

Programa de Doctorado en Biomedicina y  
Biotecnología



TESIS DOCTORAL

**“Genomic analysis of evolutionary processes  
and epidemiology of *Treponema pallidum*”**

Trabajo realizado por:  
**Marta Pla Díaz**

Supervisor:  
**Fernando González Candelas**

València, Junio de 2023



D. **Fernando González Candelas**, Doctor en Ciencias Biológicas y Catedrático del Departamento de Genética de la Universitat de València.

## CERTIFICA

Que Dña. **Marta Pla Díaz**, Graduada en Biología por la Universitat de València, ha realizado bajo mi dirección el trabajo titulado “Genomic analysis of evolutionary processes and epidemiology of *Treponema pallidum*”. La memoria de este trabajo cumple los requisitos científicos y formales para presentar la misma en depósito y proceder a la defensa de la Tesis Doctoral.

Y para que conste, en el cumplimiento de la legislación vigente, firmo el presente certificado en Valencia, a 16 de junio de 2023.

Fdo.: Dr. Fernando González Candelas



*Un plan en el que a cada momento se promete un gozo absoluto nunca puede salir bien, y la frustración general solo se conjura si puedes librarte de algún pequeño contratiempo. **Orgullo y prejuicio, de Jane Austen.***



## **Declaración**

El trabajo presentado en esta tesis ha sido desarrollado en la Unidad Mixta de Investigación “Infección y Salud Pública” FISABIO-Universitat de València, Instituto de Biología Integrativa de Sistemas (CSIC-UV), ubicada en el área de Genómica y Salud de FISABIO.

Este trabajo no hubiera sido posible sin la “Ayudas para la formación de profesorado universitario (FPU) 2017” del Ministerio de Ciencia, Innovación y Universidades [FPU17/0237].

Los fondos proceden de los proyectos BFU2014-58656-R, BFU2017-89594-R y PID2021-127010OB-100 del Ministerio de Ciencia, Innovación y Universidades y CIPROM/2021/053 de la Conselleria d’Innovació, Universitats, Ciència i Societat Digital, Generalitat Valenciana. Los fondos para las infraestructuras e instrumental para el Servicio de Secuenciación de FISABIO proceden de los fondos europeos FEDER.





## Agradecimientos

Antes de emprender cualquier viaje, pasamos mucho tiempo preparando el equipaje que llevaremos con nosotros. Le dedicamos tanto tiempo, porque creemos que será todo con lo que contaremos cuando nos alejemos de nuestro punto de partida. Sin embargo, aunque el equipaje inicial puede ser muy importante, siempre hay que dejar un hueco para equipaje extra. Y cuanto menos conveniente, estar abiertos a la posibilidad de tener que dejar atrás parte de aquello con lo que comenzamos.

En mi vida, posiblemente, la tesis doctoral haya sido uno de los viajes más largos y emocionantes que haya tenido el placer de realizar. Cuando todo lo que ahora acaba solo hacía más que empezar, creí que tenía todo lo necesario. Pensé que podría esquivar cualquier piedra en el camino, y que los obstáculos solo serían anécdotas futuras que contar. “Pero la tesis es eso que pasa mientras tú te empeñas en hacer nuevos proyectos”. Y las piedras pequeñas en el camino pueden llegar a convertirse en inmensos acantilados; y los obstáculos anecdóticos pueden transformarse en verdaderas pesadillas. Sin embargo, es en esos momentos de tambaleo y aparente precipitación donde tu equipaje te ayuda a seguir adelante y a no rendirte a mitad del camino.

Mi familia es como ese chubasquero plegable que siempre llevas contigo; independientemente de donde vayas, y sea cual sea el tamaño de la maleta. Nunca sabes qué tiempo va a hacer, pero el chubasquero siempre está ahí. Llueva o haga sol, contra viento y marea, mi familia también ha sido paraguas, paracaídas, salvavidas y todo cuanto me ha hecho falta, siempre. Sin vosotros no hubiese habido viaje, ni sueños, ni pesadillas. Sois el motor de mi barco, las velas que lo empujan y los compartimientos estancos que pase lo que pase, lo mantienen a flote. Gracias por todo, y en especial a mis padres y a mi hermana. Siempre seréis indispensables en mi equipaje.

No menos importante, en la maleta siempre tiene que haber un neceser. El neceser es imprescindible, pues en él llevas de todo para hacer frente a cualquier situación

que se presente. Me gusta llamarlo el neceser polivalente, porque es como el cajón de tu madre dónde guarda las tijeras, mecheros a medio gas, restos de velas de cumpleaños, y el recetario de donde apuntó como hacer su primer arroz al horno. Hay de todo, y de lo más diverso, pero siempre es donde buscas cuando hay una emergencia. Y ese neceser polivalente es lo que han sido mis amigos para mí durante este viaje. Vosotros habéis sido mi botón rojo de asistencia, los fuegos artificiales de celebración, y el libro que siempre vuelvo a leer cuando busco desconectar del mundo.

Gracias a les “fogones” por ser mi espada y escudo. Por brindarme siempre vuestro apoyo incondicional, y por hacer que, aunque esté lejos, siempre os sienta cerca. Vosotras sois el significado de la palabra amistad.

Hector y Lorena, hemos compartido innumerables momentos, tanto antes de emprender este viaje como durante el mismo. Pero estoy convencida de que aún nos esperan muchos más por vivir juntos. Gracias, por todo, siempre.

Y a María, mi estimada compañera de biblioteca y de vida, sabes que siempre, allá donde vaya, habrá un hueco en mi maleta para ti.

Gracias también a mis compañeros de trabajo, sabéis que también sois amigos y que nos os nombro a todos porque no me quiero dejar a ninguno y habéis sido todos muy importantes. Solo os digo, que un bioinformático no es nada sin su ordenador, y vosotros habéis sido el mejor ordenador que haya podido tener. Pero no puedo no nombrarte a ti, mi querida Mígle. Tú también eres como ese chubasquero que me llevo a todos lados, porque eres ya parte de mi familia. Eres el claro ejemplo del hueco extra que hay que guardar en todo equipaje, porque, aunque no estabas al inicio del viaje, sin ti no hubiese podido acabarlo.

Quiero hacer una especial mención a mis compañeros del grupo Epimol.

Neris y Bea, siempre seréis mis Microbios de Dios favoritos. Ha sido un placer compartir tantos viajes y poder veros emprender los nuevos que empiezan. Estoy muy orgullosa de las increíbles investigadoras en que os habéis convertido.

Carlos F. y Lidia, vuestras charlas y consejos han sido un gran empujón de energía extra para seguir. Alma, Álvaro, Leo e Irving, vuestras aventuras en este proyecto de convertirnos en científicos han sido una gran fuente de inspiración para mí. Y a Andrea, Carlos V, Paz y el resto de los estudiantes del grupo, gracias por formar parte de mi viaje. ¡Ya os queda poco para escribir las memorias del vuestro!

Desde tiempos inmemoriales, los guías han tenido una función primordial en cualquier viaje. Ya sean mapas en papel como en los viejos tiempos o digitalizados como en el moderno presente, todo el mundo sabe que siempre hay que tener a mano uno. Y por suerte, durante este viaje, he tenido la suerte de poder contar con los mejores guías del mundo, mejores incluso que el Google maps.

A Fernando, mi director y supervisor, muchas gracias por darme la oportunidad de empezar este viaje. Sin tu apoyo, probablemente este proyecto nunca hubiera existido, y sin tus valiosos consejos, es posible que no lo hubiese terminado. Ha sido un camino largo, pero he tenido la suerte de contar con un guía excepcional. Gracias por tu paciencia, la confianza y por creer en mí desde el principio.

Dr. Verena J. Schünemann and Dr. Kerttu Majander: Just as a trip takes us to new places, expands our horizons, and exposes us to diverse experiences, my time at your institution has been an enlightening expedition into the world of cutting-edge research. I am very grateful to have had the opportunity to collaborate with brilliant minds and gather invaluable knowledge.

Dr. Natasha Arora, just as a travel guide unveils hidden gems and enriches the travel experience, you have shared your expertise, insights, and opportunities, opening doors to new possibilities, and broadening my horizons. With sincere gratitude, I express my heartfelt thanks for always inspiring and empowering me to reach new heights.

Y, por último, pero no menos importante, agradecer desde el fondo de mi corazón a mi otra pieza de equipaje indispensable. Richi. Tu apoyo, al igual que el de tu familia, ha sido más que incondicional. No sabía dónde me llevaría esta aventura, al igual que tampoco sé dónde nos llevará la vida, pero lo más bonito de cualquier viaje es disfrutarlo, y eso es lo que tú me enseñas días tras día. Gracias por ser el timón en mis días malos, por ser mi brújula sin importar el destino y por llenar de color mis días más grises.

# Contents

<b>ABSTRACT .....</b>	<b>I</b>
<b>ABBREVIATIONS .....</b>	<b>III</b>
<b>INTRODUCTION .....</b>	<b>3</b>
1. SYPHILIS AND THE ENDEMIC TREPONEMATOSSES .....	3
2. METHODS FOR MOLECULAR CHARACTERIZATION, DETECTION AND DIAGNOSIS.....	12
2.1 <i>Diagnostic tests</i> .....	12
2.2 <i>Molecular characterization of T. pallidum</i> .....	14
3. METHODS AND STRATEGIES FOR GENERATING SEQUENCE DATA FROM <i>T. PALLIDUM</i> SAMPLES .....	16
3.1 <i>Working with modern samples</i> .....	16
3.2 <i>Working with ancient samples</i> .....	19
3.3 <i>Bioinformatic strategies to process T. pallidum sequencing data</i> .....	20
4. GENERAL GENOMIC CHARACTERISTICS.....	23
4.1 <i>Intragenomic rearrangements in T. pallidum</i> .....	26
4.2 <i>Variability in the tprK gene</i> .....	28
5. EVOLUTIONARY MECHANISMS .....	29
5.1 <i>Point mutation</i> .....	29
5.2 <i>Genetic recombination</i> .....	32
5.4 <i>Natural selection</i> .....	34
6. GENOMIC EPIDEMIOLOGICAL STUDIES AND PHYLOGENETIC RECONSTRUCTION OF MODERN <i>T. PALLIDUM</i> GENOMES.....	35
7. PALEOGENOMICS OF <i>T. PALLIDUM</i> .....	40
7.1 <i>Ancient DNA (aDNA) and paleogenomics</i> .....	40
7.2 <i>Different hypotheses about the origin and spread of the treponematoses</i> .....	43
7.3 <i>Ancient T. pallidum genomes</i> .....	46
<b>OBJECTIVES.....</b>	<b>55</b>
<b>MATERIAL AND METHODS .....</b>	<b>59</b>

1. ANCIENT SAMPLES RECOLLECTION, ARCHAEOLOGICAL SITE INFORMATION AND RADIOCARBON DATING.....	59
2. ANCIENT SAMPLE PROCESSING.....	60
2.1 Ancient remains sampling.....	60
2.2 DNA extraction and library preparation.....	60
2.3 Pathogen screening.....	61
2.4 Damage profiles for aDNA authentication.....	61
2.5 Whole genome capture of the ancient samples.....	61
3. DATASET SELECTION, READ PROCESSING AND MULTIPLE GENOME ALIGNMENT GENERATION	62
4. ANTIBIOTIC RESISTANCE.....	63
5. RECOMBINATION DETECTION: PHYLOGENETIC INCONGRUENCE METHOD (PIM).....	63
5.1 Phylogenetic signal test (Likelihood mapping test).....	65
5.2 Phylogenetic congruence tests.....	66
5.3 Polyphyletic SNP distribution.....	66
6. RECOMBINATION DETECTION WITH ALTERNATIVE TOOLS (CHAPTER 1).....	67
7. SELECTION ANALYSES.....	67
8. PHYLOGENY RECONSTRUCTION.....	68
9. MAPPING THE TPE AND TEN NODE-DEFINING SNPs ON THE WHOLE GENOME-BASED TREE	69
10. MOLECULAR CLOCK DATING.....	70

