Inferring patterns of recombination and divergence with ancient and modern treponemal genomes

Gülfirde Akgül*(1), Marta Pla-Díaz*(2,3), Martyna Molak (4,5), Louis du Plessis (6,7), Hanna Panagiotopoulou (5), Karolina Doan (5), Wiesław Bogdanowicz (5), Paweł Dąbrowski (8), Maciej Oziębłowski (9), Joanna Grzelak (8), Natasha Arora (10), Fernando González-Candelas† (2,3), Kerttu Majander†(1,11,12), Verena J. Schuenemann†(1,11,12).

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

* These authors contributed equally to this manuscript.

† Authors for correspondence:

Verena J. Schünemann <verena.schuenemann@iem.uzh.ch>

Kerttu Majander <kerttu.majander@gmail.com>

Fernando González-Candelas <fernando.gonzalez@uv.es>

KEYWORDS: ancient genomes; historical samples; syphilis; selection; tMRCA; Treponema pallidum

Abstract

The treponemal diseases yaws, bejel and syphilis, all caused by subspecies of the bacterium *Treponema pallidum*, are re-emerging worldwide, yet their origins and spread remain largely unresolved. Albeit still rare, reconstructed ancient genomes of various *T. pallidum* strains now exist, to complement the analyses on the modern-day diversity of treponemes. Here, we report a new high-coverage (35X) ancient genome of a historical *T. pallidum* subsp. *pallidum* strain from the 17th century. This novel addition, combined with a selection of 76 modern and historical strains, enables a new level of in-depth investigation of treponemal evolution across all human-infecting strains, with detailed analyses on recombination, positive selection, and divergence history of *T. pallidum* subspecies. Altogether 18 recombinant genes with strong evidence for effects of positive selection are identified, potentially responsible for virulence and immune evasion. The profound impact of recombination is in particular demonstrated in the diversification of the yaws- and bejel-causing clades, as excluding these recombinant genes from phylogenetic analysis causes these well-separated subspecies to cluster into a single clade. Both the involvement of ancient genomes in several recombination events, and the molecular clock dating of the subspecies' divergence history emphasize the importance of recombination in the early adaptations of all *T. pallidum* strains. These findings are crucial in resolving the evolutionary history of *T. pallidum*, and in understanding the functionalities of treponemes beyond what could be achieved with modern genomic data alone.

Introduction

Treponemal diseases, including the widespread sexually transmitted syphilis, present a continuous risk to human health throughout the world, as it leads to systemic damage and can be life-threatening when left untreated. Historically, syphilis has caused epidemics, starting from an aggressive initial outbreak in Europe at the end of 15th century, almost 500 years before antibiotics could be successfully used as treatment. The causative agent, *Treponema pallidum,* is well-adapted to avoid host immunity, can acquire resistance to macrolide antibiotics, and is re-emerging in contemporary human populations (World Health Organization 2021). The other treponematoses, yaws (*T. pallidum* subsp *pertenue*; TPE) and bejel (*T. pallidum* subsp *endemicum*; TEN), are transmitted predominantly among children in rural communities of tropical countries through skin contact (Grange et al. 2013; Noda et al. 2018; Shinohara et al. 2022), whereas syphilis (*T. pallidum* subsp. *pallidum*; TPA) persists globally and is typically transmitted either sexually or congenitally, i.e. during pregnancy or childbirth. Interestingly, treponemes have occasionally been observed to adopt different transmission routes and clinical etiology than expected from the causative agent (Noda et al. 2018; Shinohara et al. 2022). Whether these events necessitate genetic exchanges between the subspecies or represent responses to contemporary environments is as of yet unknown. However, they demonstrate the versatile adaptive behavior of treponemal strains.

The categorization of treponemal subspecies into the syphilis-causing strains and those causing the endemic treponematoses, yaws and bejel, was proposed by Pillay et al. (1998) using clinical specimens. However, the genetic factors defining the functional differences among these groups are still largely enigmatic. The distinction of *T. pallidum* infections from other diseases, and even among its own subspecies, is still prone to error in clinical settings (Fonseca et al. 1999), and extremely challenging to conduct from paleopathological remains. Neither symptoms nor serologic methods can reliably differentiate among treponemal subspecies, rendering any precise distinctions dependent on molecular typing (Mitjà et al. 2013; Grillová et al. 2018). The existing genomic sequences of each treponemal group show low variation (0.02% divergence across *Treponema* subspecies), reducing the resolution of phylogenetic signals and thwarting attempts to infer past evolutionary events. Clinical samples contain very little DNA and until recently, treponemes were considered unculturable *in vitro*. These obstacles have complicated the recovery of whole genomes and the exploration of representative treponemal diversity. Nonetheless, the increasing number of draft genomes available have yielded new findings on the roles of recombination (Strouhal et al. 2018; Grillová et al. 2019; Vrbová et al. 2020; Pla-Díaz et al. 2022) and natural selection (Giacani et al. 2012; Kumar et al. 2018) These findings underscore that key factors in treponemal evolution are still poorly understood, and new approaches are needed to disentangle the essential mechanisms behind the persistence and success of human-adapted treponemes.

Ancient treponemal genomes provide an unforeseen opportunity to integrate detailed information on evolutionary relations and diversification into the current view of treponemal subspecies. Historical strains provide phylogenetic inference with otherwise inaccessible depth and robustness. Additionally the dated historical samples provide calibration points that allow more

accurate and precise molecular dating of past evolutionary events. However, complete and representative genome reconstruction required by these methods is often constrained by the low quantity and high level of degradation of ancient DNA. The fragmentary DNA material and read sparsity, resulting in low genomic coverage, pose a challenge for genome reconstruction and the identification of single nucleotide polymorphisms (SNPs), especially for highly variable regions. The increase of the number of high-coverage ancient genomes available is therefore of paramount importance to achieving accuracy of genomic and phylogenetic analyses.

In this study, we present a new 17th century genome "W86" of a syphilis-causing strain, sequenced at 35X coverage. We analyze the W86 genome together with a set of 76 modern and ancient genomes that are representative of the diversity observed across the three *T. pallidum* subspecies. As an optimized mapping approach, we employ four different reference genomes and map each genome to its closest reference (Valiente-Mullor et al. 2021). This approach enables us to gain comprehensive insights on the roles of recombination and positive selection in the evolution of treponemal genomes. Several genes newly identified as recombinant are revealed, many of which show signals of evolution under positive selection, and found to be involved in inter-subspecies horizontal gene transfer. These genes are potentially responsible for virulence and immune evasion, aside from playing an important role in the diversification of the yaws- and bejel-causing clades. Furthermore, the inclusion of our high-coverage W86 genome provides an additional calibration point for Bayesian molecular clock dating. Our findings illuminate the role of recombination and selection as driving forces for the diversification of *T. pallidum* subspecies and help us to identify key events in the evolutionary history of *T. pallidum*.

RESULTS

Geographical origins and osteological analyses of the W86 sample

A bone sample yielding the new treponemal genome for this study was collected from individual W86, from the 17th century Ostrow Tumski cemetery in Wroclaw, Poland, most probably dating to between 1621 and 1670 (Pankiewicz and Witkowski 2012). The sample, which did not display paleopathological signs of infection, was originally collected as part of a human population genetics project. Interestingly, the presence of *T. pallidum* was detected through a routine pathogen screening. More information about the chemical analysis and a skeletal material description for the new historical sample obtained is detailed in Supplementary Notes 1 and 2.

Pathogen screening for the new historical sample

The W86 sample was subjected to a screening procedure using direct shotgun sequencing (Hodges et al. 2009; Arora et al. 2016). This process resulted in 581 unique reads that mapped to a *T. pallidum* reference, confirming the sample as positive for treponemal DNA. The reads showed 20% deaminated bases at the 5′ ends and 12% at the 3′ ends, with an average fragment length of 68 bp, signaling authenticity of ancient DNA (Briggs et al. 2007; Sawyer et al. 2012) (for the damage profile, see Supplementary Figure 1). Following these screening and authentication steps, genome-wide enrichment for *T. pallidum* DNA (Hodges et al. 2009; Arora

et al. 2016; Majander et al. 2020) and high throughput sequencing were conducted. The resulting 87.8 million raw reads were merged and duplicate reads were removed.

New ancient genome reconstruction and multiple reference-based genome alignment

As it is known that SNP calling is dependent on the choice of reference genome used for mapping (Valiente-Mullor et al. 2021), in this study we conducted a proximity evaluation to determine the most suitable reference for each strain. The reads of each of the 61 strains with raw data available were mapped to one of four representative reference genomes, according to the three different *T. pallidum* subspecies and the Nichols- and SS14-lineages clades within TPA (CDC2, BosniaA, Nichols, and SS14). After mapping, a multiple genome alignment was generated with these sequences as well as the 15 previously assembled genomes with no raw data available.

For the new ancient genome W86, mapping of reads to all four reference genomes and subsequent phylogenetic analyses consistently revealed placement within the TPA clade. Nonetheless, the position in TPA varied depending on the reference genome. When CDC2 or Bosnia were used as reference genome, the W86 genome was placed basal to the TPA clade, but when Nichols or SS14 were employed as reference then the W86 lineage was placed basal to only the SS14 lineage. In light of these results, SS14 was considered the closest reference sequence for W86 and we classified W86 as a SS14-like strain.

The W86 sample yielded 524,587 unique treponemal reads that could be mapped to the SS14 reference, resulting in 98.21% of the bases covered with a minimum of 3X and a median depth coverage of 35X (Supplementary Table 1). Variant calling resulted in the identification of 163 SNPs.

******* Figure 1 about here *******

The W86 sequence was added to the multiple genome alignment and realigned, resulting in a final alignment of 77 sequences spanning a total of 1,141,136 nucleotides (Supplementary File 1) with 4,822 SNPs detected (details on each of these genomes is provided in Supplementary Table 1).

To identify the location of genes in this multiple genome alignment, we conducted an orthology inference with the four reference genomes. This analysis revealed a total of 1161 genes, 24 of which were not shared across all four references (Supplementary Table 2). The genomic coordinates for the individual 1161 genes in the multiple genome alignment are detailed in Supplementary Table 3.

In order to analyze the sensitivity of our new genome to macrolide antibiotics, two known *T. pallidum* 23S ribosomal RNA gene mutations (A2058G and A2059G) were examined. However, neither of these mutations were present in the W86 genome.

Analysis of recombination

To obtain a reliable phylogenetic reconstruction, it is necessary to remove genomic regions that are not strictly subject to vertical inheritance, e.g. recombinant regions or loci with intra- or intergenic conversion. Previous studies examining different sets of genomes have identified and analyzed such loci (Pětrošová et al. 2012; Štaudová et al. 2014; Strouhal et al. 2018; Grillová et al. 2019; Majander et al. 2020; Pla-Díaz et al. 2022; Arora et al. 2016). Here we proceed to carry out a comprehensive investigation of recombination and its impact, using an unprecedented dataset including all ancient genomes sequenced to date. Our detailed recombination detection pipeline, the phylogenetic incongruence method (PIM), yielded 26 recombinant regions. These derived from 18 different genes, and encompassed a total of 1,024 SNPs (21.24% of the total SNPs) among the *T. pallidum* strains analyzed here (Table 1 and Supplementary Figures 3-20). Six of these recombinant genes were novel compared to previously published results. The average length of the recombinant regions was 469 bp, with a minimum length of 5 bp and a maximum of 2,097 bp. Details on the intermediate results of the likelihood mapping and topology tests from the PIM procedure applied to these genomes are given in the Supplementary Material (Supplementary Note 3 and Supplementary Tables 4-6), and each recombinant gene is described in detail in the Supplementary Figures 2-19. In addition to the identification of recombination, we also classified 11 of the the putative 26 recombination events as ancient (before 1500 CE) and 15 as modern (after 1500 CE) events, according to the phylogenetic placement and dating of the strains and/or nodes involved in the recombination events.

