**Supplementary Information for:**

**Ancient DNA from speleothems: opportunity or challenge?**

Alexis Marchesini1,6, Daniela Festi2, Matteo Girardi3, Diego Micheletti4, Cristiano Vernesi1\*§, Marc Luetscher5§

1Forest Ecology Unit, Research and Innovation Centre- Fondazione Edmund Mach, via E. Mach 1-S. Michele all’Adige (Trento), 38010 Italy

2Institute for Interdisciplinary Mountain Research, Austrian Academy of Sciences, Innrain 25, A-6020 Innsbruck Austria

3Conservation Genomics Unit, Research and Innovation Centre- Fondazione Edmund Mach, via E. Mach 1-S. Michele all’Adige (Trento), 38010 Italy

4Computational Biology Unit, Research and Innovation Centre- Fondazione Edmund Mach, via E. Mach 1-S. Michele all’Adige (Trento), 38010 Italy

5Swiss Institute for Speleology and Karst Studies (SISKA), Serre 68, CH-2300 La Chaux-de-Fond

6Institute for Sustainable Plant Protection (IPSP), The National Research Council of Italy (CNR), 50019 Sesto Fiorentino (Florence), Italy

§ these authors equally contributed

\* Alexis Marchesini Email: alexis.marchesini@ipsp.cnr.it

|  |  |
| --- | --- |
| Appendix S1. Details on the cave system and speleothem samples | Page 2 |
| Appendix S2. PCR amplification mixture and thermal conditions | Page 5 |
| Appendix S3. Bioinformatics analysis and post-processing of data | Page 6 |

**Appendix S1**. **Details on the cave system and speleothem samples**

*Milchbach speleothems: details on the cave system and sampling*

The speleothems under investigation were sampled in Milchbach cave (46°37’N, 8°05’E, 1840 m a.s.l.), a periglacial cave system that opens along the western flank of the Upper Grindelwald Glacier valley, Switzerland (Luetscher et al, 2011).

The immediate surroundings of the cave entrances are presently free of vegetation. The glacier foreland and the Grindelwald valley, however, are characterized by intense human activities and open vegetation, which confine the mixed *Picea* and *Acer* forest to the steep valley slopes. In the study area, the timberline ecotone ranges between 1900 and 2100 m asl and is mainly composed of *Picea abies*, whereas *Pinus cembra*, *Larix decidua* and *Pinus mugo* occasionally play an important role (Providoli and Kuhn, 2012). The absence of closed forests in the vicinity of the cave entrances favors a regional origin of the pollen grains transported to the cave.



Figure S1. View of the Upper Grindelwald Glacier with location of the Milchbach cave system (left; view from the North). The immediate surrounding of the cave entrance is largely free of vegetation (right; view from the South).

The cave system comprises four major horizontal levels, intersected by a vadose shaft system with a maximum altitude difference of 286 m. Each of these levels opens to the exterior with a specific cave entrance forcing air convection through the system (chimney effect). This ventilation aborts during phases of glacier advances, when the entrances get sealed by the ice. All stalagmite samples originate from the very same cave passage (BL-1750) and were deposited within a few meters apart from each other but not necessarily exposed to the same airflow. Vertical slabs were cut in the lab along the stalagmite growth axes and subsampled for U/Th dating and stable isotope analyses (Luetscher et al., 2011). Age modelling based on 26 independent ages points to uncertainties averaging ± 50 years (2s). A diamond wire saw was used to cut small sections of between 0.64 and 7.28 g calcite for pollen analysis (Festi et al., 2016) and parallel samples for eDNA metabarcoding (this study). Each of these sections was cleaned in an ultrasonic bath using deionized water and stored in a sealed sterile bag until further processing.

*Speleothem formation process, external DNA absorption and preservation*

The age of our speleothems (7,000-8,000 yr BP; Luetscher et al., 2011) is largely compatible with the possibility of retrieving endogenous DNA from pollen grains. Further confirmations come from the taphonomy and environmental features of the study site. The divalent Ca cation, which is present in the speleothem calcite (CaCO3), is well known for its capacity to adsorb DNA (Cleaves et al., 2011) and to preserve it across time, protecting from the effects of diagenesis (Freeman et al., 2020). Moreover, due to its periglacial context, Milchbach cave experiences temperatures between 0 and 2°C (Luetscher et al., 2011), which are highly favourable for DNA preservation (Lindahl, 1993). The average pH of the drip water, ranging from 7.7 to 8.2 (Luetscher et al., 2011), also favours DNA preservation (Lindahl, 1993). Finally and most importantly, the steady deposition of calcite encapsulates the pollen grains deposited on the speleothem surface, ultimately leading to overall good DNA preservation. With an estimated average deposition rate of 40 μm/yr ( Luetscher et al., 2011), a few years are enough for the calcite to completely cover even very large pollen grains, such as those of Norway spruce (ca. 100 μm). This sealing is a peculiarity of Alpine speleothems where organic compounds such as humic acids, which are known for inhibiting PCR-based molecular analyses (Wilson 1997), have less impact than in soft sediments.

