Mining ancient microbiomes using selective enrichment of damaged DNA molecules

Clemens L. Weiß[a], Marie-Theres Gansauge[b], Ayinuer Aximu-Petri[b], Matthias Meyer[b], Hernán A. Burbano[a]#

[a]Research Group for Ancient Genomics and Evolution, Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany

[b]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany

Running Head: Uracil selection for validation of ancient metagenomes

#Address correspondence to Hernán A. Burbano, hernan.burbano@tuebingen.mpg.de

Abstract

The identification of bona fide microbial taxa in microbiomes derived from historical samples is complicated by the unavoidable mixture between DNA from ante- and post-mortem microbial colonizers. One possibility to distinguish between these sources of microbial DNA is querying for the presence of age-associated degradation patterns typical of ancient DNA (aDNA). The presence of uracils, resulting from cytosine deamination, has been detected ubiquitously in aDNA retrieved from diverse sources, and used as an authentication criterion. Here, we employ a library preparation method that separates molecules that carry uracils from those that do not for a set of samples that includes Neandertal remains, herbaria specimens and archaeological plant remains. We show that this method facilitates the discovery of authentic ancient microbial taxa, as it amplifies degradation patterns that would otherwise be difficult to detect in sequences from diverse microbial mixtures.

Importance

The utility of DNA from historical specimens is being recognized in a growing number of fields, ranging from human, animal and plant genetics, to microbiology and epidemiology of infectious diseases. Providing positive evidence for the authenticity of such ancient DNA from diverse sources is instrumental for all studies that make use of this resource. This is especially challenging when studying ancient microbes, due to their high genetic diversity and the incompleteness of reference databases. The method we employ and characterize here aids this process through the selective enrichment of molecules that carry signatures of age-associated degradation.

Introduction

DNA retrieved from historical or ancient samples is a complex mixture of molecules that contains not only endogenous host DNA, but also DNA from microorganisms that were present ante-mortem or that colonized the tissue post-mortem (1). Therefore, all ancient DNA (aDNA) shotgun sequencing projects are metagenomic in nature. While earlier aDNA research has mostly focused on the evolution of animals and plants (2, 3), a growing number of studies are now centering on the identification and characterization of ancient pathogens and microbiomes (4). Ancient microbes permit the replacement of indirect inferences about the past with direct observations of microbial genomes through time. In the pathogen field, it has been possible to identify causal and/or associated agents of historical plant and animal disease outbreaks, as well as their spreading patterns throughout both space and time (e.g. (5, 6)). Another challenging endeavour is the characterization of shifts in composition of microbial communities over time. For example, dental calculus from hominids has been exploited as a source of ancient microbiomes and analyzed in the context of diet and lifestyle changes (7–9), whereas coprolites have been used to investigate ecological interactions between animals and microorganisms (10). However, this approach is at its beginnings and the influence of major selective pressures on microbiome evolution remains to be explored.

A major challenge for the study of aDNA in general, and ancient microbiomes in particular, is the presence of contaminating exogenous DNA, which makes distinction between *bona fide* ancient microbiome sequences and those of recent origin crucial. One of the most typical features of aDNA is the presence of uracils (Us) that originate from post-mortem deamination of cytosines (Cs), especially in single-stranded overhangs at molecule ends (11). Uracils are read as thymines (Ts) by most DNA polymerases, which generates a characteristic increase in C-to-T substitutions at the end of aDNA sequences ((11), Figure 1D and 2A). The presence of such C-to-T substitutions can be used as evidence for the authenticity of DNA sequences retrieved from historical material (12–14).

Recently, a single-stranded library preparation method (U-selection) was developed, which allows physical separation of uracil-containing molecules from non-deaminated ones (15). In U-selection all library molecules are initially immobilized on streptavidin beads, to which molecules without uracils remain attached (U-depleted fraction), while uracil-containing molecules (originally deaminated) are released into solution (U-enriched fraction). U-selection was originally developed with the aim of increasing the amount of ancient hominid DNA (e.g. Neandertals) from a background of present-day human and microbial DNA (15). However, the method seems to be specially suited to study microbiomes, due to the inherent difficulty to authenticate their ancient origin. This complication arises from the fact that microbes can colonize tissues at different times, resulting in different levels of deamination of microbial DNA in historical samples. Although sequences that carry terminal C-to-T substitutions can be selected *in silico* (16, 17), there are two factors that could hinder this approach. Firstly, low levels of deamination will reduce the number of molecules suitable for selection *in silico*. Secondly, high sequence divergence between samples and reference genomes can mask age-associated deamination signals thereby hinder authentication.. Consequently, enriching for deaminated molecules during library preparation is fundamental to tackling these problems. As a proof-of-principle experiment, we used here U-selection in combination with taxonomic binning of Illumina sequenced reads to characterize the microbiomes of Neandertal bones (~39,000 years old), herbaria specimens (between 41 and 279 years old) and plant archaeological remains (~2,000 years old) (Table 1).

Table 1. Provenance of herbarium specimens and archaeological remains

ID	Country of origin	Age	Species	Reference*
KM177500	UK	171+	*Solanum tuberosum*	1
KM177497	UK	170+	*Solanum tuberosum*	1
BM000815937	UK	279+	*Solanum lycopersicum*	2
BH0000061459	USA	119+	*Arabidopsis thaliana*	3
OSU13900	USA	82+	*Arabidopsis thaliana*	4
NY1365364	USA	127+	*Arabidopsis thaliana*	5
NY1365375	USA	119+	*Arabidopsis thaliana*	5
CS5	USA	1852++	*Zea mays*	6
CS6	USA	Undated	*Zea mays*	6
CS20	USA	1881++	*Zea mays*	6
El Sidrón 1253	Spain	39,000++	Neanderthal	7
Vindija 33.17	Croatia	Undated	Neanderthal	8
Vindija 33.19	Croatia	Undated	Neanderthal	8

*[1]Kew Royal Botanical Gardens; [2]Natural History Museum, London; [3]Cornell Bailey Hortorium; [4]Ohio State University Herbarium; [5]New York Botanical Garden; [6]Turkey Pen Shelter, UTAH, USA; [7]El Sidrón Cave, Spain; [8]Vindija Cave, Croatia.

+Calculated from collection dates (in years).

++B.P. (Before present years)

Results and Discussion

Our experiments were motivated by the previous observation that in some Neandertal samples, e.g. from El Sidrón, Spain, the proportion of Neandertal DNA fragments remains unchanged in both the U-depleted and U-enriched fractions, whereas in others, from Vindija Cave, Croatia, this proportion increased in the Uracil-enriched fraction (15). It was hypothesized that the latter effect could have been due to differences in deamination, and hence in age, between Neandertal- and microbial-derived DNA fragments. To explore this effect further, we re-analyzed the previously generated Neanderthal sequence data from both sites by performing taxonomic binning of reads derived from the U-depleted and U-enriched fractions, instead of aligning them only to the human reference genome, as had been done previously. Reads aligning to the two most abundant bacterial phyla (Actinobacteria and Proteobacteria) from the Vindija Neandertals were enriched in the U-depleted fraction, while hominid reads were enriched in the U-enriched fraction (Figure 1A). This is in accordance with a previous study that reported absence of DNA damage in Actinobacteria derived from a Neandertal bone from Vindija cave (18). In contrast, in reads obtained from the El Sidrón Neandertals, we found enrichment of both hominid and Actinobacteria reads in the U-enriched fraction, whereas Proteobacteria reads were enriched in the U-depleted fraction. (Figure 1B). Overall, bacteria-derived reads were dominated by the Actinobacteria *Streptosporangium roseum* (Figure 1C), which showed almost 50% deamination at the first base in the U-enriched fraction (Figure 1D), suggesting its ancient origin. The analysis of reads derived from Neandertal bones illustrates how U-selection permits distinguishing between ancient bacteria enriched in the U-enriched fraction and more recent colonizers enriched in the U-depleted fraction.

