Identification of DNA Viruses in Ancient DNA from Herbarium Samples

Gianluca Grasso, Silvia Rotunno, Régis Debruyne, Lucie Bittner, Laura Miozzi, Roland Marmeisse, Valeria Bianciotto

The manuscript appeared in Methods in Molecular Biology (2024) **2732**:221-234.
doi: 10.1007/978-1-0716-3515-5_15.

Identification of DNA viruses in ancient DNA from herbarium samples

Gianluca Grasso[1-2-3], Silvia Rotunno[3], Régis Debruyne[4], Lucie Bittner[2,5], Laura Miozzi[3], Roland Marmeisse[2-3], Valeria Bianciotto[3]

1. Dipartimento di Scienze della Vita e Biologia dei Sistemi, Università degli Studi of Turin, Viale Mattioli 125, Torino.
2. Muséum National d'Histoire Naturelle, Institut Systématique Evolution, Biodiversité, (ISYEB: UMR7205 CNRS-MNHN-Sorbonne Université-EPHE-UA), 12 rue Buffon, F-75005 Paris, France.
3. Institute for Sustainable Plant Protection (IPSP), National Research Council (CNR), Turin, Italy.
4. Muséum National d'Histoire Naturelle, Archéozoologie, Archéobotanique: Sociétés, Pratiques et Environnements (AASPE: UMR 7209 CNRS-MNHN), CP 56, 43 rue Buffon, F-75005 Paris, France.
5. Institut Universitaire de France, Paris, France

Abstract

Herbaria encompass millions of plant specimens, mostly collected in the 19th and 20th centuries, that can represent a key resource for investigating the history and evolution of phytopathogens. In the last years, the application of high-throughput sequencing technologies for the analysis of ancient nucleic acids has revolutionized the study of ancient pathogens including viruses, allowing the reconstruction of historical genomic viral sequences, improving phylogenetic based molecular dating, and providing essential insight into plant virus ecology. In this chapter, we describe a protocol to reconstruct ancient plant and soil viral sequences starting from highly fragmented ancient DNA extracted from herbarium plants and their associated rhizospheric soil. Following Illumina high-throughput sequencing, sequence data are de novo assembled and DNA viral sequences are selected, according to their similarity with known viruses.

Key words: ancient DNA, DNA viruses, herbaria, museomics

1. Introduction

Plants host numerous viruses belonging to different families, causing, for several of them, important threats to crop plants. Since the causal association between specific disease symptoms and viruses is rather recent in the history of phytopathology, little is known regarding the origin, spread and prevalence of these pathogens in a recent past.

Natural history collections, and more specifically herbaria encompass millions of plant specimens, mostly collected in the 19th and 20th centuries, that can be searched for the presence and prevalence of specific pathogens and pests. Straightforward visual observation of herbarium plants has been carried out in the case of pathogens producing unambiguous symptoms as in the case of anther smut fungi infecting *Silene* flowers [1] or of the Horse-chestnut leaf miner [2]. Alternatively, plant pathogens can be identified among degraded (ancient) aDNA molecules extracted from infected collection specimens. This approach has been reported for fungal [3] or Oomycete [4] eukaryotic pathogens and has led to the reconstruction of an entire historical genomic sequence of the Citrus canker bacterial agent [5].

A similar approach has been reported to investigate DNA virus evolution and prevalence in past human populations. For example, the systematic search for Hepatitis B virus sequences in ancient human DNA sequences has highlighted the temporal and spatial dispersal of different virus genotypes during the last ca 10,000 years [6]. Alternatively, the systematic search for any known pathogen

sequences in the DNA extracted from 5th-8th century human remains from a common settlement in Germany identified several cases of co-infection by hepatitis B, smallpox viruses and Parvovirus B, highlighting the poor health status of this local human population [7].

In the case of plants, small RNA extracted from 90-year-old Cassava herbarium specimens allowed identification of a Cassava Mosaic Geminivirus sequence that was used to estimate its evolutionary rate [8], showing how analyses of ancient viral genomic sequence data obtained from historical samples can substantially improve phylogenetic based molecular dating studies. Historical data on viral sequences can also provide essential insight into plant virus ecology, as demonstrated by the analysis of Barley yellow dwarf luteovirus sequences obtained from RNA extracted from herbarium specimens dating from the end of 19th century to the first half of the 20th century [9].

