Trends in Ecology and Evolution

Strengthening global-change science by integrating aeDNA with paleoecoinformatics

--Manuscript Draft--

Manuscript Number:	TREE-D-22-00310R2
Article Type:	Opinion
Keywords:	Ancient environmental DNA; community-curated data resources; global change ecology; metabarcoding; paleoecoinformatics
Corresponding Author:	John Warren Williams, PhD University of Wisconsin-Madison Madison, Wisconsin UNITED STATES
First Author:	John Warren Williams, PhD
Order of Authors:	John Warren Williams, PhD
	Trisha Spanbauer
	Peter Heintzman
	Jessica Blois
	Eric Capo
	Simon Goring
	Marie-Eve Monchamp
	Laura Parducci
	Jordan Von Eggers
	Inger Alsos
	Chris Bowler
	Marco Coolen
	Nicola Cullen
	Sarah Crump
	Laura Epp
	Antonio Fernandez-Guerra
	Eric Grimm
	Ulrike Herzschuh
	Alessandro Mereghetti
	Rachel Meyer
	Kevin Nota, PhD
	Mikkel Pedersen
	Vilma Perez
	Beth Shapiro
	Kathleen Stoof-Lechsenring
	Jamie Wood
Abstract:	Ancient environmental DNA (aeDNA) data are close to enabling insights into past global-scale biodiversity dynamics at unprecedented taxonomic extent and resolution. However, achieving this potential requires solutions that bridge bioinformatics and paleoecoinformatics. Essential needs include support for dynamic taxonomic inferences, dynamic age inferences, and precise stratigraphic depth. Moreover, aeDNA

	data are complex and heterogeneous, generated by dispersed researcher networks, with methods advancing rapidly. Hence, expert community governance and curation are essential to building high-value data resources. Immediate recommendations include uploading metabarcoding-based taxonomic inventories into paleoecoinformatic resources, building linkages among open bioinformatic and paleoecoinformatic data resources, harmonizing aeDNA processing workflows, and expanding community data governance. These advances will enable transformative insights into global-scale biodiversity dynamics during large environmental and anthropogenic changes.

Strengthening global-change science by integrating aeDNA with paleoecoinformatics

Author Names

Williams, John W., Department of Geography, University of Wisconsin-Madison, Madison, WI 53704 USA

corresponding author: jwwilliams1@wisc.edu

Spanbauer, Trisha L., Department of Environmental Science and Lake Erie Center, University of Toledo, Toledo, Ohio 43606 USA

Heintzman, Peter D., The Arctic University Museum of Norway, UiT The Arctic University of Norway, Tromsø, Norway / Centre for Palaeogenetics, Svante Arrhenius väg 20C, SE-10691 Stockholm, Sweden / Department of Geological Sciences, Stockholm University, SE-10691, Stockholm, Sweden

Blois, Jessica, Department of Life and Environmental Sciences, University of California -Merced, Merced, CA 95343 USA

Capo Eric, Department of Ecology and Environmental Science, Umeå University, Linnaeus väg 4-6, 907 36 Umeå, Sweden

Goring, Simon, Department of Geography, University of Wisconsin-Madison, Madison, WI 53704 USA

Monchamp, Marie-Eve, Department of Biology, McGill University, Montreal, Qc, Canada

Parducci, Laura, Department of Environmental Biology, Sapienza University of Rome, Piazzale Aldo Moro 5, 00185, Rome, Italy / Department of Ecology and Genetics, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 75236 Uppsala, Sweden

Von Eggers, Jordan M., Department of Geology and Geophysics, University of Wyoming, Laramie, WY 82071, USA

Supporting Authors

Alsos, Inger Greve, The Arctic University Museum of Norway, UiT The Arctic University of Norway, Tromsø, Norway

Bowler, Chris, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Ecole Normale Supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France

Coolen, Marco J.L., WA Organic and Isotope Geochemistry Centre (WA-OIGC), The Institute of Geoscience Research (TIGeR), School of Earth and Planetary Sciences (EPS), Curtin University, Bentley, WA 6102, Australia.

Cullen, Nicola, Department of Life and Environmental Sciences, University of California -Merced, Merced, CA 95343 USA

Crump, Sarah, Department of Geology & Geophysics, University of Utah, Salt Lake City, UT, 84112 (deceased)

Epp, Laura Saskia, Limnological Institute, Department of Biology, University of Konstanz. 78464 Konstanz, Germany

Fernandez-Guerra, Antonio, Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Oester Voldgade 5-7, DK-1350 Copenhagen, Denmark

Grimm, Eric, Department of Earth Sciences, University of Minnesota, Minneapolis, MN 55455 USA (deceased)

Herzschuh, Ulrike, Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Polar Terrestrial Environmental Systems, Potsdam, 14473, Germany

Mereghetti, Alessandro, University of Maine, Climate Change Institute and School of Biology and Ecology, Orono, Maine, United States

Meyer, Rachel Sarah, Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA 95064, USA

Nota, Kevin, Max Planck Institute for Evolutionary Anthropology, Leipzig 04103 Germany

Pedersen, Mikkel Winther, Section for GeoGenetics Globe institute. University of Copenhagen. Denmark

Pérez, Vilma, Australian Centre for Ancient DNA (ACAD)and ARC Centre of Excellence for Australian Biodiversity and Heritage (CABAH), School of Biological Sciences, University of Adelaide, Adelaide, SA 5005, Australia

Shapiro, Beth, Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA 95064 *and* Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA 95064 USA

Stoof-Leichsenring, Kathleen R., Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Polar Terrestrial Environmental Systems, Potsdam, 14473, Germany

Wood, Jamie, The Environment Institute and School of Biological Sciences and the Australian Centre for Ancient DNA, University of Adelaide, Adelaide, South Australia 5005, Australia

Key Words

Ancient environmental DNA; community-curated data resources; global change ecology; metabarcoding; paleoecoinformatics

Abstract

Ancient environmental DNA (aeDNA) data are close to enabling insights into past global-scale biodiversity dynamics at unprecedented taxonomic extent and resolution. However, achieving this potential requires solutions that bridge bioinformatics and paleoecoinformatics. Essential needs include support for dynamic taxonomic inferences, dynamic age inferences, and precise stratigraphic depth. Moreover, aeDNA data are complex and heterogeneous, generated by dispersed researcher networks, with methods advancing rapidly. Hence, expert community governance and curation are essential to building high-value data resources. Immediate recommendations include uploading metabarcoding-based taxonomic inventories into paleoecoinformatic resources, building linkages among open bioinformatic and paleoecoinformatic data resources, harmonizing aeDNA processing workflows, and expanding community data governance. These advances will enable transformative insights into global-scale biodiversity dynamics during large environmental and anthropogenic changes.

Achieving aeDNA capacity for global biodiversity research

The fast-growing field of **ancient environmental DNA** (**aeDNA**; see Glossary) from sedimentary **archives** is transforming the study of past biodiversity dynamics [1–3]. aeDNA data provides information about the distribution and diversity of species (and whole taxonomic groups) that were previously invisible in the fossil record [4]. Examples of new insights powered by aeDNA include the demonstrated persistence of taxa in formerly cryptic refugia [5–7], refined timing of arrival and extinction events [8–10], better understanding of precursors to extinction [1,11], and the responses of ecosystems to anthropogenic perturbations and high-frequency environmental variability [12,13].

However, aeDNA so far has been at the alpha stage of discovery, with primary emphasis on generating new records from a few localities at a time and advancing laboratory and data processing methods. Now, as the number of sites grows worldwide (Fig. 1), aeDNA research is at the cusp of supporting analyses of the distribution and diversity of life over broad spatial and temporal scales across terrestrial, aquatic, and marine habitats (e.g., [1,3,11,14]). The next step is to better integrate these aeDNA records with each other, other paleoecological and paleoenvironmental **proxies**, and contemporary genomic resources (Fig. 2). This integration will enable multi-proxy, multi-scale, and reproducible analyses into past ecological, evolutionary, and environmental change (Fig. 3).

Prior syntheses of networks of ‘classical’ paleoecological proxies have transformed our understanding of global-scale processes. Examples include past rates of vegetation change driven by climate and anthropogenic processes [15–17], megafaunal extinction and biodiversity loss [18,19], and the emergence of novel communities [20,21]. This work has demonstrated that community curation and expert governance are essential for robust macro-scale paleobiological research, because of the complex processes that produce the fossil record and the resultant risk of erroneous scientific inference [22]. Thus, cutting-edge macro-scale paleoecological research now relies on **community-curated data resources (CCDRs),** supported by a shared cyberinfrastructure, and governed by experts [23].

Sedimentary aeDNA occupies the intersection between biology and geology. Hence, its data infrastructure must also leverage and connect existing elements in bioinformatics and geoinformatics (Fig. 2), while supporting needs unique to aeDNA (Fig. 4). Taxonomic inferences based on aeDNA must be regularly updated against the latest genetic **reference databases**, while precise age inferences and integration with other proxies requires close links of aeDNA to other **paleoecoinformatics** data resources and services [24]. aeDNA methods are developing rapidly, so any system for the archival and macro-scale analysis of aeDNA data must be dynamic and flexible.