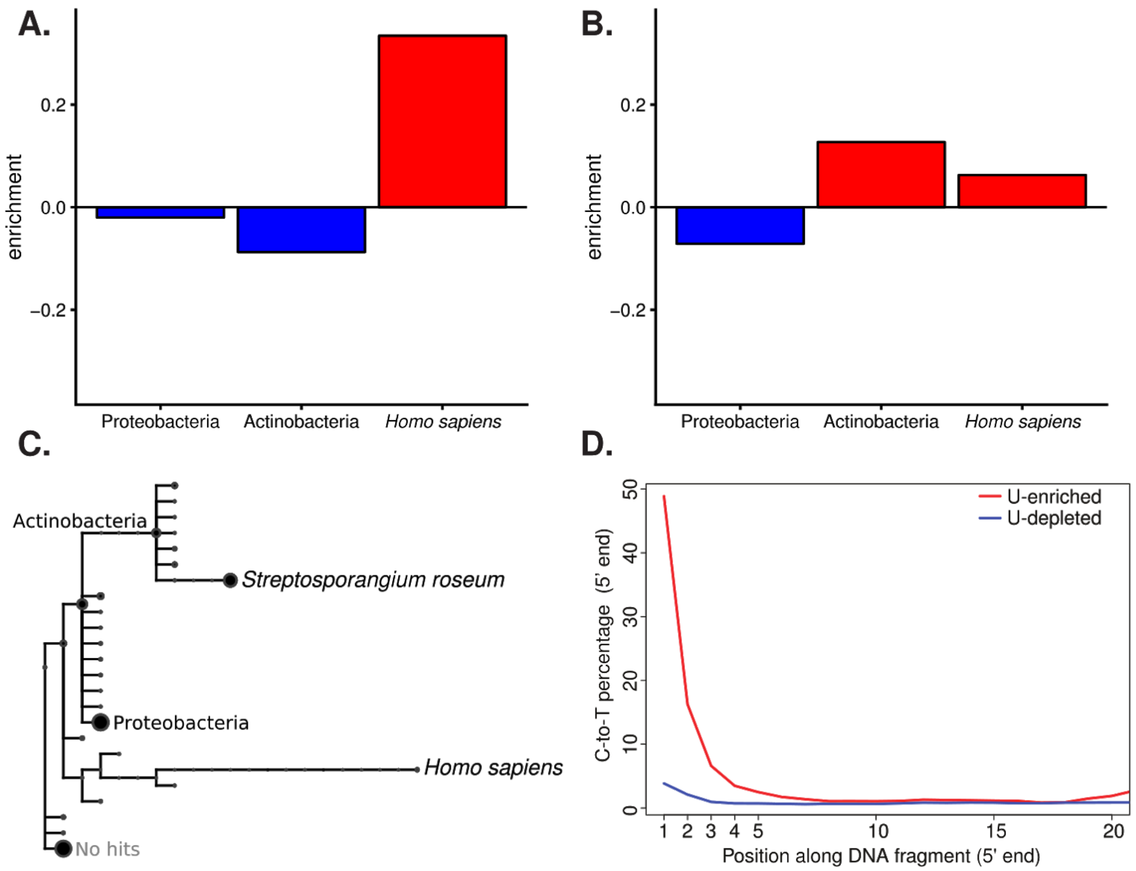

Figure 1. Relative enrichment, taxonomic assignment and substitution profiles of Neandertal-derived U-selected libraries. **A.** Relative enrichment (number of reads) in the U-enriched relative to the U-depleted fraction from Vindija Neandertal assigned to the phyla Actinobacteria and Proteobacteria, as well as to *Homo sapiens*. **B.** Relative enrichment (number of reads) in the U-enriched relative to the U-depleted fraction from Sidrón Neandertal assigned to the phyla Actinobacteria and Proteobacteria, as well as to *Homo sapiens*. **C.** Taxonomic tree of reads from Sidrón Neandertal assigned to different taxonomic levels. The size of the circle represents the amount of reads assigned to a particular part or the taxonomy. Assignments to the phyla Actinobacteria and Proteobacteria, as well as the species *Streptosporangium roseum* and *Homo sapiens* are named in the taxonomic tree. **D.** Cytosine to Thymine substitutions at the 5’ end of reads aligned to *S. roseum* from the Sidrón Neandertal U-selected library (U-enriched and U-depleted fractions).

In order to further evaluate the performance of U-selection in characterizing microbial communities, we selected a set of plant samples (both herbaria specimens and archaeological remains) with low levels of deamination. We extracted DNA from plant samples and generated libraries using both a regular double-stranded (ds) approach (19), and U-selection (15). Sequences from the dsDNA libraries were then used as a baseline to evaluate depletion and enrichment of uracil-containing molecules (Figure 2A). U-selection successfully enriched for deaminated molecules in all plant samples, as it is manifest in the much higher levels of deamination present in the U-enriched fraction compared with the dsDNA libraries and the U-depleted fraction (Figure 2A-B). The plant samples showed substantial variation in the content of endogenous DNA (2.8-91%), which was very similar between the U-depleted and U-enriched fractions, indicating similar levels of deamination between host- and microbe derived reads (Figure S1A). Assuming that plant- and microbial-derived DNA deaminate at a similar rate (20), this observation indicates that microbes found in plant tissue were present at the time of collection or colonized the tissue shortly thereafter. The percentage of reads (including host-derived reads) that could be taxonomically binned varied depending on the sample (Figure S1B) and, since the host genome was included in the nucleotide database, positively correlated with the percentage of host endogenous DNA (Figure S1C). The inability to taxonomically assign the vast majority of reads from samples with low endogenous DNA reflects the incompleteness of the reference database compared to the diversity of the microbiomes in those samples.

Additionally, single stranded DNA library preparation methods as employed during Uracil enrichment generate shorter reads (21, 22), which are more difficult to map to a reference genome and to assign taxonomically to a nucleotide database. This is reflected in the higher percentage of reads mapped and assigned from the dsDNA library compared with shorter reads derived from both the U-depleted and U-enriched fraction (Figure S2A-B). Originally, it was reported that the U-enriched fraction shows a mild increase in GC-content (15), however in the plant libraries analyzed here we did not find a significant difference in GC-content between the U-depleted and U-enriched fractions (Figure S2C). In theory, since Us originate from Cs, the U-enriched fraction would be enriched for GC-rich species and GC-rich genomic regions within a given genome. However, as the enrichment would depend on the diversity of taxa present and their relative age difference, and hence difference in deamination, GC-biases, if any, are expected to be highly sample-dependent.

