

SUPPLEMENTARY MATERIAL INFORMATION

“Insights into *T. pallidum* genomics from modern and ancient genomes using a novel mapping strategy”

Marta Pla-Díaz^{#(1,2,3)}, Gülfirde Akgül^{#(4)}, Martyna Molak^(5,6,), Louis du Plessis^(7,8), Hanna Panagiotopoulou⁽⁶⁾, Karolina Doan⁽⁶⁾, Wiesław Bogdanowicz⁽⁶⁾, Paweł Dąbrowski⁽⁹⁾, Maciej Oziembłowski⁽¹⁰⁾, Barbara Kwiatkowska⁽¹¹⁾, Jacek Szczurowski⁽¹¹⁾, Joanna Grzelak⁽⁹⁾, Natasha Arora⁽¹²⁾, Kerttu Majander^{†(1,4,13,14)}, Fernando González-Candelas^{†(2,3)}, Verena J. Schuenemann^{†(1,4,13,14)}.

(1) *Department of Environmental Sciences, University of Basel, Basel, Switzerland.* (2) *Unidad Mixta Infección y Salud Pública FISABIO/Universidad de Valencia-I2SysBio, Valencia, Spain.* (3) *CIBER in Epidemiology and Public Health, Valencia, Spain.* (4) *Institute of Evolutionary Medicine, University of Zurich, Zurich, Switzerland.* (5) *Centre of New Technologies, University of Warsaw, Warsaw, Poland* (6) *Museum and Institute of Zoology PAS, Poland.* (7) *Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland.* (8) *Swiss Institute of Bioinformatics, Lausanne, Switzerland.* (9) *Department of Anatomy, Wrocław Medical University, Wrocław, Poland.* (10) *Faculty of Biotechnology and Food Sciences, Wrocław University of Environmental and Life Sciences, Wrocław, Poland.* (11) *Department of Anthropology, Wrocław University of Environmental and Life Sciences, Wrocław, Poland.* (12) *Zurich Institute of Forensic Medicine, University of Zurich, Zurich, Switzerland.* (13) *Department of Evolutionary Anthropology, University of Vienna, Vienna, Austria.* (14) *Human Evolution and Archaeological Sciences (HEAS), University of Vienna, Vienna, Austria.*

Table of Contents

Supplementary Notes	3
Supplementary Note 1. Skeletal material description for sample W86 (OT20.1)	3
Supplementary Note 2. Chemical analysis of sample W86 (OT20.1)	6
Supplementary Note 3. Topological variations in phylogenetic trees among SRB and MRB datasets	6
Supplementary Note 4. Intermediate results of the PIM procedure: Likelihood mapping and Topology tests	7
Supplementary Note 5. Repetitive samples comparison	8
Supplementary Tables	9
Supplementary Figures	11
Supplementary Material (Files)	15
References	15

All the Supplementary Material Information is available at the DOI:
<https://doi.org/10.5281/zenodo.14535982>

Supplementary Notes

Supplementary Note 1. Skeletal material description for sample W86 (OT20.1)

Dr Paweł Dąbrowski, dr Joanna Grzelak, dr hab. Maciej Oziębłowski

The grave marked OT20 is in fact a collection of bones belonging to several individuals (possibly four). Two cardboard boxes containing bones of the postcranial skeleton and 4 boxes containing skulls or their fragments were made available for research:

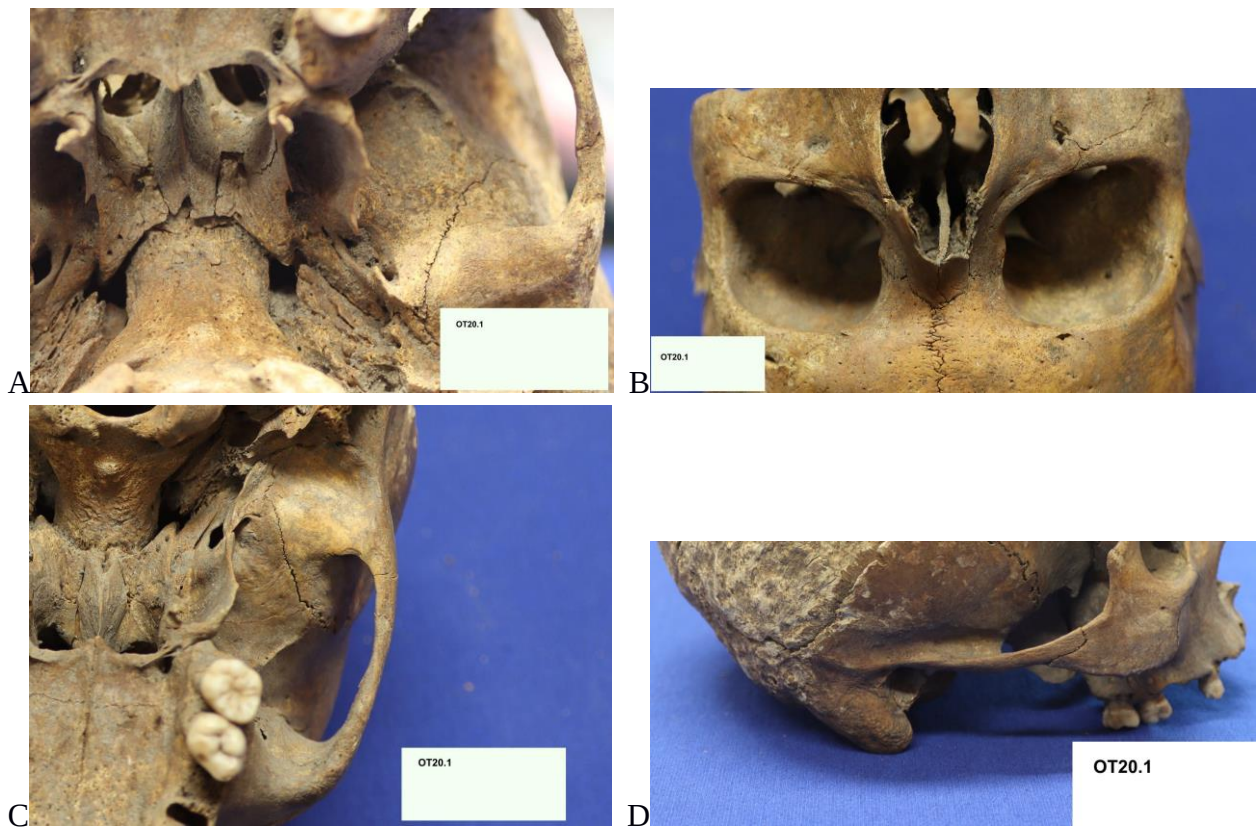
- a) **OT20.1** - skull without the lower jaw. Tooth sampled for genetic analysis and labeled W86. (Supplementary Note 1, Figure 1)
- b) **OT20.2** - skull without the lower jaw with a damaged left zygomatic arch.
- c) **OT20.3** - skull without the mandible, with both zygomatic arches damaged, with complete atrophy of the maxillary alveolar process.
- d) **OT20.4** - fragment of the skull: only the brain case preserved.



Supplementary Note 1, Figure 1. W86 (OT20.1) skull. a) Frontal, b) right side view. Photo: J. Grzelak

Samples for genetic testing were taken from the tooth of the subject designated as OT20.1.