**CHAPTER 1: “EVOLUTIONARY PROCESSES IN THE EMERGENCE AND RECENT SPREAD OF *T. PALLIDUM*” .....75**

1. BACKGROUND.....	75
2. MATERIAL AND METHODS.....	76
2.1 Dataset selection.....	77
2.2 Read processing.....	78
2.3 Recombination detection of <i>tpr</i> genes.....	79
3. RESULTS.....	80
3.1 Reference-based alignments.....	80
3.2 Recombination events in <i>T. pallidum</i> .....	80
3.3 Most recombination events have occurred between subspecies.....	85
3.4 Recombination detection with alternative tools.....	89

3.5 Selection analyses.....	89
3.6 Gene implications for the selection of vaccine candidates and the design of a broadly protective syphilis vaccine.....	98
4. DISCUSSION.....	100

**CHAPTER 2: “INFERRING PATTERNS OF RECOMBINATION AND DIVERGENCE WITH ANCIENT AND MODERN TREPONEMAL GENOMES” ..... 111**

1. BACKGROUND.....	111
2. MATERIAL AND METHODS.....	113
2.1 Sample recollection.....	113
2.2 Dataset selection.....	114
2.3 Read processing.....	114
2.4 Multiple genome alignment generation.....	115
3. RESULTS.....	116
3.1 Geographical origins and osteological analyses of the W86 sample.....	116
3.2 Pathogen screening for the new historical sample.....	116
3.3 Results of the PIM procedure.....	120
3.4 Recombinant events detected.....	120
3.5 Phylogeny reconstruction.....	124
3.6 Different possible scenarios for the detected recombination events.....	127
3.7 SNPs involved in the divergence between TPE and TEN.....	129
3.8 Selection analysis.....	132
3.9 Molecular clock dating.....	134
4. DISCUSSION.....	137

**CHAPTER 3: “REDEFINING THE TREPONEMAL HISTORY THROUGH PRE-COLUMBIAN ANCIENT GENOMES FROM BRAZIL” ..... 145**

1. BACKGROUND.....	145
2. MATERIAL AND METHODS.....	147
2.1 Dataset selection.....	147
2.2 Read processing and multiple reference-based genome alignment generation.....	148

3. RESULTS.....	151
3.1 Geographical origins and osteological analyses of samples .....	151
3.2 Preliminary pathogen screening and authenticity estimation of ancient DNA .....	153
3.3 Genome reconstruction.....	155
3.4 Multiple reference-based genome alignment.....	157
3.5 Phylogenetic analysis and genetic recombination .....	158
3.5 Molecular clock dating .....	161
4. DISCUSSION .....	165
<b>CHAPTER 4: “DEVELOPMENT AND EVALUATION OF A NEW MULTILOCUS SEQUENCE TYPING (MLST) SCHEME FOR <i>T. PALLIDUM</i>” .....</b>	<b>175</b>
1. BACKGROUND.....	175
2. MATERIAL AND METHODS.....	176
2.1 Genomic dataset generation for the new MLST scheme design .....	177
2.2. Design of the new MLST scheme .....	178
2.3. Fine-tuning of the new MLST scheme .....	179
2.4 In silico application of the new MLST scheme.....	182
2.5 Genetic diversity and population divergence in <i>T. pallidum</i> .....	183
3. RESULTS.....	183
3.1 Reference-based alignment for the design of a new MLST scheme.....	183
3.2 Selection of loci for primer design.....	184
3.3 Allelic profiles identified with the new MLST scheme .....	191
3.4 Sequence types (STs) identified among all typed samples.....	193
3.5 Phylogenetic analysis .....	201
3.6 Population genetic structure.....	211
4. DISCUSSION .....	218
<b>DISCUSSION.....</b>	<b>229</b>
<b>CONCLUSIONS.....</b>	<b>245</b>
<b>REFERENCES .....</b>	<b>249</b>
<b>APPENDICES .....</b>	<b>291</b>



SUPPLEMENTARY NOTES .....	291
SUPPLEMENTARY FILES.....	302
SUPPLEMENTARY FIGURES .....	304
SUPPLEMENTARY TABLES.....	316
<b>RESUMEN EN CASTELLANO .....</b>	<b>329</b>
INTRODUCCIÓN .....	329
OBJETIVOS .....	332
METODOLOGÍA.....	333
CAPÍTULO 1 .....	340
CAPÍTULO 2 .....	341
CAPÍTULO 3 .....	343
CAPÍTULO 4 .....	345
DISCUSIÓN .....	346
CONCLUSIONES.....	352



## Abstract

Recent advancements in high-throughput sequencing technologies have increased the availability of *T. pallidum* genomes, leading to a deeper understanding of this bacterium. This PhD thesis encompasses four studies that delve into both ancient and contemporary *T. pallidum* genomes, seeking profound insights into its evolution and genomics.

A new method called PIM was developed to detect recombination and selection in 75 contemporary *T. pallidum* genomes. PIM outperformed other tools in recombination detection, revealing the crucial roles of recombination and positive selection in *T. pallidum* evolution, particularly in defense and virulence.

Obtaining ancient *T. pallidum* genomes was considered impossible until recent advancements. This thesis successfully obtained two ancient high-coverage genomes: W86 (TPA) from Poland dating back to the 17th century, and ZH1540 (TEN) from Brazil dating back 2,000 years, the first pre-Columbian *T. pallidum* genome from the Americas. By incorporating these two new ancient genomes into diverse datasets of *T. pallidum* genomes, the study uncovered several additional novel recombinant genes using the PIM method. The identification of the strains involved in each recombination event shed light on potential recombination between TPE/TEN and TPA strains in the Old World, indicating the coexistence and circulation of these subspecies in the same region. Moreover, the divergence dates from ancient genomes were older than estimates based on modern genomes alone, enhancing evolutionary timeline accuracy through Bayesian molecular clock dating.

A new mapping approach was developed to enhance genome coverage, reduce reference bias, and improve the accuracy of phylogenetic inference and assignment. This approach eliminated the need for comparisons with multiple reference genomes, streamlining subsequent analyses. While the choice of reference genome influenced the phylogenetic placement of ancient genomes, it did not impact the classification of strains within subspecies.

Furthermore, a novel MLST scheme was devised utilizing 121 *T. pallidum* genomes, incorporating seven variable genes and 23S rRNA genes. This scheme effectively discriminated between strains across all *T. pallidum* subspecies, revealing genetic diversity and highlighting the prevalence of macrolide resistance, particularly within the SS14 sublineage. Notably, sample amplification was achieved using a single PCR instead of nested PCRs, resulting in significant time and cost savings while improving efficiency. Analysis of genetic diversity and population structure unveiled localized transmission patterns and underscored the influence of regional factors in the spread of *T. pallidum*.

This doctoral thesis represents a significant advancement in our comprehension of *T. pallidum* evolution, genomics, and epidemiology. The inclusion of ancient genomes, the innovative mapping approach, and the novel MLST scheme collectively contribute to the progress of this field.

## Abbreviations

<b>aDNA</b>	Ancient DNA
<b>ATP</b>	Adenosine triphosphate
<b>aBSREL</b>	Adaptive Branch-Site Random Effects Likelihood
<b>BAPS</b>	Bayesian analysis of population structure
<b>BCE</b>	Before the Common Era
<b>BEAST</b>	Bayesian evolutionary analysis by sampling trees
<b>BamA</b>	Barrel assembly machinery A
<b>BLAST</b>	Basic local alignment search tool
<b>BRIG</b>	BLAST Ring Image Generator
<b>BUSTED</b>	Branch-Site Unrestricted Statistical Test for Episodic Diversification
<b>BWA</b>	Burrows-Wheeler Aligner
<b>CDC</b>	United States Centers for Disease Control and Prevention
<b>CE</b>	Common Era
<b>DMSO</b>	Dimethylsulfoxide
<b>DNA</b>	Deoxyribonucleic Acid
<b>dNTP</b>	Deoxynucleotide Triphosphates
<b>ECDC</b>	European Centre for Disease Prevention and Control
<b>EIA</b>	Enzyme immunoassays
<b>ENA</b>	European Nucleotide Archive
<b>EEA</b>	European Economic Area
<b>EU</b>	European Union
<b>ELW</b>	Expected Likelihood Weights
<b>FTA-ABS</b>	Fluorescent treponemal antibody absorption test
<b>GTR</b>	Generalized time reversible
<b>GATK</b>	Genome Analysis Toolkit
<b>GC</b>	Guanine - Cytosine
<b>HIV</b>	Human Immunodeficiency Virus
<b>HPD</b>	Highest posterior density

<b>HTS</b>	High-throughput sequencing
<b>ISRs</b>	Intergenic spacer regions
<b>ID</b>	Identification
<b>I.e.</b>	For example
<b>LPS</b>	Lipopolysaccharide
<b>LM</b>	Likelihood-mapping
<b>MAFFT</b>	Multiple Alignment using Fast Fourier Transform
<b>MCMC</b>	Monte Carlo Markov chain
<b>MCC</b>	Maximum clade credibility
<b>MDA</b>	Mass drug administration
<b>MHA-TP</b>	Microhemagglutination Assay for <i>Treponema pallidum</i>
Antibodies	
<b>MLST</b>	Multilocus Sequence Typing
<b>MSA</b>	Multiple genome alignment
<b>MSM</b>	Men who have sex with men
<b>MUSIAL</b>	MUlti Sample varIant AnaLysis
<b>NA</b>	Not available
<b>NC</b>	Non computable
<b>NCBI</b>	National Center for Biotechnology Information
<b>OMP</b>	Outer membrane proteins
<b>PCR</b>	Polymerase chain reaction
<b>PGF</b>	Paralogous gene families
<b>PIM</b>	Phylogenetic Incongruence Method
<b>PP</b>	Posterior probability
<b>rDNA</b>	Ribosomal Deoxyribonucleic Acid
<b>RxML</b>	Randomized Axelerated Maximum Likelihood
<b>RDP</b>	Recombination detection program
<b>RPR</b>	Rapid Plasma Reagin
<b>rRNA</b>	Ribosomal ribonucleic acid
<b>RT-PCR</b>	Real Time-Polymerase chain reaction

<b>SMBT</b>	State Model Based Testing
<b>SNP</b>	Single Nucleotide Polymorphism
<b>S/S/Y</b>	Substitutions/site/year
<b>SPP</b>	Species
<b>ST</b>	Sequence Type
<b>TEN</b>	<i>Treponema pallidum</i> subspecies <i>endemicum</i>
<b>tMRCA</b>	Most recent common ancestor
<b>TPA</b>	<i>Treponema pallidum</i> subspecies <i>pallidum</i>
<b>TPE</b>	<i>Treponema pallidum</i> subspecies <i>pertenue</i>
<b>TPc</b>	<i>T. paraluisuniculi</i>
<b>tRNA</b>	Transfer RNA
<b>UNIV</b>	University
<b>UV</b>	Ultraviolet
<b>UniProt</b>	The Universal Protein Resource
<b>VDRL</b>	Venereal disease research laboratory test
<b>VCF</b>	Variant Call Format
<b>WB</b>	Western blot
<b>WGA</b>	Whole Genome Amplification
<b>WGS</b>	Whole-Genome Sequencing
<b>WHO</b>	World Health Organization
<b>YR</b>	Year