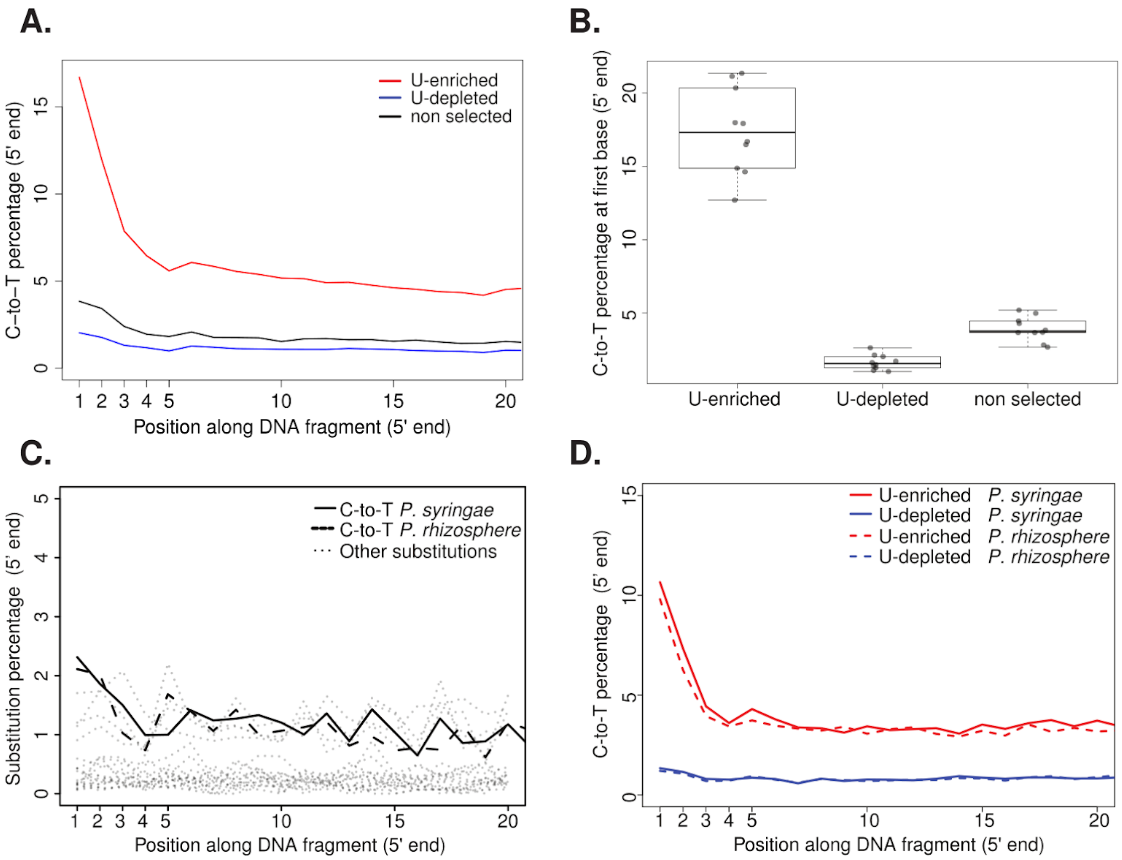

Figure 2. Patterns of cytosine to thymine (C-to-T) substitutions at the 5' end of plant- and *Pseudomonas*-derived reads. **A.** C-to-T substitutions at the 5' end of *Solanum tuberosum* sample KM177500 for a non-selected and U-selected library (U-enriched and U-depleted fractions). **B.** Distributions of C-to-T substitution percentage at first base (5' end) for non-selected and U-selected libraries (U-enriched and U-depleted fractions). Median values are denoted as black lines and points show the original value for each individual sample. **C.** Substitution patterns at the 5' end of *Pseudomonas syringae* and *Pseudomonas rhizosphere* mapped reads from a non-selected library from a *Solanum tuberosum* sample KM177500. **D.** Cytosine to Thymine substitutions at the 5' end of *P. syringae* and *P. rhizosphere* mapped reads from a U-selected library (U-enriched and U-depleted fractions) from a *Solanum tuberosum* sample KM177500.

Given the low taxonomic diversity of microorganisms in the samples included in our proof-of-principle experiment, instead of centering our analyses on the compositional assessment of microbial communities, we investigated in detail samples in which a specific microbe or group of microbes were more prevalent based on read abundance. We identified a large number of reads that were assigned to the bacterium *Pantoea vagans* in a potato (*Solanum tuberosum*) and a maize (*Zea mays*) sample (Figure S3). In both samples we found patterns of C-to-T substitutions that suggest the historical nature of the sequenced reads (Figure 3A). Since *P. vagans* is a plant epiphyte (23), it is not entirely surprising to find it in two different plant species. We compared the potato and maize *P. vagans* with publicly available genomes using single nucleotide polymorphisms (SNPs) ascertained in these modern samples. Our analysis linked the two historical strains to a distinct cluster of modern strains based on genetic similarity (Figure 3B). Based on a set of 432,891 SNPs, the two historical isolates showed 95% SNP identity between them, and an average of 92% SNP identity between historical and modern strains of the same cluster. Conversely, comparisons between historical strains and any modern strain of a different cluster showed only an average of 59% identity at variable positions.

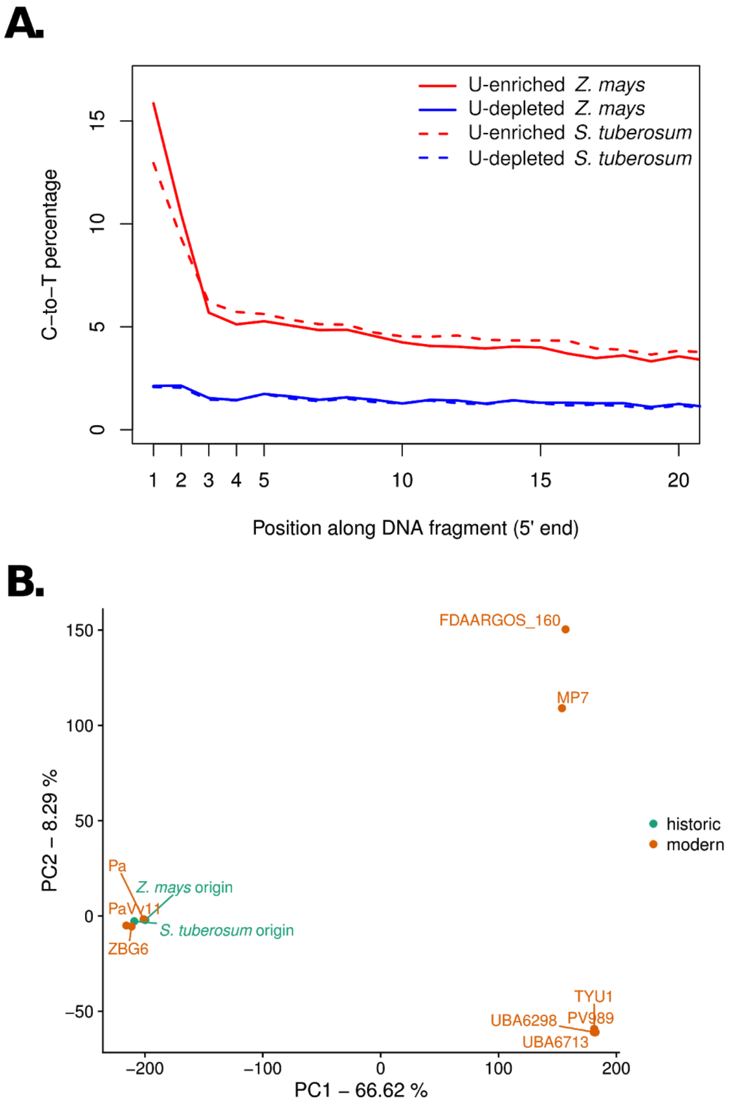

Figure 3. Substitution patterns and genetic distances of the bacterium *Pantoea vagans* identified from *Zea mays* and *Solanum tuberosum* samples (same samples as in Figure S3). **A.** Cytosine to Thymine substitutions at the 5' end of *P. vagans* for U-selected libraries (U-enriched and U-depleted fractions) from *Z. mays* and *S. tuberosum*. **B.** Principal component analysis of *P. vagans* from *Z. mays* and *S. tuberosum* samples, as well as nine publicly available genomes, based on single nucleotide polymorphisms. Numbers in axis labels indicate the percentage of the variance explained by each principal component (PC).