Here, we first describe the scientific opportunities enabled by global-scale aeDNA networks and review the paleoecoinformatics ecosystem. We then review the characteristics and informatics needs of aeDNA data and recommend solutions for meeting these needs. These recommendations represent the collective perspective of an emerging community of aeDNA researchers, data scientists, and paleoecologists, and, if enacted, will enable the next generation of cutting-edge global-scale research into past biodiversity dynamics jointly powered by the latest advances in aeDNA methods, a rapidly growing worldwide network of sites, and a shared community data architecture.

Scientific Rationale for Global, Interdisciplinary, and Integrative aeDNA Data Systems

Multiple scientific advantages accrue from integrating aeDNA into the established cyberinfrastructure for paleoecoinformatics and bioinformatics (Fig. 3). First, aeDNA, as a newer proxy, needs cross-checking against independent paleoecological proxies (Box 1). All paleoecological proxies recovered from sedimentary **archives**, including macrofossils, microfossils, biogeochemical tracers, and aeDNA, are produced by some mixture of ecological and post-depositional processes [25]. Organismal differences in preservability and transportability will cause each proxy to carry some form of taphonomic bias that causes the after-death assemblage to differ from the source communities. Prior comparisons of aeDNA inventories to other proxies (e.g., as plant pollen and macrofossils [26–28], diatom remains [29–31], or micro-algal pigments [32]) demonstrate that concomitant temporal shifts are often observed in aeDNA and other proxies [33], despite differences in detectability, apparent abundance, and sensitivity to sedimentary context.

Second, analyzing aeDNA with other proxies can reveal multiple dimensions of past environmental change and multiple levels of ecological response. For instance, long-term effects of lake eutrophication on species turnover at multiple trophic levels were revealed by combining aeDNA inventories with invertebrate remains and algal pigments [34]. In the Black Sea, salinity-driven changes in plankton communities were inferred from parallel analysis of aeDNA and hydrogen isotopes from algal biomarkers [35]. Pollen and aeDNA from the High Arctic show that the Last Interglacial period resulted in high latitude greening and northward plant range shifts over hundreds of kilometers [36]. These site-level studies show the power of multi-proxy investigations into past environmental and ecosystem change; similar capacity is needed globally.

Third, assembling aeDNA records across many sites is essential to achieving aeDNA's promise for new global-scale insights into biodiversity dynamics. Macro-scale syntheses, which integrate many kinds of data from many sites and times, are transforming our understanding of species and communities response to environmental change across scales [37,38], and are necessary to identify teleconnections, biosphere-atmosphere interactions, and other emergent phenomena. A global infrastructure for aeDNA data can help identify spatiotemporal gaps in coverage and priority areas for future research. Other 'classic' paleoecological proxies show the power of building global-scale networks of sites. Global networks of fossil pollen records have been used to assess the sensitivity of terrestrial ecosystems to global warming [16] and identify periods of rapid change [15], some of which can be attributed to human arrival [17]. Continental- to global-scale syntheses of terrestrial vertebrates are a foundation for modeling drivers of extinction [39,40] and the functional relationship to ecological traits [19]. Via syntheses of archaeological, paleoecological, and paleoclimatic data, the worldwide impact of humans on the Earth System can be detected [41,42]. Hence, the building of a global aeDNA data system can build upon lessons learned and paleoecoinformatic resources developed for these other global investigations of past biodiversity dynamics.

Fourth, integrating advances and linking resources across paleoecoinformatics and bioinformatics will help advance the harmonization of associated bioinformatic workflows and other resources, thereby helping establish best practices. Best practices now exist for sampling and laboratory protocols designed to minimize and monitor for contamination by exogenous DNA [33], but bioinformatic and data analysis standards are not yet broadly established (e.g., for raw sequence pre-processing thresholds, taxonomic assignment, contaminant removal, and downstream ecological inferences).

The Paleoecoinformatics Ecosystem: Current Resources and Recent Developments

The contemporary paleoecoinformatics ecosystem comprises a coalition of CCDRs that are loosely but increasingly interconnected, each of which supports and is supported by communities of researchers. The scientific origins of these resources can be traced to early campaigns to gather networks of proxy sites at continental to global scales to study past evolutionary, ecological, and climate dynamics [43–45]. Both the emergent structure of the paleoecoinformatics ecosystem and its deep history result from the nature of fossil and paleoenvironmental proxy data. On the one hand, these are classic ‘long tail’ data, in which millions of data points across tens of thousands of individual field sites are collected by thousands of scientists dispersed globally [46]. Knowledge is also dispersed, as each scientist is expert in particular taxonomic groups or proxies. On the other hand, paleoecological data, once collected, have long-lasting value, because they represent a unique measurement of the state of the Earth-life system at some particular spatiotemporal locus, and each measurement accrues value as it is joined to an ever-expanding network of other measurements. Hence, paleoecological data are ‘small data’ at point of collection but ‘big data’ in aggregate. Gathering and using these data effectively requires close partnerships between proxy specialists and data scientists [23].

Several major paleoecological and paleoenvironmental data resources have emerged, including the Neotoma Paleoecology Database [47], Paleobiology Database (PBDB) [48], Neptune Sandbox Berlin [49], NOAA’s National Center for Environmental Information (NCEI-Paleoclimatology) [50], the Linked Paleodata standard (LiPD and LiPDverse) [51], and PANGAEA [52]. Each resource differs in its focus, curation model, data types, and spatiotemporal domains. Some, such as PANGAEA or Dryad (https://datadryad.org/), are general-purpose repositories. Others are tailored for macro-scale paleoecological analysis, with domain-specific metadata. For example, Neotoma focuses on the Late Neogene and stores paleoecological time series, associated geochronological and paleoenvironmental data, and **surface sample** datasets for calibration. Conversely, PBDB focuses on evolutionary dynamics over the last 500 million years and stores taxonomic names and synonyms, spatiotemporal coordinates, and can handle tectonic-driven locational shifts [53]. The Linked Paleo Data (LiPD) standard is a popular data exchange format for paleoclimatic data with crowd-sourced data standards [54]. These data systems combine backend databases with a software stack and user interfaces for finding, obtaining, viewing, and analyzing data. These efforts also focus on building the scientific communities essential for high-quality global-scale data governance, curation, and analysis. PAGES, an international coordinator for past global-change research, convenes scientists into working groups to tackle macro-scale scientific challenges, while PBDB and Neotoma are led by councils of experts who set science-driven priorities for growth and development. These community data resources, which ensure that paleoecological and paleoenvironmental data are findable, accessible, interoperable, and retrievable (FAIR), remain a central priority for the paleoenvironmental research community [55]. A new and fast-growing priority is to develop ways to support the principles of Collective Benefit, Authority to Control, Responsibility, and Ethics (CARE) [56].

Efforts are underway to interlink data resources. For example, the Earth-Life Consortium built application programming interfaces (APIs) to access data from multiple paleobiological resources [57]. NCEI-Paleoclimatology now makes datasets available in LiPD format, and search engines hosted by NCEI-Paleoclimatology can now retrieve data from PANGAEA and Neotoma. DarwinCore, a data standard for biodiversity data, has been extended to include geochronological data and metadata [58]. As these data resources continue growing and interdigitating, they can support ever-more powerful joint analyses of aeDNA data with other proxies.

Sedimentary aeDNA: Data Characteristics and Informatics Needs

aeDNA data fall into three main categories: (1) non-sequenced **polymerase chain reactions (PCRs**), (2) sequenced PCR amplicons (including **metabarcoding**), and (3) metagenomic data. These three data types are generated using fundamentally different molecular biology techniques after DNA extraction (Fig. 4). Here, we focus on amplicon and metagenomic data since these data result in a **taxonomic inventory**, which can be used for biodiversity and other taxon-level analyses of aeDNA data.

Amplicon data are generated via targeted PCR followed by DNA sequencing. This approach can target a single locus across multiple taxa (metabarcoding) or multiple loci across a more restricted set of taxa (multiplex PCR). Amplicon methods are sensitive, allowing the recovery of minute quantities of DNA **template** (<10 molecules) from highly complex mixtures. However, amplicon methods require relatively long and intact template DNA molecules (often >150 base pairs, bp), whereas most preserved aeDNA molecules may be shorter (<100 bp) and damaged [59]. Analyses of amplicon data can’t differentiate aeDNA from modern DNA, because PCR amplification removes signatures of DNA damage. Metagenomic methods convert an entire pool of aeDNA molecules into a **library** that can either be sequenced directly (**shotgun metagenomics**) or enriched for molecules of interest using **target enrichment**. In this way, metagenomic approaches can recover all lengths of aeDNA molecules and retain signatures of DNA damage, thereby enabling aeDNA authentication [10]. Metagenomic datasets are, however, often dominated by microorganismal DNA. Target enrichment offers a middle ground by allowing the capture of short fragments and retaining DNA damage signatures, while reducing the recovery of off-target molecules.

Initially, aeDNA studies were restricted to just a few sites, but recent technological improvements in aeDNA recovery and the massive reduction in sequencing costs are now resulting in large-scale, multi-site studies that generate both amplicon [3,11] and metagenomic aeDNA data [1,60]. Among aeDNA data types, metabarcoding data currently represent the majority of aeDNA sequence data (e.g., in Fig. 1, 75.8% of inventoried aeDNA datasets are from metabarcoding [61]).