In this chapter, we describe a protocol for the extraction of DNA from herbarium plants and their associated rhizospheric soil. Following Illumina high-throughput sequencing of the highly fragmented ancient DNA and its *de novo* assembly, assembled DNA viral sequences are identified allowing to reconstruct the corresponding ancient plant and soil viromes.

2. Materials

2.1 Herbarium plant and soil sampling

- 2 ml sterile centrifuge tubes.
- Tweezers.
- Sterile razor blades.

2.2 DNA extraction

All experiments are performed in dedicated spaces within a cleanroom with positive pressure to prevent contamination from the outside environment, wearing laboratory coveralls and with specific cleaning procedures of equipment and spaces (with 2.6% bleach and/or 20 minutes UV-crosslinking at 256 nm).

- High Pure Viral Large Volume extraction kit (Roche Diagnostics).
- DNEasy Powersoil Pro kit (QIAGEN).
- DNA low-binding 1.5 and 2 mL microtubes.
- Sediment Lysis Buffer (SLB_{conc}): 20 mM Tris-HCl pH 9.0, 10 mM Calcium chloride, 100 mM DTT, 0.5% w/v SDS, 6.25% w/v Polyvinylpyrrolidone.
- 20 mg/mL Proteinase k solution.
- Binding Buffer (BB): 5 M Guanidinium HCl, 120 mM Sodium acetate pH 5.2, 0.05% v/v Tween-20, 40% v/v isopropanol (*see* **Note 1**).
- Elution buffer (EBT): EB buffer (from QIAGEN DNEasy Powersoil Pro kit) supplemented with 0.05% v/v Tween-20.
- Molecular Biology Grade ethanol.
- Nuclease-Free water.
- Qubit dsDNA High Sensitivity kit (Invitrogen).
- Mixer Mill (with microtube accessories).
- Hybridization oven or other heating device.
- Rotary mixer/shaker.
- Large volume refrigerated centrifuge (up to at least 3,000 g).
- Microcentrifuge (up to at least 15,000 g).
- Fluorometer (Invitrogen Qubit or equivalent).
- 15 mL Falcon tubes.

2.3 Library preparation and sequencing

- Oligonucleotides IS1, IS2 and IS3 to make-up truncated P5 and P7 adapters [10].
- Custom (7 nucleotides) indexed Illumina P5 and P7 Primers [11].

- NEBNext End-Repair module (New England Biolabs).
- NEBNext Quick Ligation Module (New England Biolabs).
- *Bst* polymerase large fragment (New England Biolabs).
- Minelute PCR purification kit (QIAGEN).
- PB_{acidic}: PB buffer (from Minelute PCR purification kit) supplemented with 60 mM Sodium Acetate pH 5.3.
- SsoAdvanced Universal SYBR supermix (Bio-Rad).
- Qubit dsDNA High Sensitivity kit (Invitrogen).
- NucleoMag cleanup and size selection beads kit (Macherey-Nagel).
- Magnet for Solid Reversible Phase Immobilization (SPRI) purification using 1.5 mL microtubes (Invitrogen Dynamag or equivalent).
- Microcentrifuge (up to at least 15,000 g).
- Incubation block with heated lid (Eppendorf thermomixer or equivalent).
- Real-Time PCR thermocycler.
- Fluorometer (Invitrogen Qubit or equivalent).
- Device for capillary electrophoresis of DNA fragments (Labchip HT Perkin-Elmer with dedicated consumables/reagents or equivalent).
- Illumina sequencer and dedicated consumables and chemistry.

2.4 Bioinformatics analyses

- fastQC (v. 0.12.0) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).
- fastp (v. 0.23.1) (https://github.com/OpenGene/fastp).
- BWA (v. 0.7.17) (https://bio-bwa.sourceforge.net/).
- samtools (v. 1.14) (http://www.htslib.org/).
- SPAdes (v. 3.15.1) (https://github.com/ablab/spades).
- QUAST (v. 5.2.0) (https://quast.sourceforge.net/).
- CAP3 (VersionDate: 02/10/15) (https://doua.prabi.fr/software/cap3).
- DIAMOND (v. 2.0.15) (https://github.com/bbuchfink/diamond).
- MEGAN6 Community Edition (v. 6.24.23) (https://software-ab.cs.uni-tuebingen.de/download/megan6/welcome.html).