******* Table 1 about here *******

We detected 15 genes with one recombinant region and 3 genes with more than one: *tp0136*, with 7 regions, *tp0621* and *tp0865*, with 2 recombinant regions each (Table 1, Supplementary Figures 2-19). Interestingly, the sequences corresponding to the putative recombination regions detected in two genes, *tp0859* and *tp1031* (Table 1), could not be found in any public database, and therefore most probably resulted from an external horizontal gene transfer event from a so far unidentified *Treponema* subspecies. Across the other 13 genes, for all 24 recombinant regions except one, the most plausible direction of genomic exchange is an inter-subspecies transfer occurring from TPE/TEN to TPA. Strikingly, one recombinant region in the *tp0136* gene corresponded to an intra-subspecies transfer within TPA (Table 1). We indicate the most parsimonious direction of transfer for the recombinant regions in Supplementary Note 5 and Supplementary Figures 2-19 while also providing several alternative scenarios that are also consistent with the data.

Notably, inclusion of all ancient genomes sequenced to date in our study allowed us to explore their role in recombination and their impact on current patterns of *T. pallidum*'s genetic diversity. Overall, we found eight recombination regions and events involving ancient genomes, four of these including W86, the new genome sequenced. One of these genes, *tp0488*, was found in previous studies to have an unusual sequence in the strain Mexico A, which clusters with TPA sequences; this sequence was identical to TPE/TEN strains. Our study indicates that this region in *tp0488* was most probably transferred from TPE/TEN to both MexicoA and the lineage of W86 as well. An event in the *tp0179* gene was detected with the W86 and PD28 strains and the modern Nichols-lineage clade as recipients and the TPE/TEN clades as putative donors.

Furthermore, W86 was found in two further recombination events detected in *tp0548* and *tp0558* genes. For the event involving *tp0548*, the W86 lineage and Nichols-lineage clade were the recipients whereas TPE/TEN were the putative donors. For the event in *tp0558*, all TPA ancient genomes and the common ancestor of the complete SS14-lineage clade were receptors and TPE/TEN the putative donors. The ancient genomes 94A and 94B were also involved in the aforementioned event detected in *tp01031*, whereas all the other strains from the SS14-lineage clade were the putative receptors from an external transfer. In the events detected for the genes *tp0179*, *tp0488*, *tp0548* and *tp1031*, the other TPA ancient genomes might also be involved but the missing data for the SNPs that define the recombination event precludes a stronger inference.

Phylogeny reconstruction

To build a vertical-inheritance genome phylogeny, we removed the 18 recombinant genes detected here, as well three genes that are hypervariable and/or subject to gene conversion (the *tp0316*, *tp0317* and *tp0897)*, from the original alignment (1,141,136 bp with 4,822 SNPs, Supplementary File 1). The resulting alignment encompassed 1,107,881 bp with 3,221 SNPs (Supplementary File 2). Both multiple alignments were used to construct maximum-likelihood trees. The removal of non-vertically inheritance genes had a notable effect on the phylogenetic reconstruction of *T. pallidum.* The topologies of the two ML trees, with and without these loci, are compared in Figure 2.

******** Figure 2 about here *******

Based on our findings, and building on previous phylogenetic classifications and nomenclature of the SS14-lineage (Beale et al. 2019; Arora et al. 2016), we defined the SS14-Ω sublineage as the clade that includes all SS14 genomes from clinical and modern samples, which was previously defined as a mostly epidemic, macrolide-resistant cluster that emerged after, and possibly prompted by, the discovery of antibiotics (Beale et al. 2019; Arora et al. 2016). Thus, the SS14 lineage is the sister clade to the Nichols-lineage according to our comprehensive phylogenetic tree.

Notably, the phylogenetic placement of the SS14 ancient genomes 94A, 94B, SJ219, PD28, and W86 was congruent in the tree reconstructed with all genes, but had stronger bootstrap support when non-vertically transmitted regions were removed (Figure 2). The difference in bootstrap support could be due to the effects of recombination. After removing the recombinant genes, the new historical genome W86 still occupies a strongly supported, basal position in the SS14 lineage, confirming its initial classification as a TPA genome (Figure 2).

By contrast, we found several topological incongruences between the two maximum likelihood trees obtained (Figure 2). The unexpected placement of Seattle81-4 in the whole-genome tree, distant from all TPA genomes and basal to the joint Nichols and SS14 lineages, changes after removal of the aforementioned 21 genes, placing it at a basal position within the Nichols lineage. A similar topological change occurs to strain CW82, originally basal to the Nichols lineage, but firmly within it in the tree obtained using the recombination-free alignment. In the TPE clade, the CDC2-2575 strain is included in a subclade with strains LMNP-1 and Gauthier in the

complete genome-based tree, but occupies a position basal to the whole TPE/TEN clade in the recombination-free alignment tree.

Nevertheless, the most dramatic change in the tree constructed from the recombination-free alignment involves the TEN genomes. In this phylogeny, both TEN genomes are nested within the TPE clade with high support and do not form a separate sister clade, which is the case in the phylogeny derived from the multiple genome alignment.

SNPs involved in the divergence between TPE and TEN

To analyze the genetic changes responsible for the differentiation between TPE and TEN in greater detail, we performed an ancestral sequence reconstruction of the complete genome-based tree with the multiple genome alignment using TreeTime.

We focused the analyses on the polymorphisms specific to either the TPE or TEN clade, and on the possible reversions occurring in the ancestral nodes of TPE and TEN strains, for the 21 genes removed from the multiple genome alignment, and also for the ancestral node of the TEN strains. The results obtained are summarized in Table 2 (see also Supplementary Tables 7-9). The nexus file with the annotated changes obtained in the ancestral sequence reconstruction of the multiple genome alignment and the reference phylogenetic tree by TreeTime are provided in Supplementary File 3. We also calculated the probability of each polymorphism present between TEN and TPE subspecies using RAxML, as detailed in Supplementary Table 8.

******* Table 2 about here *******

From the total of 2,325 SNPs present between TEN and TPE subspecies, 1,695 SNPs (Table 2) were removed to build the recombination-free tree. Mapping the TPE and TEN node-defining SNPs on the new reference phylogeny obtained, we found 185 SNPs in the 21 excluded genes (Table 2) that were involved in the differentiation of TPE and TEN from their common ancestor. Apart from the SNPs in the 21 genes removed, the branch leading to the MRCA of the for TEN genomes remains very long, caused by 402 SNPs mainly present in the genes *tp0856*, *tp0548*, *tp0856a*, *tp0858*, *tp0136* and *tp0483*, among others (Supplementary Table 8).

The marginal probabilities for the ancestral nodes of TPE and TEN subspecies calculated with RaxML (Supplementary Table 9), showed a highest posterior probability of each of the 185 SNPs detected by Treetime belonging to the recombinant genes and to loci *tp0897*, *tp0316* and *tp0317*, thus supporting the roles of these genes in the divergence between the TPE and TEN subspecies.

Selection analysis

To study the effects of positive selection, from the total of 1161 genes (Supplementary Table 3), 317 genes with three or more SNPs were analyzed by Hyphy, using aBSREL, a "branch-site" model to detect positive selection. This analysis identified 28 genes showing evidence of positive selection (Supplementary Table 10), all of which have a high number of SNPs in the strains comprising our dataset (Supplementary Table 4). These genes include 10 putatively recombinant

genes (*tp0131, tp0136, tp0462, tp0621, tp0856 tp0865, tp0859, tp0967, tp0968* and *tp1031*). Although most branches in which these recombinant genes were found to be under positive selection in the deep branches of the phylogeny, including the branch leading to the MRCA of the TPE/TEN subspecies, in some cases the signature of selection was limited only to branches leading to the specific strains found to be involved in recombination events of these genes.

Interestingly, despite poor coverage of the *tpr* genes in several genomes, apart from two other *tpr* genes that had also been detected previously as recombinant (*tp0131 and tp1031),* we found two more genes of this family (*tp0620* and *tp0117)* with evidence of positive selection. For the locus *tp0620* (*tprI),* the branch leading to the MRCA of the TPE and TEN strains in the phylogeny appears to be under positive selection. This gene has previously been described (Strouhal et al. 2018) as having a modular genetic structure in certain TPA strains. That structure differs from that of other *T. pallidum* subspecies, which might explain the detection of positive selection. Moreover, aBSREL detected the *tp0117 (tprC)* gene to be under positive selection in the branch leading to the MRCA of TPE and TEN strains, and for the recombinant *tp0131 (tprD)* gene in the branch leading to the MRCA of Nichols, Chicago, LMNP1, Gauthier, and CDC2575 strains. Previous studies have considered these two genes as paralogs (Pětrošová et al. 2012; Strouhal et al. 2018), created by a gene conversion mechanism that would have copied a portion of the *tp0117* gene into *tp0131,* thus explaining differences between these genes in some strains (Gauthier, LMNP-1, CDC_2575, Nichols and Chicago).

We also detected evidence of positive selection in loci *tp0314* and *tp0619* in the branch leading to the MRCA of TPE and TEN strains. This could be a consequence of a previously described gene duplication(Mikalová et al. 2020), in which a paralogous sequence covering the *tp0314* and *tp0619* genes was found to be almost identical to the region containing the two *tpr* genes *tp0620-tp0621* (*tpr*IJ) in all the TPE and TEN strains analyzed.

In addition, other loci (*tp0856, tp0856a,* and *tp0858*) also showed a modular structure in previous analyses (Strouhal et al. 2018), apart from the putative recombinant gene *tp0136* and *tp0620* (*tprI*), both previously described. All of these genes were found to be under positive selection in our analysis. The modular nature of these genes and the branches detected to be under positive selection together suggest that the recombination events occurred through gene duplication and gene conversion within treponemal genomes, and that they could result in substantial changes in gene and protein sequences.

Furthermore, 11 non-recombinant genes were detected to be under positive selection in the branch leading to the MRCA of TPE and TEN strains, and in the other two non recombinant genes *tp0134* and *tp0833,* the positive selection detected was associated with the branch leading to the MRCA of the historical genomes. Specifically, only the branches leading to the OKA2116 and HATO strains for the *tp0134* gene and the branches leading to the WP00227liq and WP00227scab strains for the *tp0833* gene exhibited signal of positive selection.

As we have observed a close relationship between recombination and selection, we also examined the functional roles of the proteins coded for by the recombinant and additional genes found to be under positive selection, according to Uniprot and a literature search (detailed in Supplementary Table 11). Despite some gene proteins having an unknown function, most of the

gene proteins identified here appear to play an important role in the defense of the pathogen against the host immune system and are potentially involved in virulence.

Molecular clock dating

Molecular clock dating analysis using an uncorrelated lognormal relaxed clock and a Bayesian skyline population model was performed on the dataset of 77 genomes, reduced to only variable sites, and with hypervariable and recombining genes removed. The time to the most recent common ancestor and the 95% Highest Posterior Density (HPD) intervals estimated for the major *T. pallidum* clades are detailed in Table 3. The divergence time between the TPE/TEN and TPA, i.e. the tMRCA for the entire *T. pallidum* family, was broadly estimated to fall between 7100 BCE and 550 CE (median 2015 BCE), in accordance with a previous estimate(Majander et al. 2020). Consistent with previous studies(Majander et al. 2020; Beale et al. 2021), we dated the MRCA of all TPA lineages between 190 BCE and 1290 CE (median 745 CE) (Figure 3), the TPA Nichols lineage between 1190 and 1795 CE (median 1595 CE), and between 740 and 1640 CE (median 1595) when also including Seattle 81-4 as part of the Nichols lineage. The SS14 lineage was dated to between 680 and 1415 CE (median 1150 CE) and lineage SS14-Ω to 1805 - 1965 CE (median 1910 CE) (Figure 3). The median evolutionary rate was estimated to 6.77×10^{-5} substitutions/SNPsite/year (95% HPD interval: 3.78 - 9.95×10^{-5}), which corresponds to 1.91×10^{-7} substitutions/site/year (95% HPD interval: 1.07 - 2.81×10^{-7}) for the full *Treponema* genome with hypervariable and recombining genes excluded.