The age at death was estimated on the basis of the diagnostic features of the skull: the degree of obliteration of the cranial sutures, the degree of obliteration of the spheno-occipital synchondrosis and the degree of tooth wear, as well as the condition of the metaphyses of the long bones and the condition of the ilium auricular surface of the postcranial skeleton associated to the OT20.1 skull. The age at the time of death of the OT20.1 individual is defined as juvenis / adultus (end of adolescence / beginning of adulthood). The beginning of adulthood (adultus) is considered to be the complete closure of the synchondrosis fissure. In the case under study, it is incomplete (Supplementary Note 1, Figure 2Aa).



Supplementary Note 1, Figure 2. Morphological features of skull W86 (OT20.1). A) Incomplete closure of the synchondrosis fissure, B) thickened orbit, C) massive zygomatic arch, D) massive mastoid process. Photo: J. Grzelak

The gender of the OT20.1 subject was assessed as male. The assessment was based on the following features:

- Orbital morphology: the lateral part of the upper edge of the orbit is thickened, which results from the thickening of the superciliary arch - a characteristic of the male sex (Supplementary Note 1, Figure 2B),
- Massiveness of the zygomatic arch - the lower edge is thickened, which is related to the extensive attachment of the masseter muscle in the male sex (Supplementary Note 1, Figure 2C),
- The mastoid process is very massive, which indicates a very well-developed sternocleidomastoid muscle (Supplementary Note 1, Figure 2D).



Supplemental Note 1, Figure 3. Paleopathological examination. A) scales of the occipital bone and the scales of the parietal bone, B) degree of cranial sutures fusion, C) coronal and lambdoid sutures, D) right orbit, E) left orbit, F) facial part of the cranium

Unfortunately, not all diagnostic features could be assessed: the outer surface of the occipital bone and the posterior part of the parietal bone turned out to be so mechanically damaged that it was impossible to describe the diagnostic features. In turn, the so-called pseudopathological changes, resulting from taphonomic processes, were observed; hence it is impossible to assess the sculpture of the scales of the occipital bone and the scales of the parietal bone (Supplemental Note 1, Figure 3A). Doubts as to gender assessment are also raised by the presence of parietal tumors and the proportions of the examined skull. They appear in the form characteristic for an individual of the female sex (Supplemental Note 1, Figure 1B). It should be emphasized that the uncertainty in determining the sex may be related to the young age of the examined person and the developmental changes observed in the skull. The skull is characterized by very wide, open seams and a preserved

metopic suture (Supplemental Note 1, Figure 3A and 3B). With regard to pathological changes and developmental disorders, it was found that in the coronal and lambdoid sutures there are small insertion bones, which may indicate disturbances in the growth and ossification process (Supplemental Note 1, Figure 3C). No changes in the form of *cribra orbitalia* – orbital roof lesions related to deficiency anemia, were observed (Supplemental Note 1, Figure 3D and 3E).

Similarly, the absence of hypoplastic defects in enamel may indicate that this individual was well nourished in early childhood. No carious lesions were found in the preserved tooth crowns of permanent teeth. No visible bone syphilis lesions at the anterior surface of the maxillary body and the edge of the nasal notch (Supplemental Note 1, Figure 3F). The superficial changes in the neurocranium do not present an image characteristic for the inflammatory process in the diagnosis of syphilis; in our opinion, it is the result of taphonomic processes occurring between the skeleton and the filling of the burial cavity.

The sample was a part of a human genetic analysis conducted at the Museum and Institute of Zoology, Polish Academy of Sciences as a part of the University of Wrocław research project. The DNA libraries from this project were screened for presence of a range of pathogens using PCR amplification of chosen genetic markers (i.a. 106 bp *T. pallidum* arp gene fragment [1]). As sample W86 (and two other samples from the project) tested positive by PCR for arp, it was sent to the Paleogenetics Laboratory at the University of Zurich for further *Treponema*-targeted analysis.

Supplementary Note 2. Chemical analysis of sample W86 (OT20.1)

In OT20.1, the content of calcium, phosphorus and alkaline earth elements (Sr and Ba) was determined in the bone tissue collected from the rib. Then the proportions of these elements were calculated.

The Ca / P ratio = 1.65 indicates a relatively good state of conservation of the material.

The proportions of strontium and barium to calcium are significantly higher than the average in the studied group, which may indicate a diversified diet, with the use of plant-based foods and / or dairy products.

Supplementary Note 2, Table 1. Proportions of Sr / Ca and Ba / Ca against the background of the studied group.

Proportion	Average	OT20.1	Statistical significance
Sr/Ca	-7.71	-7.35	P=0.000
Ba/Ca	-9.05	-7.64	P=0.000

Supplementary Note 3. Topological variations in phylogenetic trees among SRB and MRB datasets

The most remarkable differences affected the SS14 strains, especially the four ancient TPA genomes (including W86) and the Mexico A strain. Specifically, in the tree derived from the BosniaA-SRB genome dataset (Supplementary Figure 3B), the ancient genomes 94A and 94B were at the base of all TPA strains, and Mexico A was located within the SS14-Ω clade. This is in contrast to the MRB tree (Supplementary Figure 3A), in which both ancient genomes were basal to all SS14 strains.

Similar topological incongruences were observed in the tree obtained using CDC2-SRB with regards to the 94A and 94B ancient genomes (Supplementary Figure 2B), although in this dataset, Mexico A remained basal to all SS14 strains.

However, in the tree obtained using the Nichols-SRB genome dataset (Supplementary Figure 4B), the most notable topological change occurred with the Mexico A strain. Similarly to the tree obtained from the CDC2-SRB dataset (Supplementary Figure 2B), Mexico A was located within the SS14- Ω clade and not basal to the entire SS14 clade. Unlike the ML trees obtained from the previously detailed SRB genome datasets (Supplementary Figures 2-4), the ancient strains 94A and 94B occupied the same position in the tree as in that obtained using the MRB genome alignment. Nevertheless, the ancient strain SJ219 appeared in the ML tree obtained with the Nichols-SRB dataset (Supplementary Figure 4B) with a much longer branch, implying many more inferred substitutions in this strain.

Moreover, in all four ML trees obtained with the SRB genome datasets (Supplementary Figures 2-5) remarkable changes were also observed in the Nichols clade when they were compared to the ML tree obtained with the MRB genome alignment. In the trees obtained using the BosniaA-, Nichols-, and SS14-SRB genome datasets (Supplementary Figures 3B-5B), the Seattle 81-4 strain, which was basal to all the strains of the Nichols clade in the MRB tree, grouped among other strains of the Nichols clade. Additionally, a remarkable change was observed regarding strain CW82. Instead of grouping with other Nichols strains as in the MRB tree, it occupied a basal position to the Nichols strains in all four of the ML trees obtained using the SRB genome datasets (Supplementary Figures 2B-5B).