3. Methods

3.1 Herbarium plant and soil sampling

The protocol was successfully implemented on DNA extracted from roots and associated rhizospheric soil (**Fig. 1**) of different plant species stored in herbaria for 115-120 years. For each plant specimen, about 20 *ca* 0.5 cm-long root fragments are collected using sterile fine tweezers and scalpel blades. Soil particles aggregated around the roots (between 10-100 mg per plant specimen (**Fig. 1**)), are detached from the roots and collected using tweezers. Root and soil samples are transferred into sterile tubes and stored at room temperature before DNA extraction.

Fig. 1. Root system of herbarium specimen of *Triticum aestivum* collected in 1856 before (A) and after (B) the sampling of fragments of root and rhizospheric soil.

3.2 DNA extraction

This protocol is a modified version of [12]'s 'Cold Spin Extraction' method (*see* **Note 2**).

1. Preheat the SLB_{conc} buffer at 50°C.
2. For each sample, weigh up to 250 mg of soil, or up to 50 mg of root in a 2 mL microtube. Prepare one extraction blank (no soil or root sample) for every series of extractions (i.e., for a total of 16 extractions, consider 15 samples and one extraction blank).
3. Transfer each sample in a single PowerBead tube provided in the DNEasy Powersoil Pro kit (already containing garnet beads and 750 µL of 181 mM $NaPO_4$ and 121 mM of guanidinium isothiocyanate).
4. Rinse each sample tube with 500 µL of SLB_{conc} solution to collect the leftover of soil/root samples attached to the tube walls. Transfer the suspension to the corresponding PowerBead tube (making up a final volume of 1250 µL of digestion solution).
5. Homogenize the samples by a 5 min agitation of the PowerBead tubes in a mixer mill at a 25 beats/s mixing frequency. If necessary, repeat this step once or twice to obtain a homogeneous suspension.
6. Add 16 µL of 20 mg/mL Proteinase K solution (for a final concentration of approx. 0.25 mg/ml in the digestion solution).
7. Set the PowerBead tubes in a rotating shaker (speed 16-18 rotations/min) for an overnight digestion (20-24 h) in a hybridization oven at 35°C in the dark. Ensure that the digestion solution, sample, and PowerBeads are moving at each oscillation.
8. Remove PowerBead tubes from the oven and centrifuge 5 min at 10,000 g (the maximum speed recommended for PowerBead tubes).
9. Transfer supernatants to a DNA low-binding 2 mL tube and freeze them at -20°C. Recover as much digest solution as possible at this step. Tiny portions of the pellet can be pipetted without consequences (they will be eliminated on step 13). The extraction protocol can be stopped after this step (usually until the next day).
10. Preheat the EBT buffer at 30°C.
11. Thaw the digested supernatants and centrifuge them briefly. Pipet each of them into a 15 mL Falcon tube filled with 13 mL of BB.
12. Spin the Falcon tubes at 3,000 g for a minimum of 3 hours (up to overnight) in a refrigerated centrifuge at 4°C.
13. Decant the supernatant after centrifugation, not disturbing the dark pellet at the bottom of the tube and add it to a high-volume silica column (High Pure Extender Assembly).
14. Centrifuge the high-volume silica columns at 1,000 g for 2 min. In case the entire volume has not passed through, renew the centrifugation step.
15. Detach the silica column from the assembly and put it in a 2 mL collection tube.
16. Add 500 µL of the Inhibitor Removal Buffer of the High Pure Viral Nucleic Acid Large Volume to the column and centrifuge at 3,000 g for 1 min at room temperature.
17. After centrifugation, transfer the column to a new collection tube and add 450 µL of the Wash Buffer (High Pure Viral Nucleic Acid Large Volume) to the column. Centrifuge at 6,500 g for 1 min at room temperature.
18. Repeat step 17 for a second wash.
19. Transfer the column to a new collection tube and centrifuge to dry the silica columns at 15,000 g for 1 min.
20. Elute the DNA off the silica column with 25 µL EBT. Centrifuge at 15,000 g for 1 min.
21. Repeat step 20 for a total elution volume of 50 µL.
22. Total DNA estimate is performed via fluorometric quantitation of 1 µL of each extract.
23. Store the extracted DNA at -20°C until processed into libraries and at -80°C for long-term storage.

3.3 Library preparation and sequencing

This protocol is a modified version of the dsDNA library preparation method by [11], implementing the double indexing strategy of [12]. All specimen extracts and blank extractions are to be processed in the same way.