As the number of labs using these methods grows, the need to integrate and harmonize aeDNA data produced by different research groups has intensified. Heterogeneity among aeDNA datasets emerges during data generation (e.g., DNA extraction method used, PCR conditions, sequencing depth) and data processing (e.g. removal of amplification artifacts, duplicated sequences, or sequences with low information content) (Fig. 4). **Reference databases** used to identify recovered sequences also lead to heterogeneity, as these differ in geographic and/or taxonomic completeness (e.g., [62,63]). Community efforts are underway to establish aeDNA metadata standards. The *Standards, Precautions, and Advances in Ancient Metagenomics* (SPAAM) community (https://spaam-community.github.io/) is, for example, developing *Minimum Information for an Ancient DNA Sequence* (MInAS) standards for metagenomic data. Standardized pipelines for processing these data are emerging (e.g., OBITools [64], QIIME2 [65], SqueezeMeta [66]). Despite these efforts, the heterogeneity in methods for aeDNA data production and analysis substantially hinders global-scale integration of aeDNA data.

In order to support biodiversity science that is linked to the best-available information about taxonomic inferences, the cyberinfrastructure ecosystem for aeDNA must be able to store sequence data, the resulting **taxonomic inventory**, and metadata about the reference libraries, workflows, and parameters used to generate the taxonomic inferences. In particular, a common output of metabarcoding aeDNA studies is the **Amplicon Sequence Variant (ASV)** (e.g., [67–69]) or **Operational Taxonomic**

Unit (OTU) table. ASV tables store both the primary genetic sequence and the associated inferred taxonomic name (i.e. the taxonomic inventory, Fig. 4), while OTU tables store genetic sequences aggregated into inferred taxonomic units, with a taxon identifier assigned to one representative sequence per OTU. Because ASV and OTU tables have already gone through some initial processing (Section 4, Fig. 4), they represent an intermediate stage in aeDNA pipelines that is valuable both to experts, who can compare the original genetic sequences to the latest reference databases to update taxonomic inferences, and biodiversity scientists, who can use the taxonomic inventory as the best available information about taxon occurrences. Open repositories for storing raw sequence data and their associated metadata exist [e.g., the EMBL European Nucleotide Archive (ENA), NCBI Sequence Read Archive (SRA), EMBL European Bioinformatics Institute (MGnify)], as do community-curated databases of links to these resources for some aeDNA data (e.g., *AncientMetagenomeDir*; [70]). However, ASV and OTU tables currently have no standard data repository and are scattered across Dryad and other generic repositories with no attempt to, e.g., standardize table structures or vocabularies. Hence, in terms of FAIR standards [55], most ASV and OTU data are Findable and Accessible, but not Interoperable or Reusable. Moreover, existing bioinformatics-oriented repositories do not currently store metadata about depth and temporal position at the detail needed for the precise age-depth modeling that is necessary for multi-proxy and multi-site paleoecological research.

Building a Linked Open Ecosystem for aeDNA-Powered Global Biodiversity Research: Vision and Recommendations

Given the rapid advances in sedimentary aeDNA methods, the growing global network of sites (Fig. 1), and the on-going growth and interdigitation of the paleoecoinformatics ecosystem, all pieces are in place for the next generation of multi-proxy, global-scale research into past biodiversity dynamics, in which new insights from sedimentary aeDNA are richly contextualized by the ever-growing network of paleoecological and paleoenvironmental proxies (see **Outstanding Questions**). Global-scale biodiversity science requires high-quality data about species identifications and occurrences that are precisely positioned spatially and temporally. These needs can be met by building an open linked ecosystem for aeDNA data that bridges across existing open resources in bioinformatics (particularly the repositories for raw sequence data and the bioinformatics pipelines used for taxonomic inferences) and paleoecoinformatics (particularly data resources that support precise depth information, regeneration of age-depth models, and community data governance), and emerging community data standards (e.g. MInAS).

Because the taxonomic inventories available from aeDNA data are essential to biodiversity science and have no standard data home, a first priority should be to develop standardized informatics solutions for the storing and sharing of ASV and OTU tables, sourced from both metabarcoding and metagenomics studies. Because the species identifications associated with aeDNA data are changeable, as reference databases improve, the taxonomic inventories available from aeDNA data must be accessioned in a way that allows direct linked to the primary sequence archives maintained by EMBL and NCBI. Datasets should also include all minimally essential metadata (e.g. MInAS standards). These linkages will enable any given aeDNA-based species inventory to be critically assessed, and, ideally, updatable as reference libraries improve. While both metabarcoding and metagenomics can produce a taxonomic inventory, metabarcoding aeDNA projects are recommended for initial efforts because they are currently the most common form of aeDNA data (Fig. 1).

In this envisioned open and hybrid bioinformatic/geoinformatic ecosystem, the paleoecoinformatic components are employed to store the taxonomic inventories represented by ASV and OTU tables, along with the necessary metadata about stratigraphic deposition and age controls that are needed for the best-available age inferences, as age-depth models and geochronological parameterizations improve. Following best practices developed for other paleobiological data resources (e.g. PBDB, Neotoma), these systems should include mechanisms for expert community data governance, to ensure that data systems are designed to meet the needs of user communities. Within this broad vision, we recommend the following next steps forward:

Integration and upload of aeDNA-derived ASV and OTU tables into Neotoma, LiPD, and other paleoecoinformatics resources. Pilot efforts are already underway (Box 2), and based on these experiments, three next steps are ready for immediate action. First, to update paleodata schemas and associated software services to better align with the particular needs of sedimentary aeDNA (e.g., supporting derived taxonomic inferences with linkages to reference databases and analytical pipelines). Second, to appoint and train Data Steward experts in aeDNA who can help establish and implement the community standards (e.g., controlled vocabularies) necessary for data harmonization. Third, to engage in a broad-scale, community-supported data mobilization campaign, in which participating research groups send their data to appointed Data Stewards for curation and upload, in order to establish a well-curated suite of aeDNA datasets that can serve as the backbone for further macro-scale research.

Harmonization and integration of transparent workflows for lab processing and bioinformatics standards. Informed interpretation of aeDNA results depends critically on knowledge of how the data were generated and analyzed, e.g., the use of negative and positive controls, replicates, and other lab processing steps, as well as the choice of reference database(s) (Section 4, Fig. 4). Not all of this information can or should be stored in paleoecoinformatics data resources. Rather, the analytical pipelines are themselves a primary form of process documentation and transparency [65]. There exists a tension between methodological innovation and standardization, and while aeDNA has been in its early stages, innovation has been paramount. Hence, the immediate need is to enhance transparency by setting community norms that laboratory and analytical workflows should be published as reproducible protocols (e.g., https://protocols.io) or code (e.g., https://github.com), while the next step is to establish standard community pipelines and protocols wherever possible.

Integrate emerging metagenomics standards into this open, linked, bioinformatics and paleoecoinformatics cyberinfrastructure. Although metabarcoding data are currently the most common (Fig. 1b), shotgun metagenomics and targeted enrichment methods are rapidly growing in popularity and likely will surpass metabarcoding soon [42]. These aeDNA data types will require their own somewhat customized informatics and curation solutions, given large data volumes and reads from a broader set of genomic regions than for metabarcoding. The emerging standards for metagenomic aeDNA and eDNA (https://spaam-community.github.io/) should be integrated into the genomics and paleoecoinformatics ecosystems that support aeDNA.

Building open, ethical, and global communities of practice and community data governance. The aeDNA community of researchers is growing quickly with a high preponderance of early career researchers, and new communities of practice are rapidly forming (e.g., PaleoEcoGen Working Group, the SedaDNA Scientific Society). CCDRs help advance these efforts by serving as boundary organizations [71], whereby specialists from different communities (e.g., aeDNA specialists, data scientists, biogeographers, educators) can convene and exchange knowledge across disciplinary boundaries. In this effort, enhancing inclusivity and accessibility is essential, because paleobiogeographic data are rife with biases caused by past and present inequities in scientific practice [72]. Similarly, management and sharing of aeDNA should support CARE principles for Indigenous data governance [56]. As examples of initial efforts here, the SedaDNA Scientific Society launched the African sedaDNA Working Group in 2021, while recent data mobilization campaigns for Neotoma have focused on improving data representation across the Southern Hemisphere.

Concluding Remarks

Building cyberinfrastructure is a means, not an end; the ultimate goal is to power the next generation of question-driven macro-scale integrative research and new insights into the processes governing biodiversity dynamics over space and time (see Outstanding Questions). The scale is too vast and the data too heterogeneous for any single researcher or research lab to unilaterally conduct global-scale analyses effectively, so well-curated, harmonized datasets are essential. Our experience in this era of open science has been that as soon as new high-quality data resources are built and openly shared, they are immediately used to advance discovery. The steps described here are essential for conducting the next generation of integrative global-change science.

Box 1: A Biogeographic Multi-proxy and Multi-site Case Study: Where Was *Cedrus* (Cedar) at the Last Glacial Maximum?

Understanding species' past distribution and diversity relies on accurate inferences of species' presence and absence. However, each type of paleoecological proxy is affected differentially by taphonomic and biological processes that affect the probability of detecting a species, precision of taxonomic identification, and spatial source area represented by a given fossil occurrence. Inferences based upon multiple paleoecological proxies reduce uncertainty and carry more power.