*Analyzed samples and previously characterized pollen assemblages*

**MB-3** is a c. 230 mm long stalagmite which deposited between 9076 ± 140 and 2037 ± 24 a b2k. The stalagmite comprises contrasting petrographies with pollen rich layers characterized by white, opaque calcite fabrics associated with bacterial mediation (Luetscher et al, 2011). The calcite subsample (MB3-208) used for eDNA metabarcoding was deposited between 7883 ± 46 and 7475 ±50 a b2k. Analysis of the adjacent calcite has revealed a pollen concentration of 17 pal/g (the term palynomorphs (pal) is here used to refer to pollen grains and spores). Arboreal pollen (AP) represent 65% of the pollen assemblage, with *Pinus* and *Corylus* being the most frequent; Poaceae are the most abundant non-arboreal pollen (NAP) type.

**MB-4** is a c. 60 mm long stalagmite which deposited between 7455 ± 60 and 4820 ± 30 a b2k. The calcite subsample for eDNA metabarcoding was cut from a section comprising mainly columnar calcite fabrics. It was deposited between 6750 and 6350 ka, i.e. during a phase of glacier advance and thus with only short phases of active cave ventilation (Luetscher et al., 2011). In absence of noticeable pollen-rich layers, this stalagmite has not been analysed for its pollen content.

**MB-5** is a c. 130 mm long stalagmite which deposited between 9076 ± 140 and 2037 ± 24 a b2k. The sample under investigation (MB5-D) deposited between 7601 ± 236 and 7272 ± 158 a b2k. Similar to MB-3 the speleothem petrography comprises distinct white layers, rich in pollen and spores inclusions (i.e. 564 pal/g). The pollen assemblage is dominated by arboreal pollen (81%) with taxa typical of a deciduous forest mixed with Pine, i.e. *Pinus* (25%), *Corylus* (10%), *Quercus* (10%), *Ulmus* (10%), *Tilia* (4%), *Alnus* (8%). Poaceae (9%) and Cyperaceae (5%) are the prevailing taxa among the NAP but a wider range has been reported (Festi et al., 2016).

*Sample preparation and pre-processing*

Sample preparation, DNA extraction and PCR (Polymerase Chain Reaction) amplification were performed in three different rooms, separated by a decontamination room. Every step of the workflow was performed following the strict procedures required for ancient DNA studies to avoid contaminations, e.g. overnight UV-irradiation of rooms before each work session, autoclaving and/or UV-light exposure of the equipment, frequent bleach sterilization of surfaces, use of special equipment (whole disposable coats, hair caps, double gloves, full face masks, etc.). Prior to sample preparation, speleothem specimens were immersed in 3.0% (w/v) sodium hypochlorite (NaClO; approximately equal parts of commercial bleach and water) as an additional step for the elimination of surface contaminations (Kemp & Smith, 2005). Prior to sample preparation, speleothem specimens were immersed in 3.0% (w/v) sodium hypochlorite (NaClO; approximately equal parts of commercial bleach and water) as an additional step for the elimination of surface contaminations (Kemp & Smith, 2005).

*References*

Festi D., Hoffmann D.L., Luetscher M., 2016. Pollen from accurately dated speleothems support alpine glacier low-stands during the Early Holocene. *Quaternary Research*, 86, 45-53.

Cleaves, H. J. et al., 2011. The adsorption of short single-stranded DNA oligomers to mineral surfaces. *Chemosphere* 83, 1560–1567.

Freeman, C. L., Dieudonne, L., Collins, M. J. & Sand, K. K., 2020. Survival of environmental DNA in natural environments: Surface charge and topography of minerals as driver for DNA storage. *BioRxiv*, doi.org/10.1101/2020.01.28.922997.

Kemp, B. M. and Smith, D. G., 2005. Use of bleach to eliminate contaminating DNA from the surface of bones and teeth. Forensic Science International 154, 53–61.

Lindahl, T., 1993. Instability and decay of the primary structure of DNA. *Nature* 362, 709–715.

Luetscher M., Hoffmann D., Frisia S., Spötl C., 2011. Holocene glacier history from alpine speleothems, Milchbach cave, Switzerland. *Earth and Planetary Science Letters*, 302, 95–106.

Providoli, T., & Kuhn, N. J. (2012). Climate and land use effects on forest cover in the Bernese Alps during the 20th century. *Geographica Helvetica*, 67, 15-25.