3.3.1 Truncated adapter preparation

1. Incubate the 200 µM truncated P5 and P7 mixes 10 min at 95°C and let them cool down to 12°C slowly at -0.1°C /s.
2. Mix equal amounts of truncated P5 and P7 to obtain a 100 µM P5+P7 adapter mix. Vortex mix and centrifuge briefly.
3. Aliquot the adapter mix and freeze. Use each aliquot only once and then discard.

3.3.2 Universal library preparation

1. Prepare an End-Repair reaction premix containing: 5 µL of 10X NEBNext End-Repair buffer, 2.5 µL of NEBNext End-Repair Enzyme mix, and 27.5 µL of Nuclease-Free water per library.
2. Use 35 µL of the premix for each individual library. Add 15 µL of DNA extract (it is not necessary to use a specific quantity of DNA for library construction) for a final reaction volume of 50 µL.
3. Incubate 15 min at 25°C. Transfer at least 5 min at 4°C.
4. Proceed with a Minelute purification using 6 volumes (300 µL) of PB_{acidic} for one volume of repaired DNA. Elute the DNA off the silica column twice with 15 µL of EBT for a final volume of approximately 29 µL.
5. Dilute a P5+P7 adapter mix aliquot to 20 µM. Make up a ligation premix containing: 10 µL of 5X NEBNext T4 quick ligation buffer, 7.5 µL of Nuclease-Free water, and 2.5 µL of the diluted adapter mix.
6. Add 20 µL of that premix to each library. Vortex mix and centrifuge briefly.
7. Add 1.5 µL of NEBNext T4 ligase to each individual library for a reaction volume of 50 µL (final concentration of 12 U of ligase / µL and 1 µM of P5+P7 adapter mix).
8. Incubate 90 minutes at 22°C.
9. Proceed with a Minelute purification using 5 volumes (250 µL) of PB_{acidic} for binding DNA. Elute each library twice with 15 µL of EBT for a final volume of approximately 29 µL.
10. Prepare a fill-in reaction premix with the following reagents: 4 µL of Nuclease-free water 4 µL of 10X Thermopol buffer, 1 µL of 10 mM dNTPs, and 2 µL of *Bst* Polymerase large fragment (16 U per reaction).
11. Dispense 11 µL of the fill-in premix into each library tube (for a total reaction volume of 40 µL containing 1X Thermopol buffer and 250 µM of each dNTP). Vortex and quick spin.
12. Incubate 20 min at 37°C and then transfer at 80°C for another 20 min incubation.
    Store libraries at -20°C.

3.3.3 Library indexing and characterization

1. Amplify 10 µL of each library using a unique pair of custom P5 and P7 indexing primer in a PCR reaction using 1X SsoAdvanced Supermix and 500 nM of each primer. Due to various levels of inhibitions between samples, the total reaction volume can be adjusted from 40 (by default) up to 100 µL.
2. Perform the PCR amplification of the libraries in Real-Time with the following conditions: 2 min of hot start denaturation at 98°C, followed by 20 cycles of 10 s at 98°C, 20 s at 60°C and 20 s at 72°C.
3. Address the level of fluorescence for each library separately at each cycle: remove individual libraries from the thermocycler when they show a start in the plateau phase of the amplification.
4. Quantify each amplified library via fluorometric quantification.
5. Dilute the library according to the requirements of your capillary electrophoresis equipment. For a Labchip HT characterization, libraries are diluted to a 0.2-2.0 ng/µL range and evaluated with the NGS 3k chip and consumables.
6. Based on both fluorometric quantification and electrophoretic size distribution of the fragments, calculate the molarity of each individual library.

3.3.4 Pooling and Sequencing

1. Make an equimolar pool of the libraries to sequence, based on their estimated molarity.
2. Make a SPRI purification of the pool using 1.25X volume of NucleoMag purification beads.
3. Estimate the molarity of this final pool (via fluorometric and electrophoretic analyses) and prepare the Illumina sequencing accordingly.
4. Perform the shotgun sequencing of the pooled libraries on the relevant Illumina platform/chemistry. In our case, we sequenced the pool of DNA libraries made from either the roots or the soils herbarium samples on a Novaseq 6000 platform in Paired-End sequencing (2*50 bp). We aimed at generating approximately 20 million reads for each library.