******* Figure 3 about here *******

******* Table 3 about here *******

DISCUSSION

In this study, we report the sequencing and subsequent analysis of a new, high coverage (35X) ancient syphilis genome, W86, integrated in a large set of complete and nearly complete modern and ancient *T. pallidum* genomes from previous studies and publically available datasets. The joint analysis of nine ancient genomes, four belonging to TPE and five to TPA, the implementation of a new methodology to overcome the problems derived from using one single reference for mapping HTS reads , and the availability of genomes with higher coverage, have allowed gaining more insight into the evolutionary processes shaping *T. pallidum* genomes. We have been able to identify new recombination events and targets of positive selection, including the identification of loci playing a significant role in the differentiation between TPE and TEN, and improved date estimates for relevant split events in the history of the species.

Usually, in genomic studies of *T. pallidum*, and of many other bacterial species, a single reference genome is selected and used to map the high-throughput sequencing (HTS) reads of all samples,

despite the divergence of the isolates into well-defined subspecies and lineages. This can result in various problems, such as more undetermined positions, less confident SNP calls, and the effect of reference genome attraction (Kushwaha et al. 2015; Valiente-Mullor et al. 2021; Yang et al. 2019). We diverged from this tradition and selected four reference genomes and subsequently assigned each reconstructed genome to its closest reference, according to the best mapping result. We selected one reference genome from each of the *T. pallidum* subspecies, TPE and TEN, and one from each of the major TPA lineages, Nichols and SS14. The importance of selecting the closest reference for mapping was clearly demonstrated according to the different mapping results obtained for the new ancient genome W86 (Supplementary figure 2), where, although the subspecies identification was concordant between the four mapping results, the phylogenetic placement of the new ancient genome was different depending on the reference genome used for mapping its reads. With this novel mapping approach we increased the genome coverage (especially important when using sparse ancient DNA reads), reduced the reference bias, and improved the phylogenetic inference and assignment. This new mapping approach has also enabled us to ensure the consistency of the results, regardless of the reference chosen, and avoid laborious comparisons in the downstream analyses with several reference genomes.

Despite the challenges in identifying the mechanisms of recombination for this bacterium to date, several studies (Giacani et al. 2012; Pětrošová et al. 2012; Štaudová et al. 2014; Kumar et al. 2018; Strouhal et al. 2018; Grillová et al. 2019; Maděránková et al. 2019; Addetia et al. 2020; Mikalová et al. 2020; Pla-Díaz et al. 2022) have discussed the occurrence of recombination in *T. pallidum,* and the potential action of natural selection on the transferred genetic material. Using the PIM method (Sánchez-Busó et al. 2014; Majander et al. 2020; Pla-Díaz et al. 2022; Arora et al. 2016), we were able to identify 26 recombinant regions in 18 different genes, including five new recombinant genes (*tp0131*, *tp0346*, *tp0621*, *tp0859* and *tp0896*), not previously detected with this method (Sánchez-Busó et al. 2014; Majander et al. 2020; Pla-Díaz et al. 2022; Arora et al.) but detected as recombinant in other previous studies (Brinkman et al. 2008; Kumar et al. 2018; Marks et al. 2018; Strouhal et al. 2018; Grillová et al. 2019) with the exception of *tp0346*. These new detections are probably due to the nature of the dataset employed here, where previous comparisons have included a smaller number of genomes, mostly compiled of modern genomes, and with a lower resolution of the variation between the three *T. pallidum* subspecies than presented here. Moreover, in this study, all ancient *T. pallidum* genomes available to date (Schuenemann et al. 2018; Giffin et al. 2020; Majander et al. 2020) are integrated and analyzed together for the first time, with a further inclusion of several genomes sampled from non-human primates. Nevertheless, despite detecting the five new recombinant genes not previously detected by PIM, but reported in other studies (Brinkman et al. 2008; Kumar et al. 2018; Marks et al. 2018; Strouhal et al. 2018; Grillová et al. 2019), the improved quality of the draft genomes obtained with the novel mapping approach does not fully mend the large number of missing positions in some loci resulting from mapping short reads, especially for paralogous/duplicated genes with repetitive sequences, as observed here in some of the *tpr* genes. This highlights the difficulty of working with low-quality genomes due to the impossibility of using a standard culture system with the modern samples of this bacterium, and especially for the ancient samples, in which DNA is very degraded and sparse.

The point that all recombination events detected correspond to inter-subspecies transfers (TPE/TEN to TPA) with the exception of the recombination event in the *tp0136* gene observed between the Nichols and SS14 TPA lineages, is particularly striking, because modern-day TPE and TEN are geographically restricted to specific world regions. For recombination between TPE/TEN and TPA strains to occur, we need to assume that the diversified clades coexisted sympatrically, in order to simultaneously infect a common host, despite no evidence of human coinfection to date. Moreover, the ancient genomes of both TPE and TPA lineages are shown to be involved in the recombination events detected in eight different genes. The most intriguing questions are where and when these events took place.

Recent, indirect evidence from ancient genomes suggested a complex and lengthy coexistence of syphilis and yaws with overlapping geographical habitats in historical Europe (Schuenemann et al. 2018; Giffin et al. 2020; Majander et al. 2020). Yet, the origin and differentiation of *T. pallidum* subspecies is still perplexing and debated by scientists and historians. Improved dating with our approach sets the tMRCA for the entire *T. pallidum* family at 2015 BCE, with a HPD of 7100 BCE to 550 CE, concordant with previous estimates that placed the last common ancestor of all groups in the prehistoric era, at least 2000 BCE (de Melo et al. 2010; Majander et al. 2020; Rothschild 2005). The tMRCA for all TPA strains was estimated between 190 BCE - 1290 CE, which points to an earlier emergence than previously obtained for this subspecies and suggests that its origin predated the aggressive initial European syphilis outbreak at the end of 15th century. Furthermore, this dating, together with the divergence time estimates for TPE, between 490 - 1305 CE, and TEN, between 1650 - 1930 CE, suggest a common history of these diseases in the Old World, which could point to a novel diversification concurrent with the early spread of venereal syphilis. The coexistence of all three lineages in Europe prior to the first syphilis outbreak known from historical records is supported by the 10 ancient recombination events (all inferred to have occurred before 1500 CE) indicated in Table 1.

These findings lead to several novel conclusions regarding key events in treponemal evolutionary history. Molecular clock dating, combined with the putative recombination events involving modern TPA lineages (Table 1, Supplementary Figures 2-19), where lineages share variable genomic regions from a common TPE/TEN donor, together suggest that both TPA and TPE/TEN lineages were present in the Old World during pre-Columbian times. The alternative hypothesis, that these recombination events occurred in the New World and that only TPA was newly introduced to the European continent, implies that historical TPE/TEN strains from Europe dating from around the time of the Columbian expeditions would not be involved in these recombination events. As this runs counter to what is observed, we believe it more likely that TPA and TPE/TEN strains co-circulated in the Old World. Moreover, the diversity and wide geographical span of mutually contemporary lineages in early modern Europe, again decreases the likelihood of yaws and syphilis being simultaneously newly introduced, but does not discredit the possibility that some of these exchanged genes would in fact facilitate a rapid adaptation of certain strains to pressure from changing environment or host behavior. Indeed, according to the positive selection analysis, most of the genes detected as recombinant (along with others), appear to have also experienced evolution by positive selection. Most of the genes detected as recombinant or under positive selection pressure further appear to play an important role in

virulence, evasion of the host immune system, or defense against it. Yet, as gene variants functionally responsible for the transmission type and virulence of this bacterium still remain unclear, ancient variation cannot as of now satisfactorily resolve the reasons behind these observations (Supplementary Table 11).

Our novel mapping approach has revealed a dramatic change in the tree topology of the endemic treponematoses once recombinant and highly variable loci were removed (Figure 2B). Despite the yaws- and bejel-causing subspecies being well known and differentiated clades, in this strictly vertical phylogeny the TEN strains cluster within the TPE clade, giving clues about which genes might have played essential roles in the differentiation of these subspecies. It seems fair to suggest that the recombinant genes detected, together with the hypervariable genes *tp0316* and *tp0897*, played a key role in the divergence between the endemic treponematoses. Indeed, according to the Treetime results, 72.9% of the SNPs between the ancestral nodes of TPE and TEN are concentrated in those genes, underlining their important contribution to the differentiation between these two subspecies. Our analysis has also expanded the number of identified recent recombination events. Apart from the previous intra-subspecies event in *tp0136* between Seattle81-4 and CW82, CW83, the detection of transfer events in *tp0896* from non-human primate genomes to NL14 and CW59, both belonging to the Nichols sublineage of TPA, indicates that there has been recombination after the divergence of this sublineage and its differentiation from the SS14 and other ancient strains in TPA.

Before modern eradication projects, the endemic treponematoses were likely sympatric with the globally-distributed, sexually-transmitted strains and in the pre-antibiotic era co-infections may have provided opportunities for genetic exchange between strains. Moreover, as the strains found in African primates fall within the modern human diversity, it is also plausible that the genomic transfers between them were recent (Knauf et al. 2018; Šmajs et al. 2018). As some of the latest findings on historical yaws genomes suggest (Barquera et al. 2020; Giffin et al. 2020), it would not be unreasonable to presume that the African continent was the source for the European treponemal variants, considering the historical and prehistoric connections of these parts of the Old World. Wherever the ancestors of the modern strains existed, it seems likely that they did so in each other's vicinity.

Our results point to a common history of the three treponemal diseases in the Old World. The significant role of recombination and selection in the evolution and diversification of the three different lineages is shown especially for TPE and TEN, in which the role of the recombinant genes and *tp0897* and *tp0316* genes, mostly coding for outer membrane proteins, is imperative. As a practical measure, we showcase the benefits mapping and downstream analyses gain from selecting the most phylogenetically appropriate reference genomes. The findings here stress the importance of the ancient pathogen genomics in studying the bacterial evolutionary history by providing in-depth information based on a single, well-covered historical treponema strain. However, fundamental questions like ascertaining the origin of syphilis, still await more tangible evidence from ancient genomes dated to before the start of the 16th century.

MATERIALS AND METHODS

Sample processing

The upper-left premolar tooth sample was collected from human remains at the Wrocław University of Environmental and Life Sciences archaeological collection as part of the study project focusing on Wrocław's 17th century population genetics. Treponemal infection was not identified anthropologically. The sample was pretreated in clean room facilities, dedicated to state-of-the-art ancient DNA work, at the Museum and Institute of Zoology, Polish Academy of Sciences, in Warsaw. To avoid possible human and environmental DNA contamination, the surface of the tooth was sanded off with a hand rotary tool. The tooth was then washed with 5% sodium hypochlorite, molecular-grade water and 75% ethanol, UV-irradiated for 10 minutes on each side and then pulverized using a Retsch MM200 mixer mill. All sampling tools and all reusable items were regularly cleaned with diluted bleach and UV irradiated between each use to minimize the chance of contamination. As a part of the project, screening for specific pathogens was conducted, which revealed a probable treponemal infection in the W86 individual. A tooth powder sample was then sent to the Institute of Evolutionary Medicine in Zurich for *Treponema*-oriented in-depth analysis.