In a potato sample, in which the pathogenic oomycete *Phytophthora infestans* was previously identified (6), we found a large portion of reads assigned to the bacterial genus *Pseudomonas*. Reads were assigned in particular to the species *Pseudomonas syringae* and *Pseudomonas rhizosphere* in different proportions (Figure S4). We performed de novo assembly using reads assigned to the genus *Pseudomonas* and aligned the contigs to the reference genomes of *P. syringae* and *P. rhizosphere* covering about 80% of both reference genomes (Figure S5). We subsequently filtered for contigs that aligned uniquely to either *P. syringae* and *P. rhizosphere* genomes and found different k-mer coverage distributions in contigs aligning uniquely to each genome (Figure S6), an observation that reinforced our confidence in the presence of the two *Pseudomonas* species in this sample. Due to the high level of sequence divergence between the *Pseudomonas* in our sample and the reference genomes present in the database, it is difficult to assess typical deamination patterns in the dsDNA library (Figure 2C). However, we were able to examine damage patterns in both *Pseudomonas* species using the U-enriched fraction (Figure 2D), since the C-to-T signal is amplified and is much higher than the basal level of substitutions.

In summary, we showed here that the U-selection method selectively enriches for authentic microbial aDNA molecules in samples from plant and animal tissues with a wide-distribution of ages and deamination levels. For instance, in *P. vagans*, U-selection increases the fraction of molecules carrying a terminal C-to-T substitution at the 5'-end 2-3 fold over the library without enrichment, relative to the total number of molecules sequenced. We think that the application of U-selection for ancient microbiome research will be particularly useful in both samples with minute levels of deamination, where the nucleotide divergence between samples and reference genomes will obscure the identification of the C-to-T pattern typical of aDNA, as well as in moderately or heavily deaminated samples which carry modern contaminants, since in those samples ancient taxa would be efficiently enriched. Since it is extremely difficult to differentiate between ante-mortem and early post-mortem colonizers based only on deamination patterns, it is fundamental to also evaluate the biological relevance of detected taxa by comparing them with reference modern microbiomes.

Materials and Methods

Plant Samples

We used herbarium specimens from three different plant species (*Arabidopsis thaliana*, *Solanum tuberosum*, and *Solanum lycopersicum*) with ages ranging from 41 to 279 years (Table 1). *S. tuberosum* herbarium specimens were documented to be infected by *Phytophthora infestans* (6). We used also *Zea mays* archaeological remains excavated in the Turkey Penn shelter in Utah, USA (24). The *Zea mays* samples were dated using accelerator mass spectrometry and have ages ranging between 1852 and 1881 years BP (Before Present) (Table 1). The sequencing data for these samples is available on the European Nucleotide Archive under study number PRJEB30666.

Neanderthal samples

We used Neanderthal samples (Table 1) prepared by (15), which were sequenced deeper for this study.

DNA extraction and library preparation

DNA from all herbarium specimens and plant archaeological remains was performed as previously described (6).

For each plant sample two libraries were produced, one using a double-stranded library preparation (19, 25) and the second using the single-stranded U-selection protocol (15) without enzymatic removal of uracils (26).

Sequencing and initial data processing

Since the length of aDNA molecules is in most of the cases shorter than the read length of the sequencing platform, it is possible that a fragment of the aDNA molecule is sequenced by both the forward and reverse read, and also that a part of the adaptor is sequenced (27). Therefore, it is recommended to merge sequences based on the overlapping fraction sequenced by both forward and reverse reads (27). We remove adaptors and merged sequences using the software leeHom with the "--ancientdna" option (28). Putative chimeric sequences were flagged as failing quality.

Mapping of sequenced reads to their host genome

Merged reads were mapped as single-ended reads to their respective or most closely relative genome: *Zea mays* (29), *Arabidopsis thaliana* (30, 31), *Solanum tuberosum* (32), *Solanum lycopersicum* (33), *Homo sapiens* (Genome Reference Consortium Human Build 37). The mapping was performed using BWA-MEM (version 0.7.10) with default parameters, which includes a minimum length cutoff of 30 bp (34).

Metagenomics assignment of sequenced reads

Reads were aligned to the full non-redundant NCBI nucleotide collection (nt) database (downloaded January 2015) using MALT (version 0.0.12, (35)) in BlastN mode. The resulting RMA files were analyzed using MEGAN (version 5.11.3, (36)). The reads were assigned to the NCBI taxonomy using a lowest common ancestor algorithm (36).

Mapping of sequenced reads to microbial genomes

Libraries were mapped to microbial reference genomes of interest, after the presence of certain taxa was detected during metagenomic assignment. Specifically, the references of *Streptosporangium roseum* (37), *Pseudomonas syringae* pv. *syringae* B728a (38), *Pseudomonas rhizosphaerae* (39) and *Pantoea vagans* (23) were used. Since mapping metagenomic libraries to bacterial reference genomes is very prone to false alignments, we used a different mapping strategy for these genomes. The mappings were performed with bowtie2 (version 2.2.4, (40)), with the settings “--score-min 'L,-0.3,-0.3' --sensitive --end-to-end” to increase stringency.

Assessment of nucleotide substitution patterns

All types of nucleotide substitutions relative to the reference genome were calculated per library using mapDamage 2.0 (v. 2.0.2–12, (41)). The percentages of C-to-T substitutions at the 5’ end were extracted from the output file 5pCtoT_freq.txt produced by mapDamage.

Pantoea vagans genomic variation

In order to reduce the effect of aDNA-associated C-to-T substitutions on variant discovery, we used exclusively the U-depleted fraction of libraries where *P. vagans* was detected in the metagenomic screening. The libraries were mapped to the *P. vagans* reference genome using BWA-MEM, to reduce reference bias and increase SNP discovery. False alignments from the metagenomic libraries posed a lesser problem here, as variants were ascertained based on modern material. Variants for historical samples were called for both libraries together using the bcftools (version 1.8, (42)) utilities mpileup (“bcftools mpileup -q 1 -I -Ou -f $REF $IN1 $IN2”) and call (“bcftools call --ploidy 1 -m -O -z”). Additionally, 11 assemblies of different contiguity were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/genome/genomes/2707). These assemblies were aligned to the reference genome using minimap2 (version 2.10-r764, (43)) and its “asm20” parameter preset. Only strains with at least 80% reference coverage were kept for subsequent analysis (9/11, average reference coverage: 91%). The paftools utility, which is distributed with minimap2, was used to call variants from these alignments, with the parameter set “-l 2000 -L 5000”. All resulting VCF files from modern samples were merged using bcftools’ merge utility with the parameter “--missing-to-ref”, assuming that those positions not called by paftools in any one sample were indeed reference calls. The merged VCF from modern material was then merged with the VCF from the two historical samples using bcftools (version 1.8, (42)), and filtered to include only full information, biallelic SNPs. This approach discovers sites, which are segregating in modern material, and have read data (be it reference, alternative or segregating sites) in both historical samples. The resulting VCF file was loaded into R using vcfR (version 1.7.0, (44)), and a PCA was produced by converting the information into a genlight object using adegenet (version 2.0.1, (45)) in R (version 3.3.3, (46)).

Pseudomonas spp. assembly and evaluation

To evaluate the presence of *Pseudomonas* spp. strains in a *Solanum tuberosum* historic herbarium sample, we extracted from this library all reads that were taxonomically assigned to the *Pseudomonas* genus or to inferior taxonomic levels within it. These reads were then assembled using SPAdes (version 3.5.0) with default parameters (47). The resulting contigs were filtered for a minimum length of 2Kb, which yielded 3,314 contigs with a total length of 16Mb. We used the lastz (version 1.03.66, (48)) and Circos (version 0.64, (49)) interface of AliTV (50) to align these contigs to either the *P. syringae* or *P. rhizosphaerae* reference genome. We were able to align 72% of contigs to either one or both of these reference genomes in alignments of at least 1Kb. We then extracted all contigs which had alignments of at least 10Kb in length and were unique to one of the reference genomes. These sets of contigs were again aligned to their corresponding reference using AliTV as described above. Additionally, we used these uniquely aligning contigs to assess their average kmer coverage during the assembly, as reported by SPAdes.