— **INTRODUCTION** —



## Introduction

### 1. Syphilis and the endemic treponematoses

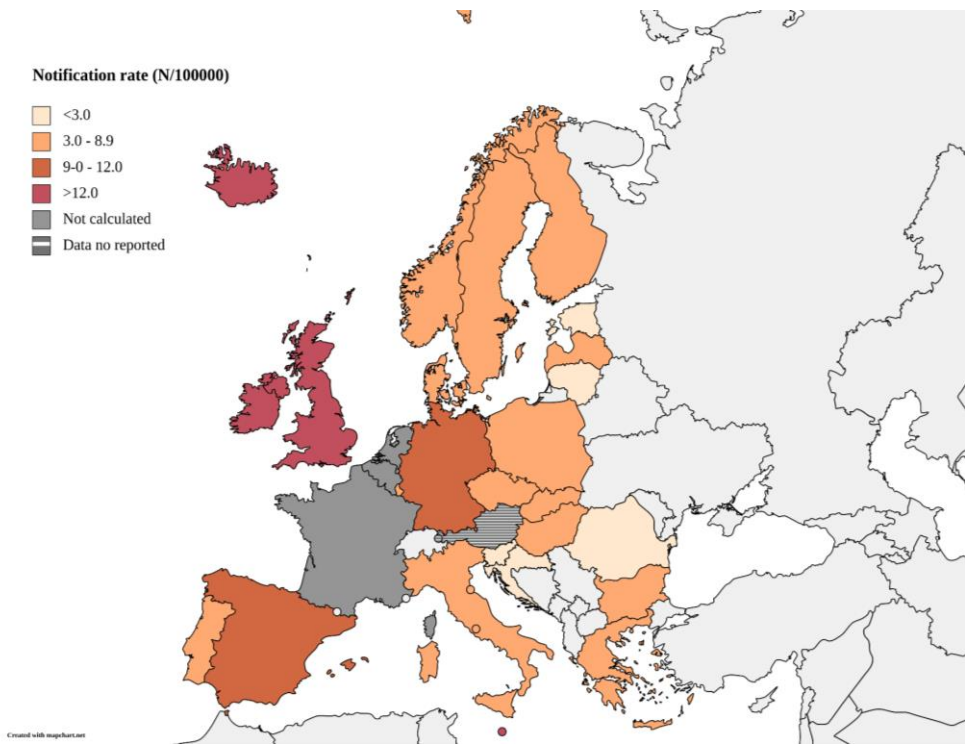
One of the most mysterious and challenging pathogens to study is *Treponema pallidum*, Gram-negative bacteria that belong to the order *Spirochaetales*, family *Spirochaetaceae*, and genus *Treponema*. This bacterium is the etiological agent of the treponemal diseases syphilis, yaws and bejel, caused by three closely related subspecies, *T. pallidum* subsp. *pallidum* (TPA), *T. pallidum* subsp. *pertenue* (TPE) and *T. pallidum* subsp. *endemicum* (TEN), respectively.

The causative agents of syphilis, yaws and bejel were originally classified as separate species, but are now considered subspecies of *T. pallidum* due to their high genetic relatedness (99.7%) proven by DNA hybridization and whole genome sequencing [1,2]. There is a similar, less well-known treponematosis called pinta that is caused by *T. carateum* [3,4]. However, genetic analyses of this organism have been prevented by the lack of an isolation of the agent of pinta, so it is not currently classified as a subspecies of *T. pallidum* and it retains its separate name [4–6]. Moreover, there is an additional *Treponema* species, *T. paraluisuniculi*, which is the causative agent of rabbit venereal spirochetosis but it is not infectious to humans [7,8]. This lagomorph pathogen shares a 99.2% genome identity with *T. pallidum* [9].

Yaws, along with bejel and pinta, are designated as endemic treponematoses that are typically spread by skin-to-skin contact through scrapes or cuts especially among children in poor and rural communities, in warm and humid tropical regions where they are endemic [6,10,11]. In contrast, syphilis persists worldwide and is typically spread either sexually or congenitally, i.e., during pregnancy or childbirth [6,12].

Despite the availability of an effective treatment, syphilis has been re-emerging worldwide, with more than 7 million cases reported each year, and featuring epidemics that particularly affect networks of men who have sex with men (MSM) in high-income countries [13].

According to the latest report of the European Centre for Disease Prevention and Control (ECDC), in 2019, 29 EU/EEA Member States reported 35,039 confirmed syphilis cases, for a crude notification rate of 7.4 cases per 100 000 people [14] (see Figure 1). Men reported nine times more cases of syphilis than women, with the highest rates in the male age group of 25 to 34 years (31 cases per 100,000 population). MSM account for the vast majority (74%) of syphilis cases with information on transmission categories. Syphilis notifications among men increased between 2010 and 2017, primarily as a result of an increase in cases among MSM; however, this increase appears to have slowed down in 2018 and 2019. There were very slight variations in syphilis notifications among heterosexuals at the EU/EEA level throughout the same time period. Pregnancy-related syphilis is the second most common reason for stillbirths worldwide and also causes preterm, low birthweight, neonatal mortality, and infections in neonates [13,15].



**Figure 1.** Distribution of confirmed syphilis cases per 100 000 population by country, EU/EEA, 2019. Data from [14]. Figure generated with <https://www.mapchart.net/world.html> [16]

Yaws is the most prevalent endemic treponematoses (Figure 2). It is widespread in Africa, Asia, Latin America and the Pacific, and was one of the first diseases targeted for eradication in the 1950s by the World Health Organization (WHO) [16,17]. Renewed eradication efforts started in 2012 and have significantly reduced the incidence of yaws in some areas [18–21]. However, so far, only India has been certified as free of yaws transmission, in 2016. In 2021, yaws was thought to be endemic in 15 countries, but the current status of 82 other nations, regions, and territories is uncertain [16]. Only 1,102 of the 123,866 suspected yaws cases reported to WHO in 2021 from 13 countries were confirmed in those same countries [13]. The WHO region reporting the most suspected cases is the Western Pacific, particularly with 92,856 instances in Papua New Guinea and 13,599 cases in the

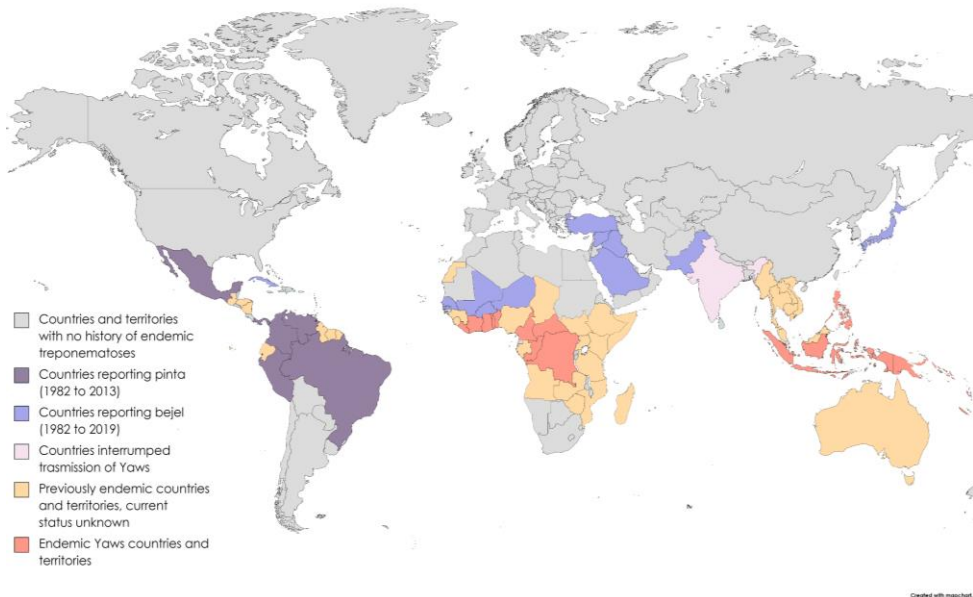
## Introduction

---

Solomon Islands. Unfortunately, laboratory testing is not yet frequently used to confirm cases due to the low country's resources where it is endemic [13].

In the 20th century, bejel was eradicated from Europe, where it had previously been common since the 16th century [22]. However, in arid and hot regions, such as the Sahel region of western Africa, part of Botswana, Zimbabwe, and the Arabian Peninsula, bejel is still widely prevalent but the exact number of cases is unknown [4,13]. Only a few cases of bejel have been reported in non-endemic areas since 1999, including France [23], Canada [24], and Cuba [5,24], mostly corresponding to imported cases (Figure 2). Although bejel can be transmitted sexually, this route has not been thoroughly investigated because bejel primarily affects children, but recent studies from France, Cuba and Japan that examined cases of TEN infection among sexually-active males suggested its transmission through sexual contact [6,11,12].

Pinta may still be endemic in Mexico and Central and South America, where it was common in the 1980s, but treatment campaigns and presumably better living conditions, access to healthcare, and cleanliness standards have led to a drop in the disease's incidence [4,25] (Figure 2). Currently, the prevalence of pinta is unknown due to lack of surveillance data [26] .



**Figure 2.** Endemicity status and geographic distribution of yaws, bejel, and pinta worldwide. Data from: WHO Global Health Observatory: Endemic Treponematoses. Available at: [http://apps.who.int/neglected\\_diseases/ntddata/treponematoses/treponematoses.html](http://apps.who.int/neglected_diseases/ntddata/treponematoses/treponematoses.html) and [https://www.who.int/images/default-source/maps/yaws\\_2021\\_endemicity.png?sfvrsn=748acf73\\_3](https://www.who.int/images/default-source/maps/yaws_2021_endemicity.png?sfvrsn=748acf73_3). Figure generated with <https://www.mapchart.net/world.html>

The human treponematoses have remarkably similar clinical symptoms (Table 1). With the exception of pinta, all can develop into serious lesions that damage cartilages, bones, and the skin [4,27]. The early signs of syphilis are localized and include the development of a chancre at the site of entry of the microorganism [3,17]. It normally causes no pain and has a diameter of 1 to 2 cm; thus, it might go unnoticed. Although it is often one lesion, persons who also have HIV typically have several lesions [27,28]. Following this first phase, a second phase starts 6 to 8 weeks after the initial infection. It displays a more complex clinical picture with blisters and rashes along with general symptoms like fatigue, fever, or weight loss. While the symptoms of the two phases frequently overlap in HIV patients, they typically do not coexist in time with the initial phase's chancres [28,29]. The primary

and secondary phase symptoms of syphilis may disappear on their own even if the infected person is not treated, leaving latent syphilis. Classically, the first year following infection is referred to as the early latent period, and the second year as the late latent phase. The significance of this distinction lies in the fact that transmission is more likely to occur in patients in the early latent phase than in the late latent phase, which necessitates a longer course of treatment [30]. Patients in the early latent phase may also have the typical mucocutaneous lesions of secondary syphilis, which are potentially contagious. Following this (between 2 and 40 years after infection), a tertiary stage might develop and up to one-third of the persons infected may experience the disease's most severe symptoms, such as cardiovascular involvement, granulomatous subcutaneous lesions, or nervous system involvement if left untreated (neurosyphilis) [17]. Interestingly, the amount of treponemes in the tertiary syphilis stage is low in contrast to primary and secondary syphilis, making these lesions less contagious [31,32].

The clinical features of the endemic treponematoses are typically split between an early stage (comprising primary and secondary signs) and a late stage, similar to venereal syphilis. Early-stage lesions can last for weeks, months, or even years after they first develop and are very contagious. Moreover, as a result of the host's immune reaction to the infection, the early symptoms regress spontaneously and the infection enters a state of latency that frequently lasts a lifetime. Nevertheless, the infection can sometimes advance from latency to tertiary illness, which is marked by tissue destruction.