Supplementary Note 4. Intermediate results of the PIM procedure: Likelihood mapping and Topology tests

After selecting 317 genes with more than 3 SNPs present in at least one reference genome (Supplementary Table 4), a likelihood mapping [4] (LM) test was performed using IQTREE to ascertain which genes had some phylogenetic signal. Briefly, the test constructs unrooted phylogenetic trees for each quartet (subsets of four sequences) in the data. Quartet likelihoods are then mapped to a triangle, where the location represents the "treelikeness" of the quartet. Quartets in corners are fully resolved, on the sides partially resolved and in the central zone unresolved. From the 317 genes, there were 53 genes with too much missing data to perform the LM test. In order to include them in the PIM pipeline, for each individual gene, problematic sequences with more than 50% of positions with missing data were removed. Then, the likelihood-mapping test was performed and 160 genes were retained (Supplementary Table 5) and the remaining genes were discarded, because for all quartets considered, the distribution of the corresponding likelihoods fell in the central zone of the triangle, a key graphical representation in the test used to visualize phylogenetic signal. When likelihoods cluster in this central zone, it indicates a weak phylogenetic signal, suggesting that the gene does not provide clear evolutionary information. Next, we tested the phylogenetic congruence between trees, comparing the tree obtained from each retained gene and the provisional reference tree obtained from the complete genome alignment using the SH and ELW topology tests. From the 91 genes that showed phylogenetic incongruence, 18 were further verified to contain at least three consecutive SNPs, thus supporting a recombination event (Supplementary table 6).

Subsequently, we detailed the recombination results by PIM obtained for each one of the SRB genome datasets:

Out of the 1125 genes extracted from the CDC2-SRB genome dataset according to the CDC2 genomic reference coordinates, 302 with more than 3 SNPs underwent recombination analysis through the PIM pipeline. However, extensive missingness of data limited the topology test to 230 genes, identifying 47 with reciprocal incongruence, but only 10 as recombinant genes. Those 10 genes were among those obtained with the MRB genome alignment, while the remaining 10 recombinant genes detected using the MRB genome alignment could not be tested in the CDC2-SRB genome dataset due to a significant proportion of missing data.

Similarly, from the 1122 genes extracted from the BosniaA-SRB genome dataset according to the BosniaA genomic reference coordinates, 301 genes with more than 3 SNPs were analyzed for recombination. However, due to substantial missingness of data, the topology test could only be conducted for 229 genes, revealing 46 with reciprocal incongruence, of which 11 were recombinant. Those 11 genes were among those obtained with the MRB genome alignment, while the remaining 9 recombinant genes detected using the MRB genome alignment could not be tested in the BosniaA-SRB genome dataset due to a significant proportion of missing data.

For the 1035 genes extracted from the Nichols-SRB genome dataset according to the Nichols genomic reference coordinates, 306 genes with more than 3 SNPs were analyzed for recombination. Despite substantial missingness of data, the topology test considered 238 genes, indicating 48 with reciprocal incongruence and 12 of those as recombinant. Those 12 genes were among those obtained with the MRB genome alignment, while the remaining 8 genes detected employing the MRB genome alignment could not be tested in the Nichols-SRB genome dataset due to a significant proportion of missing data.

Finally, among the 1032 genes extracted from the SS14-SRB genome dataset according to the SS14 genomic reference coordinates, 289 genes with more than 3 SNPs underwent recombination analysis. However, due to significant missingness of data, the topology test was limited to 226 genes, revealing 43 with reciprocal incongruence, of which 11 were recombinant. Those 11 genes were among those obtained with the MRB genome alignment, while the remaining 9 recombinant genes detected in the MRB genome alignment analysis could not be tested in the SS14-SRB genome dataset due to a significant proportion of missing data.

Supplementary Note 5. Repetitive samples comparison

For this study, we included in the new dataset generated (see Supplementary Table 1), four samples, IND1, K363, Nichols, and NIC2, originally derived from the same two clinical samples. The inclusion of these sequences aimed to discern potential variations resulting from the distinct methods employed in their acquisition. These four samples have undergone sequencing in different studies, employing different technologies and processing methods.

IND1, obtained through the enrichment of treponemal DNA based on the hybridization capture technique, directly originated from the original clinical sample and was featured in Arora *et al.* [2].

In contrast, K363 was acquired using the Whole Genome Amplification (WGA) technique, as detailed by Strouhal *et al.* [3].

Nichols and NIC2 both originated from the same original sample, which underwent rabbit culture to enhance DNA content. Nichols, featured in Petrosova *et al.* [2], was sequenced using the Whole Genome Amplification (WGA) technique. In contrast, NIC2, featured in Arora *et al.* [2], was sequenced following enrichment of treponemal DNA based on hybridization capture.

When we compared the sequences of the two samples sequenced twice, we obtained 86 SNPs between IND1 and K363, and 18 SNPs between Nichols and NIC2.

Regarding the placement of these four strains in the reference phylogeny, as depicted in the whole genome alignment with and without genes showing a signal of recombination (Figure 2), it is noteworthy that Nichols and NIC2 maintain the same position in both trees. However, IND1 and K363 exhibit more divergent positions in the whole genome reference tree compared to the reference tree when recombination was removed. This discrepancy arises from the identification of 84 out of the 86 SNPs found between IND1 and K363, primarily located in the tRNAs *tp0012* and *tp0015*. These tRNAs were flagged as potential recombinant genes, leading to their removal from the phylogenetic tree without considering recombination (see Figure 2B). Although both genes were labeled as recombinant, we cannot completely dismiss the possibility that the SNPs in these genes may be artifacts or the result of contamination.

Supplementary Tables

Additional File 2; Supplementary Table 1. Genome reconstruction details for all published samples and genomes used in the analyses. Type of file: excel file

Additional File 3; Supplementary Table 2. Orthology analysis results obtained by Proteinortho showing the detected orthologous genes in the four reference genomes employed. Type of file: excel file

Additional File 4; Supplementary Table 3. The new genomic coordinates of each gene present in at least one of the four reference genomes, calculated according to their corresponding location in the final merged alignment. The new genes were named according to the Proteinortho results obtained, calling them using the acronyms of TPASS, TPANIC, TENBA or TPECDC2 depending on the orthology found and their presence or absence in each reference genome. Type of file: excel file

Additional File 5; Supplementary Table 4. The number of SNPs detected per gene present in at least one of the four different reference genomes employed in the mapping analysis. Type of file: excel file

Additional File 6; Supplementary Table 5. The results of the likelihood-mapping test performed. 160 genes showed some phylogenetic signal and were retained for the subsequent analyses. (Zones 1-3 represent cases in which one topology has a significantly higher likelihood than the two

alternative topologies; zones 4-6 represent cases in which one topology has a significantly lower likelihood than the other two, and zone 7 represents cases in which all the topologies have similar likelihoods, hence the corresponding gene does not carry enough phylogenetic signal to differentiate between the 3 evolutionary hypotheses tested in each case.). Type of file: excel file

Additional File 7; Supplementary Table 6. Topology test results for the genes retained in previous PIM steps. The table presents the results of the two tests (p-values for SH, and a posterior weight for ELW). All these genes were examined in detail *a posteriori* because the two tests rejected the reference tree topology with the gene alignment ($p < 0.20$, weight value close to 0) and the complete genome alignment rejected the gene topology (reciprocal incongruence, $p < 0.05$ and weight value close to 0). Type of file: excel file