Wilson, I. G., 1997. Inhibition and facilitation of nucleic acid amplification. *Applied and Environmental Microbiology* 63, 3741.

**Appendix S2. PCR amplification mixture and thermal conditions**

The amplification mixture contained 2.5 U GoTaq Polymerase (Promega, Madison, WI, USA), 1x of GoTaq Flexi Buffer, 3 mM of MgCl2, 0.2 mM each dNTPs, and 0.2 μM of each primer. All PCR amplifications were carried out on a Veriti 96 well thermal cycler (Applied Biosystems), with the following program: 2 min at 95 °C and 50 cycles of 15 s at 95 °C, 15 s at 52 °C and 30 s at 72 °C, followed by 5 min at 72 °C.

During the set-up phase, PCR trials were performed for several samples using different template DNA quantities and dilutions (undiluted; 1:10; 1:20; 1:40), to test for the effect of low concentration and inhibitors. Dilution generally did not improve yield and better results were obtained using high quantities of undiluted DNA. Therefore, the DNA amplification was carried out for all samples in a final volume of 50 μl using 10 μl of undiluted template DNA.

**Appendix S3. Bioinformatics analysis and post-processing of metabarcoding data**

*Bioinformatics analysis*

Raw Illumina data were processed using the MICCA v. 1.7 (Albanese et al., 2015) pipeline and VSEARCH (Rognes et al., 2016). The read pairs were merged using the command *mergepairs* of MICCA, setting a minimum overlap length of 40 and a maximum number of allowed mismatches in the overlap region of 8. The merged reads were trimmed to remove the forward and reverse primers and then filtered setting a minimum length of 20 and an allowed error rate of 1%. Sequence Variants (SVs) were identified using the de novo un-noise command of MICCA that implement the UNOISE3 protocol (Edgar, 2016a) for denoising (i.e. error-correcting) Illumina amplicon reads. The identified SVs represent the Operational Taxonomic Units (OTUs) that were then classified using the Sintax algorithm (Edgar, 2016b) as implemented in VSEARCH. Taxonomic assignment of OTUs was performed by comparison with a custom-made reference database including trnL DNA sequences for most of the Alpine seed plants (Spermatophyta). This local database was built starting from a previously developed trnL database (Leontidou et al., 2018), which was focused on the flora of a south-eastern Italian Alpine region, subsequently complemented with public sequences (NCBI Genbank database; National Center for Biotechnology Information, www.ncbi.nlm.nih.gov/genbank) for the other most frequent species of the Alpine flora (Aeschimann et al., 2004).

*Methods for contamination identification and removal: rationale and multi-step procedure*

After the metabarcoding analysis, a post-processing procedure was implemented to the dataset in order to remove negligible OTUs (i.e. recorded only with negligible No. of reads) and identify major and minor (potential) contaminations. For each OTU, decision for removal (i.e. assignment to contaminants) was taken by means of a multiple approach based on the following parameters: (a) the ratio between the total No. of reads in samples and the total No. of reads in negative controls (S/NC ratio); (b) minimum No. of reads in each single replicate (sample or negative control), with threshold for exclusion set to 100; (c) maximum No. of reads in each single negative control replicate, with a threshold for contamination level set to 100.

S/NC ratio can be viewed as a first way to identify major contaminants and rare OTUs which were recorded mainly in negative controls, while the adopted 100 reads minimum threshold for exclusion was identified from the dataset as clearly separating negligible OTUs (OTUs which were recorded always in traces and often of difficult classification; see Results).

For each OTU, we proceeded as follow.

Step 1. When S/NC ratio was < 1 AND/OR No. of reads in a single sample replicate was always < 100, the OTU was removed from the dataset (negligible).

Step 2. When S/NC ratio was < 1 AND No. of reads in a single sample replicate was > 100 (in at least one replicate), the corresponding taxon was considered as a major contaminant, i.e. a taxon which was recorded with higher abundance in the negative controls than in the samples and with non-negligible quantities. These taxa were considered of potential external origin, that is contaminations in the reagents or from the laboratory environment and each of the corresponding PCR reactions and extractions was investigated for highlighting potential common patterns (and eventually try to roughly identify the step at which contamination occurred; e.g. during DNA extraction or during PCR reaction). Conversely, only when S/NC ratio was > 1 AND No. of reads >100 in at least one sample replicate, the OTU was included in the dataset as true endogenous presence, i.e. DNA for the corresponding plant taxa can be considered to be derived from the speleothem samples with high confidence.

Step 3. The 100 reads maximum threshold for negative controls is used to identify minor and non-systematic contaminations, even when S/NC ratio > 1 (i.e. OTUs/taxa which are present mainly in samples, but that are occasionally recorded with non-negligible quantities also in negative controls).