For example, a persistent question has been whether the conifer *Cedrus* (cedar) survived in southern Italy across glacial-interglacial cycles. It has been suggested, based on fossil pollen, that climate changes between 0.9 and 0.7 Ma extirpated *Cedrus* from the Italian Peninsula, while it persisted longer in Greece [73]. Palynologists have interpreted the few pollen grains of *Cedrus* found in Late Pleistocene lake sediments from southern Italy as sourcing from populations in north Africa (Fig. I; [74,75]).

Because aeDNA in lake sediments is believed to source locally from plants growing in the watershed and not from windblown pollen from more distant sources [76–78], aeDNA can be used to explore hypotheses about local refugia. However, aeDNA itself needs to be carefully checked to rule out the possibility of false positives due to laboratory contamination or other factors [5,7]. At Lago Grande di Monticchio in southern Italy, prior work has reported occasional *Cedrus* pollen grains from glacial-aged sediments, at levels too low to confidently establish local presence [79]. Metabarcoding aeDNA data from an investigation aimed at reconstructing the flora at Monticchio suggest that *Cedrus* was present at this site during the last glacial and the late Holocene period (Fig. I). *Cedrus* aeDNA was reported in 12 samples from Monticchio, yet was undetected in the extraction and PCR negative controls, nor in samples from the other lakes analyzed in the same sequencing run, which argues against a false positive caused by cross-sample contamination.

To further explore the Monticchio aeDNA findings , we mapped pollen data from Neotoma, which indicate widespread but low abundances across late-glacial samples from southern Europe (Fig. I). The combination of local aeDNA presence at Monticchio and trace quantities of *Cedrus* pollen across southern Europe reinforce the hypothesis that *Cedrus* was present in southern Italy during the last glacial period, showing how biogeographic inferences can be strengthened by combining aeDNA data with regional networks of other paleoecological proxies.

Box 2: Putting recommendations into action: Pilot uploads of aeDNA data into Neotoma

As a first step towards integrating aeDNA site-level data with paleoecoinformatics resources, we have launched pilot uploads of metabarcoding-derived ASV tables into the Neotoma Paleoecology Database. Neotoma carries several advantages as a home for taxonomic inferences sourced from metabarcoding and metagenomic analyses. First, most of Neotoma's data span the last 10^2 to 10^6 years – a timescale that matches well with the temporal duration of aeDNA data [80–82]. Second, Neotoma already stores much of the spatial and temporal metadata needed to analyze past species distributions, such as site location, depositional context, radiometric and other age controls, and multiple age-depth models and associated age inferences. Third, Neotoma contains other paleoecological proxies from both terrestrial and marine archives. Fourth, Neotoma stores samples from modern depositional contexts (e.g., [83]), which is essential for aeDNA ground truthing [76,84] and building statistical inferences about past ecosystems and environments [85,86]. Other paleoenvironmental resources such as LiPD are also expanding support for aeDNA data (McKay, pers. comm.).

In this pilot effort, a metabarcoding dataset was uploaded from Lake Naleng on the Tibetan Plateau [87]. This effort revealed a generally close but imperfect match between Neotoma's data schema and the metadata needs associated with aeDNA. Some mismatches could be quickly resolved, by expanding controlled vocabularies in Neotoma to accommodate key metadata needs associated with aeDNA. For example, 'Metabarcoding aeDNA' is a newly added dataset type. Similarly, the Elements field in Neotoma is intended to indicate which part of the organism a fossil comes from, but we have expanded its usage to also store information about the genetic locus used in metabarcoding research, e.g., '18S rRNA' or '*trn*L p6-loop'.

Other mismatches will need deeper modifications to Neotoma's data schema. For example, in Neotoma, just the taxonomic name is stored, while an ASV table stores both the primary genetic sequence and derived taxonomic identification, so Neotoma's data model needs to be expanded to hold both pieces of information. Similarly, Neotoma needs better linking capacity to other components of the emerging informatics ecosystem for aeDNA data, including repositories for raw sequence data and reference databases. All these points are resolvable, however, so this pilot effort shows both how conceptual and semantic misalignment can create hidden barriers to building global-scale, multi-proxy, and multi-disciplinary community data resources, and how these barriers can be overcome.

Glossary

Amplicon sequence variant (ASV). A unique DNA sequence generated by metabarcoding analysis. ASV methods seek to identify true sequences and discard putative sequencing and PCR errors. ASVs are increasingly replacing clustering methods based only on similarities among sequences (i.e., OTUs).

Ancient environmental DNA (aeDNA). Ancient DNA is any DNA that is recovered from a non-living tissue, organism, or environmental sample; the latter is aeDNA. To clearly differentiate aDNA from modern DNA, aDNA is any DNA that has degraded into short fragments and exhibits post-mortem damage signatures.   Common examples of aeDNA include DNA extracted from sedimentary archives, such as soil samples from caves or archaeological sites or samples from lake or marine sediments.

Archive: a sedimentary record or other geological medium from which aeDNA or other paleoecological and paleoenvironmental proxies are retrieved.

Community-Curated Data Resources (CCDRs): an active database in which data are added and stewarded by experts drawn from the community that initially generated the data.

Library (or Sequencing Library). DNA molecules that have been prepared for high-throughput sequencing by adding readable adapters (artificial DNA sequence) to their ends. In a metagenomic library, the DNA molecules are prepared directly from a DNA extract, whereas an amplicon library is prepared from PCR amplicons.

Metabarcoding. Taxonomic identification of aeDNA molecules through sequencing of selected short (typically ~30-600 bp) regions of DNA called barcodes, which are standardized markers that are sufficiently conserved to target a higher taxonomic group but variable enough to discriminate species or genera.

Operational Taxonomic Unit (OTU). DNA sequences recovered from a metabarcoding analysis that are clustered together based on sequence similarity. The clustering of DNA sequences into OTUs is done from processed reads. OTU identification of taxa typically requires long barcodes with multiple substitutions.

Paleoecoinformatics. The intersection of the information, Earth, and biological sciences in which biological data are collected from geohistorical archives and are stored, integrated, and analyzed through an informatic pathway.

Polymerase chain reaction (PCR). A laboratory technique to increase the concentration of a genomic fragment of interest from a DNA template by performing multiple rounds of amplification. This technique requires a pair of short synthetic DNA fragments (primers) that bind to either side of the genomic region of interest.

Proxies. Physical, chemical, or biological data that preserve some signal of past environments and ecosystems to provide information about the unobservable past states of the variable(s) of interest.

Reference database. An inventory of identified DNA sequences that unidentified aeDNA sequence data can be compared against. Reference databases differ in their marker representation, completeness, scope, and quality of curation.

Shotgun metagenomics. Direct sequencing of a metagenomic library without any enrichment that offers a randomized sampling of aeDNA present within a sample.

Surface samples. Sediments containing recently deposited and therefore usually well-preserved eDNA. Surface samples, together with independent observations of contemporary environments, are used to understand the taphonomic processes governing the relationships between living and ancient assemblages and to constrain proxy-based quantitative inferences.

Taxonomic inventory. A list of taxa identified from an aeDNA sample by matching sequence data to a reference database.

Template. Extracted DNA used to perform a molecular assay, such as a PCR, qPCR, ddPCR reaction, or to prepare a shotgun metagenomics library. For PCR analyses, template molecules must be long enough to include the primer binding sites and genomic region of interest.

Target enrichment. The enrichment of a metagenomic library for genomic regions of interest using pre-designed DNA or RNA probes. The probes hybridize with genomic library fragments of interest, which are then immobilized. Non-hybridized library molecules are then removed, resulting in an enrichment of data from the targeted genomic regions. Probe sequences can target a wide range of taxa and single loci, organellar genomes/exomes, and/or low-copy nuclear regions.

Acknowledgments

Support for the Sedimentary aeDNA-Informatics Consortium was provided by NSF (2011295). The aeDNA data compilation shown in Figure 1 [61] was led by members of the PAGES PaleoEcoGen Working Group and the SedaDNA Scientific Society. Other paleoecological proxy data were obtained from the Neotoma Paleoecology Database (http://www.neotomadb.org) and its constituent databases. The work of data contributors, data stewards, and the Neotoma community is gratefully acknowledged. Sadly, two co-authors, Drs. Eric Grimm and Sarah Crump passed away during the development of this manuscript. Dr. Grimm passed away in November 2020, at an early stage of community building and manuscript planning, while Dr. Sarah Crump, one of the original manuscript leaders, passed away in November 2022. Both will be missed, and this manuscript is dedicated in their honor.