DNA extraction and library preparation

DNA extraction (Dabney et al. 2013) and double-stranded Illumina library construction were performed according to an established protocol (Kircher et al. 2012; Meyer and Kircher 2010) in the clean room facilities dedicated to ancient DNA processing at the Institute of Evolutionary Medicine in Zurich. Library pools were shotgun sequenced with an Illumina Nextseq platform using a NestSeq 500 Mid Output Kit (75 cycles paired-end) for the first pathogen screening.

UDG treated libraries

Subsequently, to remove ancient DNA specific DNA damage, the Uracil-DNA glycosylase (UDG) (Briggs et al. 2010) was used to treat the additional libraries prepared from the same extraction.

Whole genome capture

A custom target enrichment kit (Arbor Biosciences) was used for the whole genome capture. For this purpose, 60 bp long RNA baits with a 4bp tiling density and 99% identity were designed based on a selection of representative genomes (Nichols, GenBank: CP004010.2, Fribourg Blanc: GenBank CP003902, SS14: Gen-Bank CP000805.1) from each *Treponema pallidum* subspecies or clade. UDG-treated libraries were pooled in equimolar concentration, and five hundred ng final pools were hybridized in 60°C for 48 hours following the manufacturer's instructions. 10nM capture pools were sequenced using an Illumina NextSeq 500 High-throughput Kit (75 bp paired-end). Sequencing reads of the UDG treated, enrichment data were processed with the EAGER pipeline (v.1.92.55) (Peltzer et al. 2016). After removing the adapter sequence using AdapterRemoval v. 2.2.1a (Schubert et al. 2016), libraries from the same individual were merged and processed as paired-end sequencing reads.

Dataset selection and read processing

We generated a genomic dataset comprising 77 modern *T. pallidum* draft genomes (47 TPA, 19 TPE and 2 TEN) obtained from previously published studies, in addition to 8 historical genomes, and one new ancient genome we obtained (W86). The modern genomes were selected as a representative set of the diversity patterns observed in the phylogenetic trees reconstructed at the start of this study in 2021. The historical ancient draft genomes selected were those with a minimum of 5X coverage among those published to date(Schuenemann et al. 2018; Barquera et al. 2020; Giffin et al. 2020; Majander et al. 2020). Out of these 77 genomes, raw sequencing data was available for 62 of these, while for the other 15 only the consensus sequences files could be obtained. The data was downloaded from the NCBI database. The authenticity of ancient DNA was assessed by EAGER analyzing C to T deamination at the terminal base of the DNA fragment.

To reconstruct the individual genomes from the raw short-read data, we carried out raw read quality control and preprocessing, remove duplicates and variant identification using the programmes implemented in the EAGER pipeline (v.1.92.55) (Peltzer et al. 2016), as also done in previous studies (Peltzer et al. 2016; Majander et al. 2020; Arora et al. 2016). After processing the de-multiplexed sequencing reads, sample sequencing quality was analyzed with FastQC version 0.11.5 (Wingett and Andrews 2018). Following processing by AdapterRemoval ver. 2.2.1a (Schubert et al. 2016), the mapping was carried out using CircularMapper version 1.0 (Peltzer et al. 2016), with default BWA (-l 32, -n 0.04, q 37) parameters (Li and Durbin 2009) and Nichols (NC_021490.2) and SS14 (NC_010741.1) genomes, which represent the two main groups of TPA, and the CDC2 (NC_016848.1) and BosniaA (NZ_CP007548.1) genomes, which are well-studied TPE and TEN strains, respectively, as reference. Each of the genomes in the dataset was mapped to each of these four references. The *MarkDuplicates* method provided by Picard Toolkit (2019) was applied to remove duplicate reads and DamageProfiler version v0.3.12 was utilized to estimate the DNA damage parameters (Neukamm et al.). Indel realignments were performed using GATK (version 3.6) (McKenna et al. 2010) and single nucleotide polymorphisms (SNPs) for the resulting mappings were called using GATK UnifiedGenotyper with the following parameters for SNP calling: -nt 30, -stand_call_conf 30, --sample_ploidy 2, -dcov 250, --output_mode EMIT_ALL_SITES; and the following parameters for SNP filtering: DP>5, QUAL>30. The reconstructed W86 genome and its main features were represented graphically using BRIG (Alikhan et al. 2011).

Antibiotic resistance

Two mutations on the 23S ribosomal RNA operon, A2058G and A2059G (Beale et al. 2019; Beale et al. 2020), were investigated to assess macrolide azithromycin resistance in the new ancient genome W86 obtained. For this purpose, 23S rRNA gene sequences for operons 1 and 2 with 200 bp added to the 5’ and 3’ flanking regions were extracted from the Nichols reference genome and aligned to the W86 genome. Subsequently, the presence or absence of each of the two mutations was assessed with variant calling.

Multiple genome alignment

As it is known that the SNP calling in a genome is dependent on the choice of the reference used for mapping, we carried out a proximity evaluation to determine the most appropriate reference for each of the genomes (Valiente-Mullor et al. 2021). This procedure involved selecting among four representative reference genomes, according to the three different *T. pallidum* subspecies and the Nichols and SS14 clades (CDC2, BosniaA, Nichols, and SS14). In accordance with the strain classification in a *T. pallidum* subspecies or clade obtained in the previous studies from where the genomes were selected for this study, the reads for each strain were mapped against the corresponding closest reference genome. The resulting sequences were aligned to produce a multiple genome alignment. We added the 15 previously assembled genomes with no raw data available, and realigned all sequences, resulting in a new alignment of 76 *T. pallidum* genomes. The new ancient genome W86 was mapped to each of the four reference genomes, obtaining four different genome mapped versions of W86 which were merged with the previous multiple genome alignment to build a phylogenetic tree using IQ-TREE (version 1.6.10) (Nguyen et al. 2015), in order to determine which of the four references was closer to the W86 genome. The selected sequence mapping version for W86 genome was added to the multiple genome alignment and realigned resulting in a final alignment of 77 sequences.

An orthology analysis was carried out by Proteinortho (version V6.0b) (Lechner et al. 2011) to detect orthologous genes in the four reference genomes employed. The genomic coordinates of each gene present in at least one of the four reference genomes were then calculated according to their corresponding location in the final merged alignment.

The protein translations for all genes present in at least one reference genome were compared to the original gff3 files of each of the four references, to ensure that the final multiple genome alignment was correct, and that no protein was accidentally truncated (Supplementary files 4-5).

Recombination detection: PIM

The presence of recombination in the complete genomes of *T. pallidum* could potentially interfere with the inference of phylogenetic tree topologies, as described in Pla-Díaz et al. (2022). We thus used the PIM pipeline (Pla-Díaz et al. 2022) to investigate putatively recombining genes. In summary, the procedure included the following steps (with details provided in Supplementary Note 3): 1) A maximum likelihood (ML) tree was computed for the multiple genome alignment using IQ-TREE (Nguyen et al. 2015). 2) The 1061 genes present in at least one of the reference genomes were extracted, and the number of SNPs per gene was computed using an in-house script (Supplementary File 6), discarding genes with less than 3 SNPs. 3) For each of the genes retained, the phylogenetic signal in each gene alignment was tested using likelihood-mapping (Strimmer and von Haeseler 1997) in IQ-TREE, retaining only those genes that showed some phylogenetic signal (see Supplementary Note 3). 4) For the remaining genes, an ML tree was computed using IQ-TREE. Next, topology tests for each candidate gene were performed by IQ-TREE using two different methods: Shimodaira–Hasegawa (SH, Shimodaira and Hasegawa 1999) and Expected Likelihood Weights (ELW, Strimmer and Rambaut 2002). First, using the corresponding gene alignment, we compared the likelihood of each individual gene tree to the reference genome-wide tree (the phylogeny constructed from the multiple genome alignment). Then, using the entire genome alignment, we compared the same likelihoods. When both tests

reject the topology that is not derived from the corresponding alignment, this is referred to as a reciprocal incongruence (individual gene in the first comparison, the complete genome in the second). 5) For all genes passing the previous filtering steps, the presence of a minimum of three consecutive homoplasic SNPs congruent with a recombination event were checked using MEGAX (Kumar et al. 2018) to classify a gene as recombinant.

Several genes have a large proportion of sites that have missing data due to the challenges in mapping short read data from regions containing repetitive DNA. Most of these genes pertain to the *tpr* family, which comprises groups of paralogous genes. For each of these genes, those strains with a high proportion of missing data were removed, in order to be able to still analyze these interesting genes with the PIM pipeline. The hypervariable gene *tprK (tp0897)* was completely discarded from the recombination analysis following the results in Pla-Díaz et al. (2022) because its seven hypervariable regions undergo intrastrain gene conversion which have been studied in detail elsewhere (Pinto et al. 2016; Liu et al. 2019; Addetia et al. 2020)

Phylogeny reconstruction

Prior to the final phylogenetic analyses, genes identified by PIM as recombinant were excluded from the multiple genome alignment. To avoid introducing a potential bias into the phylogenetic signal, 3 genes (*tp0897, tp0316, tp0317*), which contain repetitive regions and have previously been reported as hypervariable and/or under gene conversion (Pinto et al. 2016; Liu et al. 2019; Addetia et al. 2020); (Strouhal et al. 2018; Vrbová et al. 2022) and thus induce recombination-like effects in mapping, were also removed. As the *tp0317* gene is embedded inside the *tp0316* gene, and moreover the coordinates from the BosniaA reference genome for *tp0316* spanned a longer region compared to the other reference genomes, *tp0316* and *tp0317* were removed according to the BosniaA reference genome coordinates for *tp0316*. A reference phylogenetic tree was then constructed with IQ-TREE, using GTR+G+I as the evolutionary model.

Mapping the TPE and TEN node-defining SNPs on the complete genome-based tree

To analyze the genetic changes between the ancestral nodes of the TPE and TEN subspecies in the phylogeny, an ancestral reconstruction of the multiple genome alignment and the reference phylogenetic tree was carried out with TreeTime (version 0.8.4) (Sagulenko et al. 2018). We used IcyTree (Vaughan 2017) to view the annotated tree with the changes obtained in the ancestral sequence reconstruction. We extracted from this file the annotation of non-shared polymorphisms between the ancestral nodes of TPE and TEN subspecies, analyzing in detail the genes removed from the multiple genome alignment before the phylogenetic reconstruction, because they are recombinant, hypervariable (*tp0897*) or considered to be under gene conversion (*tp0316* and *tp0317*). The gene *tp0316* was analyzed according to the annotation of this locus for the BosniaA reference genome as explained above. Moreover, the gene *tp0317*, which is embedded within *tp0316*, was analyzed independently.

To infer ancestral states, it is necessary to calculate the conditional (posterior) probabilities given the data (Molecular Evolution: A Statistical Approach, Ziheng Yang, pp 126). The probability of each of these polymorphisms was calculated using RAxML (version 8.2.11)(Stamatakis 2014).