Despite many similarities in their clinical manifestations, some differences among the three treponematoses have been described [3,4,17] and are detailed in Table 1. While cardiovascular, neurological, and ophthalmological manifestations during syphilis infection are well recognised, these manifestations are rarely or not reported for endemic treponematoses (Table 1).



## Introduction

---

The inability to grow these spirochetes in vitro presents a significant challenge. This limitation requires expensive and time-consuming continuous propagation in lab animals. Additionally, the remarkably fragile cellular ultrastructure of the pathogen further restricts mechanical manipulation, thus preserving its integrity becomes a feasible but limited option. Furthermore, the bacterium's limited viability outside a host adds another obstacle. These factors combined greatly hinder our overall understanding of the pathogenesis of human treponematoses. [3].

## Introduction

**Table 1.** Characteristics of the different human treponematoses. Modified from: [4].

<b>Feature</b>	<b>Yaws</b> <i>T. pallidum subsp pertenuae</i>	<b>Bejel</b> <i>T. pallidum subsp endemicum</i>	<b>Pinta</b> <i>T. caretum</i>	<b>Syphilis</b> <i>T. pallidum subsp pallidum</i>
Geographical distribution	Western/Central Africa, Southeast Asia, Pacific Islands	Sahelian Africa, Saudi Arabia	Central and South America	Global
Climatic conditions	Tropical (hot and humid)	Hot and dry (semiarid/arid)	Warm (semiarid)	All
Age group (peak incidence of lesions)	Children (<15 yr)	Children (2–15 yr)	Children and adults	Adults
Common mode of transmission	Skin-to-skin contact	Mucous membrane and skin-to-skin contact (sharing of eating and drinking utensils )	Skin-to-skin contact	Sexual and congenital; occasionally nonsexual contact
Congenital transmission	Commonly unrecognized but affirmed [33]	Commonly unrecognized but affirmed [6,11-12]	No	Frequent
Most common location of primary lesion	Lower extremities	Oral mucosa (rarely seen)	Extremities	Genitalia, anal, and oral mucosae
Most affected organs	Skin and bone	Oral and nasal mucosae, intertriginous areas and bone	Skin	Skin; systemic involvement, including central nervous system and fetus
Late complications in the absence of treatment	Destructive osteitis, saddle nose, destruction of the palate and nasal septum, painful lesions on soles	Destructive osteitis of the nose, palate, and nasal septum	Depigmented lesions over hands, wrists, elbows, feet, and ankles	Neurological (optic atrophy, paresis, tabes dorsalis), cardiovascular (aortitis, aortic aneurysm), gummatous (skin, lung, liver, brain, other organs)
Central nervous system involvement	Believed to be rare	Believed to be extremely rare or absent	Not recognized	Frequent
Animal models	Rabbit $\cong$ hamster	Rabbit $\cong$ hamster	Primate only	Rabbit > primate > guinea pig > hamster

## Introduction

Genome size (kbp)	1139.3–1139.7	1137.7	NA	1138.0–1140.0
Genomic identity with <i>T. pallidum</i> subsp. <i>pallidum</i> (%)	99.8	99.7	NA	NA

ND, no data.

NA, not applicable.

Treponematoses are treated with single-dose antibiotic therapy using benzathine penicillin but, in certain situations, such as in cases of allergy to penicillin, macrolides like azithromycin are the second line option. Nevertheless, macrolide resistance, especially in syphilis, has been associated with two point mutations in the 23S rDNA gene (A2058G and A2059G) [17,34]. Macrolide resistance mutations in TPA have a global prevalence of 90% , but there are significant regional differences: Cuba, the US, and several European nations report high prevalence rates (61-100%), while Canada, Russia, South Africa, Madagascar, Taiwan, Peru, and Argentina report low to no macrolide resistance (0-14%) [35]. However, there is only one study reporting macrolide resistance in bejel in the recently sequenced four Japanese strains [36]. Unfortunately, there are several reports of macrolide resistance in yaws [19,34,37], which highlighted a significant link between macrolide resistance prevalence and macrolide consumption in *T. pallidum*, further supporting this observation. These collective results underscore the detrimental consequences of the widespread use and misuse of macrolides, as it has created selective pressure, subsequently leading to an alarming rise in macrolide-resistant strains. Consequently, this scenario greatly facilitates the dissemination of such strains within the population [19,34,37,38]. These results highlight the necessity of monitoring the affected communities following mass drug administration (MDA) in order to quickly identify the emergence of new azithromycin resistance and to consider alternative treatments for cases discovered following MDA.

## **2. Methods for molecular characterization, detection and diagnosis**

### **2.1 Diagnostic tests**

The challenging identification of these treponematoses is another reason why the study of these diseases has also been so perplexing. As all treponematoses are similar multi-stage infections, it is not always possible to accurately identify patients with the disease by their clinical symptoms, because the symptoms are easily confused with other conditions or go unnoticed, especially the painless lesions of primary stages, which may develop in inaccessible areas such as the cervix or rectum. Nonetheless, in the absence of confirmatory laboratory tests, the traditional diagnosis of yaws, pinta, bejel or syphilis is based on the analysis of symptoms and signs, together with the epidemiological context of each treponematoses [4].

There are direct and indirect tests for laboratory diagnosis, which have been classically divided into "treponemal" [25] and "non-treponemal" tests [39] (Table 2). Non-treponemal tests are indicated in the screening process, while treponemal tests would be the confirmatory tests.

**Table 2.** Tests available for laboratory diagnosis of *T. pallidum* and their limitations.

Tests for laboratory diagnosis		
Type	Name	Limitations
Treponemal test (direct)	VDRL (Venereal Disease Research Laboratory)	<ul style="list-style-type: none"> <li>- False positives. Associated with age unrelated infections or autoimmune disorders.</li> <li>- False negatives in very early disease stages</li> </ul>
	RPR (Rapid Plasma Reagin)	
Non-treponemal tests (indirect)	FTA-ABS (Fluorescent Treponemal Antibody Absorption Test)	<ul style="list-style-type: none"> <li>- Offer positive results even after the infection has been overcome</li> <li>- They cannot be used to monitor or assess the success of the treatment used [40].</li> </ul>
	Double-stained FTA-ABS	
	MHA-TP (Microhemagglutination Assay for <i>T. pallidum</i> Antibodies)	
	Enzyme immunoassays (EIA)	
	Western blot (WB)	

There are other methods employed for the diagnosis of syphilis, like direct observation of *Treponema* by darkfield microscopy [39] and a battery of quick screening tests [41]. The latter consist of straightforward tests that can be carried out outside a lab setting, with little training of staff and without the need for specialized equipment. This helps to solve the common issues in less developed countries, where conditions are precarious due to lack of access to a laboratory and also with low patient return rates.

Nevertheless, all of these methods have a low detection sensitivity (70–80%) and serological methods are unable to differentiate these diseases from each other, being more difficult to clearly define them and to obtain an accurate diagnosis [42].

In 1990, the polymerase chain reaction (PCR) was introduced as a new syphilis diagnosis tool [43]. Since its creation, this method has undergone improvements that have boosted its sensitivity and specificity as well as enabled the use of other variants, such as Real Time-PCR (RT-PCR), which enables simultaneous amplification and quantification of the desired amplification product. PCR and RT-PCR typically target either the DNA polymerase I gene (*polA*) or the 47-kilodalton membrane protein gene (*tp0047*) [44].

### **2.2 Molecular characterization of *T. pallidum***

With the development of molecular technologies, all *T. pallidum* subspecies may now be identified utilizing genotyping tools and DNA-sequencing-based techniques. Nevertheless, these bacteria are still very difficult to grow and obtaining genomic sequences requires an expensive and time-consuming enrichment technique, highlighting the need for genotyping tools to expand our understanding of the epidemiology of these bacteria.

Multilocus Sequence Typing (MLST) is a molecular technique that allows the characterization of bacterial isolates using the sequences of housekeeping genes (usually six or seven genes) [45–47]. In general, internal fragments of approximately 400-600 bp of each gene are used for this purpose, which can then be accurately sequenced by different techniques (Sanger, high-throughput sequencing technologies, etc.). With the sequences obtained, an allele number is assigned to each locus by comparison with previously identified variants and with the combination of the six or seven numbers (depending on the MLST scheme used) the Sequence Type (ST) is obtained. The number of nucleotide differences between alleles is not taken into account in MLST, and sequences are assigned different allele numbers regardless of how many nucleotide sites they differ at [47,48]. Most bacterial species have sufficient variation within house-keeping genes to provide many alleles per locus, allowing billions of distinct allelic profiles to be distinguished using six or seven house-keeping loci. For example, an average of 30

alleles per locus allows about 20 billion genotypes to be resolved [45]. For this reason, there are many MLST schemes available and well standardized for different pathogens of great epidemiological interest such as *Neisseria* spp. [45], *Staphylococcus aureus* [49], *Campylobacter jejuni* [50], *Streptococcus pneumoniae* [50,51], etc.

There are several molecular typing schemes currently available for *T. pallidum* (Table 3). However, these schemes do not meet the characteristics of an MLST typing scheme and will be referred to as typing schemes. They use few or too many loci and are also affected by significant technical difficulties when applied to clinical samples, such as the requirement to do a nested PCR for the sample sequencing, because the amplicon size is too large to obtain good sequencing results using Sanger's methodology. Some of these typing schemes have even been questioned in several researches due to potential intra-strain variability at the loci (*arp* and *tpr* genes) employed [52–54]. Furthermore, these typing systems were not designed considering the information from a complete whole genome dataset with representative genomes from the three subspecies and their variation, causing a limited ability of these typing systems to reveal genetic variability within and to differentiate among them. They are also mostly designed to type and characterize in detail only one of the three subspecies of *T. pallidum* at a time. The ease of access to the typing results obtained by MLST schemes is one of the main reasons to continue using this epidemiological tool in comparison with other genomic techniques. Moreover, they have a low economic cost compared to other genomic methodologies applied for *T. pallidum*, in addition to a high sensitivity and reproducibility. They also have, for the most part, public databases easily accessible worldwide where the results obtained are deposited [45]. However, only the typing system for TPA was introduced in 2018 by Grillová *et al.* [55], and its results obtained from its application so far are available in a public database (PubMLST) [56]. All these precedents highlight the need for a novel approach.

**Table 3.** Different molecular typing strategies currently available for *T. pallidum*.

Molecular typing scheme	Target	Loci	Reference
CDC	TPA	<i>arp</i> , <i>tprE</i> , <i>tprG</i> , and <i>tprJ</i> + <i>rpsA</i>	[52,53]
SMBT	TPA	<i>tp0136</i> , <i>tp0548</i> and 23S rDNA	[57,58]
ECDCT	TPA	previous CDC typing scheme + <i>tp0548</i>	[59]
MLST	TPA	<i>tp0136</i> , <i>tp0548</i> , and <i>tp0705</i> + 23S rDNA	[55]
MLST	TPE	<i>tp0548</i> , <i>tp0136</i> and <i>tp0326</i>	[60]
MLST	TPE	<i>tp0488</i> , and <i>tp0548</i>	[61]
CDC_TPE	TPE	CDC and CDCE typing schemes + <i>tp0279</i>	[62]
MLST	TEN	<i>tp0136</i> , <i>tp0326</i> , <i>tp0367</i> , <i>tp0488</i> , <i>tp0548</i> , <i>tp0859</i> , <i>tp0861</i> , <i>tp0865</i> and 16S rDNA + 23S rDNA	[5]

### 3. Methods and strategies for generating sequence data from *T. pallidum* samples

#### 3.1 Working with modern samples

Until recently, *T. pallidum* could not be cultured *in vitro* continuously. As a result, the inoculation of rabbits or other experimental animals was used to spread this infection and those of other similar spirochetes, which is a time-consuming, labor-intensive, and expensive process [63,64]. However, recent studies have allowed a significant achievement by creating a tissue culture system using a medium called TpCM-2 that promotes the long-term proliferation of TPA and TEN in *in vitro* culture. The possibility of sustained, constant culture creates new opportunities for investigation, [65–67], although it does not eliminate the issue of host DNA contamination as cultures are still characterized by the presence of cottontail rabbit epithelial cells (Sf1Ep). Since there is not yet available a standardized protocol that allows to culture this bacterium routinely as other pathogens, this system cannot be used to obtain complete *T. pallidum* genomes, at least for now. Interestingly, it has been demonstrated that *T. pallidum* can be genetically modified [68], though further studies are needed to standardize this technique and apply it for future functional studies, or even to help improve the previously mentioned cultivation systems of this bacterium.