Acknowledgements

We thank Verena Schuenemann for help in the laboratory; Patricia Lang, Claudia S. Burbano, members of the Research Group for Ancient Genomics and Evolution, and specially Talia Karasov for input on data analysis and comments on the manuscript; Janet Kelso for help with processing the raw sequencing data; Ivan Gusic, Zeljko Kucan, Carles Lalueza-Fox, Marco de la Rasilla, Antonio Rosas, Pavao Rudan, Sandra Knapp, Bruce Benz, Michael Blake, R.G. Matson, Bryn Dentinger, Anna Stalter, Robert Capers, John Peter for providing samples. This work was funded by the Max Planck Society and its Presidential Innovation Fund.

References

1. Poinar HN, Schwarz C, Qi J, Shapiro B, Macphee RDE, Buigues B, Tikhonov A, Huson DH, Tomsho LP, Auch A, Rampp M, Miller W, Schuster SC. 2006. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science 311:392–394.

2. Gutaker RM, Burbano HA. 2017. Reinforcing plant evolutionary genomics using ancient DNA. Curr Opin Plant Biol 36:38–45.

3. Orlando L, Gilbert MTP, Willerslev E. 2015. Reconstructing ancient genomes and epigenomes. Nat Rev Genet 16:395–408.

4. Warinner C, Herbig A, Mann A, Fellows Yates JA, Weiß CL, Burbano HA, Orlando L, Krause J. 2017. A Robust Framework for Microbial Archaeology. Annu Rev Genomics Hum Genet 18:321–356.

5. Bos KI, Schuenemann VJ, Golding GB, Burbano HA, Waglechner N, Coombes BK, McPhee JB, DeWitte SN, Meyer M, Schmedes S, Wood J, Earn DJD, Herring DA, Bauer P, Poinar HN, Krause J. 2011. A draft genome of Yersinia pestis from victims of the Black Death. Nature 478:506–510.

6. Yoshida K, Schuenemann VJ, Cano LM, Pais M, Mishra B, Sharma R, Lanz C, Martin FN, Kamoun S, Krause J, Thines M, Weigel D, Burbano HA. 2013. The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine.

Elife 2:e00731.

7. Adler CJ, Dobney K, Weyrich LS, Kaidonis J, Walker AW, Haak W, Bradshaw CJA, Townsend G, Sołtysiak A, Alt KW, Parkhill J, Cooper A. 2013. Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. Nat Genet 45:450–5, 455e1.

8. Warinner C, Rodrigues JFM, Vyas R, Trachsel C, Shved N, Grossmann J, Radini A, Hancock Y, Tito RY, Fiddyment S, Speller C, Hendy J, Charlton S, Luder HU, Salazar-García DC, Eppler E, Seiler R, Hansen LH, Castruita JAS, Barkow-Oesterreicher S, Teoh KY, Kelstrup CD, Olsen JV, Nanni P, Kawai T, Willerslev E, von Mering C, Lewis CM Jr, Collins MJ, Gilbert MTP, Rühli F, Cappellini E. 2014. Pathogens and host immunity in the ancient human oral cavity. Nat Genet 46:336–344.

9. Weyrich LS, Duchene S, Soubrier J, Arriola L, Llamas B, Breen J, Morris AG, Alt KW, Caramelli D, Dresely V, Farrell M, Farrer AG, Francken M, Gully N, Haak W, Hardy K, Harvati K, Held P, Holmes EC, Kaidonis J, Lalueza-Fox C, de la Rasilla M, Rosas A, Semal P, Soltysiak A, Townsend G, Usai D, Wahl J, Huson DH, Dobney K, Cooper A. 2017. Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. Nature.

10. Boast AP, Weyrich LS, Wood JR, Metcalf JL, Knight R, Cooper A. 2018. Coprolites reveal ecological interactions lost with the extinction of New Zealand birds. Proc

Natl Acad Sci U S A 115:1546–1551.

11. Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prüfer K, Meyer M, Krause J, Ronan MT, Lachmann M, Pääbo S. 2007. Patterns of damage in genomic DNA sequences from a Neandertal. Proceedings of the National Academy of Sciences 104:14616–14621.

12. Krause J, Briggs AW, Kircher M, Maricic T, Zwyns N, Derevianko A, Pääbo S. 2010. A Complete mtDNA Genome of an Early Modern Human from Kostenki, Russia. Curr Biol 20:231–236.

13. Prüfer K, Meyer M. 2015. Comment on “Late Pleistocene human skeleton and mtDNA link Paleoamericans and modern Native Americans.” Science 347:835–835.

14. Weiß CL, Dannemann M, Prüfer K, Burbano HA. 2015. Contesting the presence of wheat in the British Isles 8,000 years ago by assessing ancient DNA authenticity from low-coverage data. Elife 4.

15. Gansauge M-T, Meyer M. 2014. Selective enrichment of damaged DNA molecules for ancient genome sequencing. Genome Res 24:1543–1549.

16. Skoglund P, Malmström H, Raghavan M, Storå J, Hall P, Willerslev E, Gilbert MTP, Götherström A, Jakobsson M. 2012. Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. Science 336:466–469.

17. Meyer M, Fu Q, Aximu-Petri A, Glocke I, Nickel B, Arsuaga J-L, Martínez I, Gracia A, de Castro JMB, Carbonell E, Pääbo S. 2014. A mitochondrial genome sequence of a hominin from Sima de los Huesos. Nature 505:403–406.

18. Zaremba-Niedźwiedzka K, Andersson SGE. 2013. No ancient DNA damage in Actinobacteria from the Neanderthal bone. PLoS One 8:e62799.

19. Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc 2010:db.prot5448.

20. Weiß CL, Schuenemann VJ, Devos J, Shirsekar G, Reiter E, Gould BA, Stinchcombe JR, Krause J, Burbano HA. 2016. Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. R Soc Open Sci 3:160239.

21. Gansauge M-T, Meyer M. 2013. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. Nat Protoc 8:737–748.

22. Gansauge M-T, Gerber T, Glocke I, Korlević P, Lippik L, Nagel S, Riehl LM, Schmidt A, Meyer M. 2017. Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. Nucleic Acids Res.

23. Smits THM, Rezzonico F, Kamber T, Goesmann A, Ishimaru CA, Stockwell VO, Frey JE, Duffy B. 2010. Genome sequence of the biocontrol agent Pantoea vagans

strain C9-1. J Bacteriol 192:6486–6487.

24. Matson RG, Chisholm B. 1991. Basketmaker II Subsistence: Carbon Isotopes and Other Dietary Indicators from Cedar Mesa, Utah. Am Antiq 56:444–459.

25. Kircher M, Sawyer S, Meyer M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res 40:e3.

26. Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S. 2010. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Res 38:e87.

27. Kircher M. 2012. Analysis of high-throughput ancient DNA sequencing data. Methods Mol Biol 840:197–228.

28. Renaud G, Stenzel U, Kelso J. 2014. leeHom: adaptor trimming and merging for Illumina sequencing reads. Nucleic Acids Res 42:e141.

29. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado

B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh C-T, Emrich SJ, Jia Y, Kalyanaraman A, Hsia A-P, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia J-M, Deragon J-M, Estill JC, Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK. 2009. The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115.

30. Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796–815.

31. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, Radenbaugh A, Singh S, Swing V, Tissier C, Zhang P, Huala E. 2008. The Arabidopsis Information Resource (TAIR): gene

structure and function annotation. Nucleic Acids Res 36:D1009–14.