References

1. Wang, Y. *et al.* (2021) Late Quaternary dynamics of Arctic biota from ancient environmental genomics. *Nature* 600, 86–92
2. Epp, L.S. (2019) A global perspective for biodiversity history with ancient environmental DNA. *Mol. Ecol.*, 28, 2456–2458
3. Rijal, D.P. *et al.* (2021) Sedimentary ancient DNA shows terrestrial plant richness continuously increased over the Holocene in northern Fennoscandia. *Sci. Adv.* 7, eabf9557

4. Parducci, L. *et al.* (2019) Shotgun environmental DNA, pollen, and macrofossil analysis of Lateglacial lake sediments from southern Sweden. *Front. Ecol. Evol.* 7, 189

5. Alsos, I.G. *et al.* (2020) Last Glacial Maximum environmental conditions at Andøya, northern Norway; evidence for a northern ice-edge ecological “hotspot.” *Quat. Sci. Rev.*, 239106364

6. Nota, K. *et al.* (2022) Norway spruce postglacial recolonization of Fennoscandia. *Nat. Comm.* 13, 1333

7. Clarke, C.L. *et al.* (2019) Persistence of arctic-alpine flora during 24,000 years of environmental change in the Polar Urals. *Sci. Rep.*, 919613

8. Alsos, I.G. *et al.* (2022) Postglacial species arrival and diversity buildup of northern ecosystems took millennia. *Sci. Adv.* 8, eabo7434

9. Pedersen, M.W. *et al.* (2016) Postglacial viability and colonization in North America’s ice-free corridor. *Nature*, 53, 745–49

10. Graham, R.W. *et al.* (2016) Timing and causes of mid-Holocene mammoth extinction on St. Paul Island, Alaska. *Proc. Natl. Acad. Sci.* 113, 9310–9314

11. Willerslev, E. *et al.* (2014) Fifty thousand years of Arctic vegetation and megafaunal diet. *Nature* 506, 47–51

12. Capo, E. *et al.* (2017) Tracking a century of changes in microbial eukaryotic diversity in lakes driven by nutrient enrichment and climate warming. *Env. Microbio.* 19, 2873–2892

13. Monchamp, M.-E. *et al.* (2018) Homogenization of lake cyanobacterial communities over a century of climate change and eutrophication. *Nat. Ecol. Evol.* 2, 317–324

14. Liu, S. *et al.* (2020) Holocene vegetation and plant diversity changes in the north-eastern Siberian treeline region from pollen and sedimentary ancient DNA. *Front. Ecol. Evol.* 8, 560243

15. Mottl, O. *et al.* (2021) Global acceleration in rates of vegetation change over the past 18,000 years. *Science* 372, 860–864

16. Nolan, C. *et al.* (2018) Past and future global transformation of terrestrial ecosystems under climate change. *Science* 361, 920–923

17. Nogué, S. *et al.* (2021) The human dimension of biodiversity changes on islands. *Science*, 372, 488–491

18. Barnosky, A.D. *et al.* (2004) Assessing the causes of Late Pleistocene extinctions on the continents. *Science* 306, 70–75

19. Smith, F.A. *et al.* (2018) Body size downgrading of mammals over the late Quaternary. *Science* 360, 310–313

20. Burke, K.D., .J.W. Williams, S. Brewer, W. Finsinger, T. Giesecke, D.J. Lorenz and Ordonez, A. (2019) Differing climatic mechanisms control transient and accumulated vegetation novelty in Europe and eastern North America. *Phil. Trans. R. Soc. B Biol.* 374, 20190218

21. Finsinger, W. *et al.* (2017) Emergence patterns of novelty in European vegetation assemblages over the past 15 000 years. *Ecol. Lett.* 20, 336–346

22. Price, G.J. *et al.* (2018) Big data little help in megafauna mysteries. *Nature*, 558, 23–25

23. Williams, J.W. *et al.* (2018) Building open data: Data stewards and community-curated data resources. *PAGES Mag.* 26, 50–51

24. Brewer, S. *et al.* (2012) Paleoecoinformatics: applying geohistorical data to ecological questions. *Trends Ecol. Evol.* 27, 104–112

25. Behrensmeyer, A.K. *et al.* (2000) Taphonomy and paleobiology. *Paleobio.* 26, 103–147

26. Alsos, I.G. *et al.* (2016) Sedimentary ancient DNA from Lake Skartjørna, Svalbard: Assessing the resilience of arctic flora to Holocene climate change. *The Holocene* 26, 627–642

27. Courtin, J. *et al.* (2022) Pleistocene glacial and interglacial ecosystems inferred from ancient DNA analyses of permafrost sediments from Batagay megaslump, East Siberia. *Envir. DNA* 9, 625096

28. Clarke, C.L. *et al.* (2020) A 24,000-year ancient DNA and pollen record from the Polar Urals reveals temporal dynamics of arctic and boreal plant communities. *Quat. Sci. Rev.* 247, 106564

29. Huang, S. *et al.* (2020) Genetic and morphologic determination of diatom community composition in surface sediments from glacial and thermokarst lakes in the Siberian Arctic. *J. Paleolim.* 64, 225–242

30. Stoof-Leichsenring, K.R. *et al.* (2014) A combined paleolimnological/genetic analysis of diatoms reveals divergent evolutionary lineages of Staurosira and Staurosirella (Bacillariophyta) in Siberian lake sediments along a latitudinal transect. *J. Paleolimnol.* 52, 77–93

31. Dulias, K. *et al.* (2017) Sedimentary DNA versus morphology in the analysis of diatom-environment relationships. *J. Paleolimnol.* 57, 51–66

32. Picard, M. *et al.* (2022) Molecular and pigment analyses provide comparative results when reconstructing historic cyanobacterial abundances from lake sediment cores. *Microorganisms* 10, 279

33. Capo, E. *et al.* (2021) Lake sedimentary DNA research on past terrestrial and aquatic biodiversity : Overview and recommendations. *Quaternary* 4, 1–61

34. Domaizon, I. *et al.* (2017) DNA-based methods in paleolimnology: new opportunities for investigating long-term dynamics of lacustrine biodiversity. *J. Paleolimnol.* 58, 1–21

35. Coolen, M.J.L. *et al.* (2013) Evolution of the plankton paleome in the Black Sea from the Deglacial to Anthropocene. *Proc. Natl. Acad. Sci.* 110, 8609–8614

36. Crump, S.E. *et al.* (2021) Ancient plant DNA reveals High Arctic greening during the Last Interglacial. *Proceedings of the National Academy of Sciences* 118, e2019069118

37. Heffernan, J.B. *et al.* (2014) Macrosystems ecology: understanding ecological patterns and processes at continental scales. *Frontiers in Ecology and the Environment* 12, 5–14

38. McGill, B.J. (2019) The what, how and why of doing macroecology. *Global Ecology and Biogeography* 28, 6–17

39. Fordham, D.A. *et al.* (2021) Process-explicit models reveal pathway to extinction for woolly mammoth using pattern-oriented validation. *Ecology Letters* 25, 125-137

40. Saltré, F. *et al.* (2016) Climate change not to blame for late Quaternary megafauna extinctions in Australia. *Nat. Comm.* 7, 10511

41. Ruddiman, W.F. *et al.* (2020) The early anthropogenic hypothesis: A review. *Quat. Sci. Rev.* 240, 106386

42. Dussex, N. *et al.* (2021) Integrating multi-taxon palaeogenomes and sedimentary ancient DNA to study past ecosystem dynamics. *Proc. R. Soc. B Biol. Sci.* 288, 20211252

43. COHMAP Members (1988) Climatic Changes of the Last 18,000 Years: Observations and Model Simulations. *Science* 241, 1043–1052

44. FAUNMAP Working Group (1996) Spatial response of mammals to late Quaternary environmental fluctuations. *Science* 272, 1601–1606

45. Raup, D.M. and Sepkoski, Jr., J.J. (1982) Mass extinctions in the marine fossil record. *Science* 215, 1501–1503

46. Williams, J.W. *et al.* (2017) Cyberinfrastructure in the paleosciences: Mobilizing long-tail Data, building distributed community infrastructure, empowering individual geoscientists. Authorea, https://www.authorea.com/users/152134/articles/165940-cyberinfrastructure-in-the-paleosciences-mobilizing-long-tail-data-building-distributed-community-infrastructure-empowering-individual-geoscientists

47. Williams, J.W. *et al.* (2018) The Neotoma Paleoecology Database: A multi-proxy, international community-curated data resource. *Quat. Res.* 89, 156–177

48. Peters, S.E. and McClennen, M. (2016) The Paleobiology Database application programming interface. *Paleobio.* 42, 1–7

49. Renaudie, J. *et al.* (2020) NSB (Neptune Sandbox Berlin): An expanded and improved database of marine planktonic microfossil data and deep-sea stratigraphy. *Palaeontol. Electron.* 23, a11

50. Gross, W. *et al.* (2018) New advances at NOAA's World Data Service for Paleoclimatology - Promoting the FAIR principles. *PAGES Mag.* 26, 58

51. McKay, N.P. and Emile-Geay, J. (2016) Technical note: The Linked Paleo Data framework – a common tongue for paleoclimatology. *Clim. Past* 12, 1093–1100

52. Diepenbroek, M. (2018) PANGAEA - data publisher for Earth and Environmental Sciences. *PAGES Mag.*, 26, 59.

53. Uhen, M.D. *et al.* (2013) From card catalogs to computers: databases in vertebrate paleontology. *J. Vert. Paleo.* 33, 13–28

54. Khider, D. *et al.* (2019) PaCTS 1.0: A crowdsourced reporting standard for paleoclimate data. *Paleocean. and Paleoclim.* 341, 570–1596

55. Wilkinson, M.D. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 160018

56. Carroll, S.R. *et al.* (2020) The CARE principles for indigenous data governance. *Data Sci. J.* 19, 43

57. Uhen, M.D. *et al.* (2021) The EarthLife Consortium API: An extensible, open-source service for accessing fossil data and taxonomies from multiple community paleodata resources. *Frontiers Bio.*, e50711