The first genome sequence of *T. pallidum* was completed in 1998 [69] and it cost US\$1.1 million approximately over the course of more than a year and 1,000 working hours. Since then and with the advent of HTS (High-Throughput Sequencing), more than 1400 of new draft *T. pallidum* genomes have been generated [2,21,36,70–77]. However, due to the low concentrations of treponemes in the samples and the strong background DNA from the host, direct sequencing of clinical samples has been hampered [73,78,79] and researchers had to overcome a number of obstacles. Samples from non-human primates are likewise impacted by this problem [75,80], and historical and ancient samples are impacted even more.

In order to solve all these problems, researchers have resorted to using treponemal DNA purification and enrichment techniques as a development that made direct sequencing of *T. pallidum* samples from patient specimens to obtain genome sequences easier. Treponemal genome sequencing necessitates either treponemal cell isolation, treponemal DNA amplification or a combination of the two techniques. It is therefore not surprising that the amplification of treponemal DNA was one of the most important steps for the whole genome sequencing of treponemes, and for which four main strategies have been followed:

1. An entire genome can be amplified robustly using the Whole Genome Amplification (WGA) technique [81], which uses primers with random sequence, to obtain a good number of DNA copies for the following sequencing step, starting with nanograms of DNA and ending with micrograms of amplified products. Nevertheless, the quantity of *T. pallidum* DNA copies produced by WGA directly depends on the quantity and presence of contaminating DNA, so this method is only appropriate for samples with an excess of treponemal DNA in comparison with the contaminating DNA on the sample, because it significantly reduced TPA amplification when a very small amount of human DNA was mixed with treponemal DNA [81].

2. Semi-specific or specific amplification of treponemal DNA using primers that recognize treponemal DNA, as well as the capture and removal of CpG-methylated host DNA [82]. Nonetheless, the primary difficulty of this method is the exclusion of divergent sections from the final genome sequence due to the inability to amplify such regions, which can be originated by differences in the genome used as reference for primer design among others.
3. Enrichment of bacterial DNA based on methylation pattern using the restriction enzyme *DpnI* [73], which opens the door to future studies on *Treponema* methylation patterns.
4. Enrichment of treponemal DNA based on hybridization capture [34,71,75,78–80,83,84]. This is the most popular and possibly most effective method for selective enrichment and amplification of treponemal DNA until now. Even so, similar to the case of the specific amplification of treponemal DNA, a handicap of the hybridization capture technique is also the exclusion of progressively divergent sections from the final genome sequence due to their insufficient amplification.

Nevertheless, the genomes obtained by the use of these techniques are not complete, because these approaches do not fill up the gaps in areas with high genetic complexity. Therefore, to provide entire genome sequences it is still necessary a combination of short read (by Illumina) and long read (by PacBio or Oxford Nanopore Technologies) techniques or the use of Sanger sequencing after PCR amplification [2,85,86]. Although in recent years it has been possible to improve and optimize all these techniques to reduce the economic cost and number of working hours needed to obtain whole genomes of *T. pallidum*, it is still a very expensive process compared to the price of obtaining genomes of other easily cultivated bacteria [36,83,87,88].

### 3.2 Working with ancient samples

Obtaining complete modern genomes of *T. pallidum* is already challenging, but acquiring ancient genomes of this bacterium poses an even greater difficulty. The degradation and scarcity of ancient DNA (aDNA) makes it necessary to create working conditions that reduce the possibility of contamination and inaccurate results [89,90]. Although it is often impossible, it is recommended that sterile and clean conditions begin at the archaeological site where the remains are located. It is important to collect samples and extract aDNA in a specific laboratory where no modern DNA is used. If this step is carried out in a structure that does not function with contemporary DNA, it is even more reliable. The lab should be designated as a "clean lab" and before and after each use, surfaces should be cleaned and decontaminated using ethanol and UV radiation. In addition to a specific space only designated for sample processing and DNA extraction, the whole laboratory must have an enclosed ventilation system [89].

When working inside a clean lab, workers should utilize sterile materials, gloves, gowns, and masks [91]. Moreover, it is required to employ solely negative controls for library formation and enrichment and both positive and negative controls for PCR [92]. The analysis of environmental samples taken from the reservoir with an environmental control is also beneficial to ensure that the results are non contaminated [93].

In the case of *T. palladium*, the extraction of aDNA is usually done from bone remains where the characteristic lesions produced by this bacterium have been previously located by anthropological specialists [92]. Before handling the bone, documentation is first performed. This consists of photographing, weighting, and measuring the bone remains in order to have as accurate a description as possible. Then, the biological material is irradiated with UV light to remove any trace of modern DNA contamination on the surface [92]. For *T. pallidum*, DNA extraction is based on the well-established extraction protocol for ancient DNA [94], which is

divided into four steps: homogenization/pulverization, lysis, purification, and elution. The pulverization and homogenization step is done by a hand rotary tool. Afterwards, the extracted DNA is purified, being a key and very delicate step for which a protocol based on columns is used [94].

The most widely used HTS technique in paleogenomics is sequencing by synthesis offered by Illumina. For library construction for subsequent HTS sequencing of the samples, two protocols can be used: double-stranded DNA (dsDNA) and single-stranded DNA (ssDNA). Which of the two types of libraries is more suitable for aDNA research has been the subject of discussion and comprehensive comparison [95–97]. ssDNA libraries are more expensive and time-consuming than dsDNA libraries, but they recover more fragments, particularly fragments of less than 70 bp, than dsDNA libraries. The difference between the two libraries is not great enough in samples with high levels of preservation to justify the costs of ssDNA. Hence, if after executing a dsDNA library it is discovered that the DNA is substantially degraded or has an average length of less than 70 bp, it is often advised to perform ssDNA libraries. Furthermore, to enhance the recovery of *T. pallidum* aDNA, hybridization capture is employed for the enrichment of treponemal DNA [92,98]. Additionally, the libraries are treated with UDG (uracil-DNA glycosylase) to eliminate uracil residues from the DNA, creating abasic sites. This treatment has proven effective in reducing C/G→T/A misincorporations associated with ancient DNA [99].

### **3.3 Bioinformatic strategies to process *T. pallidum* sequencing data**

After HTS sequencing of the enriched DNA, to reconstruct the individual genomes from the raw data, the resulting reads are merged sample-wise, following different general steps [100]: 1) check the reads quality, 2) remove sequencing adapters, 3) mapping or *de novo* assembly, and 4) variant calling and filtering.

As mentioned above, there are two methodological strategies to obtain the final sequences from the sequencing data: mapping or *de novo* assembly. During mapping, the algorithm locates reads with nucleotide similarities to the reference genome and maps them to the reference genome. Without a reference to compare reads to, *de novo* assembly tries to piece reads together like a jigsaw puzzle, resulting in either one or several contigs, which are fragments of a single genome or several genomes. Contigs must next be assembled using additional computational procedures to obtain a whole genome. Given the nature of the samples, *de novo* assembly is ruled out for ancient samples, so mapping is the only available methodology to obtain the sequences of these samples for which there are specific and widely used pipelines such as EAGER [101].

However, in general terms for modern samples, depending on the characteristics of the microorganism and above all on the availability of a good reference genome, one strategy or another is better. *De novo* assembly is more accurate because it does not need a reference genome to generate the final sequence [102]. This is important because there might be a bias dependent on the reference genome used for mapping, which could affect the SNP calling and the subsequent phylogenetic analyses [103]. However, *de novo* assembly has higher computational and economic costs, because the bioinformatics processing time is longer and a high coverage sequencing is needed to guarantee the acquisition of a single contig per genome (a whole genome) [104].

In the particular case of *T. pallidum*, as obtaining quality genomes is very difficult, mapping is the most frequently used strategy to process the sequencing data and to obtain the final sequence even for modern samples.

According to the different subspecies/lineage of *T. pallidum*, Nichols (CP004010.2) from TPA (Nichols lineage), SS14 (NC\_021508.1) from TPA (SS14 lineage), CDC2 (CP002375.1) from TPE, and BosniaA (CP007548.1) from TEN, would be the best genomic references available. They are the best candidate strains because

all of them are whole genomes (no draft genomes), with very little missing data (172 Ns at most) [86], and the gaps between the contigs obtained in the *de novo* assembly of these samples have been closed by PCR and subsequent sequencing with Sanger. However, as highlighted in section 5.1 of the Introduction, a recent study has demonstrated the occurrence of point mutations due to the successive passages in rabbits [105]. These passages were conducted to amplify the DNA quantity in the sample and facilitate their sequencing. While further investigations are required to assess the impact of these mutations when utilizing these strains as reference genomes, their significance cannot be understated and must be duly considered.

The Nichols strain is the most widely used reference genome in genomic studies of this bacterium [70,78,79,106], although there are some studies using the SS14 strain [71,83,107]. Some studies use CDC2 as the reference genome because the genomes generated and/or employed in those studies are from TPE [34,72,80]. Interestingly, the few TEN genomes available have been obtained by *de novo* assembly, except four strains from Japan obtained by Lieberman *et al.* [36]. These four samples were obtained by a hybrid strategy explained below.

Few studies have investigated how the bias introduced by the use of one *T. pallidum* genome reference could affect the results obtained [83,92,108]. For that purpose, those studies compared the results generated using different *T. pallidum* genomic references and concluded that the differences were not significant and did not affect the conclusions obtained. Nonetheless, the number of genomes included in these studies for each *T. pallidum* subspecies was not equal or was reduced to just one subspecies since the number of *T. pallidum* genomes available was much lower, especially for TPE and TEN. Therefore, further studies incorporating more genomes of each subspecies which represent adequately the variability of *T. pallidum*, are needed to ascertain the effect of the choice of one reference or another on the results obtained.

Interestingly, there are two studies using a hybrid strategy of *de novo* assembly and mapping [36,73]. First, they do *de novo* assembly, and subsequently, contigs longer than 200 bp obtained were mapped back to a TPA reference genome, in order to merge contigs and fill gaps with the reference genome used to generate a hybrid assembly. However, Lieberman *et al.* [36], used SS14 as the genomic reference for this hybrid strategy for all *T. pallidum* samples, instead of using the closest one for each one, which may still introduce a bias affecting the subsequent results [103].

All these studies and information underscore the need for further studies focusing on the implications of all methodological aspects on the best strategies for analyzing whole genome sequences of *T. pallidum* derived from HTS experiments.

#### **4. General genomic characteristics**

*T. pallidum* strains have a high level of genetic similarity, because there is only a 0.03% of divergence across all *T. pallidum* subspecies [2]. All treponemal subspecies form a compact genetic cluster of obligatory human and animal pathogens that naturally infect also nonhuman primates (TPE) [109] and domestic and wild-living lagomorphs [110,111].