32. Potato Genome Sequencing Consortium, Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R, Wang J, Orjeda G, Guzman F, Torres M, Lozano R, Ponce O, Martinez D, De la Cruz G, Chakrabarti SK, Patil VU, Skryabin KG, Kuznetsov BB, Ravin NV, Kolganova TV, Beletsky AV, Mardanov AV, Di Genova A, Bolser DM, Martin DMA, Li G, Yang Y, Kuang H, Hu Q, Xiong X, Bishop GJ, Sagredo B, Mejía N, Zagorski W, Gromadka R, Gawor J, Szczesny P, Huang S, Zhang Z, Liang C, He J, Li Y, He Y, Xu J, Zhang Y, Xie B, Du Y, Qu D, Bonierbale M, Ghislain M, Herrera M del R, Giuliano G, Pietrella M, Perrotta G, Facella P, O'Brien K, Feingold SE, Barreiro LE, Massa GA, Diambra L, Whitty BR, Vaillancourt B, Lin H, Massa AN, Geoffroy M, Lundback S, DellaPenna D, Buell CR, Sharma SK, Marshall DF, Waugh R, Bryan GJ, Destefanis M, Nagy I, Milbourne D, Thomson SJ, Fiers M, Jacobs JME, Nielsen KL, Sønderkær M, Iovene M, Torres GA, Jiang J, Veilleux RE, Bachem CWB, de Boer J, Borm T, Kloosterman B, van Eck H, Datema E, Hekkert B te L, Goverse A, van Ham RCHJ, Visser RGF. 2011. Genome sequence and analysis of the tuber crop potato. Nature 475:189–195.

33. Tomato Genome Consortium. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635–641.

34. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [q-bioGN].

35. Herbig A, Maixner F, Bos KI, Zink A, Krause J, Huson DH. 2016. MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman. bioRxiv.

36. Huson DH, Auch AF, Qi J, Schuster SC. 2007. MEGAN analysis of metagenomic data. Genome Res 17:377–386.

37. Nolan M, Sikorski J, Jando M, Lucas S, Lapidus A, Del Rio TG, Chen F, Tice H, Pitluck S, Cheng J-F, Others. 2010. Complete genome sequence of Streptosporangium roseum type strain (NI 9100 T). Stand Genomic Sci 2:29.

38. Feil H, Feil WS, Chain P, Larimer F, DiBartolo G, Copeland A, Lykidis A, Trong S, Nolan M, Goltsman E, Thiel J, Malfatti S, Loper JE, Lapidus A, Detter JC, Land M, Richardson PM, Kyrpides NC, Ivanova N, Lindow SE. 2005. Comparison of the complete genome sequences of Pseudomonas syringae pv. syringae B728a and pv. tomato DC3000. Proc Natl Acad Sci U S A 102:11064–11069.

39. Kwak Y, Jung BK, Shin J-H. 2015. Complete genome sequence of Pseudomonas rhizosphaerae IH5 T (= DSM 16299 T), a phosphate-solubilizing rhizobacterium for bacterial biofertilizer. J Biotechnol 193:137–138.

40. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.

41. Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L. 2013.

mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics 29:1682–1684.

42. Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993.

43. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics.

44. Knaus BJ, Grünwald NJ. 2017. vcfr: a package to manipulate and visualize variant call format data in R. Mol Ecol Resour 17:44–53.

45. Jombart T, Ahmed I. 2011. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics 27:3070–3071.

46. R Development Core Team. 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

47. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477.

48. Harris RS. 2007. Improved pairwise alignment of genomic DNA. The Pennsylvania State University.

49. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. Genome Res 19:1639–1645.

50. Hohlfeld S, Ankenbrand M, Förster F, Hackl T. 2016. AliTV: Version 0.4.1. Zenodo.

Supplementary Figures

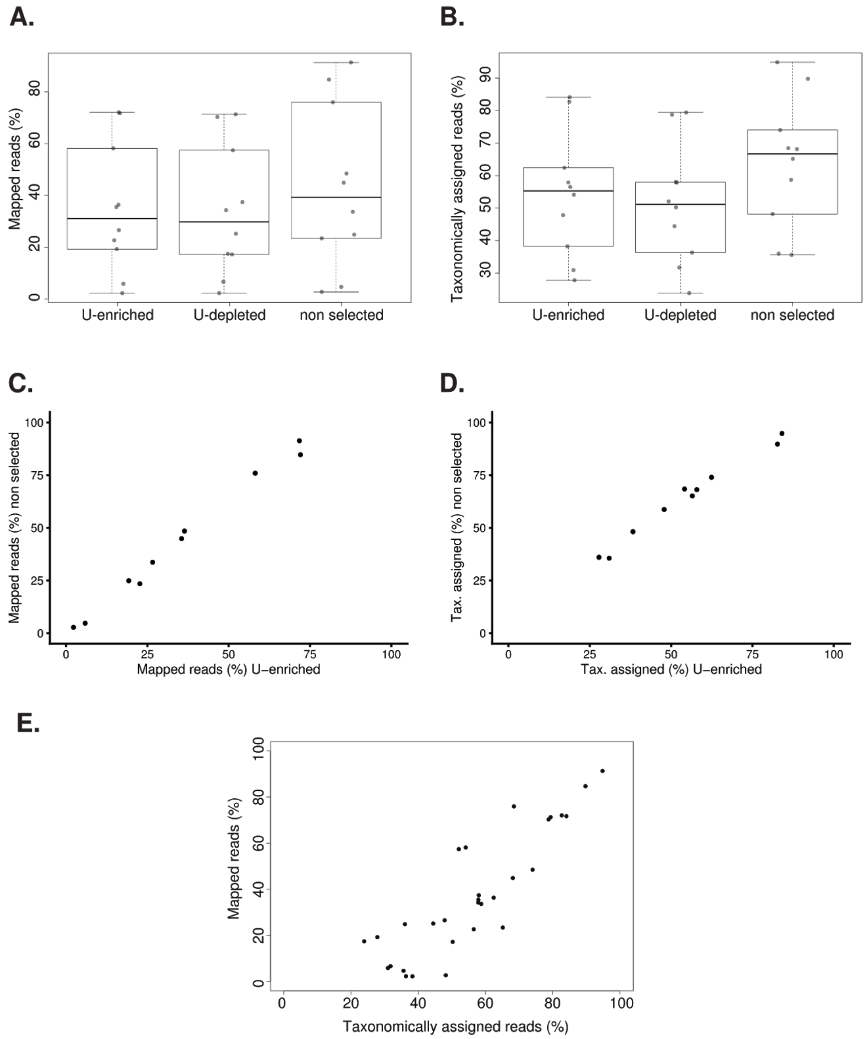

Figure S1. Mapped and taxonomically assigned reads of plant historical specimens. **A.** Distributions of percentage of mapped reads for non-selected and U-selected libraries (U-enriched and U-depleted fractions). **B.** Distributions of percentage of taxonomically assigned reads for non-selected and U-selected libraries (U-enriched and U-depleted fractions). **C.** Correlation of the percentage of mapped reads between the U-enriched and the non-selected library **D.** Correlation of the percentage of taxonomically assigned reads between the U-enriched and the non-selected library **E.** Relation between percentages of mapped and taxonomically assigned reads from U-selected libraries (U-enriched fraction).