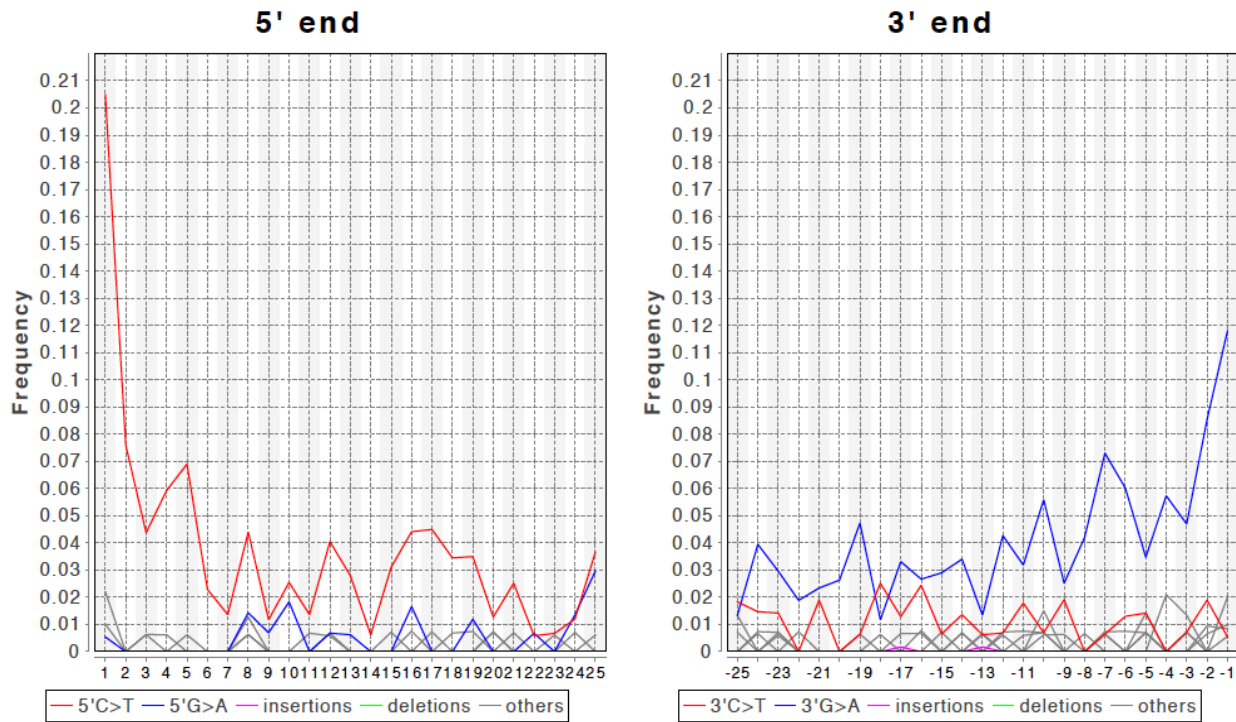
Additional File 8; Supplementary Table 7. Number of SNPs for each gene extracted from the alignment of each genome-mapping dataset, based on the coordinates of those genes in the respective genomic reference. Type of file: excel file

Additional File 9; Supplementary Table 8. Topology test results for genes retained in previous PIM steps, using four different reference genomes in the mapping analysis. The tables display p-values for the SH test and posterior weights for ELW. Genes were examined further if both tests rejected the reference tree topology with the gene alignment ($p < 0.20$, weight value close to 0), and the complete genome alignment rejected the gene topology (reciprocal incongruence, $p < 0.05$, and weight value close to 0). Type of file: excel file

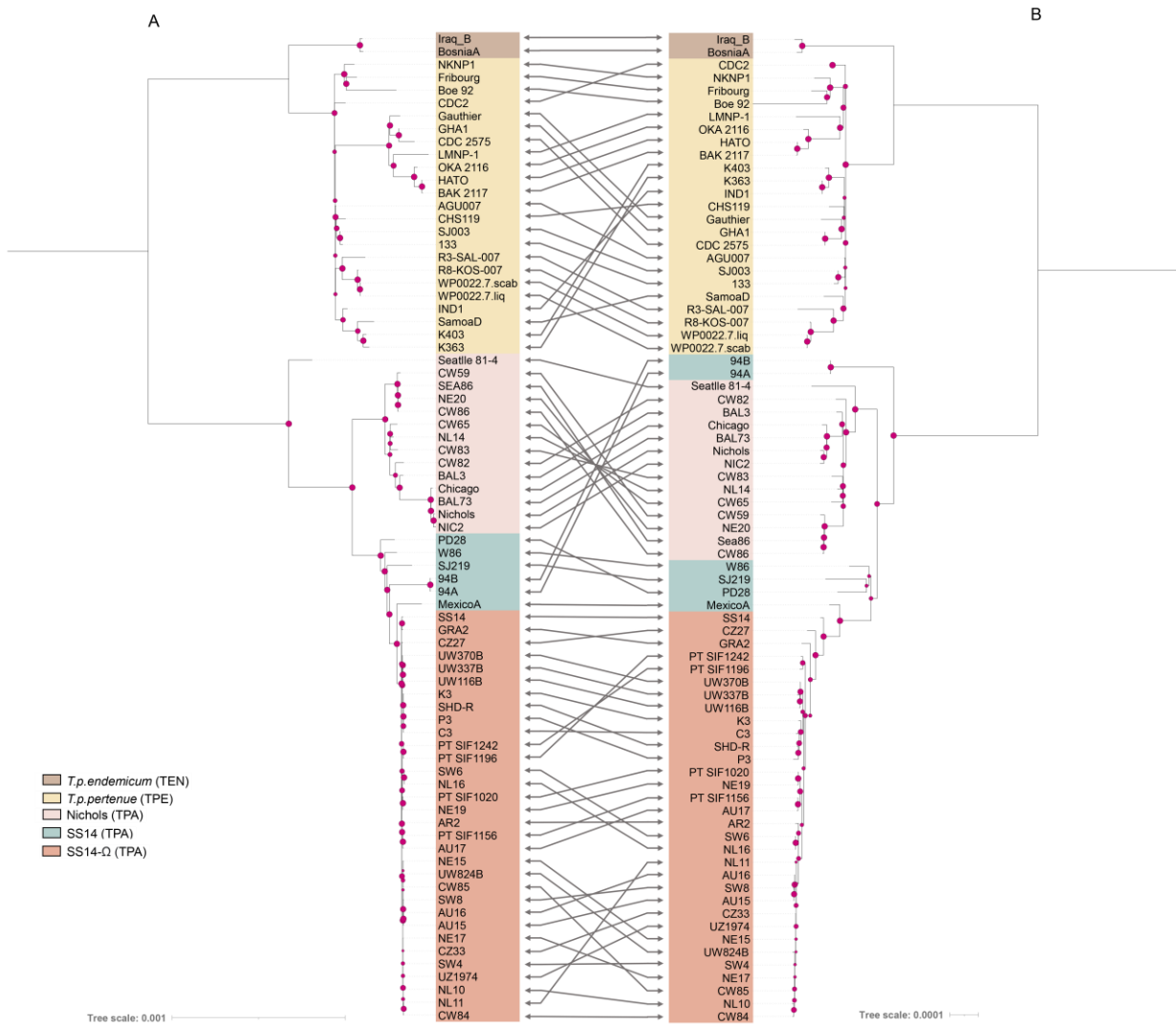
Additional File 10; Supplementary Table 9. Results of the aBSREL test, a "branch-site" model implemented in HyPhy to study the effects of positive selection along all different lineages on the phylogeny of the 18 putative recombinant genes (Table 1) on the phylogeny of genes with 3 or more SNPs present in at least one reference genome. The total number of branches tested and the number of branches under positive selection are indicated in the table plus the strains inside in each node or branch and the p-value obtained. The color legend is detailed below. Type of file: excel file

Additional File 11; Supplementary Table 10. Functional significance of the genes detected as recombinant and/or under positive selection according to Uniprot or in the literature. Type of file: excel file