The first *T. pallidum* genome sequence obtained in 1998 was Nichols (a TPA genome) and revealed a total of 1,041 genes [69]. With the significant increase in the number of whole genomes available, the number of predicted genes has varied in a range of a few dozens, due to the different prediction gene criteria employed, the increasing numbers of annotated genes in the public databases, and more nucleotide differences in the determined genomes.

In 2010, four TPE genomes were amplified for the first time in 133 overlapping amplicons [112]. The four genomes obtained corresponded to four strains isolated from human patients: SamoaD (isolated in Samoa in 1953), CDC-2 (isolated in Ghana in 1980), Gauthier (isolated in Nigeria in 1960), and the Fribourg-Blanc

strain isolated in 1966 from baboons (*Papio cynocephalus*) in West Africa. A few years later, to precisely define genetic differences between TPA and TPE, high-quality whole genome sequences of three previously sequenced TPE strains (SamoaD, CDC-2 and Gauthier) [113] plus the Fribourg-Blanc strain [114] were determined using next-generation sequencing techniques. It was possible to identify 1,039 genes for Gauthier and CDC2 strains, 1,036 for SamoaD, and 1,040 for Fribourg-Blanc. The first TEN genome was obtained in 2014 [86] from a human patient sample, which was isolated in 1950 in Bosnia, southern Europe; for which 1,039 different genes could be identified.

There is one study about the pan-genome (entire set of genes within all *T. pallidum* subspecies) of *T. pallidum* [115]. They estimated the pan-genome of all *T. pallidum* (2,112 genes), and the pan-genome of TPA (1,049 genes) and TPE (982 genes). They use Heap's law ( $n = k \cdot N^\gamma$ ) (number of genes for a given number of genomes) [116] to infer the alpha value ( $\alpha = 1 - \gamma$ ) and estimate if the pangenome is open or not, obtaining an  $\alpha$  of 0.9435. An  $\alpha < 1$  means an open pan-genome, and  $\alpha > 1$  a closed one. Then, according to the results, *T. pallidum* is still increasing the number of new gene families and this increment does not seem to be asymptotic regardless how many new genomes are added to the pangenome. However, the estimation of the level of openness of the *T. pallidum* genome, as well as the number of genes that constitute the pangenome of this bacterium, may be erroneous or biased by different causes: a) The genomes used are draft genomes (not completed) and have missing data that may bias orthology analysis [79,118]. Some stricter genome quality filtering criteria should have been applied to avoid this possible bias besides to improve the current genome annotations. b) Heap's law is not the most appropriate for pan-genome inference, because according to its formula it is mathematically impossible to obtain a value of  $\alpha > 1$  [119], so the inferred pan-genomes will always be open. All of this explains the results obtained in Jaiswal *et al.* [115] since, according to the



genomic characteristics of *T. pallidum*, it should not have an open pan-genome, and it is hard to believe that such a large number of genes (2,112 genes) make up its pan-genome. However, more studies on pan-genomes and in particular on the pan-genome of *T. pallidum* are needed to better account for these results.

Despite the slight variations in gene prediction numbers, which may be due to problems or differences in the gene annotation methodology used, *T. pallidum* genomes are among the smallest bacterial pathogens, especially in comparison to other extracellular pathogens [120–122]. According to Radolf *et al.* [3], *T. pallidum* have just a small number of genetic capacities, and their need for animal and human hosts (or their eukaryotic cells) to multiply appears to be correlated with the set of roughly a thousand genes which they own.

*T. pallidum* has 42 different paralogous gene families (PGF), with varying numbers of paralogs in each [69]. A total of 129 predicted genes (around 12% of the total of *T. pallidum* genes) can be assigned to the 42 paralogous gene families [123]. Interestingly, small bacterial genomes, and especially intracellular pathogens, have a modest number of paralogous genes [124,125]. However, despite the fact that there is an overall association between the number of paralogous genes discovered and total genome size, *T. pallidum* has a larger number of paralogous genes than expected for its small genome size. This may be due to *T. pallidum* being an extracellular bacteria and the significant role assigned to these paralogous genes, as extracellular pathogens have created methods for antigenic variation employing gene paralogs of surface proteins [3,126,127]. For all these reasons, and despite the difficulty for obtaining sequences of these genes, they are widely studied at epidemiological and genomic levels for their role in the pathogenesis and evolution of *T. pallidum* [73,108,128,129], but especially for their possible involvement in the development of an effective syphilis vaccine [130]. The most significant PGF genes in *T. pallidum* are PGF2 [131], which has 12 *tpr* genes (Table 4), and PGF34, with

six genes (Table 4) encoding for outer membrane proteins that are structurally similar to the FadL protein family of *Escherichia coli* [132].

Treponemes also lack plasmids, transposons, and phages [3]. More small-scale changes such as different subsets of gene paralogs, pseudogenes, and other genomic differences as indels and nucleotide variants are what cause the differences in the invasiveness of obligatory human and animal pathogenic treponemes (defined as the ability to infect different host tissues) and animal tropism [133].

**Table 4.** The main paralogous gene families (PGF) in the different *T. pallidum* subspecies.

Paralogous gene families (PGF)	Gene names
<b>PGF2</b>	<i>tp0009 (tprA), tp0011 (tprB), tp0117 (tprC), tp0131 (tprD), tp0313 (tprE), tp0316 (tprF), tp0317 (tprG), tp0610 (tprH), tp0620 (tprI), tp0621 (tprJ), tp0897 (tprK), tp1031 (tprL)</i>
<b>PGF34</b>	<i>tp0548, tp0856, tp0858, tp0859, tp0860, tp0865</i>

The genome sizes of *T. pallidum* strains range between 1,137,653 bp (TEN, strain BosniaA), [86] and 1,140,481 bp (TPE, strain Fribourg-Blanc) [114], representing a maximum known difference of 2,828 nucleotides and 0,25 % of the genome size. The treponemes' genome size appears to decrease in the following order TPE-TPA-TEN and this decrease is often accompanied by an increasing number of pseudogenes and lack of infectivity to humans [3,9]. Additionally, there are no known exceptions to the general gene synteny, which is consistent with the shared origin of these treponemal subspecies, with the highly adapted genomes to animal and human hosts, and with the slow evolution of the genome sequences.

#### 4.1 Intragenomic rearrangements in *T. pallidum*

The paralogous gene sequences identified in *T. pallidum* could be copied by gene conversion mechanisms into related paralogs, and this mechanism has been shown to operate in TPA and TPE strains (in TEN there are no sufficient available finished

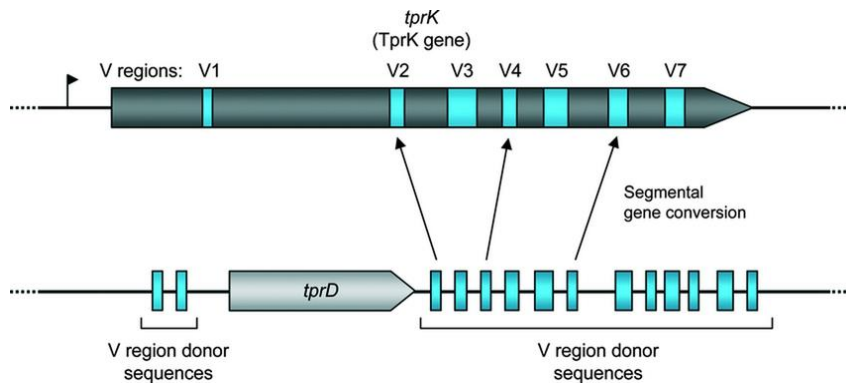
genomes to explore it). Moreover, the thorough investigation of several repetitive sequences (modules) shared by various *Treponema* strains revealed more evidence for potential intra-strain recombination events [74]. These particular gene sections were discovered to be modular structures through the investigation of 16 *T. pallidum* strains and *T. paraluisuniculi* (TPc), particularly of the *tp0136*, *tp0856*, and *tp0896* genes. Additional investigations uncovered regions with a modular genetic structure that are unique to one or more strains over others. These genes include *tp0126*, *tp0126b*, *tp0126c*, *tp0127b*, *tp0128*, *tp0130*, *tp0898*, and the *tprCDFI* (*tp0117*, *tp0131*, *tp0316*, and *tp0620*) family, indicating that this mechanism, which enables genetic diversification, is quite common in treponemal genomes. In addition, this study also identified additional genes including other *tpr* genes (*tprEGJK*) and several loci from two different gene paralogs families such as *tp0133*, *tp0134*, *tp0462*, *tp0548*, *tp0858*, *tp0859*, and *tp0865* genes, with direct or inverted repeats and the potential for genetic recombination. Some of those patterns were also identified in Grillová *et al.* [73] and in Pla-Díaz *et al.* [108].

rRNA genes are co-localized in rRNA (*rrn*) operons. In addition, *rrn* operons contain intergenic spacer regions (ISRs) and may also contain tRNA genes and regulatory regions. ISRs are more heterogeneous among bacterial species and strains because they are less conserved than rRNA genes [134,135]. The genomes of *T. pallidum*, both for strains causing human and animal treponematoses, contain two *rrn* operons, each encoding the 16S-23S-5S rRNA genes and ISRs. There are two different *rrn* spacer patterns (AUU/GCC and GCC/AUU) which are distributed randomly across TPA and TPE subspecies groups while in TEN there are no sufficient available finished genomes to assess the random character of *rrn* spacer patterns [134]. There is further evidence that the random distribution of *rrn* spacer patterns in *T. pallidum* is generated by reciprocal translocation of *rrn* operons mediated by a *recBCD*-like system found in the intergenic spacer regions (ISRs) by

Čejková *et al.* [134]. The identity of the 5S, 16S, and 23S rRNA gene sequences in both *rrn* operons, as determined by sequence analysis of all completed whole genomes [2,74,81,136] and by analysis of amplified 23S rDNA from 144 patients [39], is another intriguing characteristic of *rrn* operons.

### **4.2 Variability in the *tprK* gene**

Since the immune evasion and disease progression of *T. pallidum* are linked with sequence variability in the hypervariable outer membrane protein expressed by the *tprK* gene (*tp0897*), it is thought that this gene plays a crucial role in the pathogenesis of these bacteria. Several investigations found significant heterogeneity at both the inter- and intra-strain levels [79,137–139]. The sequence diversity of this gene is restricted to seven variable (V) regions (Figure 3), namely V1 to V7, which are separated by conserved sequences [79,140–142]. Each variable region is predicted to form a loop exposed at the host-pathogen interface, according to the putative *tprK* beta-barrel structure [143]. Variation in these regions is produced through nonreciprocal segmental gene conversion, which involves joining sections from donor sites flanking the *tprD* (*tp0131*) gene [144,145]. However, the overall number of distinct sequences that can be produced in *T. pallidum* has not been determined yet [146], despite the identification of 47 probable donor locations [108,143]. The remarkable variation and redundancy observed in the *tprK* gene, both within and between *T. pallidum* subspecies, point to continuous parallel adaptive diversification during human infection.



**Figure 3.** Structure of the *T. pallidum* hypervariable *tprK* gene and the mechanism of gene conversion through antigenic variation of the variable (V) regions of *TprK* was generated. Extracted from [142].

## 5. Evolutionary mechanisms

Mutation and recombination drive the evolution of most pathogens by generating genetic variation upon which natural selection operates [147,148]. This genetic variation enables adaptation, including mechanisms to evade the host immune system, and drug resistance, among others. Hence, investigating the presence and consequences of these forces acting as a source for change and adaptation in *T. pallidum* is very important for understanding the evolutionary dynamics of this bacterium.