3.4 Bioinformatics analyses

3.4.1 Preprocessing (quality control, trimming and merging)

1. Check quality of the reads using FastQC [13] and visualize the HTML reports generated by FastQC using a web browser.
   >fastqc path_to_sample_name.fq.gz
2. Trim the reads to remove primer sequences and short reads and merge paired reads with Fastp [14] using the following command. For the trimming, use the default commands except for the *-l* option (length of reads to trim) which has to be set to 25 nucleotides and the *--overlap_len_require* option (overlap needed between different pair reads, default value 30) which is here lowered to 10 nucleotides. These parameters are here adapted to the short length of the reads, i.e., usually less than 100 bp for ancient/historical DNA obtained from herbarium samples. As *-l* value was chosen 25 nucleotides to facilitate merging, mapping and assembly, in order to achieve high accuracy, since the trimmed sequences are longer than the default parameter of fastp (15 nucleotides) (*see* **Note 3**).
   >fastp -h -g -l 25 --adapter_fasta list_adapter.fasta --overlap_len_require 10 -m -I sample_name_R1.fq -I sample_name_R2.fq -o sample_name_R1_trim.fastq.gz -O sample_name_R2_trim.fastq.gz --merged_out sample_name_trim_merge.fastq.gz --unpaired1 sample_name_R1_trim_unpaired.fastq.gz --unpaired2 sample_name_R2_trim_unpaired.fastq.gz

3.4.2 Removal of reads mapping to reference plant genomes

Sequences are then mapped to the corresponding plant reference genome (using available genomes present in the NCBI genomes database, https://www.ncbi.nlm.nih.gov/home/genomes/) to remove the plant nuclear, mitochondrial, and plastid genomes from the metagenomic datasets. The tool BWA aln [15], appropriate for processing short sequences, is used for mapping Illumina reads to the plant genomes. Sequences that do not map to plants are then extracted using Samtools [16]. The hypothesis is that they correspond to the host plant microbiome (i.e., eubacteria, archaea, fungi, viruses).

1. Index the reference plant genome downloaded from the NCBI genomes database.
   >bwa index -a bwtsw reference_genome.fasta
2. Align the merged sequences to the plant reference genome.
   >bwa aln reference_genome.fasta sample_name_trim_merge.fastq.gz> sample_name_trim_merge.sai
3. Convert alignment in SAM format.
   >bwa samse reference_genome.fasta sample_name_trim_merge.sai sample_name_trim_merge.fastq.gz >sample_name_trim_merge.sam
4. Convert SAM file into BAM file.
   >samtools view -b -S sample_name_trim_merge.sam >sample_name_trim_merge.bam
5. Sort the BAM file.
   >samtools sort sample_name*_trim_merge.bam -o sample_name_trim_merge_sort.bam*
6. Index the BAM file.
   >samtools index sample_name*_trim_merge_sort.bam*
7. Calculate final statistics.

>samtools flagstat sample_name_trim_merge_sort.bam > sample_name_statistic_mapping.txt

8. Select unmapped sequences.
   >samtools view -b -f 4 sample_name_trim_merge_sort.bam > sample_name_trim_merge_sort_unmapped.bam
9. Recovery fastq files for the unmapped sequences.
   >samtools fastq sample_name_trim_merge_sort_unmapped.bam > sample_name_unmapped.fq
10. To exclude any contamination from sample manipulation in the herbaria, during sampling and analysis, repeat the same alignment procedure on the unmapped reads using the most recent version of the human genome (GCF_000001405.26) as reference genome.

3.4.3 *De novo* assembly

1. Perform the *de novo* assembly of the reads using the software SPAdes with default parameters (*see* **Note 4**).
   > spades.py -s sample_name_unmapped.fq.gz -o spades_results_sample_name/
2. Evaluate the assembly (i.e., number of assembled scaffolds, N50, N90, scaffold length distribution) with the tool QUAST, using default parameters (*see* **Note 5**).
   > quast.py -k spades_results_sample_name/contigs.fasta spades_results_sample_name/scaffolds.fasta --single clean_reads.fq.gz -o quast_results_sample_name/
3. Collapse redundant scaffolds with CAP3 using default parameters.
   > cap3 spades_results_sample_name/scaffolds.fasta -x sample_name
4. Concatenate in one file the obtained contigs and singlets sequences.
   > cat spades_results_sample_name/scaffolds.fasta.sample_name.contigs spades_results_sample_name/scaffolds.fasta.sample_name.singlets > sample_name_cap3.fasta

3.4.4 Identification of viral sequences

To identify viral sequences, contigs are aligned against a protein reference database using the DIAMOND tool. A non-redundant protein database needs to be downloaded and formatted according to the following steps.