Selection analysis

In addition to recombination processes, natural selection also plays an important role in genetic diversity patterns. In prior research, Pla-Díaz et al. (2022) demonstrated a close relationship between recombination and selection in *T. pallidum*, suggesting an important role of these processes in the evolution of this bacterial genus, especially in TPA lineages. Here we tested for positive selection in a subset of 317 (out of 1161) genes with three or more SNPs. Problematic sequences with high proportions of missing data were removed. Additionally, *tp0897*, *tp0316* and *tp0317* were excluded from this analysis because of hypervariable regions and the gene conversion signal present in these genes (Pětrošová et al. 2012; Pinto et al. 2016; Strouhal et al. 2018; Mikalová et al. 2020). Then, to test if positive selection occurred along the different lineages on a phylogeny (Yang and dos Reis 2011), we employed Hyphy (version 2.5.32) (Pond and Muse 2005), using the aBSREL model (adaptive Branch-Site Random Effects Likelihood), which is an is an improved version of the commonly-used "branch-site" models (Yang and dos Reis 2011; Lu and Guindon 2014). We used default settings and the ML phylogenies of each gene. We assessed statistical significance using a Likelihood-ratio test (LRT) .

Molecular clock dating

Molecular clock analysis was performed for the dataset comprising all 77 genomes. The alignment had genes with signals of recombination or hypervariable regions removed and was reduced to variable sites only (1,107,881 bp with 3,221 SNPs). The timescale of *T. pallidum* molecular evolution was estimated using time-calibrated Bayesian phylogenetic inference implemented in BEAST v2.6.3(Bouckaert et al. 2019). An uncorrelated lognormal relaxed molecular clock was calibrated using the ages of the samples (see Supplementary Table 1) with a diffuse prior (uniform 10^{-12} - 10^{4} for the mean rate and a gamma distribution (α=0.5396, β=0.3819) for the rates' standard deviation). Historical samples, for which only age ranges (based on archaeological contexts or radiocarbon dating) rather than exact ages were available, were assigned age priors spanning uniformly across the entire range. A strict clock was rejected based on the estimated coefficient of variation for the relaxed clock model, as the 95% HPD did not include zero (Ho and Duchêne 2014). A coalescent Bayesian Skyline tree prior with 5 groups was used as a simple model that is sufficiently flexible to fit many different kinds of dynamics. We used bModeltest v1.2.1 (Bouckaert and Drummond 2017) to average across all possible reversible substitution models. According to the results a model with rac=rgt, no gamma rate variation and no invariable sites received the most support. The MCMC chain was run for 2 billion steps with every 50,000th step sampled. The first 15% of samples were discarded as burn-in. Convergence and mixing was inspected using Tracer v1.7.1 (Rambaut et al. 2018); the ESS of all parameters exceeded 100. The maximum clade credibility tree was generated using TreeAnnotator, a part of the BEAST v.2.6.1 software package, and visualized using FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

A summary of the workflow used in the analysis of the 77 *T. pallidum* genomes is shown in Figure 4.

******* Figure 4 about here *******

Acknowledgements and funding

This work was supported by the Swiss National Science Foundation: grant number 188963 - "Towards the origins of syphilis" (V.J.S., K.M.), the University of Zurich's University Research Priority Program "Evolution in Action: From Genomes to Ecosystems" (V.J.S, G.A.), the National Programme for the Development of Humanities Poland grant no. 8121/MH/IH/14 and by projects BFU2017-89594R and PID2021-127010OB-100 from Spanish Ministerio de Ciencia e Innovación (FGC). MPD was funded by program FPU17/02367 from Spanish Ministerio de Educación. We thank Anna Lipowicz, Barbara Kwiatkowska and Jacek Szczurowski from the Department of Anthropology at Wrocław University of Environmental and Life Sciences for sharing the study material.

AUTHOR CONTRIBUTIONS

Conceptualization: FGC, KM, VS

Data curation: MPD, GA

Formal Analysis: MPD, MM, LdP, GA

Funding acquisition: VS, KM

Investigation: MPD, GA, KM

Methodology: FGC, MPD

Project administration: KM, VS

Resources: VS, GA, HP, KD, WB, PD, MO, JG

Supervision: FGC, KM, VS

Validation: FGC, KM, VS

Visualisation: MM, MPD

Writing – original draft: MPD, KM, GA

Writing – review & editing: FGC, MM, LdP, NA, VS

ADDITIONAL INFORMATION

Supplementary Material is available for this paper from the MBE website and Zenodo (https://doi.org/10.5281/zenodo.7291180). Reprints and permissions information is available at MBE website. Correspondence and requests for materials should be addressed to Verena J. Schünemann (verena.schuenemann@iem.uzh.ch), Kerttu Majander (kerttu.majander@gmail.com), or Fernando González-Candelas (fernando.gonzalez@uv.es).

COMPETING INTERESTS

The authors declare no competing financial interests.

REFERENCES

Addetia A, Lin MJ, Phung Q, Xie H, Huang M-L, Ciccarese G, Dal Conte I, Cusini M, Drago F, Giacani L, et al. 2020. Estimation of full-length tprK diversity in *Treponema pallidum* subsp. *mBio* 11: 02726-20

Alikhan N-F, Petty NK, Ben Zakour NL, Beatson SA. 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12:402.

Arora N, Schuenemann VJ, Jäger G, Peltzer A, Seitz A, Herbig A, Strouhal M, Grillová L, Sánchez-Busó L, Kühnert D, et al. 2016. Origin of modern syphilis and emergence of a pandemic *Treponema pallidum* cluster. *Nat Microbiol* 2:16245.

Barquera R, Lamnidis TC, Lankapalli AK, Kocher A, Hernández-Zaragoza DI, Nelson EA, Zamora-Herrera AC, Ramallo P, Bernal-Felipe N, Immel A, et al. 2020. Origin and health status of first-generation Africans from early colonial Mexico. *Curr. Biol.* 30:2078–2091.e11.

Beale MA, Marks M, Cole MJ, Lee M-K, Pitt R, Ruis C, Balla E, Crucitti T, Ewens M, Fernández-Naval C, et al. 2021. Global phylogeny of *Treponema pallidum* lineages reveals recent expansion and spread of contemporary syphilis. *Nat Microbiol* 6:1549–1560.

Beale MA, Marks M, Sahi SK, Tantalo LC, Nori AV, French P, Lukehart SA, Marra CM, Thomson NR. 2019. Genomic epidemiology of syphilis reveals independent emergence of macrolide resistance across multiple circulating lineages. *Nat. Commun.* 10:1–9.

Beale MA, Noguera-Julian M, Godornes C, Casadellà M, González-Beiras C, Parera M, Kapa Jnr A, Houinei W, Wangi J, Corbacho-Monne M, et al. 2020. Yaws re-emergence and bacterial drug resistance selection after mass administration of azithromycin: a genomic epidemiology investigation. *Lancet Microbe* 1:e263–e271.

Bouckaert RR, Drummond AJ. 2017. bModelTest: Bayesian phylogenetic site model averaging and model comparison. *BMC Evol. Biol.* 17:42.

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N, et al. 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 15:e1006650.

Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prüfer K, Meyer M, Krause J, Ronan MT, Lachmann M, et al. 2007. Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci. U.S.A.* 104:14616–14621.

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

Brinkman MB, McGill MA, Pettersson J, Rogers A, Matejková P, Smajs D, Weinstock GM, Norris SJ, Palzkill T. 2008. A novel *Treponema pallidum* antigen, TP0136, is an outer membrane protein that binds human fibronectin. *Infect. Immun.* 76:1848–1857.

Cooper A. 2000. Ancient DNA: Do it right or not at all. *Science* 289:1139b – 1139.

Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.-L., et al. 2013. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci.* USA 110:15758–15763.

Fonseca E, García-Silva J, del Pozo J, Yebra MT, Cuevas J, Contreras F. 1999. Syphilis in an HIV

infected patient misdiagnosed as leprosy. *J. Cutan. Pathol.* 26:51–54.

Giacani L, Chattopadhyay S, Centurion-Lara A, Jeffrey BM, Le HT, Molini BJ, Lukehart SA, Sokurenko EV, Rockey DD. 2012. Footprint of positive selection in *Treponema pallidum* subsp. *pallidum* genome sequences suggests adaptive microevolution of the syphilis pathogen. *PLoS Negl. Trop. Dis.* 6:e1698.

Giffin K, Lankapalli AK, Sabin S, Spyrou MA, Posth C, Kozakaitė J, Friedrich R, Miliauskienė Ž, Jankauskas R, Herbig A, et al. 2020. A treponemal genome from an historic plague victim supports a recent emergence of yaws and its presence in 15 century Europe. *Sci. Rep.* 10:9499.

Grange PA, Allix-Beguec C, Chanal J, Benhaddou N, Gerhardt P, Morini JP, Deleuze J, Lassau F, Janier M, Dupin N. 2013. Molecular subtyping of *Treponema pallidum* in Paris, France. *Sex. Transm. Dis.* 40: 641-644.

Grillová L, Bawa T, Mikalová L, Gayet-Ageron A, Nieselt K, Strouhal M, Sednaoui P, Ferry T, Cavassini M, Lautenschlager S, et al. 2018. Molecular characterization of *Treponema pallidum* subsp. *pallidum* in Switzerland and France with a new multilocus sequence typing scheme.. *PLoS One* 13:e0200773.

Grillová L, Oppelt J, Mikalová L, Nováková M, Giacani L, Niesnerová A, Noda AA, Mechaly AE, Pospíšilová P, Čejková D, et al. 2019. Directly Sequenced Genomes of Contemporary Strains of Syphilis Reveal Recombination-Driven Diversity in Genes Encoding Predicted Surface-Exposed Antigens. *Front. Microbiol.* 10:1691.

Hodges E, Rooks M, Xuan Z, Bhattacharjee A, Benjamin Gordon D, Brizuela L, Richard McCombie W, Hannon GJ. 2009. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat. Protoc.* 4:960–974.

Ho SYW, Duchêne S. 2014. Molecular-clock methods for estimating evolutionary rates and timescales. *Mol Ecol* 23:5947–5965.

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

Knauf S, Gogarten JF, Schuenemann VJ, De Nys HM, Düx A, Strouhal M, Mikalová L, Bos KI, Armstrong R, Batamuzi EK, et al. 2018. Nonhuman primates across sub-Saharan Africa are infected with the yaws bacterium *Treponema pallidum* subsp. *pertenue*. *Emerg. Microbes Infect.* 7:1–4.

Kumar S, Caimano MJ, Anand A, Dey A, Hawley KL, LeDoyt ME, La Vake CJ, Cruz AR, Ramirez LG, Paštěková L, et al. 2018. Sequence Variation of Rare Outer Membrane Protein β-Barrel Domains in Clinical Strains Provides Insights into the Evolution of *Treponema pallidum* subsp. *pallidum*, the Syphilis Spirochete. *mBio* 9: 01006-18

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* 35:1547–1549.

Kushwaha B, Kumar R, Agarwal S, Pandey M, Nagpure NS, Singh M, Srivastava S, Joshi CG, Das P, Sahoo L, et al. 2015. Assembly and variation analyses of *Clarias batrachus* mitogenome retrieved from WGS data and its phylogenetic relationship with other catfishes. *Meta Gene* 5:105–114

Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. 2011. Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics* 12:124.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.

Bioinformatics 25:1754–1760.

Liu D, Tong M-L, Lin Y, Liu L-L, Lin L-R, Yang T-C. 2019. Insights into the genetic variation profile of *tprK* in *Treponema pallidum* during the development of natural human syphilis infection. *PLoS Negl. Trop. Dis.* 13:e0007621.

Lu A, Guindon S. 2014. Performance of standard and stochastic branch-site models for detecting positive selection among coding sequences. *Mol. Biol. Evol.* 31:484–495.

Maděránková D, Mikalová L, Strouhal M, Vadják Š, Kuklová I, Pospíšilová P, Krbková L, Koščová P, Provazník I, Šmajs D. 2019. Identification of positively selected genes in human pathogenic treponemes: Syphilis-, yaws-, and bejel-causing strains differ in sets of genes showing adaptive evolution. *PLoS Negl. Trop. Dis.* 13:e0007463.