### 5.1 Point mutation

Several studies have attempted to study the evolutionary rates of *T. pallidum* using Bayesian inference of molecular evolution [149] (Table 5). The evolutionary rates obtained are congruent in most of them, especially when the number of *T. pallidum* genomes incorporated in these studies has increased.

## Introduction

**Table 5.** Different molecular clock rate estimates by the BEAST program [149] obtained in several published studies. The table shows the number of genomes included in each dataset, the molecular clock model employed and the *T. pallidum* subspecies studied.

<i>T. pallidum</i> subspecies	Number of genomes used	BEAST molecular clock model	Molecular clock rate	References
ALL	39	Relaxed clock model	$6.6 \times 10^{-7}$ s/s/y (95% highest posterior density (HPD) of $1.86 \times 10^{-4}$ to $5.73 \times 10^{-4}$ )	[78]
TPA	109	Strict Constant model	$1.78 \times 10^{-7}$ (95% Highest Posterior Density (HPD) $1.15 \times 10^{-7}$ – $2.44 \times 10^{-7}$ ), or 0.20 sites/genome/year	[107]
TPA	138	Strict Clock model	$1.28 \times 10^{-7}$ (95% highest posterior density $1.07 \times 10^{-7}$ – $1.48 \times 10^{-7}$ )	[83]
ALL	233	Uncorrelated relaxed clock	$3.02 \times 10^{-7}$ s/s/y (median $2.99 \times 10^{-7}$ , 95% HPD $2.18 \times 10^{-7}$ – $3.89 \times 10^{-7}$ )	[36]
TPA	948	Combination of two: strict and relaxed lognormal	$2.14 \times 10^{-7}$ s/s/y (95% highest posterior density [HPD] $1.78 \times 10^{-7}$ to $2.56 \times 10^{-7}$ )	[71]

While most studies are focused on the estimation of mutation rates in TPA, there are no studies about the specific evolutionary rates of TEN [2] and just one study used two TPE samples to estimate  $4.1 \times 10^{-10}$  substitutions per site per generation or lower [150]. Nevertheless, the evolutionary rate ( $\mu$ ) per site per year obtained in Strouhal *et al.* [150], was not obtained by BEAST [149] but using the following formula:  $\mu = n/t \cdot gl$ , where  $n$  means the number of nucleotide differences,  $t$  represents the time since the samples have diverged measured in years, and  $gl$  represents the number of nucleotide positions (genome length), so it is not possible to compare this estimate with the obtained by BEAST [149] in other studies. The limited number of samples available (mainly TPA) and especially the relative lack of older isolates makes the model predictions difficult, even for TPA. Further studies including representative datasets of the three subspecies are needed to clarify a more precise evolutionary rate by BEAST [149].

The recent development of a system for long-term *in vitro* culture of TPA has introduced the possibility of detailed genetic analysis of this bacterium [65–67,105]. In Edmonson *et al.* [105], the *in vitro* culture system was used to isolate and characterize clonal populations of the Nichols strain (TPA). The sequences obtained revealed sequence heterogeneity, despite the strains having indistinguishable morphology and motility, highly similar *in vitro* multiplication rates, and comparable infectivity in the rabbit model. Most of the genes affected by these mutations are *omp* genes with important functions for host defense, virulence and immune system response [108]. The genetic heterogeneity at these locations were not detected in > 280 TPA genomes available in public databases, suggesting their origin related to the culture system and highlighting the importance of direct sequencing of clinical samples to avoid this bias. However, further investigations into the impact of genetic heterogeneity and mutation accumulation in cultured strains are necessary to establish reliable conclusions. This could involve culturing a larger number of strains and examining how these changes affect gene expression and function using genetic engineering techniques [68]. It is also crucial to assess the evolutionary rates of these mutations and explore other relevant factors. While maintaining stable cultures of six Nichols isolates over an extended period is indeed commendable, the findings cannot be extrapolated to all cultivated strains. Therefore, it is imperative to study a broader range of strains to determine the true influence of genomic heterogeneity on the phylogenetic and evolutionary inferences of this bacterium.

Additionally, there are two well described macrolide resistance-encoding point mutations (A2058G, A2059G) in the 23S rRNA gene in *T. pallidum* [151,152]. These point mutations are under natural selection pressures and appeared on several occasions independently but with a relatively low frequency because of the administration of macrolide antibiotics in single or repeated doses to thousands of patients [19,34,153]. All of this is described in more detail below in section 5.4 of the Introduction regarding natural selection.

### 5.2 Genetic recombination

In comparison to vertically inherited point mutations, which serve as the signal of shared common ancestry, considerable stretches of sequence transferred via diverse recombination mechanisms can significantly affect genome-wide measurements of sequence similarity [154]. The earliest attempts to locate genomic exchanges within a pathogen used techniques that searched for indications of unexpected similarity between divergent sequences, which is thought to indicate that a sequence was transferred from a donor into a recipient or that both independently received it from a common donor [155]. Such genomic interchanges lead to homoplasies, which can make it challenging to determine reliable phylogenetic trees and evolutionary inferences.

Some of the first tools to detect recombination, such as SimPlot [156], RDP [157] and TOPALi [158], look for changes in patterns of genetic diversity. Others, like Reticulate [159], calculate compatibility matrices to detect recombination, and PIST [160] explores excessive convergent evolution. LARD [161], PLATO [162] and BOOTSCAN [163] search for incongruent phylogenetic trees. However, the capability of these methods was limited to the knowledge of the donor, recipient and recombinant sequences, which is not frequently possible in bacteria where such sequences may belong to non-sampled donors [164]. For this reason, new methods to detect recombination were developed such as ClonalFrame [165], which searches closely-related isolates for evidence of exchanges of divergent sequence from a distantly related source. The algorithm of this program builds a tree based on the point mutations outside the recombinant portion candidates while concurrently using a Bayesian Monte Carlo Markov chain (MCMC) technique to discover recombination as locations with a noticeably enhanced density of polymorphisms. This tool was later updated and renamed as ClonalFrameML to be employed to detect recombination based on the ClonalFrame model and by maximum likelihood inference in large datasets [166].



Another available tool is Gubbins [164], a rapid method to detect sequence imports through recombination in bacterial populations, which is also successfully used for large genomic datasets as ClonalFrameML. Gubbins detects, by statistical analysis, regions with elevated densities of substitutions originated by a recombination event and builds a maximum phylogenetic tree after excluding them.

These two programs are widely used for the detection of recombination based on genome-wide data [167–170]. In fact, Gubbins has been the most frequently employed program to detect recombinant regions in *T. pallidum* genomes, which are subsequently removed from the whole genome alignment used to build reliable and robust phylogenies [34,36,70,83,107]. However, other studies, as some detailed below, detect the recombination patterns and/or events manually by direct observation of the sequences. Interestingly, two of these studies using direct sequence observation, applied first phylogenetic incongruence to detect the genes whose sequence would be interesting to review in detail, comparing trees of individual gene alignments versus the tree obtained with the alignment of whole genomes of interest [73,78]. The results obtained in this study highlight the necessity of surpassing the manual evaluation conducted and transitioning towards automating this methodology for its widespread application.

Although no mechanism for recombination in *T. pallidum* has been discovered so far, numerous studies have documented patterns of variation that are indicative of recombination and that provide support for the role of recombination in the evolution of *T. pallidum*. For instance, it was discovered that several loci, most of which code for outer membrane genes (*tp0117/tp0131* (*tprC/tprD*), *tp0119*, *tp0317* (*tprG*), *tp0621* (*tprJ*), *tp0856*, and *tp0858*), have a gene region that is identical (or nearly identical) to the gene region contained in other genes, primarily corresponding to intragenomic transfers [171–176].

Although the presence of recombination in *T. pallidum* is already accepted, more detailed studies on the evolutionary processes underlying this bacterium are needed to improve its phylogenetic inference and to better understand how it has evolved and adapted over time, as well as how its different subspecies have diverged and expanded globally.

### 5.4 Natural selection

Natural selection plays an important role for genetic diversity, especially maintaining or driving some alleles to fixation or extinction [177]. It is well studied in other monomorphic pathogens such as *Mycobacterium tuberculosis* [178,179], *Yersinia pestis* [180,181], among others. However, there are few studies available focused on studying the action of natural selection in the evolution of *T. pallidum* and in deciphering which specific genes are influenced under this evolutionary mechanism [129,182,183], and are more focused in TPA and/or employed small datasets. Some of the genes identified in the aforementioned studies are paralogous genes and/or encode putative or *bona fide* outer membrane proteins (OMP). These genes have been suggested as potentially involved in virulence, with an important role in the defense of the pathogen against the host and the evasion of the immune system [173]. These findings underscore the importance of conducting further in-depth studies on recombination and the role of natural selection in this bacterium.

Moreover, the influence of natural selection as a driving evolutionary force with a significant impact in the spread of *T. pallidum*, was exemplified in Beale *et al.* [107]. They demonstrated how the most recent spread of the TPA lineage may be connected to significant changes in the epidemiology of syphilis, including immune evasion or antibiotic selection pressures due to the macrolide resistance present in the SS14 lineage, the most recent clade of syphilis. Moreover, according to another genomic epidemiology investigation by Beale *et al.* [34] about yaws re-emergence and bacterial drug resistance selection after mass administration of azithromycin (MDA), it was concluded that yaws re-emergence after MDA was driven by

multiple sources, including natural selection pressure in the evolution of azithromycin resistance. Even so, it is true that the population size of this project was not enough to obtain more significant results and to affirm robustly the relationship of natural selection with the emergence and maintenance of the macrolide resistance in TPE. Considering also the findings from recent epidemiological studies [36,71] about the detection of additional macrolide resistance genotypes in newly collected samples (with a specific focus on TPA), it becomes clear that there is a strong probability of further independent evolution of azithromycin resistance. This evolutionary process is actively driven by natural selection, thereby carrying significant implications, particularly for the ongoing efforts aimed at eradicating yaws. Furthermore, it poses noteworthy challenges to the broader field of *T. pallidum* epidemiology. Worryingly, applying azithromycin-based mass drug administration to populations infected with *T. pallidum* could promote resistance in this bacterium [19]. All these results highlight the importance of studying in more detail natural selection in *T. pallidum* together with other evolutionary processes that may shed light on the evolutionary history of this bacterium.