1. Download the non-redundant (nr) protein database from the NCBI website.
   > wget ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz
2. Download the file that maps NCBI protein accession numbers to taxon ids form the NCBI website.
   > wget ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/accession2taxid/prot.accession2taxid.FULL.gz
3. Download the nodes.dmp and names.dmp files from the NCBI taxonomy website.
   > wget ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/taxdmp.zip
4. Create a DIAMOND-formatted database file (see **Note 6**).
   > diamond makedb --in nr.gz -d nr.dmnd --taxonmap prot.accession2taxid.FULL.gz --taxonnodes nodes.dmp --taxonnames names.dmp
5. Run DIAMOND in blastx-mode and save the output in a tabular file (*see* **Note 7**).
   > diamond blastx -f 6 –sensitive –quiet -d nr.dmnd -q sample_name_cap3.fasta -o sample_name.csv
6. Run DIAMOND in blastx-mode and save the output into DIAMOND alignment archive (DAA) supported by MEGAN.
   > diamond blastx -d nr.dmnd -q sample_name_cap3.fasta -a sample_name.daa

3.4.5 Visualize DIAMOND results

The taxonomic distribution of the contigs annotated with DIAMOND can be then visualized with the tool MEGAN6.

1. Download the file mapping NCBI-nr accessions “megan-map-Feb2022.db.zip” to taxonomic and functional classes (NCBI, GTDB, EC, eggNOG, InterPro2GO, SEED) from *https://software-ab.cs.uni-tuebingen.de/download/megan6/welcome.html* and unzip it.
2. Open MEGAN6, click on the “File” menu, and select “Import From BLAST”. In the tab “Files”, at the point n.1, import your DAA file, and in the tab “Taxonomy”, click on “Load MeganMapDB mapping file” and load the file “megan-map-Feb2022.db”. Click on “Apply” to visualize your data (**Fig.2**).
3. In order to extract viral contig sequences, select the node of interest, open the drop-down menu by clicking the right mouse button and select the “Extract reads” option.

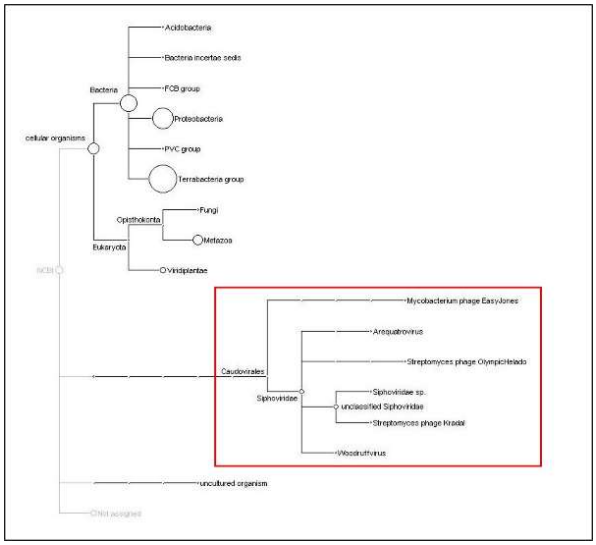

Fig. 2. Visualization of DIAMOND results using MEGAN6. The red square indicates viral taxonomic nodes.

