Building cyberinfrastructure systems to support integrative, macroscale analyses of sedimentary ancient DNA records: current resources, needs, and opportunities

John Williams\(^1\) and the Cyberinfrastructure for Ancient Sedimentary DNA Working Group\(^*\)

\(^1\)University of Wisconsin-Madison, Geography, United States of America (jwwilliams1@wisc.edu)

\(^*\)A full list of authors appears at the end of the abstract

The number and extent of ancient DNA records from sedimentary environments (sedaDNA) is rapidly increasing, which creates new opportunities for integrative and macroscale investigations into past population, community, and environmental dynamics at unprecedented taxonomic resolution and spatiotemporal extent. However, fully achieving this potential requires a robust cyberinfrastructure that supports the joint analysis of many sedaDNA records with each other and with genomic reference libraries, the latest geochronological controls and age-depth models, complementary paleoecological and paleoenvironmental proxies, and the most recent and updated DNA reference library for taxonomic identifications. Any cyberinfrastructure for macroscale data synthesis must address the variety of ancient DNA records (e.g. taxonomic groups, analytical approaches, depositional contexts) and leverage existing resources and standards such as the Neotoma Paleoecology Database, the MGnify and MG-RAST resources for environmental genomics, and the MixS standard for genetic sequences. In response, a Cyberinfrastructure for Ancient Sedimentary DNA working group has been meeting regularly since summer 2020 to assess the current state of science and informatics, assess needs and gaps, and establish recommendations for next steps forward. An initial survey found over 420 sites worldwide with published or in-development sedaDNA records, with greatest densities in Eurasia. Metabarcoding records, including Amplicon Sequence Variant data and derived taxonomic inferences, are a top priority for trial uploads to Neotoma, with pilot uploads underway, because of the relatively small dataset volumes, the widespread application of metabarcoding assays, and potential of integrating these records with other paleoecological data holdings in Neotoma and linked paleodata resources such as Linked Earth and paleoclimatic data at NOAA's National Centers for Environmental Informatics. Because taxonomic inferences are heavily conditioned by choice of bioinformatics pipeline and reference databases, a major unmet need is a repository for minimally processed output from raw sequences. In general, no existing genomics or paleoecological resource meets all needs of the sedaDNA community, although each covers key elements, so there is a good potential of advancing macroscale data syntheses by leveraging and linking existing resources.

Cyberinfrastructure for Ancient Sedimentary DNA Working Group: Inger G. Alsos, Jessica L.