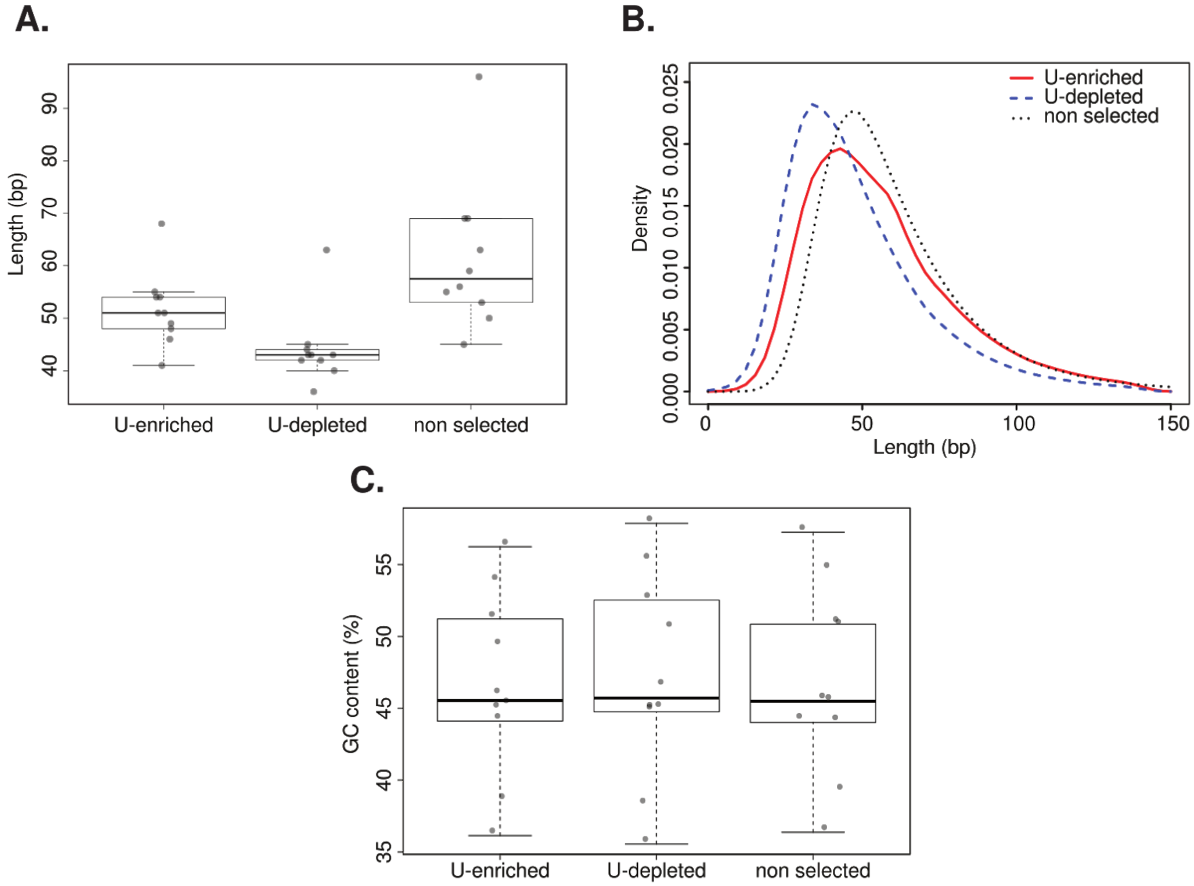

Figure S2. Length and GC content of plant historical specimens. **A.** Distributions of mean length for non-selected and U-selected libraries (U-enriched and U-depleted fractions). Median values are denoted as black lines and points show the original value for each individual sample. **B.** Length distribution of *Arabidopsis thaliana* sample NY1365375 for a non-selected and U-selected library (U-enriched and U-depleted fractions). **C.** Distributions of mean GC content for non-selected and U-selected libraries (U-enriched and U-depleted fractions).

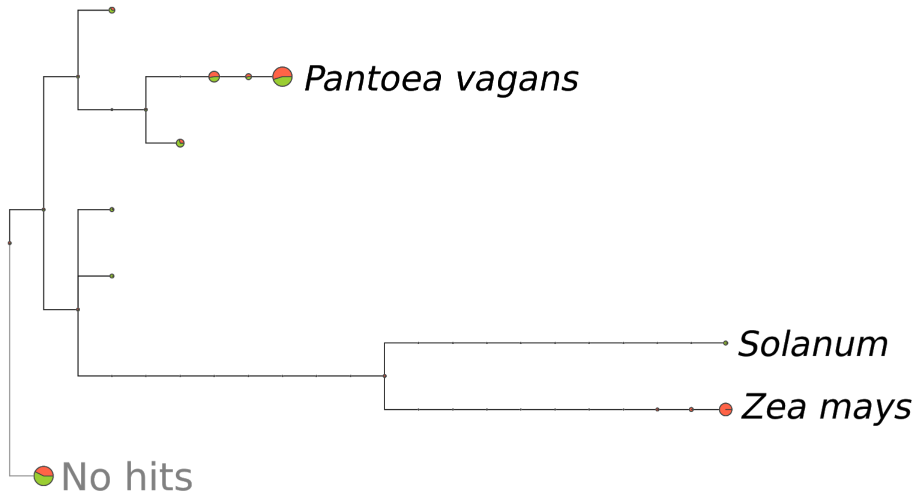

Figure S3. Taxonomic tree of reads from *Solanum tuberosum* and *Zea mays* assigned to different taxonomic levels. The size of the circle represents the amount of reads assigned to a particular part or the taxonomy. *S. tuberosum*- and *Z. mays*-derived reads are shown in green and orange, respectively.

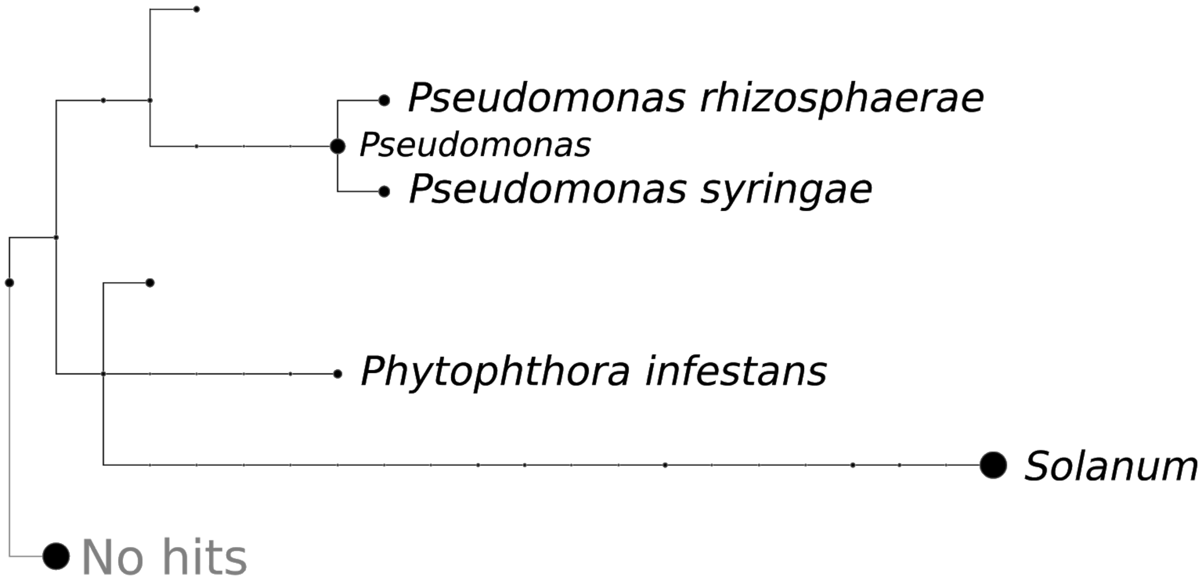

Figure S4. Taxonomic tree of reads from a *Solanum tuberosum* library assigned to different taxonomic levels. The size of the circle represents the amount of reads assigned to a particular part of the taxonomy. Reads assigned to some species *Phytophthora infestans, Pseudomonas syringae and Pseudomonas rhizosphere*, as well as the genera Pseudomonas and Solanum are named in the taxonomic tree.