58. Brenskelle, L. *et al.* (2022) A community-developed extension to Darwin Core for reporting the chronometric age of specimens. *PLOS ONE* 17, 1–13

59. Orlando, L. *et al.* (2021) Ancient DNA analysis. *Nat. Rev. Meth. Primer* 1, 1–26

60. Murchie, T.J. *et al.* (2021) Collapse of the mammoth-steppe in central Yukon as revealed by ancient environmental DNA. *Nat. Comm.* 12, 7120

61. Von Eggers, J. *et al.* (2022) Inventory of ancient environmental DNA from sedimentary archives: locations, methods, and target taxa (Version 1). *Zenodo* https://doi.org/10.5281/zenodo.6847522

62. Alsos, I.G. *et al.* (2020) The treasure vault can be opened: Large-scale genome skimming works well using herbarium and silica gel dried material. *Plants* 9, 432

63. Quast, C. *et al.* (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596

64. Boyer, F. *et al.* (2016) obitools: a unix-inspired software package for DNA metabarcoding. *Mol. Ecol. Resour.* 16, 176–182

65. Bolyen, E. *et al.* (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857

66. Tamames, J. and Puente-Sánchez, F. (2019) SqueezeMeta, A highly portable, fully automatic metagenomic analysis pipeline. *Front. Microbio.* 9, 3349

67. Stoof-Leichsenring, K.R. *et al.* (2022) Sedimentary DNA identifies modern and past macrophyte diversity and its environmental drivers in high-latitude and high-elevation lakes in Siberia and China. *Limn. Ocean.* 67, 1126–1141

68. Gauthier, J. *et al.* (2022) Sedimentary DNA of a human-impacted lake in Western Canada (Cultus Lake) reveals changes in micro-eukaryotic diversity over the past ~200 years. *Envir. DNA* 4.5, 1106-1119.

69. Curtin, L. *et al.* (2021) Sedimentary DNA and molecular evidence for early human occupation of the Faroe Islands. *Comm. Earth Env.* 2, 253

70. Fellows Yates, J.A. *et al.* (2021) Community-curated and standardised metadata of published ancient metagenomic samples with AncientMetagenomeDir. *Sci. Data* 8, 31

71. Goring, S.J. *et al.* (2018) The Neotoma Paleoecology Database: A research-outreach nexus. In *Paleobiology Short Course* Cambridge: Cambridge University Press. doi:10.1017/9781108681582

72. Raja, N.B. *et al.* (2022) Colonial history and global economics distort our understanding of deep-time biodiversity. *Nat. Ecol. & Evol.* 6. 145–154

73. Magri, D. *et al.* (2017) Quaternary disappearance of tree taxa from southern Europe: Timing and trends. *Quat. Sci. Rev.* 163, 23–55

74. Magri, D. and Parra, I. (2002) Late Quaternary western Mediterranean pollen records and African winds. *Earth Plan. Sci. Lett.* 200, 401–408

75. Magri, D. (2012) Quaternary history of Cedrus in south Europe. *Ann. Bot.* 2, 57-66

76. Alsos, I.G. *et al.* (2018) Plant DNA metabarcoding of lake sediments: How does it represent the contemporary vegetation. *PLOS ONE*, 13e0195403

77. Parducci, L. *et al.* (2017) Ancient plant DNA in lake sediments. *New Phytol.* 214, 924–942

78. Sjögren, P. *et al.* (2017) Lake sedimentary DNA accurately records 20th Century introductions of exotic conifers in Scotland. *New Phytol.* 213, 929–941

79. Allen, J.R.M. and Huntley, B. (2000) Weichselian palynological records from southern Europe: correlation and chronology. *Quat. Int.* 73–4, 111–125

80. Kirkpatrick, J.B. *et al.* (2016) Fossil DNA persistence and decay in marine sediment over hundred-thousand-year to million-year time scales. *Geology* 44, 615–618

81. Willerslev, E. *et al.* (2007) Ancient biomolecules from deep ice cores reveal a forested southern Greenland. *Science* 317, 111–114

82. Armbrecht, L. *et al.* (2022) Ancient marine sediment DNA reveals diatom transition in Antarctica. *Nat. Comm.* 13, 5787

83. Amesbury, M.J. *et al.* (2018) Towards a Holarctic synthesis of peatland testate amoeba ecology: Development of a new continental-scale palaeohydrological transfer function for North America and comparison to European data. *Quat. Sci. Rev.* 201, 483–500

84. Stoof-Leichsenring, K. *et al.* (2020) Plant diversity in sedimentary DNA obtained from high-latitude (Siberia) and high-elevation lakes (China). *Biodiv. Data J.* 8, e57089

85. Raiho, A. *et al.* (2022) 8000-year doubling of Midwestern forest biomass driven by population- and biome-scale processes. *Science* 376, 1491–1495

86. Chevalier, M. *et al.* (2020) Pollen-based climate reconstruction techniques for late Quaternary studies. *Earth-Sci. Rev.* 210, 103384

87. Liu, S. *et al.* (2021) Sedimentary ancient DNA reveals a threat of warming-induced alpine habitat loss to Tibetan Plateau plant diversity. *Nat. Comm.* 12, 2995

88. Caudullo, G. *et al.* (2017) Chorological maps for the main European woody species. *Data Brief* 12, 662–666

89. Nota, K. (2022) Palaeoenvironmental reconstructions from modern and ancient plant DNA. Thesis

Figure Captions

Figure 1. A mapped inventory of published ancient environmental DNA (aeDNA) datasets and other paleoecological proxies, compiled as of July 18, 2022 [61], shown for the purpose of comparing the spatial and taxonomic coverage of aeDNA to classic paleoecological data types. (a) All aeDNA datasets, (b) metabarcoding aeDNA datasets only, and (c) other paleoecological proxies from Neotoma. In (a) and (b), sites are color-coded by four broad categories of taxonomic groups targeted in aeDNA analyses (animals, plants, fungi, and microorganisms), while shape indicates type of sedimentary archive. The ‘All’ category is used for shotgun metagenomics studies, given the untargeted nature of this data type. The number of sites representing marine or lake surface sediments is 436 (a) and 393 (b). In (c), Neotoma datasets are organized into similarly broad taxonomic and functional groups: aquatic organisms (diatoms, dinoflagellates, ostracods), vertebrates, macroinvertebrates, testate amoebae, and plants (pollen and macrofossils).

Figure 2. A schematic of the knowledge domains for aeDNA to be supported by cyberinfrastructure. Initial collection requires tracking of metadata from field to lab, where information about processing of samples and controls must be tracked along with metadata associated with bioinformatics pipelines. These pipelines act to reduce data volume, compare to reference databases, and infer taxonomic identities (Fig. 4). Further extraction of ecological and evolutionary insights from aeDNA requires precise temporal

positioning through geochronological controls and age depth-modeling, understanding of other environmental and ecosystem dynamics from other proxies at the same site, placement in the tree of life through phylogenetics, and linking to paleoecological and paleoenvironmental records at other sites.

Figure 3. A site-level schematic of how the interpretation of the biodiversity dynamics recorded by aeDNA can be further enriched by other proxies that provide indicators of past climate variations (e.g., biomarkers) and independent indicators of past community dynamics (e.g., pollen, diatoms, sterols). Ecological interpretation of aeDNA can be based upon simple presence, abundance based on relative read counts, and/or frequency of occurrence across PCR replicates (either collected as 'technical replicates' from the same extract or as 'biological replicates' as multiple samples from the same spatiotemporal location), combined with phylogenetic position. Temporal position is based on an age-depth model that infers time as a function of depth, with uncertainty, based on age controls such as radiocarbon (^{14}C) dates.

Figure 4. Schematic overview of typical aeDNA workflows, from sediment sample to published data. Purified DNA is first isolated from sediment samples via DNA extraction. Negative controls are monitored for contamination (represented by tubes without DNA molecules). A non-sequenced PCR workflow (red boxes) estimates the abundance/quantity of a DNA template but does not generate sequence data that can be taxonomically identified. The amplicon pipeline (blue boxes) includes metabarcoding and multiplex PCR. PCR products can either be individually converted into a library (left; as done during a two-step library preparation) or pooled before library preparation

(right). For the sequencing runs boxes, each smaller box represents one library within a sequencing run. In metagenomic approaches (yellow boxes), a library can be either directly shotgun sequenced or enriched for a target of interest before sequencing. The sequence reads are post-processed for quality control (‘read quality control’) by removing short and/or low-quality sequences and other artifacts and by collapsing identical sequences. After quality control, sequences are aligned with external reference databases to enable taxonomic identification. Refinement of alignments includes the removal of contaminants and/or curation of taxonomic assignments. The resulting data include a taxonomic inventory and information about abundance based on counts of reads or frequency of presence across replicates, haplotypic variation within species, and, for metagenomic approaches, information about ancient DNA damage and population genomic variation. HRM: High-resolution melt; qPCR: quantitative PCR; ddPCR: droplet digital PCR.