Majander K, Pfrengle S, Kocher A, Neukamm J, du Plessis L, Pla-Díaz M, Arora N, Akgül G, Salo K, Schats R, et al. 2020. Ancient Bacterial Genomes Reveal a High Diversity of *Treponema pallidum* Strains in Early Modern Europe. *Curr. Biol.* 30:3788–3803.e10.

Marks M, Fookes M, Wagner J, Butcher R, Ghinai R, Sokana O, Sarkodie Y-A, Lukehart SA, Solomon AW, Mabey DCW, et al. 2018. Diagnostics for Yaws Eradication: Insights From Direct Next-Generation Sequencing of Cutaneous Strains of *Treponema pallidum*. *Clin. Infect. Dis.* 66:818–824.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20:1297–1303.

de Melo FL, de Mello JCM, Fraga AM, Nunes K, Eggers S. 2010. Syphilis at the crossroad of phylogenetics and paleopathology. *PLoS Negl. Trop. Dis.* 4:e575.

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

Mikalová L, Janečková K, Nováková M, Strouhal M, Čejková D, Harper KN, Šmajs D. 2020. Whole genome sequence of the *Treponema pallidum* subsp. *endemicum* strain Iraq B: A subpopulation of bejel treponemes contains full-length *tprF* and *tprG* genes similar to those present in *T. pallidum* subsp. *pertenue* strains. *PLoS One* 15:e0230926.

Mikalová L, Strouhal M, Oppelt J, Grange PA, Janier M, Benhaddou N, Dupin N, Šmajs D. 2017. Human Treponema pallidum 11q/j isolate belongs to subsp. endemicum but contains two loci with a sequence in TP0548 and TP0488 similar to subsp. pertenue and subsp. pallidum, respectively. *PLoS Negl. Trop. Dis.* 11(3): e0005434.

Mitjà O, Šmajs D, Bassat Q. 2013. Advances in the diagnosis of endemic treponematoses: yaws, bejel, and pinta. *PLoS Negl. Trop. Dis.* 7:e2283.

Neukamm J, Peltzer A, Nieselt K. DamageProfiler: Fast damage pattern calculation for ancient DNA. Bioinformatics 37: 3652-3653.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* 32:268–274.

Noda AA, Grillová L, Lienhard R, Blanco O, Rodríguez I, Šmajs D. 2018. Bejel in Cuba: molecular identification of *Treponema pallidum* subsp. *endemicum* in patients diagnosed with venereal syphilis. *Clin. Microbiol. Infect.* 24:1210.e1–e1210.e5.

Pankiewicz A, Witkowski J. 2012. Dewocjonalia barokowe odkryte na cmentarzysku przy kościele św.

Piotra i Pawła na Ostrowie Tumskim we Wrocławiu, Wroclavia antiqua 17:1621–1670.

Peltzer A, Jäger G, Herbig A, Seitz A, Kniep C, Krause J, Nieselt K. 2016. EAGER: efficient ancient genome reconstruction. *Genome Biol* 17: 60.

Pětrošová H, Zobaníková M, Čejková D, Mikalová L, Pospíšilová P, Strouhal M, Chen L, Qin X, Muzny DM, Weinstock GM, et al. 2012. Whole Genome Sequence of *Treponema pallidum* ssp. *pallidum*, Strain Mexico A, Suggests Recombination between Yaws and Syphilis Strains. *PLoS Negl. Trop. Dis.* 6:1832

Picard Toolkit. 2019. Broad Institute, GitHub Repository. https://broadinstitute.github.io/picard/; Broad Institute

Pillay A, Liu H, Chen CY, Holloway B, Sturm AW, Steiner B, Morse SA. 1998. Molecular subtyping of *Treponema pallidum* subspecies *pallidum*. *Sex. Transm. Dis.* 25:408–414.

Pinto M, Borges V, Antelo M, Pinheiro M, Nunes A, Azevedo J, Borrego MJ, Mendonça J, Carpinteiro D, Vieira L, et al. 2016. Genome-scale analysis of the non-cultivable *Treponema pallidum* reveals extensive within-patient genetic variation. *Nat Microbiol* 2:16190.

Pla-Díaz M, Sánchez-Busó L, Giacani L, Šmajs D, Bosshard PP, Bagheri HC, Schuenemann VJ, Nieselt K, Arora N, González-Candelas F. 2022. Evolutionary Processes in the Emergence and Recent Spread of the Syphilis Agent, *Treponema pallidum*. *Mol. Biol. Evol.* 39: msab318.

Pond SLK, Muse SV. 2005. HyPhy: Hypothesis Testing Using Phylogenies. In: Statistical Methods in Molecular Evolution. New York: Springer-Verlag. p. 125–181.

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst Biol* 67:901–904.

Rothschild BM. 2005. History of Syphilis. *Clin Infect Dis* 40: 1454–1463.

Sagulenko P, Puller V, Neher RA. 2018. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol* 4:vex042.

Sánchez-Busó L, Comas I, Jorques G, González-Candelas F. 2014. Recombination drives genome evolution in outbreak-related *Legionella pneumophila* isolates. *Nat. Genet.* 46: 1205-1211.

Sawyer S, Krause J, Guschanski K, Savolainen V, Pääbo S. 2012. Temporal Patterns of Nucleotide Misincorporations and DNA Fragmentation in Ancient DNA. *PLoS ONE* 7:e34131.

Schubert M, Lindgreen S, Orlando L. 2016. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* 9:88.

Schuenemann VJ, Kumar Lankapalli A, Barquera R, Nelson EA, Iraíz Hernández D, Acuña Alonzo V, Bos KI, Márquez Morfín L, Herbig A, Krause J. 2018. Historic *Treponema pallidum* genomes from Colonial Mexico retrieved from archaeological remains. *PLoS Negl. Trop. Dis.* 12:e0006447.

Shinohara K, Furubayashi K, Kojima Y, Mori H, Komano J, Kawahata T. 2022. Clinical perspectives of *Treponema pallidum* subsp. *endemicum* infection in adults, particularly men who have sex with men in the Kansai area, Japan: A case series. *J. Infect. Chemother.* 28:444–450.

Šmajs D, Strouhal M, Knauf S. 2018. Genetics of human and animal uncultivable treponemal pathogens. *Infect. Genet. Evol*. 61:92–107.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.

Štaudová B, Strouhal M, Zobaníková M, Čejková D, Fulton LL, Chen L, Giacani L, Centurion-Lara A, Bruisten SM, Sodergren E, et al. 2014. Whole Genome Sequence of the *Treponema pallidum* subsp. *endemicum* Strain Bosnia A: The Genome Is Related to Yaws Treponemes but Contains Few Loci Similar to Syphilis Treponemes. *PLoS Negl. Trop. Dis.* 8:e3261.

Strimmer K, von Haeseler A. 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proc. Natl. Acad. Sci. U.S.A.* 94:6815–6819.

Strouhal M, Mikalová L, Haviernik J, Knauf S, Bruisten S, Noordhoek GT, Oppelt J, Čejková D, Šmajs D. 2018. Complete genome sequences of two strains of *Treponema pallidum* subsp. *pertenue* from Indonesia: Modular structure of several treponemal genes. *PLoS Negl. Trop. Dis.* 12:e0006867.

Valiente-Mullor C, Beamud B, Ansari I, Francés-Cuesta C, García-González N, Mejía L, Ruiz-Hueso P, González-Candelas F. 2021. One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads. *PLoS Comput. Biol.* 17:e1008678.

Vaughan TG. 2017. IcyTree: rapid browser-based visualization for phylogenetic trees and networks. *Bioinformatics* 33:2392–2394.

Vrbová E, Mikalová L, Grillová L, Pospíšilová P, Strnadel R, Dastychová E, Kojanová M, Kreidlová M, Vaňousová D, Rob F, et al. 2020. A retrospective study on nested PCR detection of syphilis treponemes in clinical samples: PCR detection contributes to the diagnosis of syphilis in patients with seronegative and serodiscrepant results. *PLoS One* 15:e0237949.

Vrbová E, Noda AA, Grillová L, Rodríguez I, Forsyth A, Oppelt J, Šmajs D. 2022. Whole genome sequences of *Treponema pallidum* subsp. *endemicum* isolated from Cuban patients: The non-clonal character of isolates suggests a persistent human infection rather than a single outbreak. *PLoS Negl. Trop. Dis.* 16:e0009900.

Wingett SW, Andrews S. 2018. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res.* 7:1338.

World Health Organization. 2021. Global progress report on HIV, viral hepatitis and sexually transmitted infections, 2021.

Yang X, Lee W-P, Ye K, Lee C. 2019. One reference genome is not enough. *Genome Biol.* 20:104.

Yang Z, dos Reis M. 2011. Statistical Properties of the Branch-Site Test of Positive Selection. *Mol Biol Evol* 28:1217–1228.

TABLES

Table 1. Recombination events detected in *T. pallidum*. The gene ID names correspond to the general gene nomenclature for *T. pallidum.* For each recombination event, coordinates for the start and end position in the multiple genome alignment are provided (Supplementary File 1). The strains involved are detailed, with an arrow separating the donor strains from the recipient strains. Events with an asterisk may represent more than one recombinant transfer depending on the placement of the involved strains in the reference tree (See Supplementary Note 5 and Supplementary Figures 3-20). Bold type is used to highlight novel genes identified as recombinant in this study.

Gene ID	Event	Start	End	Minimal size	SNPs	Strains Involved	Event classification
tp0131	1	152624	153390	766	333	Gauthier, LMNP-1, CDC_2575 → Nichols, Chicago, BAL3,BAL73, NIC2	Modern
tp0136	1	158235	158247	12	4	TPE → NL16	Modern
	2*	158281	158292	11	4	Seattle 81-4 → CW82, CW83	Modern
	3	158292	158310	18	4	TPE clade, excluding HATO, OKA_2116 and LMNP-1 → Seattle 81-4	Modern
	4	158414	158418	4	4	TPE → Nichols-lineage clade excluding CW82	Ancient
	5	158483	158507	24	8	TEN → Nichols-lineage clade	Ancient
	6*	159058	159119	61	4	TEN/TPE→ PD28 and Nichols-lineage clade excluding CW82	Ancient
	7	159452	159466	14	6	TEN/TPE → Nichols-lineage clade	Ancient
tp0164	1*	187135	187320	185	5	TEN/TPE → Seattle 81-4, CW86, CW59, NE20, 94A, 94B	Modern
tp0179	1*	198183	198571	388	9	TEN/TPE→ Nichols-lineage clade, W86, PD28	Ancient
tp0326	1	346656	348753	2097	59	TEN → SS14-lineage clade, excluding MexicoA and syphilis ancient genomes	Modern
tp0346	1*	373266	373282	16	3	TEN/TPE → 94A, 94B	Ancient
tp0462	1	493382	494407	1025	55	TEN/TPE → CW86, Seattle86, NE20, CW59	Modern
tp0488	1*	524108	524906	798	54	TPE/TEN → MexicoA, W86	Ancient
tp0515	1	557164	559063	1899	24	TEN/TPE → Nichols-lineage clade	Ancient
tp0548	1*	596164	596828	664	57	TEN/TPE → Nichols-lineage clade, W86	Ancient
tp0558	1	607533	607953	420	4	TEN/TPE → SS14-lineage clade and syphilis ancient genomes	Ancient
tp0621	1	676649	676897	248	137	TEN/TPE → Seattle 81-4	Modern
	2	677030	677215	185	125	TEN/TPE→ Seattle 81-4	Modern

