Benchmarking metagenomics classifiers on ancient viral DNA: a simulation study

BACKGROUND

Thanks to technological advances it is now **possible to retrieve and sequence DNA from ancient samples.** This genomic data includes DNA from ancient microbes.

AIM

Several **classifiers have been developed** to characterise the microbiota

Which classifier is best suited to screen ancient samples for ancient viruses?

Tested classifiers: Centrifuge[1], Kraken2[2], DIAMOND[3] and MetaPhlAn2[4]

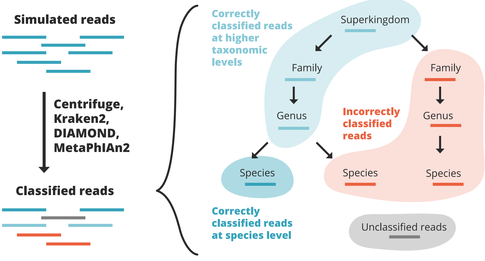
METHODS

1. Select reference genomes

2. Simulate sequencing reads

- Initial simulation: 60 bp reads
- Varying read length
- Adding deamination damage
- Adding sequencing errors

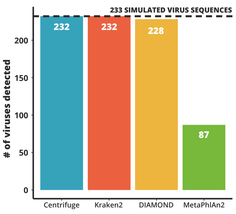
3. Read classification



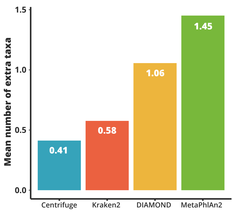
RESULTS

Compute precision and sensitivity
For 60 bp initial simulations sets:

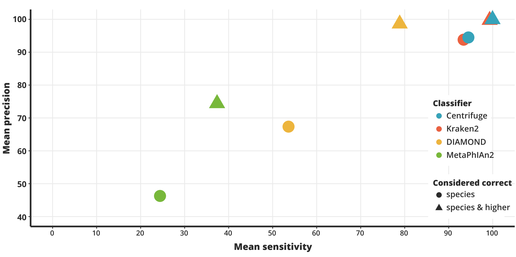
A Total number of correctly detected viruses



B Mean number of spurious extra taxa



C Sensitivity vs. precision



CONCLUSIONS

- **Centrifuge** has:
  - The **highest sensitivity and precision** values
  - The best performance **with short (30 bp) reads**
- **Most of the viruses were detected** by almost all the classifiers tested.
- **Longer reads are better classified**, while increased sequencing error and increased deamination damage worsen the classification slightly.
- **We recommend:**
  - using **strong filters** to remove human DNA
  - verifying that the **genomes of interest are included** in the classifiers' databases

1. Centrifuge: Kim D, Song L, Breitwieser FP, Salzberg SL. 2016. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Research 26:1721 1729 DOI 10.1101/gr.210641.116.
2. Kraken2: Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. Genome Biology 20:257 DOI 10.1186/s13059-019-1891-0.
3. DIAMOND: Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nature Methods 12:59 60 DOI 10.1038/nmeth.3176.
4. MetaPhlAn2: Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods 12:902 903 DOI 10.1038/nmeth.3589.
5. ART: Huang W, Li L, Myers JR, Marth GT. 2012. ART: a next-generation sequencing read simulator. Bioinformatics 28:593 594 DOI 10.1093/bioinformatics/btr708.
6. Gargammel: Renaud G, Hangh j K, Willerslev E, Orlando L. 2017. gargammel: a sequence simulator for ancient DNA. Bioinformatics 33:577 579.

PeerJ
Benchmarking metagenomics classifiers on ancient viral DNA: a simulation study
PeerJ 10:e12784 DOI: 10.7717/peerj.12784
http://peerj.com/articles/12784