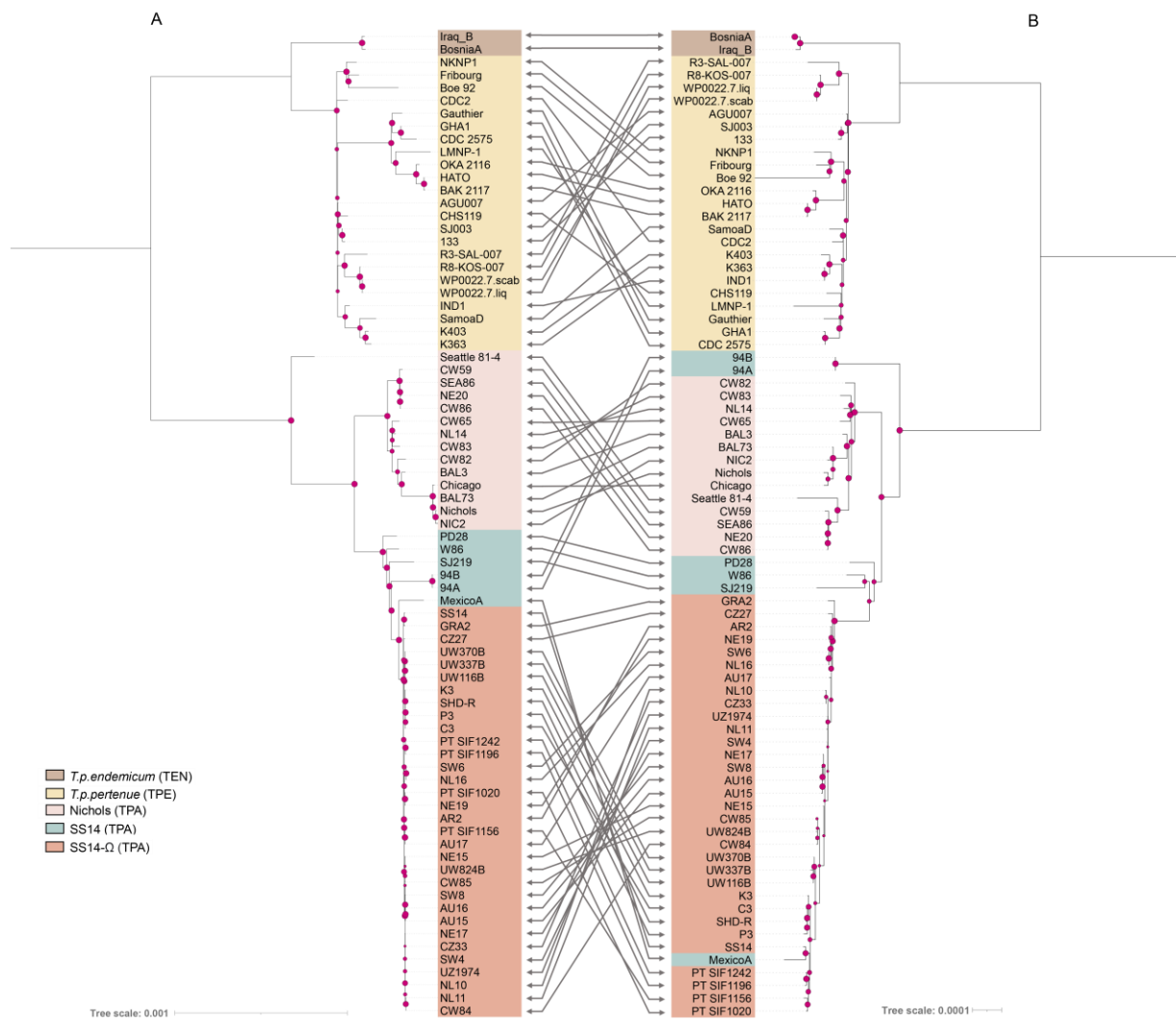
Supplementary Figures



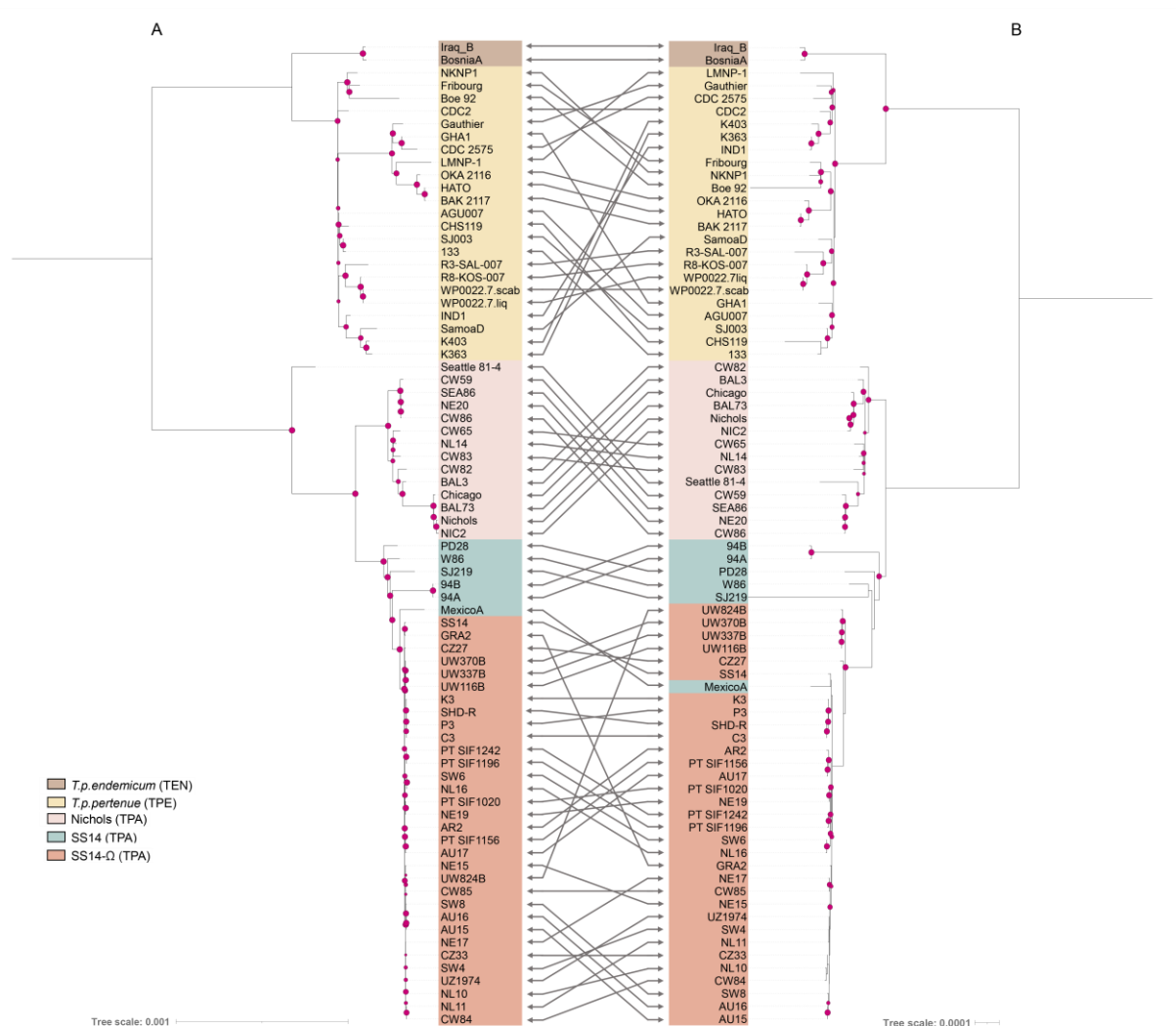
Supplementary Figure 1. Damage profile obtained by MapDamage [5] showing the misincorporation patterns and the damage at the end of sequencing reads of the new historical W86 genome obtained before the capture process. A pattern of cytosine-to-thymine base misincorporation accumulated at the end of the reads is indicative of authentic ancient DNA in the sample.