Gene ID	Event	Start	End	Minimal size	SNPs	Strains Involved	Event classification
tp0131	1	152624	153390	766	333	Gauthier, LMNP-1, CDC_2575 → Nichols, Chicago, BAL3,BAL73, NIC2	Modern
tp0136	1	158235	158247	12	4	TPE → NL16	Modern
	2*	158281	158292	11	4	Seattle 81-4 → CW82, CW83	Modern
	3	158292	158310	18	4	TPE clade, excluding HATO, OKA_2116 and LMNP-1 → Seattle 81-4	Modern
	4	158414	158418	4	4	TPE → Nichols-lineage clade excluding CW82	Ancient
	5	158483	158507	24	8	TEN → Nichols-lineage clade	Ancient
	6*	159058	159119	61	4	TEN/TPE→ PD28 and Nichols-lineage clade excluding CW82	Ancient
	7	159452	159466	14	6	TEN/TPE → Nichols-lineage clade	Ancient
tp0859	1	939420	939442	22	11	External sources →TPE	Ancient
tp0865	1*	946548	946950	402	19	TEN/TPE → CW86, Sea86, NE20, CW59 and Seattle 81-4	Modern
	2*	947238	947556	318	26	TEN → CW86, Sea86, NE20, CW59 and Seattle 81-4	Modern
tp0896	1*	977297	977302	5	4	All primate strains → NL14, CW59	Modern
tp0967	1	1052520	1053803	1283	22	TEN/TPE → Seattle 81-4	Modern
tp0968	1	1053915	1055118	1203	33	TEN → Seattle 81-4	Modern
tp1031	1	1128223	1128350	127	19	External sources →SS14-lineage clade	Modern

Table 2. Summary of SNPs included in the 21 removed genes that affect the phylogenetic placement of the TEN genomes and the number of possible reversions detected on them. The gene ID names correspond to the general gene nomenclature for *T. pallidum*. For each gene, coordinates for the start and end position in the multiple genome alignment are provided. The total of SNPs in each gene, the number of SNPs fixed belonging to the 21 excluded genes and involved in the differentiation of TPE and TEN from their common ancestor, and the number of possible reversions detected in each gene are detailed.

Gene ID	Ini	End	Number of total SNPs in each gene	Fixed SNPs between TPE/TEN	Number of possible reversions
tp0131	152453	154252	372	4	3
tp0136	158086	159573	176	46	21
tp0164	186980	187780	6	0	0
tp0179	197925	199808	11	2	0
tp0316	332459	336618	53	2	3
tp0317	334348	336618	52	1	0
tp0326	346280	348793	60	13	2
tp0346	373237	373938	30	1	0
tp0462	493325	494605	69	0	0
tp0488	523745	526282	79	19	3
tp0515	556216	559191	29	3	0
tp0548	595868	595884	91	28	7
tp0558	607120	608028	8	1	0
tp0621	675919	678195	283	1	0
tp0859	938709	940211	47	26	2
tp0865	946375	947820	67	20	2
tp0896	977201	977353	12	0	1
tp0897	977335	978924	134	6	2
tp0967	1052318	1053883	23	1	0
tp0968	1053880	1055502	35	10	1
tp0131	1126974	1129023	58	2	0

Table 3. Time to the most recent common ancestor and 95% HPD interval (Highest Posterior Density, HPD) estimated by molecular clock dating for the major clades.

	Posterior Date Estimate	
Sample/Clade	**Median**	**95% HPD interval**
SS14 (TPA)	1150 CE	680 - 1415 CE
SS14-Ω (TPA)	1910 CE	1805 - 1965 CE
Nichols, excl. Sea81-4 (TPA)	1595 CE	1190 - 1795 CE
Nichols, incl. Sea81-4 (TPA)	1265 CE	740 - 1640 CE
TPA	745 CE	190 BCE- 1290 CE
TPE	1000 CE	490 - 1305 CE
TEN	1820 CE	1650 - 1930 CE
TPA/TPE/TEN	2015 BCE	7100 BCE - 550 CE

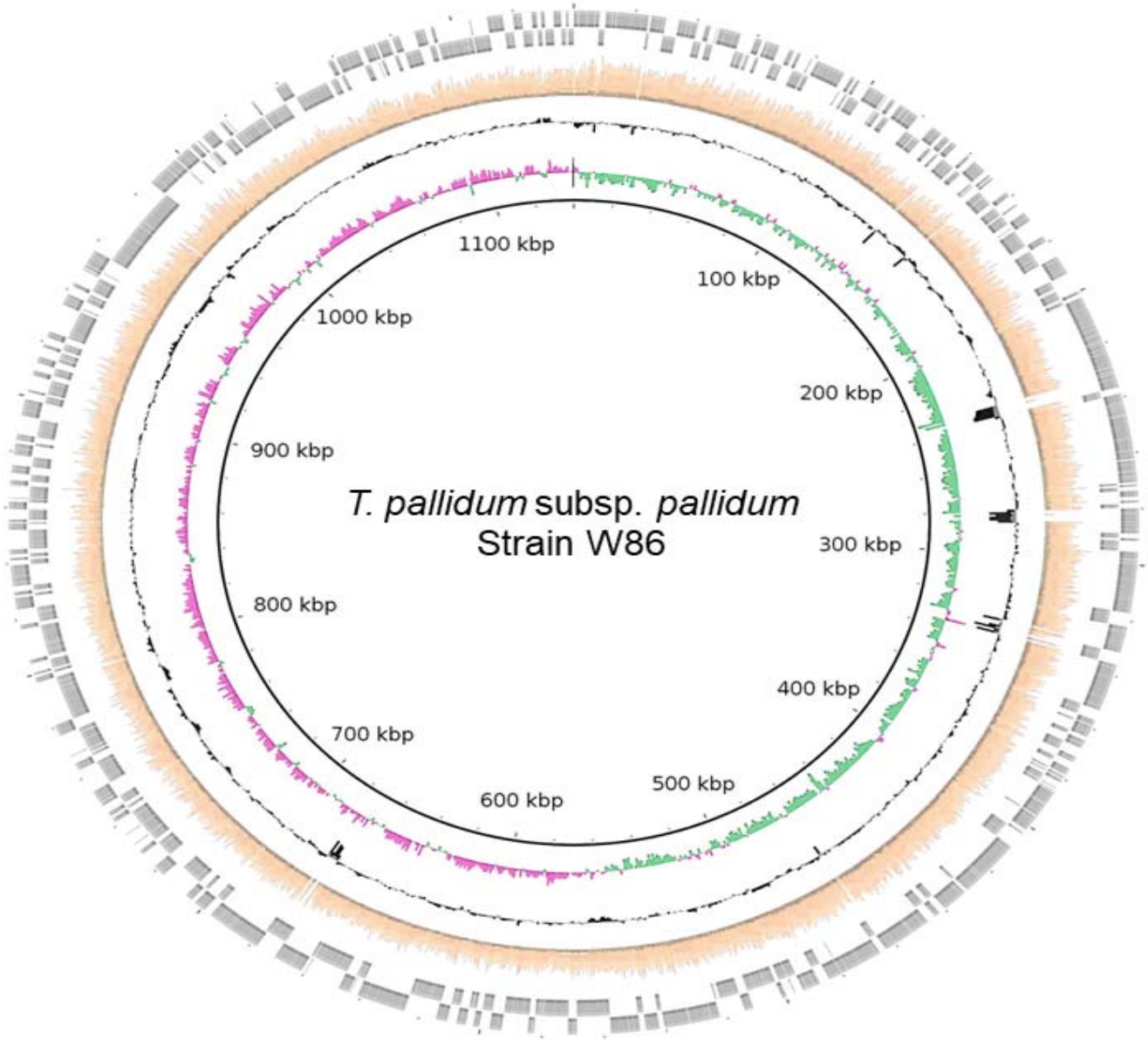
FIGURE LEGENDS

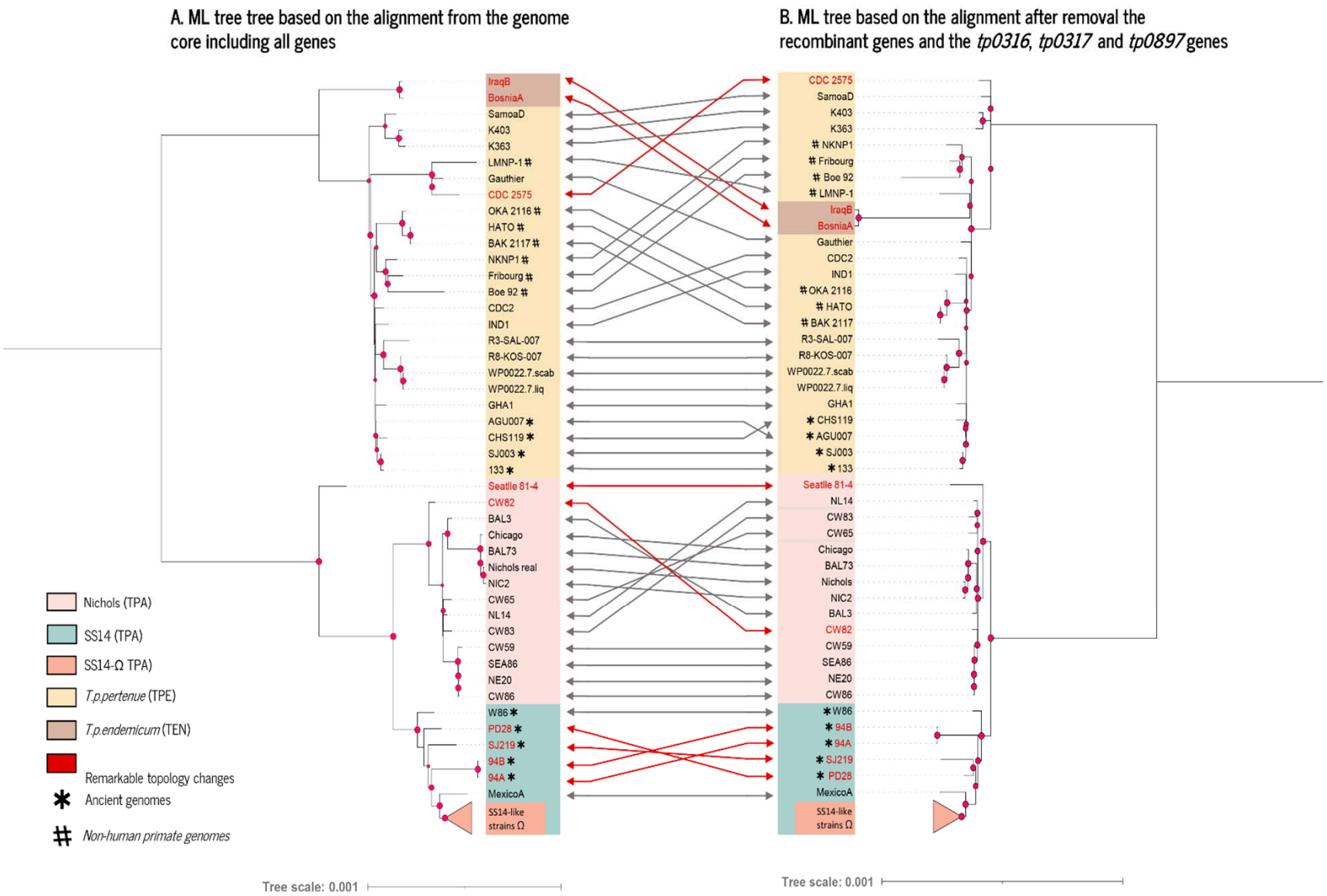
Figure 1. Circular plot of the W86 genome. Circles indicate, from inside outwards: genomic position, GC skew (pink and green); GC content (black) and coverage (orange). The outer rim (gray) shows the direction of protein-coding regions according to the annotation of the SS14 reference genome (CP000805.1): forward, outermost circle.