Box Figure I. Detections of *Cedrus* (cedar) aeDNA from glacial sediments are intriguing but, in isolation, provide an incomplete understanding of the refugial distribution of this taxon. Conversely, detections of *Cedrus* from pollen are widespread, but wind-dispersed pollen is a non-definitive indicator of local presence. The multipanel figure at left shows pollen sites from Neotoma where at least one *Cedrus* pollen grain is found, for three time periods: 32 to 11.7 thousand calendar years before present (ka BP), 11.7 to 3 ka BP, and 3 to 0 ka BP. (Open circles indicate pollen sites with no *Cedrus* pollen for that time window, while filled circles indicate presence of *Cedrus* pollen, with the color of the fill varying by time period.) Pink- and blue-colored regions show the current ranges of *C. atlantica* (Atlas cedar) and *C. libani* (Lebanese cedar) [88]. The plot at right reports preliminary metabarcoding DNA results for *Cedrus* for the 14-m Lago Grande di Monticchio core spanning the last 31 ka [89]. Each bar indicates a *Cedrus* detection in a PCR technical replicate and is colored by the number of reads recorded as *Cedrus*. For this plot, the time scale is linear for each time period but differs among time periods.

Highlights

- The pace and scale of aeDNA-powered biodiversity research is growing rapidly, and the field is now at the cusp of supporting past global-scale biodiversity research at unprecedented taxonomic resolution and temporal extent .
- In parallel, the paleoecoinformatics ecosystem is quickly growing and interdigitating, enabling support of multi-proxy and broad-scale research into past ecological and environmental change.
- Because aeDNA-derived species inferences are dynamic, as are estimated ages, a global data system for aeDNA must interlink and leverage existing resources in bioinformatics and paleoecoinformatics.
- Prior experience has shown that open and community-governed data resources are essential for high-quality global paleodata syntheses and for empowering the next generation of scientists.

Outstanding Questions

These questions are organized into two categories: Scientific and Socio-Informatic

Scientific

How were past changes in biodiversity, as revealed by aeDNA records, shaped by past environmental change, human activities, and biotic interactions?

How sensitive are species and ecosystems to climate change, at local to global scales?

What processes drive abrupt changes in ecological systems, and can early warnings of abrupt change be detected in advance?

What were the causes and consequences of past population declines and extinctions?

Where do inferences based on aeDNA agree or disagree with those based on other paleoecological proxies, and why?

Socio-Informatic

Where do existing paleoecoinformatics data systems need to be modified to support the storage and informed reuse of aeDNA data, with respect to e.g. data structure, controlled vocabularies, or supporting software services?

What community governance systems are needed to ensure high-quality and open data resources with high levels of shared social trust?

How can we best harmonize and integrate existing workflows, bioinformatic standards, and data resources to maximize data access, transparency, and reusability?

How can we best support open and equitable access to aeDNA data and knowledge for the next generation of scientists?

How can we better streamline metadata and data transfers from individual labs to community data repositories, to reduce effort and shorten time to discovery?