4. Notes

1. When preparing the BB buffer, note that the reaction is endothermic and that Guanidinium will only go into solution once the nuclease-free water is added to the mix.
2. DNA extraction was performed using the former PowerBead tubes from the DNA Powersoil kit (QIAGEN). It can readily be applied using the current version of the DNEasy Powersoil Pro kit (QIAGEN) by modifying step 3 in section 3.2. The tubes in the new kit contain only the beads; it is thus necessary to add, before step 4, 750 µL of the CD1 solution (from the same kit).
3. Use *-h* option for obtaining result report on HTML, *-g* for removing polyG, *-m* for merging the reads, *-adapter_fasta* for indicating the fasta list of adapters (that contained adapters and their reverse complement). As indicated in the main text, the *-overlap_len_require* option was also introduced to reduce the overlap length during merging (from 30, default, to 10 nucleotides). Finally, with *-q* option, it is also possible to eliminate all the reads below a certain quality threshold. In our case, the raw data (in the FastQC reports) showed high quality, and the *-q* option was not used.
4. Use *-t* option for setting the number of threads and *-m* option to set memory limit in Gb, according to the characteristics of your IT infrastructure.
5. Use *-t* option for setting the number of threads and *-m* option to set memory limit in Gb, according to the characteristics of your IT infrastructure. Use *–silent* option if you do not want printed on screen detailed information about each step; the information will be stored anyway in the log file.
6. Use *-p* option for setting the number of threads according to the characteristics of your IT infrastructure. By default, DIAMOND uses all available threads.
7. Use *-p* option for setting the number of threads according to the characteristics of your IT infrastructure. By default, DIAMOND uses all available threads. Use *-e* option to set the e-value threshold to report an alignment. Use *-f* option to format the output file. Use *–sensitive* option to enable the sensitive mode designed for full sensitivity for hits of >40% identity. Use *-k* option

to set the maximum number of target sequences per query to report alignments for. Use –*quiet* option to disable all terminal output.

Acknowledgements

Work on ancient plant-associated microbiota was supported by the Muséum National d'Histoire Naturelle grant ATM 2021 (HoloHerbier). GG was supported by a PhD grant from the University of Turin and a mobility grant from the French-Italian University (program da Vinci 2022). We would like to thank the curators and staff of the herbaria of the Muséum National d'Histoire Naturelle in Paris, of the Lyon botanical garden and of the Universities of Montpellier, Lyon, and Clermont-Auvergne to give us access to herbarium specimens for implementation of the protocols described in this chapter. We thank the "Plateau de Paléogénomique et Génétique Moléculaire" (P2GM, MNHN, Paris), for granting access to their resources and facility.

References

1. Antonovics J, Hood ME, Thrall PH, et al. 2003. Herbarium studies on the distribution of anther-smut fungus (Microbotryum violaceum) and Silene species (Caryophyllaceae) in the eastern United States. Am J Bot. Oct; 90(10):1522-31.
2. Lees DC, Lack HW, Rougerie R et al. 2011. Tracking origins of invasive herbivores through herbaria and archival DNA: the case of the horse-chestnut leaf miner. Frontiers in Ecology and the Environment, 9: 322-328.
3. Bradshaw M, Tobin PC. 2020. Sequencing Herbarium Specimens of a Common Detrimental Plant Disease (Powdery Mildew). Phytopathology. 110(7):1248-1254.
4. Yoshida K, Schuenemann VJ, Cano LM, et al. 2013. The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine. Elife. 2:e00731.
5. Campos PE, Groot Crego C, Boyer K, et al. 2021. First historical genome of a crop bacterial pathogen from herbarium specimen: Insights into citrus canker emergence. PLoS Pathog. 17(7):e1009714.
6. Kocher A, Papac L, Barquera R, et al. 2021. Ten millennia of hepatitis B virus evolution. Science. 374(6564):182-188.
7. Bonczarowska JH, Susat J, Mühlemann B, et al. 2022. Pathogen genomics study of an early medieval community in Germany reveals extensive co-infections. Genome Biol. 23(1):250.
8. Rieux A, Campos P, Duvermy A, et al. 2021. Contribution of historical herbarium small RNAs to the reconstruction of a cassava mosaic geminivirus evolutionary history. Sci Rep. 11(1):21280.
9. Malmstrom CM, Shu R, Linton EW, et al. 2007. Barley yellow dwarf viruses (BYDVs) preserved in herbarium specimens illuminate historical disease ecology of invasive and native grasses. Journal of Ecology, 95: 1153-1166.
10. Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc. 2010(6):pdb.prot5448.
11. Kircher M, Sawyer S, Meyer M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. 40(1):e3.
12. Murchie TJ, Kuch M, Duggan AT, et al. 2021. Optimizing extraction and targeted capture of ancient environmental DNA for reconstructing past environments using the PalaeoChip Arctic-1.0 bait-set. Quat. Res. 99:305-328.
13. Andrews, S. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
14. Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 34(17):i884-i890.
15. Li H. and Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754-1760.
16. Danecek P, Bonfield JK, Liddle J, et al. 2021. Twelve years of SAMtools and BCFtools. Gigascience, 10(2):giab008.