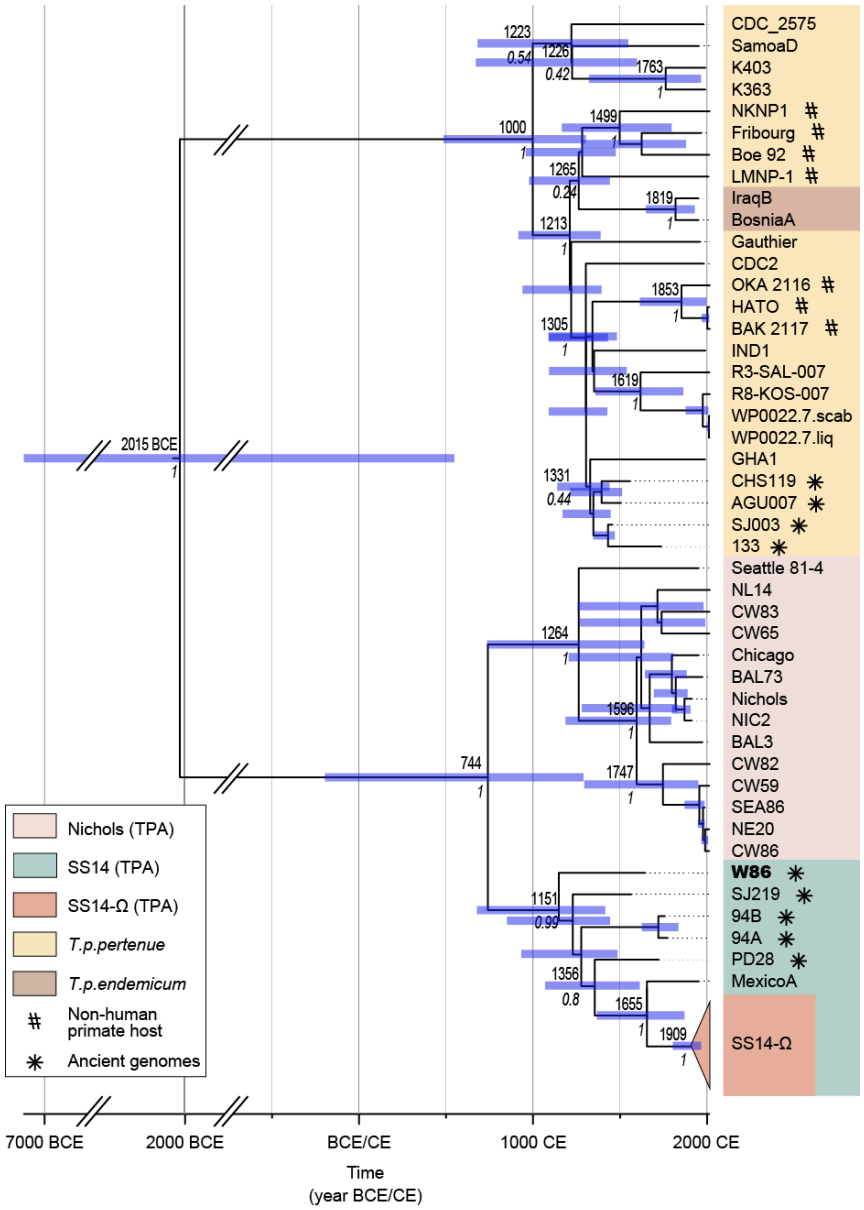
Figure 2. Comparison of the topologies of two maximum likelihood trees, A) obtained with all genes included in the multiple genome alignment, and B) obtained after excluding *tp0897*, *tp0316* and *tp0317* and recombinant genes from the multiple genome alignment. The different clades corresponding to TPE and TEN, and the Nichols and SS14 lineages of TPA are colour-coded according to the legend. Bootstrap support values higher than 70% are indicated by red circles, with circle size proportional to bootstrap support percentage.

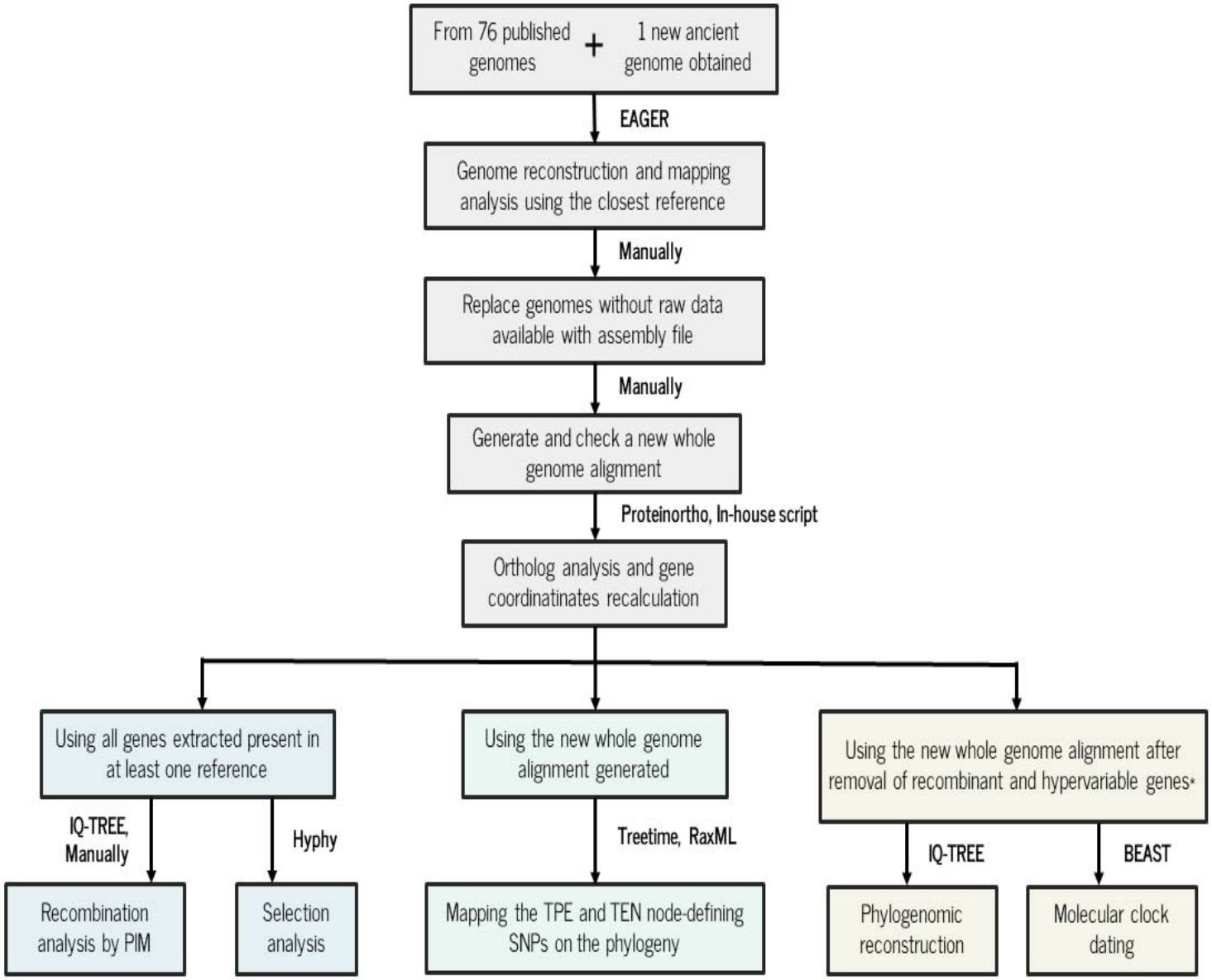
Figure 3. Maximum clade credibility (MCC) tree of the dataset consisting of 66 modern and 9 historical genomes estimated in BEAST under an uncorrelated lognormal relaxed clock model and a Bayesian skyline plot demographic model. Median age and Posterior support are shown for specific nodes. Blue bars show 95% HPD of the node age estimates . An expanded tree with the details of sequences encompassed in clade SS14-Ω is shown in Supplementary Figure 20.

Figure 4. Analysis workflow for the genomic and phylogenomic analysis of the 76 previously published *T. pallidum* genomes and one new historical genome obtained in this study (W86). The hypervariable genes indicated by * are *tp0897* and *tp0316*.


1100 kbp
100 kbp
1000 kbp
200 kbp
900 kbp
T. pallidum subsp. pallidum
Strain W86
300 kbp
800 kbp
400 kbp
700 kbp
500 kbp
600 kbp


A. ML tree tree based on the alignment from the genome core including all genes
B. ML tree based on the alignment after removal the recombinant genes and the *tp0316*, *tp0317* and *tp0897* genes
IraqB
BosniaA
SamoaD
K403
K363
LMNP-1 #
Gauthier
CDC 2575
OKA 2116 #
HATO #
BAK 2117 #
NKNP1 #
Fribourg #
Boe 92 #
CDC2
IND1
R3-SAL-007
R8-KOS-007
WP0022.7.scab
WP0022.7.liq
GHA1
AGU007 *
CHS119 *
SJ003 *
133 *
Seattle 81-4
CW82
BAL3
Chicago
BAL73
Nichols real
NIC2
CW65
NL14
CW83
CW59
SEA86
NE20
CW86
W86 *
PD28 *
SJ219 *
94B *
94A *
MexicoA
SS14-like strains Ω
Nichols (TPA)
SS14 (TPA)
SS14-Ω TPA)
T.p.pertenue (TPE)
T.p.endemicum (TEN)
Remarkable topology changes
* Ancient genomes
# *Non-human primate genomes*
Tree scale: 0.001
Tree scale: 0.001


CDC_2575
SamoaD
K403
K363
NKNP1
Fribourg
Boe 92
LMNP-1
IraqB
BosniaA
Gauthier
CDC2
OKA 2116
HATO
BAK 2117
IND1
R3-SAL-007
R8-KOS-007
WP0022.7.scab
WP0022.7.liq
GHA1
CHS119
AGU007
SJ003
133
Seattle 81-4
NL14
CW83
CW65
Chicago
BAL73
Nichols
NIC2
BAL3
CW82
CW59
SEA86
NE20
CW86
W86
SJ219
94B
94A
PD28
MexicoA
SS14-Ω
1223
0.54
1226
0.42
1763
1
1000
1
1499
1
1265
0.24
1819
1
1213
1
1853
1
1305
1
1619
1
2015 BCE
1
1331
0.44
1264
1
1596
1
744
1
1747
1
1151
0.99
1356
0.8
1655
1
1909
1
Nichols (TPA)
SS14 (TPA)
SS14-Ω (TPA)
T.p.pertenue
T.p.endemicum
Non-human primate host
Ancient genomes
7000 BCE
2000 BCE
BCE/CE
1000 CE
2000 CE
Time
(year BCE/CE)


From 76 published genomes
+
1 new ancient genome obtained
EAGER
Genome reconstruction and mapping analysis using the closest reference
Manually
Replace genomes without raw data available with assembly file
Manually
Generate and check a new whole genome alignment
Proteinortho, In-house script
Ortholog analysis and gene coordinatinates recalculation
Using all genes extracted present in at least one reference
Using the new whole genome alignment generated
Using the new whole genome alignment after removal of recombinant and hypervariable genes*
IQ-TREE, Manually
Hyphy
Treetime, RaxML
IQ-TREE
BEAST
Recombination analysis by PIM
Selection analysis
Mapping the TPE and TEN node-defining SNPs on the phylogeny
Phylogenomic reconstruction
Molecular clock dating