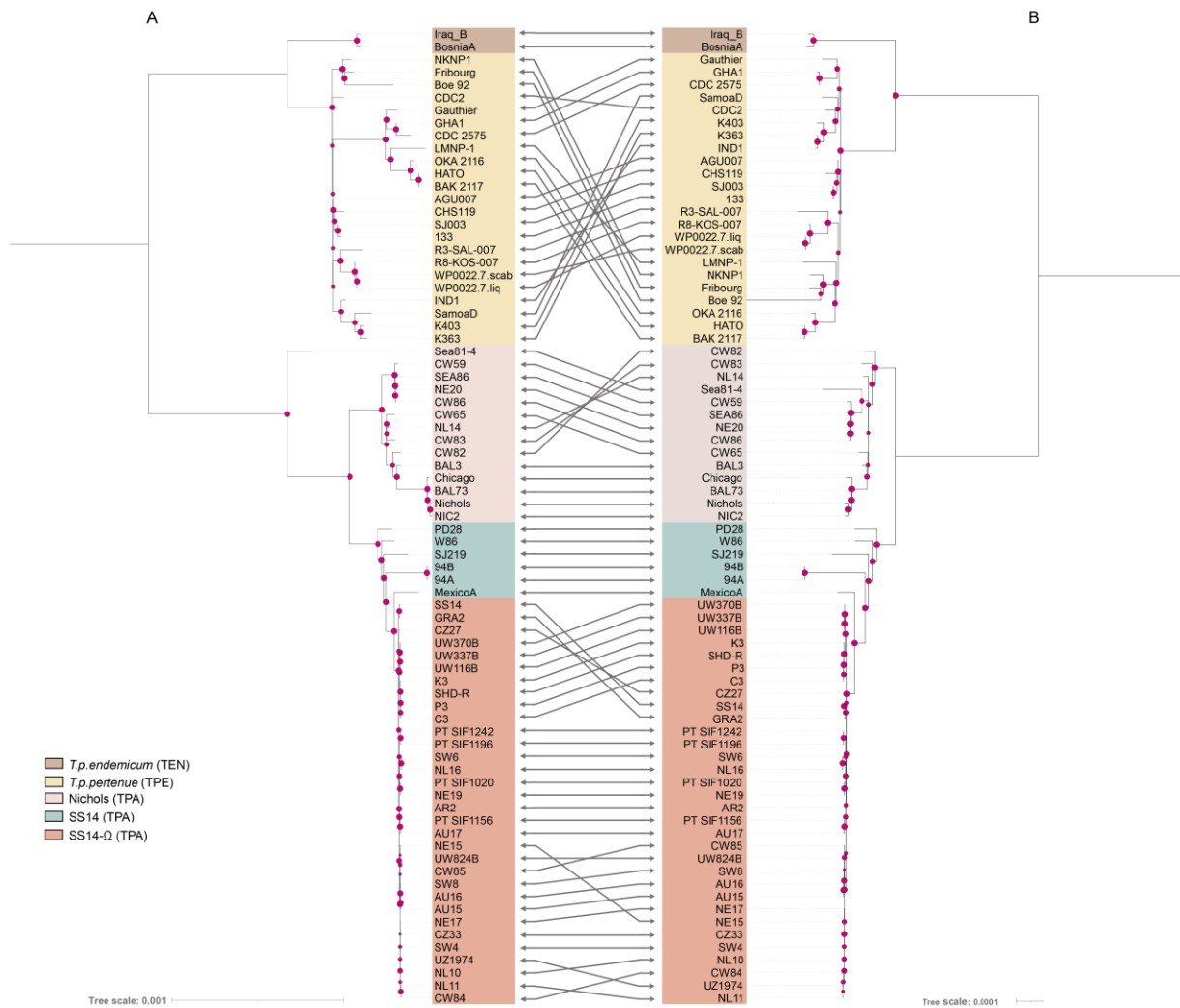
Supplementary Figure 2. A) Maximum likelihood tree (ML) obtained by IQ-TREE using GTR+G+I as the evolutionary model with 1000 bootstrap replicates using the final whole-genome multiple-reference-based alignment generated. B) Maximum likelihood tree (ML) obtained by IQ-TREE using GTR+G+I as the evolutionary model with 1000 bootstrap replicates using the CDC2-mapping genome alignment generated. Type file: pdf file



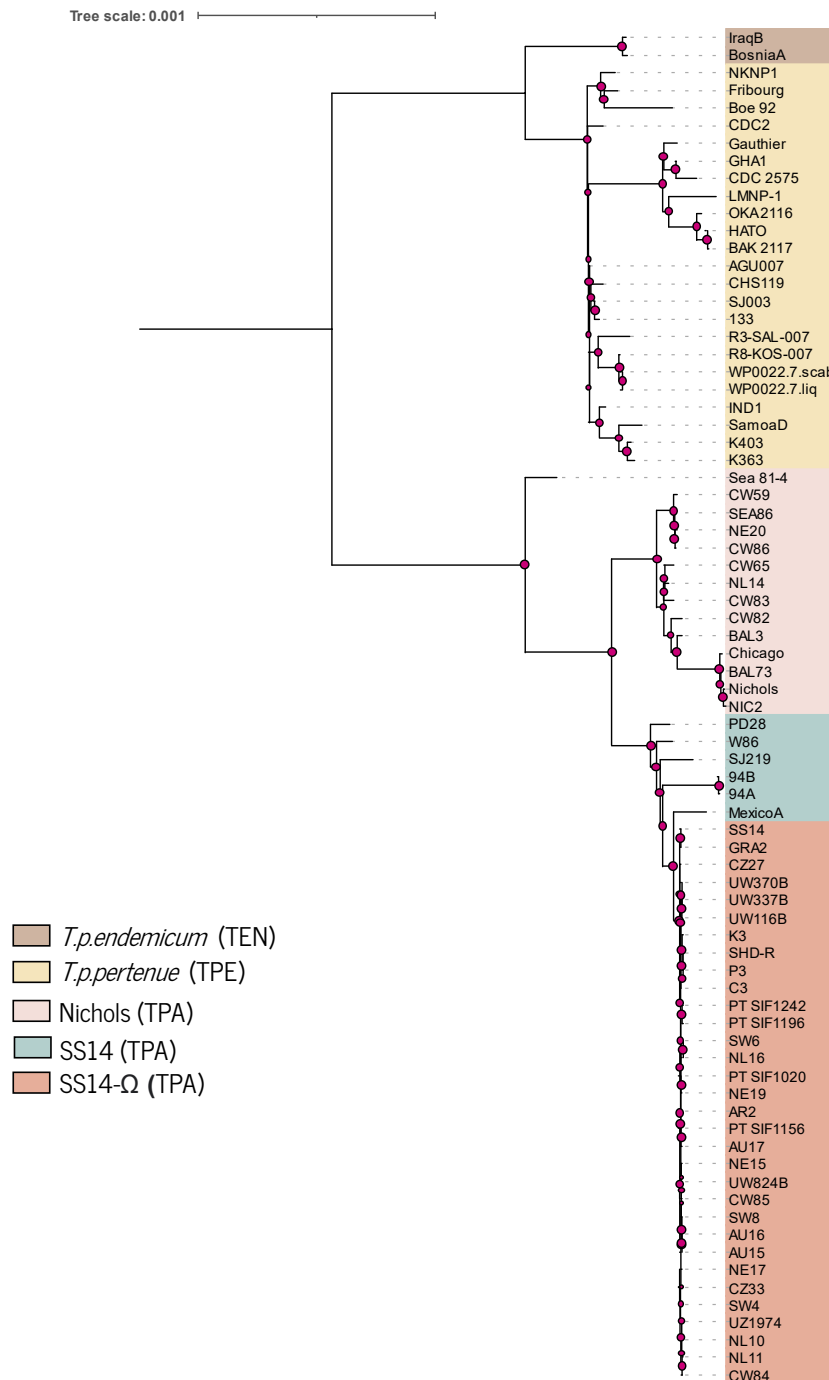
Supplementary Figure 3. A) Maximum likelihood tree (ML) obtained by IQ-TREE using GTR+G+I as the evolutionary model with 1000 bootstrap replicates using the final whole-genome multiple-reference-based alignment generated. B) Maximum likelihood tree (ML) obtained by IQ-TREE using GTR+G+I as the evolutionary model with 1000 bootstrap replicates using the BosniaA-mapping genome alignment generated. Type file: pdf file



