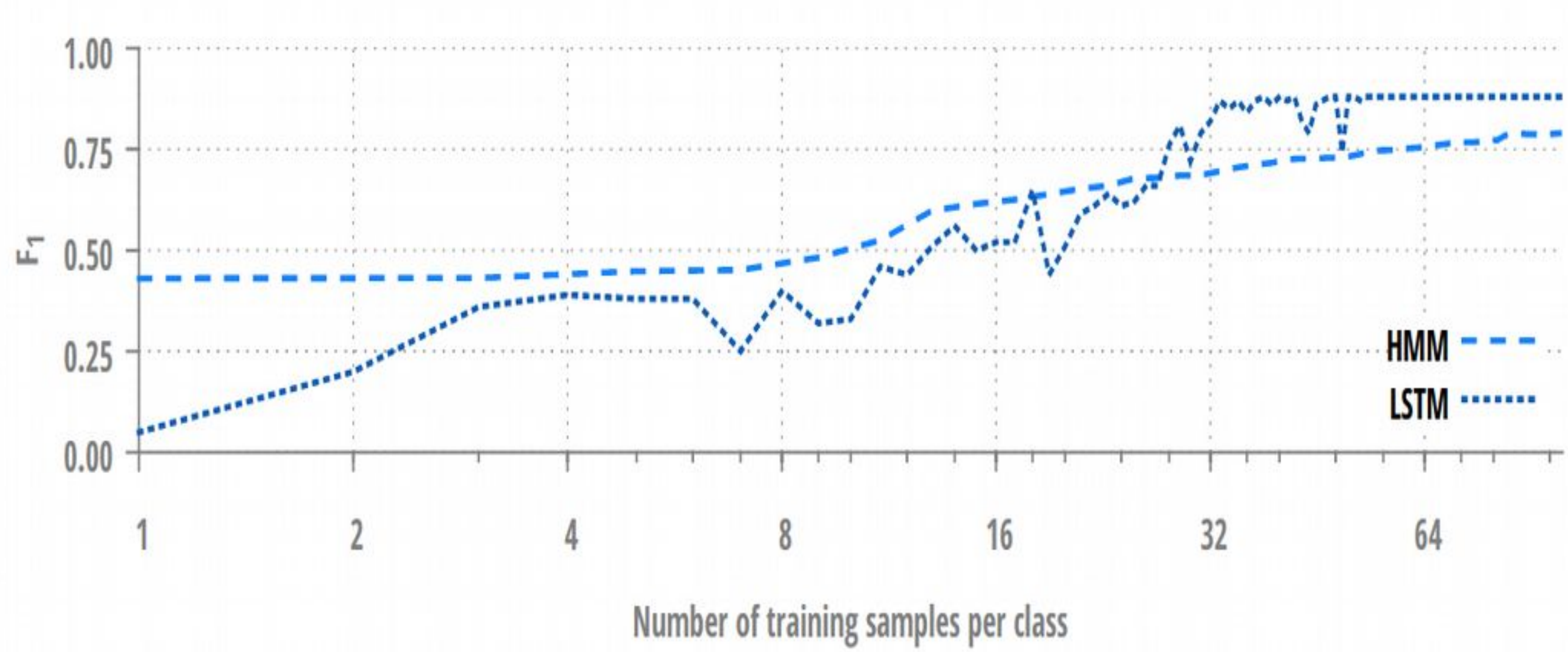




- Authenticity inference for each individual sequence
- Reference genome free approach (no alignment needed)

84% accuracy

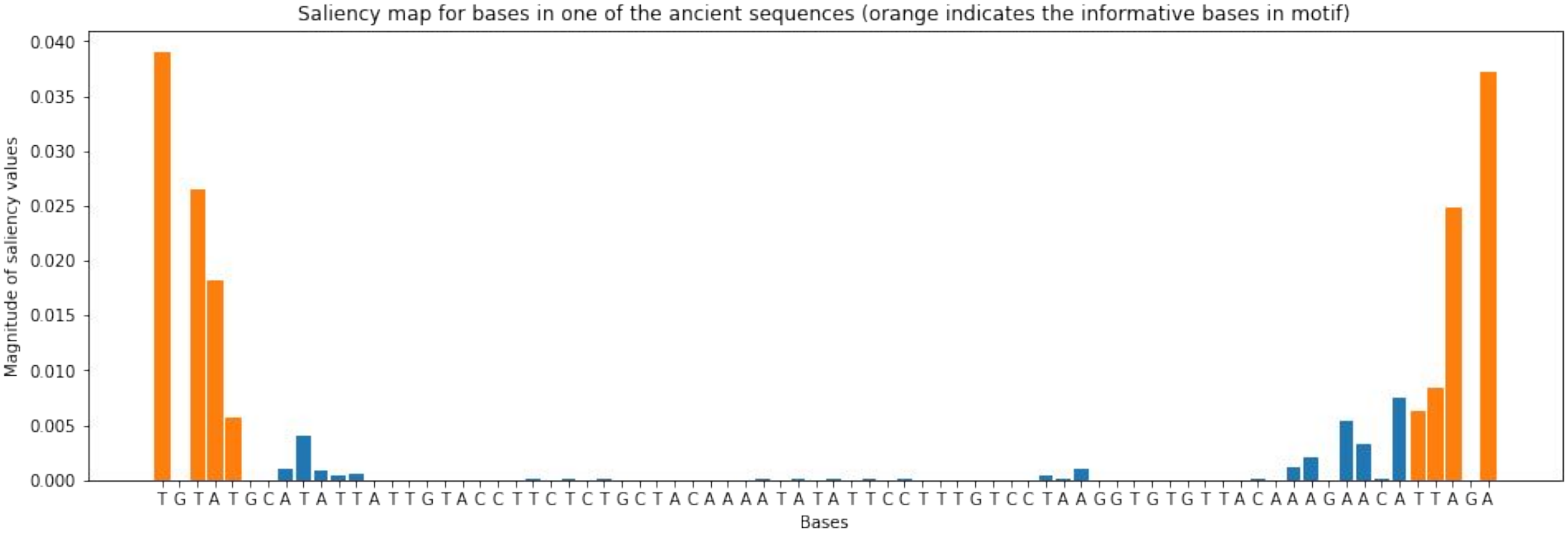
Why would you want to analyze aDNA with Neural Networks?



Panzner and Cimiano, Machine Learning, Optimization, and Big Data, 2016

- CNNs and LSTMs can keep DNA context information (long-memory algorithms)
- We know that there is a correlation (Linkage Disequilibrium) along DNA sequence

Interpretation: where does the signal come from?



- Convolutional Neural Network (CNN) takes the whole sequence as input
- It learns K-mer composition of ancient and modern reads and uses this “vocabulary” for making prediction of authenticity

https://github.com/NikolayOskolkov/DeepLearningAncientDNA

Short Questions (2 mins!)

Discussion

- Can and should we standardise comparative datasets for reference-based authentication?
- What other characteristics of ancient metagenomics could we use for reference-free methods?
- Can we define minimal authentication criteria that should always be reported?
  - Other methods? Tools?
- How to disseminate inside or outside these criteria?

Pre-sent Questions

- None 😭