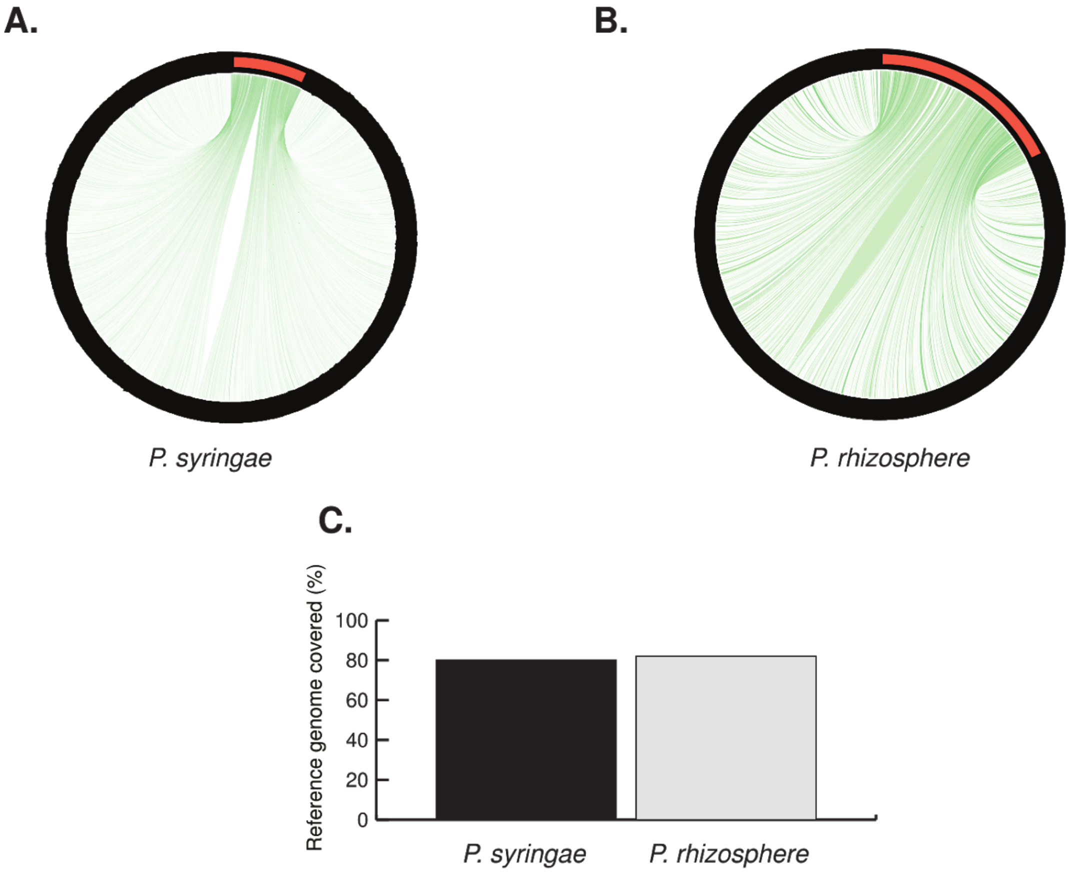

Figure S5. De novo assembly and genomic coverage of *Pseudomonas syringae* and *Pseudomonas rhizosphere* from a *Solanum tuberosum* sample. **A.** Alignments (represented as green lines) between all de novo assembly contigs (black) and *P. syringae* reference genome (red). **B.** Alignments (represented as green lines) between all de novo assembly contigs (black) and *P. rhizosphere* reference genome (red). **C.** Percentage of reference genome of *P. syringae* and *P. rhizosphere* covered by de novo assembled contigs from A. and B., respectively.

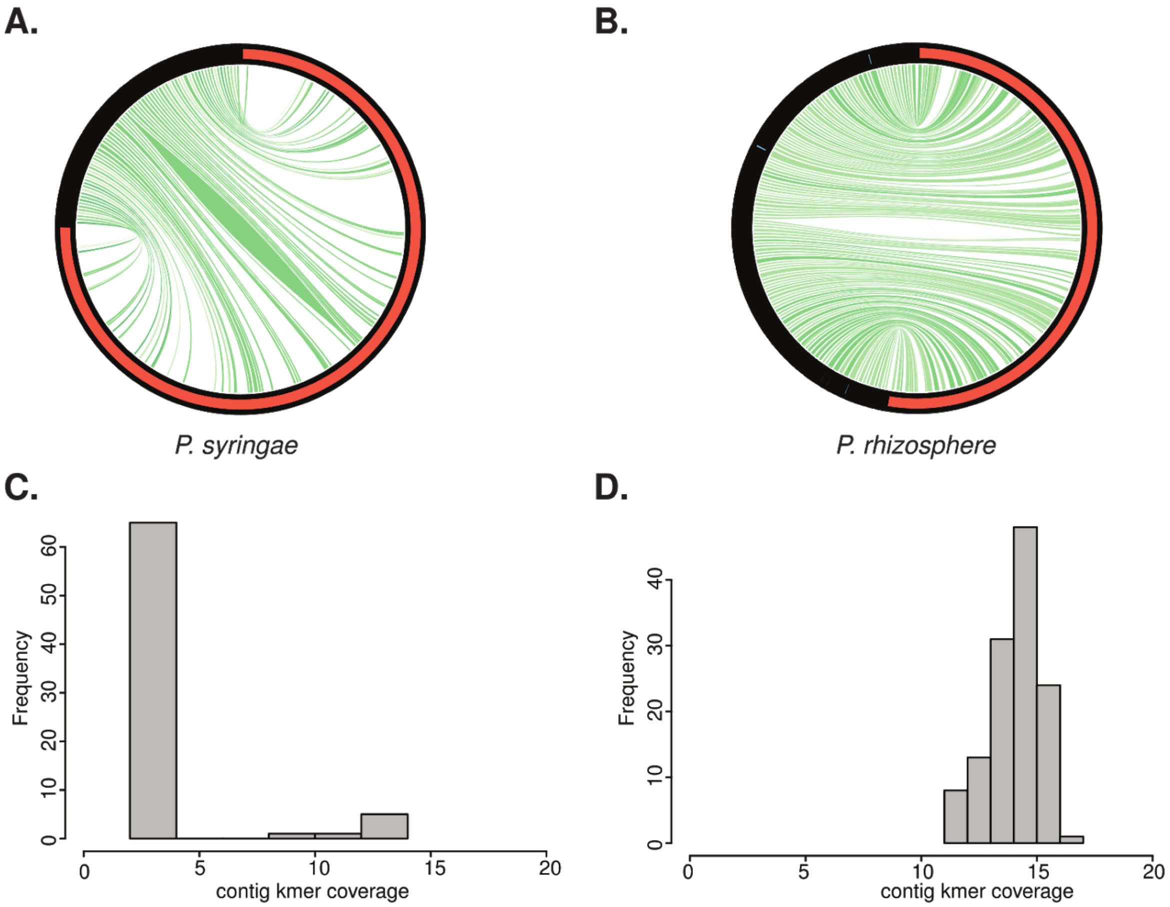

Figure S6. De novo assembly (uniquely mapped contigs) and contig k-mer coverage of *Pseudomonas syringae* and *Pseudomonas rhizosphere* from a *Solanum tuberosum* sample. **A.** Alignments (represented as green lines) between de novo assembly contigs (black) uniquely mapped to *P. syringae* reference genome (red). **B.** Alignments (represented as green lines) between de novo assembly contigs (black) uniquely mapped to *P. rhizosphere* reference genome (red). **C.** Histogram of contig k-mer coverage from de novo assembled contigs uniquely mapping to *P. syringae*. **D.** Histogram of contig k-mer coverage from de novo assembled contigs uniquely mapping to *P. rhizosphere*.