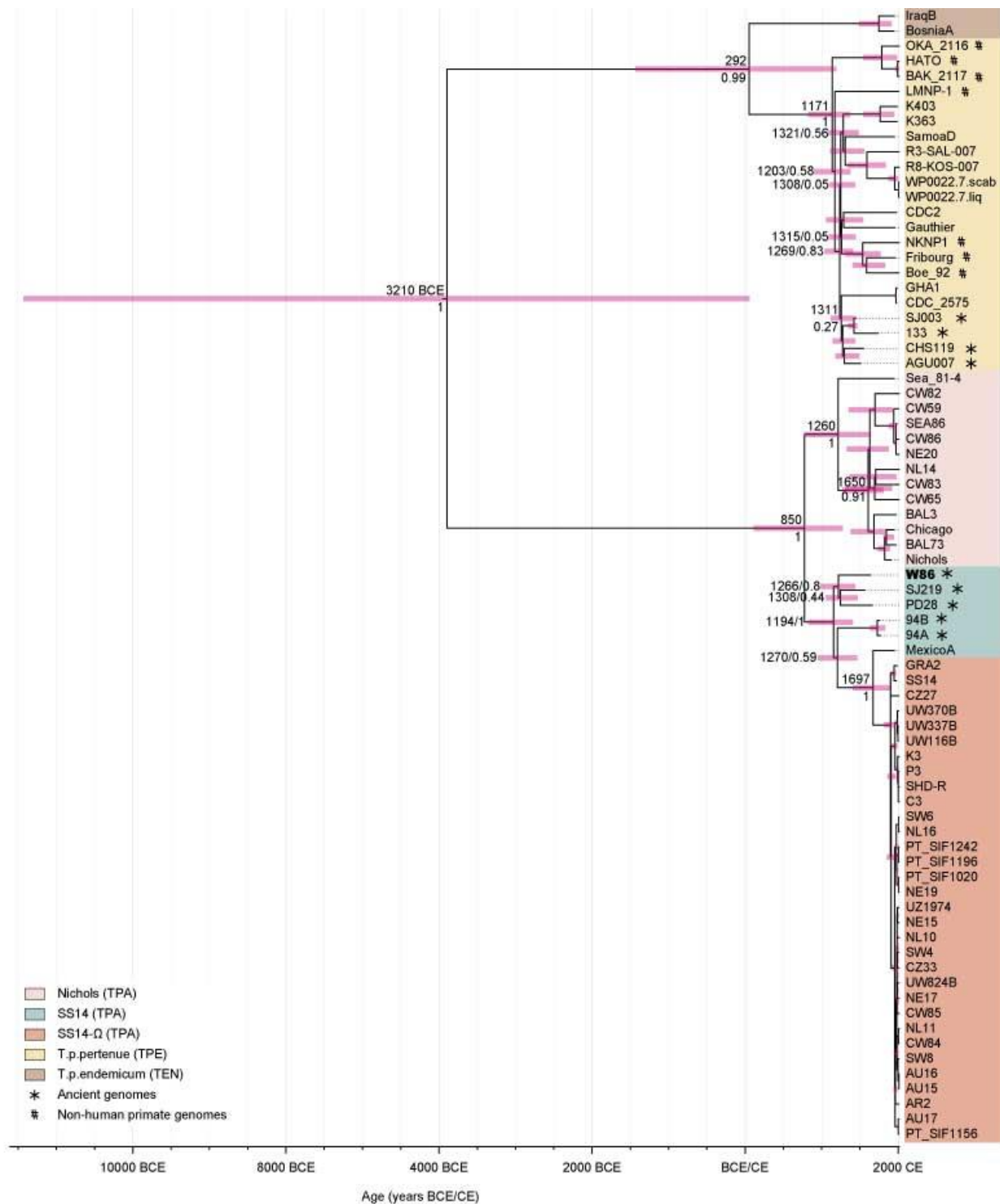
Supplementary Figure 4. A) Maximum likelihood tree (ML) obtained by IQ-TREE using GTR+G+I as the evolutionary model with 1000 bootstrap replicates using the final whole-genome multiple-reference-based alignment generated. B) Maximum likelihood tree (ML) obtained by IQ-TREE using GTR+G+I as the evolutionary model with 1000 bootstrap replicates using the Nichols-mapping genome alignment generated. Type file: pdf file