*Results of bioinformatics analysis and post-processing (contamination identification)*

A total of 4,175,496 pairs of reads were produced from the sequencing and, after filtering and denoising (i.e. removing sequence errors), a total 1,353,919 merged reads were kept: of these, 1,099,305 were present in the samples (average: 61,072 ± 17,484) and 254,614 in the negative controls (average: 6,062 ± 19,062).

MICCA (using the UNOISE3 protocol) identified 42 Operational Taxonomic Units (OTUs). Following our post-processing protocol (see Materials and Methods for details), 26 OTUs were removed for being more represented in negative controls than samples and for being always represented with No. reads < 100 (negligible OTUs: S/NC ratio < 1 AND/OR No. reads in a single sample replicate always < 100). These OTUs were always represented by a very low number of reads per sample (often < 20) and were generally of difficult classification (below to the family level), sometimes corresponding to very short sequences (two of them < 35 bp). One additional OTU, although being relatively abundant in a few samples, was removed for being a short sequence (74 bp) of ambiguous classification (highest similarity in the NCBI Genbank database for Bacteria sequences).

Three OTUs where found to be more represented in negative controls than samples, with No. reads/sample > 100 (non-negligible quantities; see Materials and Methods) and were therefore identified as major contaminants. These OTUs were classified into the following corresponding taxa: Mediterranean cypress (*Cupressus sempervirens* L.), European yew (*Taxus baccata* L.) and alder (*Alnus sp*. Mill.). Specifically, Mediterranean cypress was recorded in three replicates of negative control extraction, all corresponding to the same extraction (E3). Alder was recorded in two replicates of negative controls extraction (Figure 1), all corresponding to the same extraction (E3). These two taxa were absent from all PCR negative controls. European yew was recorded only in one PCR negative control and was never recorded from negative control extraction replicates. In summary, only 1/24 PCR negative controls resulted to be contaminated (with one taxa: European yew), while 3/18 PCR replicates of negative control extraction were affected by contamination from two taxa: Mediterranean cypress and alder. These contaminations derived from the same extraction batch. Notably, the three taxa recorded from negative controls were absent from speleothem samples when subject to the 100 reads threshold (negligible presence). Excluding these three OTUs from the dataset (and the short Bacteria-like sequence), the average number of reads for negative controls dropped down to 77 ± 86, showing no other relevant contaminants except for the three discussed taxa.

The remaining 12 OTUs (length range: 138-159 bp) were almost completely absent from negative controls (< 100 reads; except for the one corresponding to *Fagus sylvatica*, which was recorded with No. of reads > 100 in a few negative control replicates, but always reaching a maximum value of 150 reads) and were therefore considered to be reliable endogenous sequence reads.

The 12 OTUs were classified (VSEARCH Sintax algortithm) into 11 plant taxa (each one reported at the lowest reliable taxonomic level): (1) Pinaceae, *Picea sp*. (spruce); (2) Pinaceae, *Pinus sp*. (pine); (3) *Abies sp*. (fir); (4) Cupressaceae, *Juniperus communis* (common juniper); (5) Urticaceae, *Parietaria sp.* (pellitory; two OTUs); (6) Poaceae (grasses); (7) Fagaceae, *Fagus sylvatica* (European beech); (8) Plantaginaceae (the plantain family); (9) Oleaceae (the olive and ash family); (10) Betulaceae, subfam. Coryloideae; and (11) Caprifoliaceae, subfam. Valerianoideae. For Magnoliophyta, we followed the APG IV classification system: Betulaceae, subfam. Coryloideae includes the following genera: *Corylus* (hazel), *Carpinus* (hornbeam) and *Ostrya* (hop-hornbeam); while Caprifoliaceae, subfam. Valerianoideae includes the following genera: *Valeriana* (valerian), *Valerianella* and *Centranthus* (only considering the study region, i.e. the European Alps).

*References*

Aeschimann, D., Lauber, K., Moser, D. M., Theurillat, J.P., 2004. Flora Alpina: Atlas des 4500 Plantes Vasculaires des Alpes. Belin, Paris.

Albanese, D., Fontana, P., De Filippo, C., Cavalieri, D., Donati, C., 2015. MICCA: a complete and accurate software for taxonomic profiling of metagenomic data. Scientific Reports 5, 1–7.

Edgar, R. C., 2016a. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. BioRxiv doi.org/10.1101/081257.

Edgar, R. C., 2016b. SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. BioRxiv doi.org/10.1101/074161.

Leontidou, K. et al., 2018. DNA metabarcoding of airborne pollen: new protocols for improved taxonomic identification of environmental samples. Aerobiologia 34, 63–74.

Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F., 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ 4, e258