Figure 1 site map

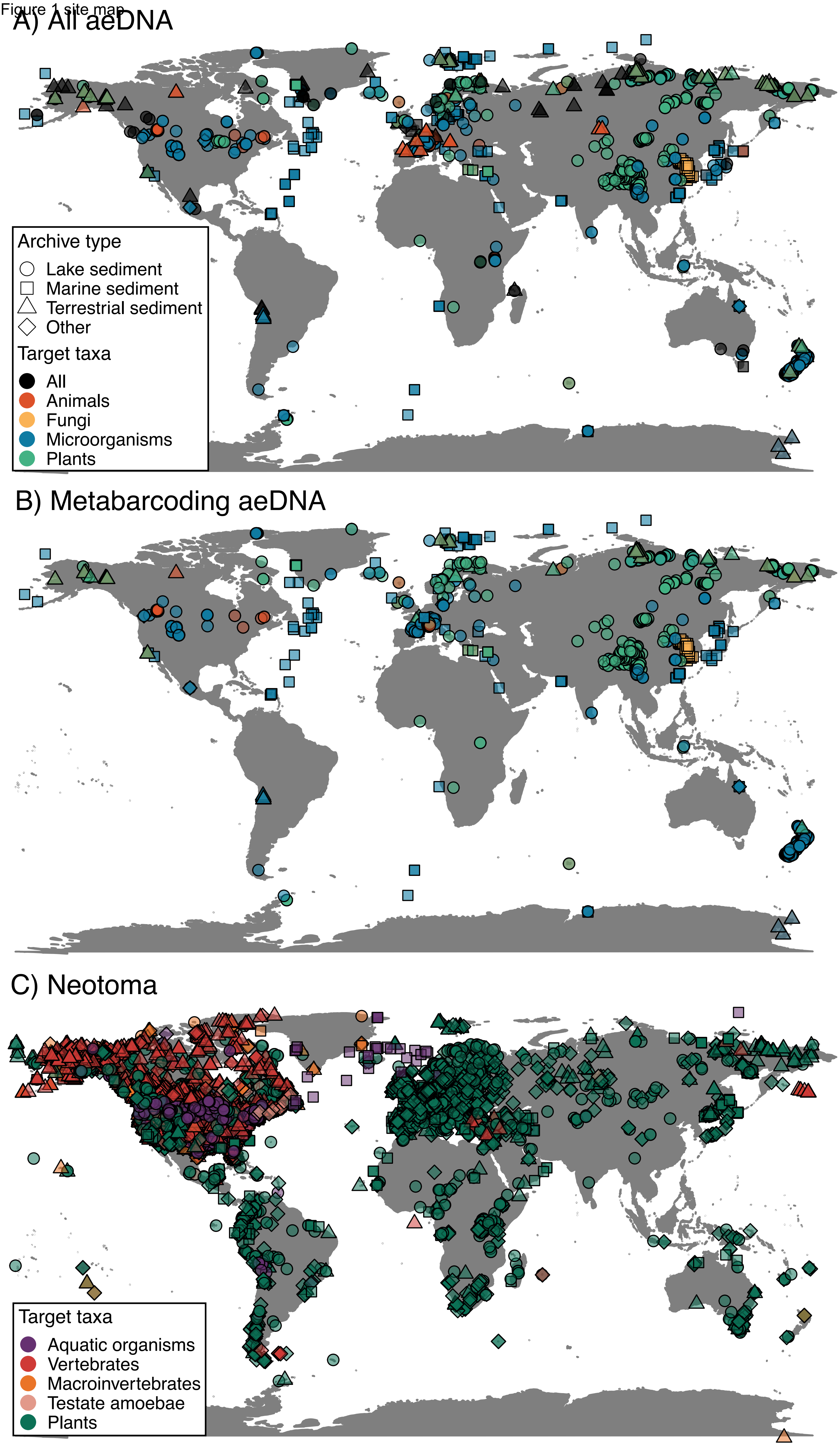


Figure 2 schematic overview

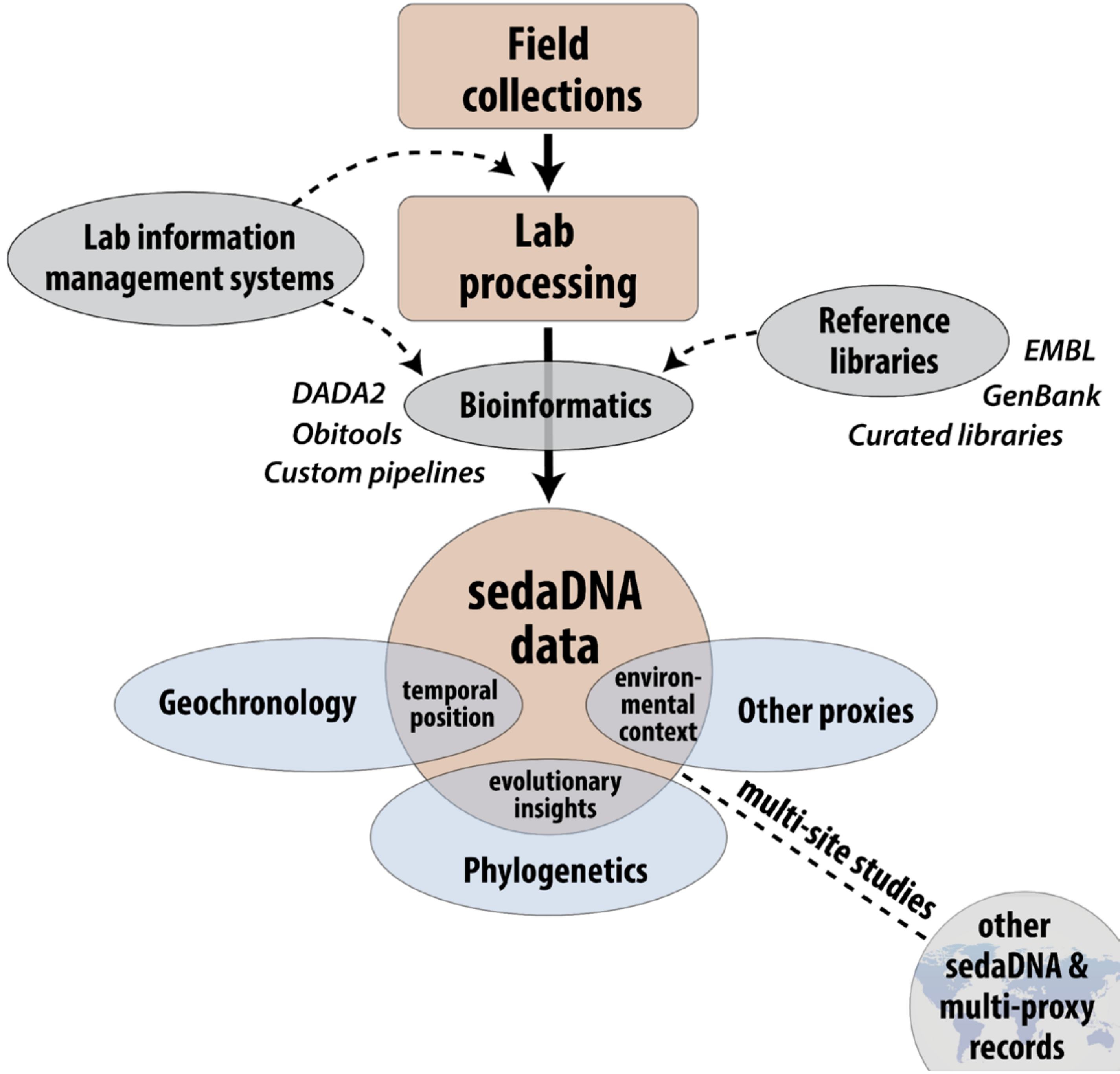


Figure 3 schematic multiproxy

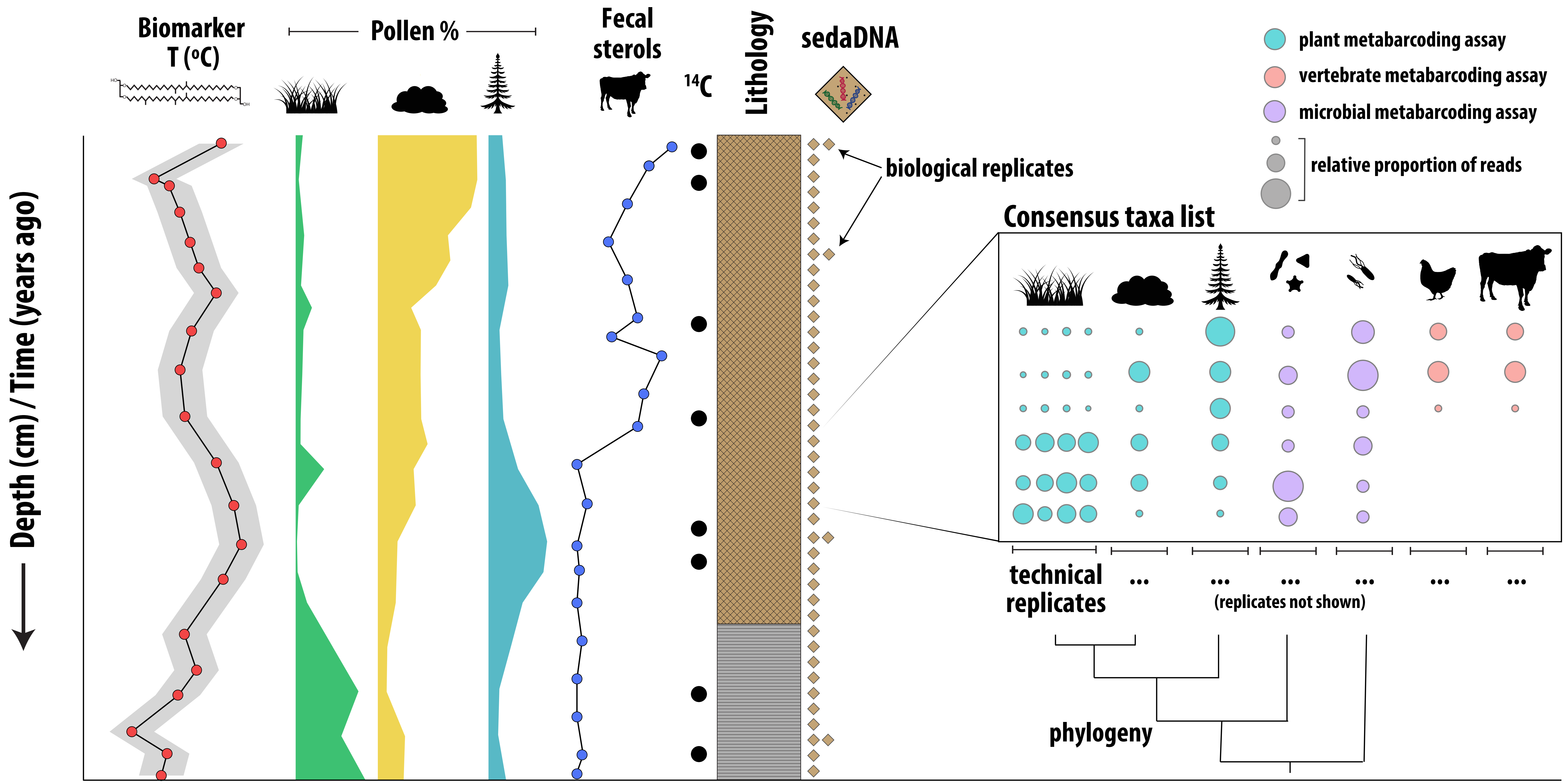
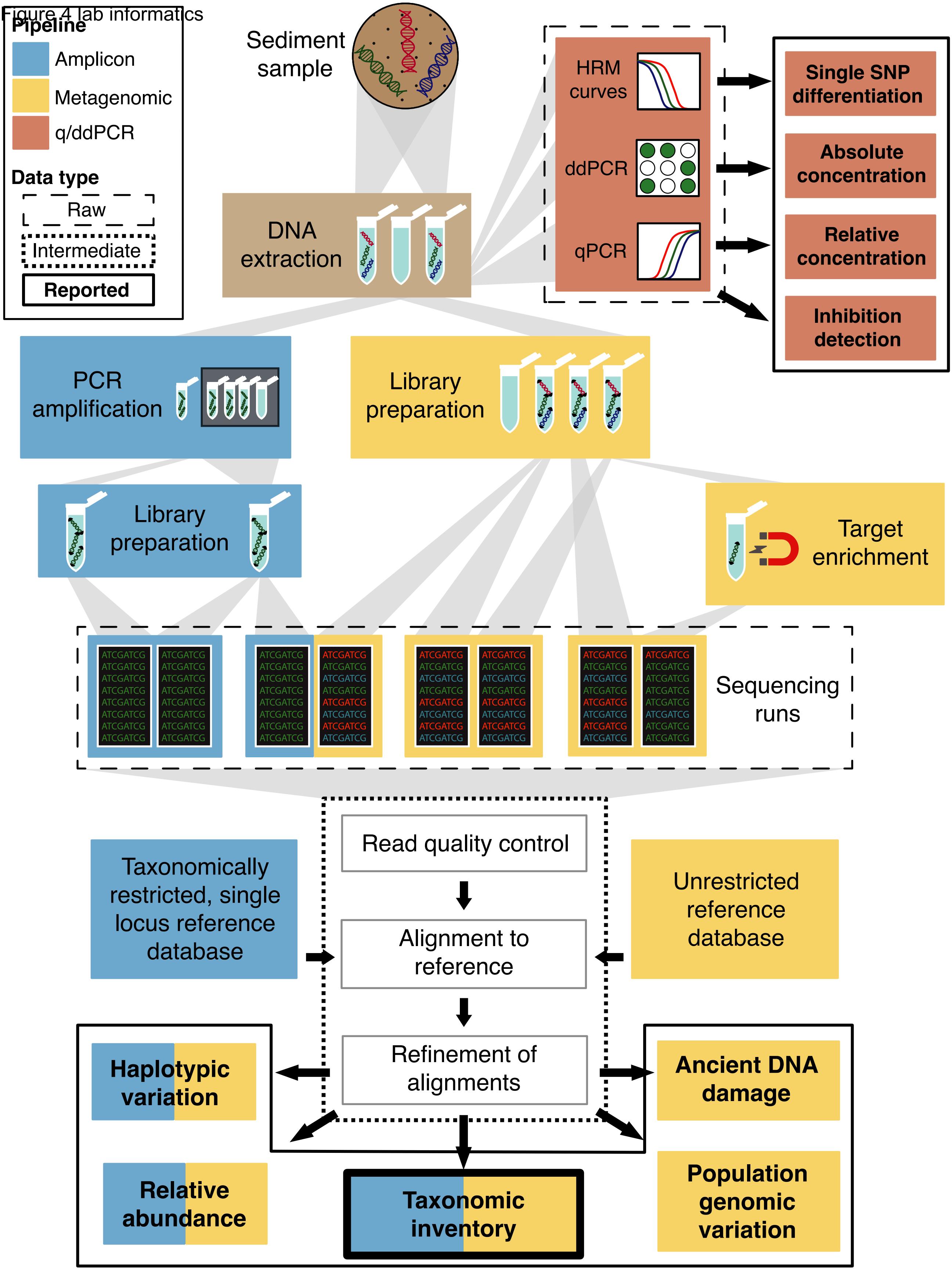


Figure 4 lab informatics



Box Fig 1 Cedrus