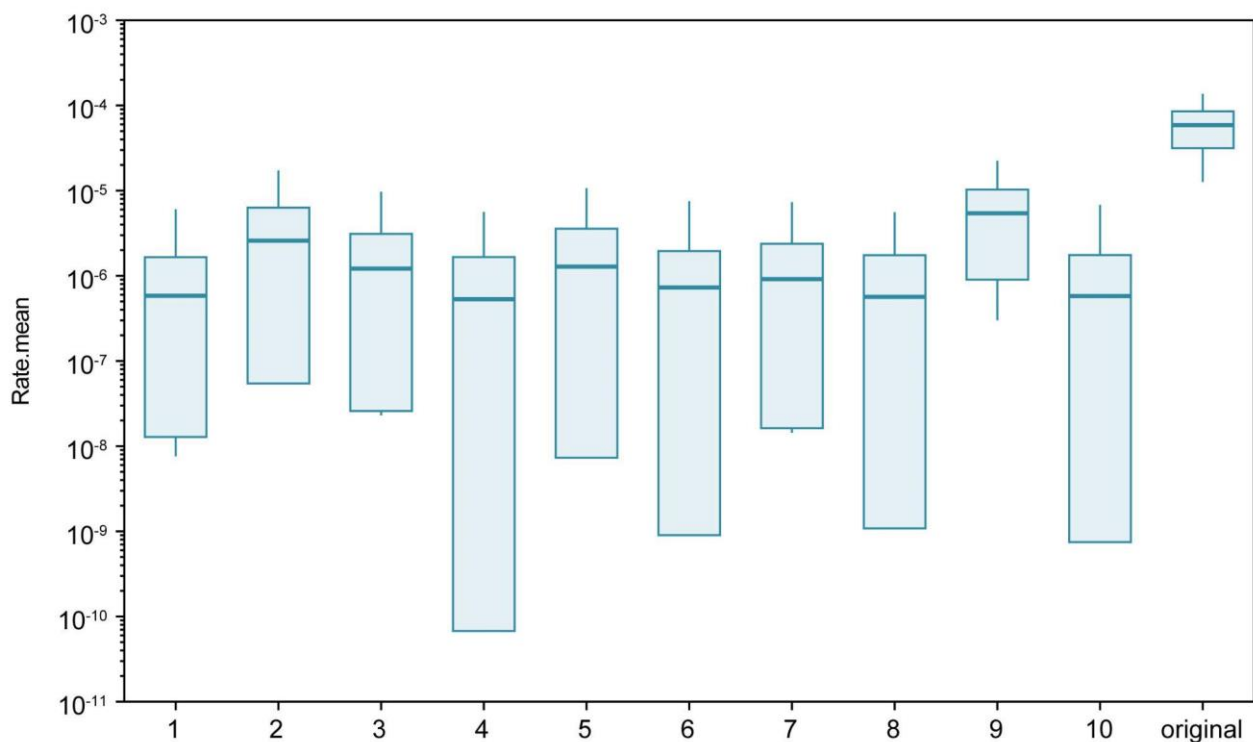
Supplementary Figure 5. A) Maximum likelihood tree (ML) obtained by IQ-TREE using GTR+G+I as the evolutionary model with 1000 bootstrap replicates using the final whole-genome multiple-reference-based alignment generated. B) Maximum likelihood tree (ML) obtained by IQ-TREE using GTR+G+I as the evolutionary model with 1000 bootstrap replicates using the SS14-mapping genome alignment generated. Type file: pdf file



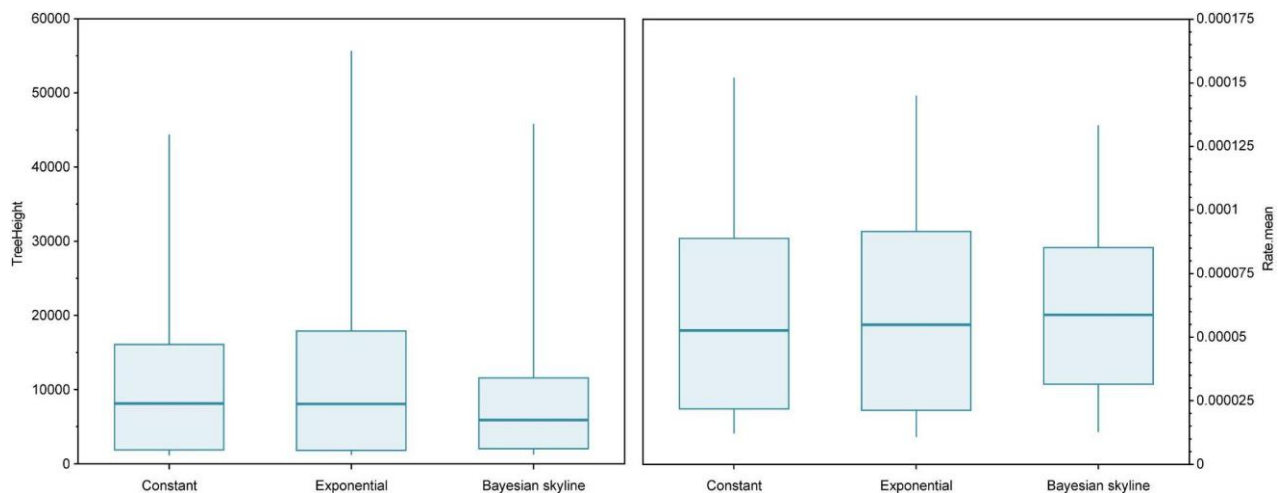
Supplementary Figure 6. Maximum likelihood tree obtained with all genes included in the whole-genome multiple-reference-based alignment without collapsing SS14-Ω. The different clades corresponding to yaws (TPE) and bejel (TEN) subspecies, and the Nichols and different SS14 lineages of the syphilis clade (TPA) are indicated in the figure according to the corresponding color legend. Bootstrap support values higher than 70% are indicated by red circles, sized in respect to the support of the node. Type file: pdf file



Supplementary Figure 7. Maximum clade credibility (MCC) tree of the dataset consisting of 66 modern (IND1 and NIC2 samples not included) and 9 ancient genomes (newly sequenced sample, W86, is in bold) with tp0897, tp0316 and tp0317 and recombinant genes excluded, estimated in BEAST2 v2.6.3 under an uncorrelated lognormal relaxed clock model and a Bayesian skyline plot demographic model. Median age and Posterior Bayesian support estimates are provided for selected nodes. Pink bars show node age 95% HPD. Type file: pdf file



Supplementary Figure 8. Date randomization test of temporal signal in the dataset. The samples were randomly reassigned ages in ten iterations and the BEAST v2.6.3 phylogenetic estimation was re-run for each of them. The 95% HPDs of the posterior estimates of the substitution rates for the datasets with randomized dates do not overlap with the rate estimate for the original dataset, indicating that the estimated evolutionary timescale is informed by the ages of the samples that were used as calibration for the molecular clock. Type file: pdf file



Supplementary Figure 9. Comparison of tree models for BEAST v2.6.3 phylogenetic estimation of the dated phylogeny. Tree height and mean rate parameter posterior is shown for constant, exponential and Bayesian skyline population size models. The Bayesian skyline model was used for the final analysis. Type file: pdf file

Supplementary Material (Files)

All Supplementary Material files are available at the DOI: <https://doi.org/10.5281/zenodo.14535982>

Supplementary Material 1. Mapping-genome alignment of the 77 *T. pallidum* genomes using CDC2 genome as reference. Type file: fasta file

Supplementary Material 2. Mapping-genome alignment of the 77 *T. pallidum* genomes using BosniaA genome as reference. Type file: fasta file

Supplementary Material 3. Mapping-genome alignment of the 77 *T. pallidum* genomes using SS14 genome as reference. Type file: fasta file

Supplementary Material 4. Mapping-genome alignment of the 77 *T. pallidum* genomes using Nichols genome as reference. Type file: fasta file

Supplementary Material 5. Multiple sequence alignment of 77 *T. pallidum* genomes using the MRB strategy. Type file: fasta file

Supplementary Material 6. Multiple sequence alignment of 77 *T. pallidum* genomes using the MRB strategy after removing the 20 recombinant genes along with the *tp0316*, *tp0317* and *tp0897* genes from the multiple genome alignment. Type file: fasta file

Supplementary Material 7. In-house script A to ensure that the final Multiple sequence alignment obtained was correct. Type file: fasta file

Supplementary Material 8. In-house script B to ensure that the final Multiple sequence alignment obtained was correct. Type file: fasta file

Supplementary Material 9. In-house script to compute the number of SNPs per gene. Type file: fasta file