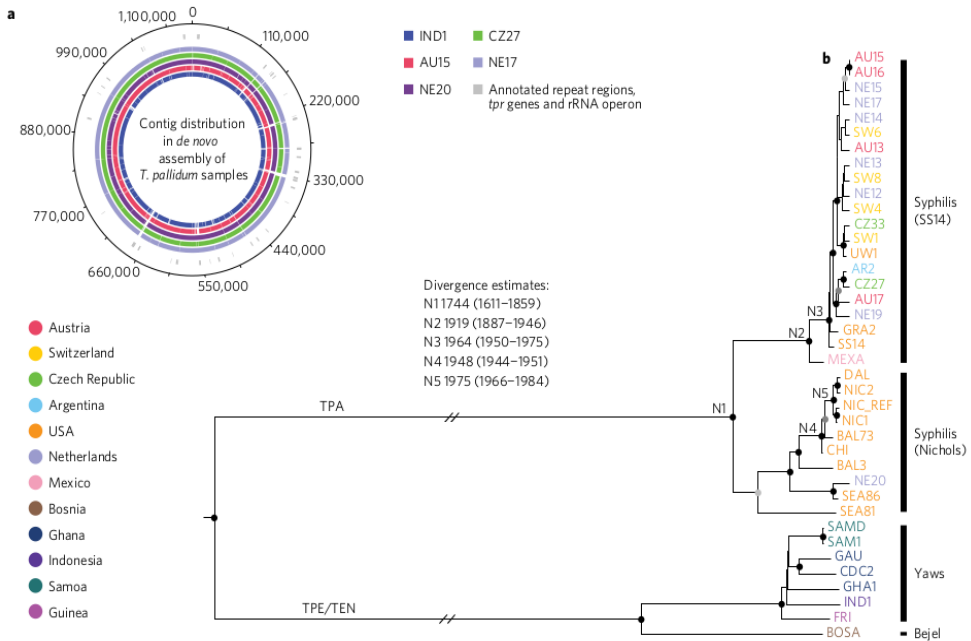
## **6. Genomic epidemiological studies and phylogenetic reconstruction of modern *T. pallidum* genomes**

With the improved application of HTS techniques, more than 1,400 of *T. pallidum* genomes have been generated and are available in public databases by the end of 2022 [2,21,36,70–77], allowing not only the classification of new *T. pallidum* strains, but also their genomic and epidemiological study. However, until 2016, the number of *T. pallidum* genomes accessible was less than 20 [78], despite the fact that new HTS technologies were already well established and the number of available genomes of other bacteria was more than 20 times larger [184]. In addition, most of these genomes belonged to TPA, and there was only one genome

from TEN [86] and four from TPE [112], which made the study of these two subspecies even more difficult.

Pinto *et al.* [79] and Arora *et al.* [78] were pioneers in the application of a new culture-independent targeted whole-genome sequencing (WGS) strategy to successfully recover novel TPA genomes. In their study, Pinto *et al.* [114] sequenced 25 genomes from Portugal, providing the first evidence of within-patient genetic variation in TPA. They also explored the variability and redundancy of the seven variable regions of the *tprK* gene, both intra- and interpatient, making significant contributions to the understanding of these genomic features of *T. pallidum*. Arora *et al.* [78] reconstructed the poorly studied evolutionary relationships among the three different subspecies of *T. pallidum* through the study of a dataset with 39 whole genomes of different strains (Figure 4). They analyzed the diversification from a common ancestor of the SS14 lineage, dated it in the mid-twentieth century, subsequent to the discovery of antibiotics, and named this recently spread resistance cluster as SS14- $\Omega$  (Figure 4). Moreover, this work explored the presence of recombination and its effects on the phylogenetic inference of this bacterium, although recombination in *T. pallidum* was a major debate at the time.

## Introduction



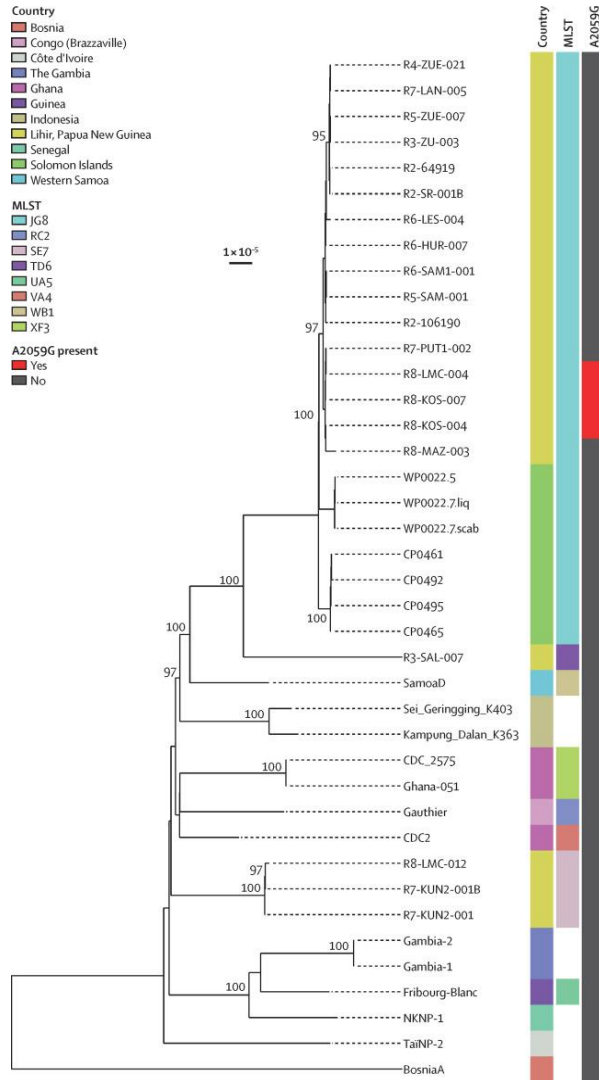
**Figure 4.** Bayesian inference tree for the 39 genomes, showing black circles for nodes with posterior probability (PP) greater than 96%, dark grey circles for nodes with PP between 91 and 95%, and light gray circles for nodes with PP between 80 and 90%. For significant well-supported TPA nodes, divergence date estimates (mean and 95% greatest posterior density) are provided in the legend. The countries where samples were taken are indicated in different colors. Extracted and modified from [78].

The results of these two investigations opened the door to new genomic research on the spread and diversification of *T. pallidum*, especially thanks to the new treponemal DNA enrichment technique they both employed. Beale’s *et al.* [107] study about the global epidemiology and genomics of TPA corroborated the preliminary results of Arora *et al.* [78], demonstrating the independent establishment of the SS14-Ω sublineage as a macrolide resistance clade from TPA.

TPE was so far a very understudied subspecies due to the scarcity of available genomes but it was well characterized in Beale *et al.* [34]. In this study, 20 new TPE genomes were obtained and contextualized with other 20 strains from previous studies revealing a greater diversity of TPE than expected (Figure 5). This allowed us to learn more about the multiple epidemiological sources of yaws re-emergence

## Introduction

as well as its susceptibility and detection of macrolide resistance cases after mass administration of azithromycin (MDA).



**Figure 5.** Extracted from [34]. Maximum likelihood of whole genome sequencing phylogeny of 40 TPE genomes employed in this study, including the 20 new genomes recovered from Lihir, Papua New Guinea.

At the end of 2022, more than 1400 *T. pallidum* genomes were available, allowing its epidemiological study from a global perspective through the analysis of genomes from different countries around the world [2,21,36,70–77]. New genomes were generated from samples obtained in previously unexplored countries such as China, Japan, Africa, Asia, Spain, Oceania, Cuba, Indonesia, Madagascar, Senegal, Australia, among others. These studies made it possible to determine the presence of the Nichols sublineage (TPA) around the world, which is still circulating and expanding [36,71,83] and, for the first time, some Nichols strains showed also macrolide resistance, which was thought to be exclusive, within TPA, to the SS14-Omega cluster [83]. Nevertheless, there still seems to be a clear overall dominance of the emerging SS14- $\Omega$  cluster around the world [36,71,83]. In a notable development, Lieberman *et al.* [36] successfully conducted direct sequencing on two previously unexamined samples from 1998 and 2002. Surprisingly, these samples were found to be part of the same clade as the MexicoA strain, which was originally isolated in 1953 in Mexico [133]. Previously, the MexicoA strain was believed to be a result of multiple passages in rabbits, and the only available sample representing that particular clade.

Six new TEN genomes have also been very recently added to the only two genomes available for this clade, 2 from Cuba [2] and 4 from Japan [36]. The latter are the first genomes available of this subspecies to possess resistance to macrolides. Although there are still very few TEN genomes available compared to the other two *T. pallidum* subspecies, it has been possible to clarify the phylogenetic relationships of this subspecies and to know that TEN has more variation than expected (as described above with TPE).

Many of these studies have tried to establish different classification systems in sublineages, clusters or hierarchical Bayesian analysis of population structure (BAPS) groups, which makes it very difficult to compare the results of the different studies with each other, since the classification systems are different [36,71,83,107].

Other bacteria use STs obtained from MLST or whole-genome MLST (wgMLST), and/or core genome MLST (cgMLST) schemes as a more specific classification criterion, which has great epidemiological utility [45,185–187]. However, although it has been used in some of these epidemiological studies [34,36], there is no unified MLST system available for the 3 subspecies of *T. pallidum* that can be used and compared for this type of studies, and much less a wgMLST or cgMLST schemes, which once again highlights the need to design a good typing scheme for *T. pallidum*.

## **7. Paleogenomics of *T. pallidum***

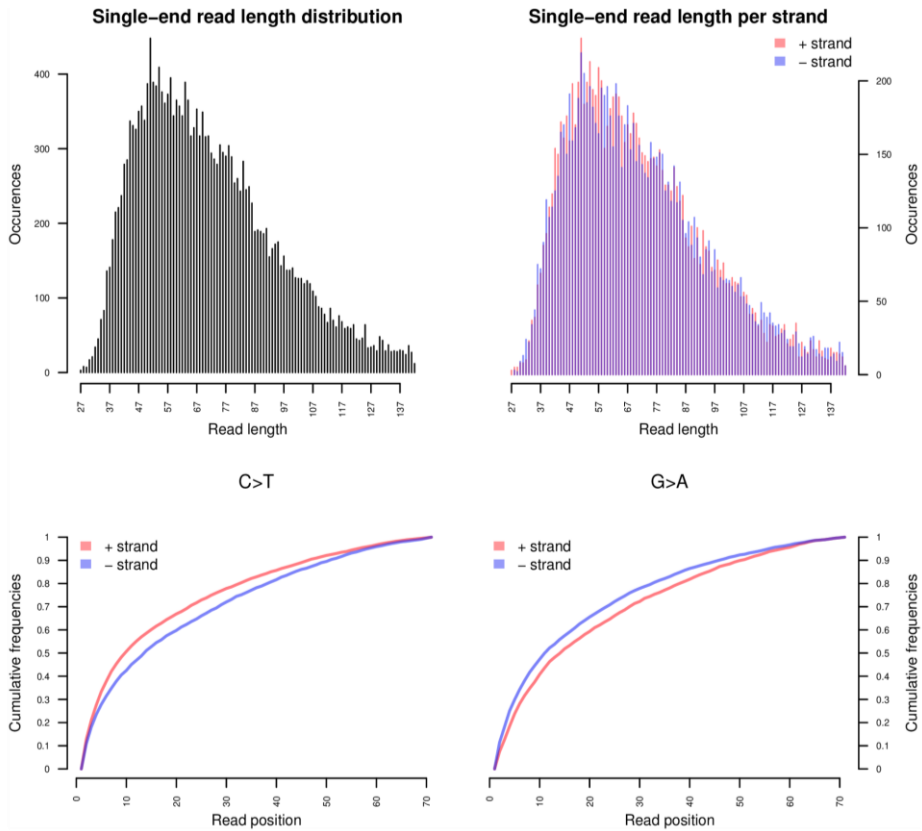
### **7.1 Ancient DNA (aDNA) and paleogenomics**

The field of palaeogenomics examines and analyzes DNA derived from ancient remains, including bone deposits and museum holdings. This kind of DNA is known as ancient DNA (aDNA), which is a DNA molecule that has undergone post-mortem degradation and is taken out of ancient remnants. Although it might be difficult to define what constitutes an ancient sample, these are typically between a few decades and thousands of years old [188].

There are several recognizable aDNA traits due to the DNA degradation that starts right after death. The primary causes of this breakdown are DNAses and the microbiome of living beings. Later, the activity of environmental microbes and chemical reactions such as condensation, oxidation, hydrolysis, and alkylation would continue to degrade it [189].

In fact, it is believed that the primary cause of aDNA's severe fragmentation is hydrolytic damage to the phosphodiester chain or damage created following depurination [95]. Because of this fragmentation, DNA chains between 30 and 70 bp (Figure 6) are one of aDNA's distinguishing characteristics [189].

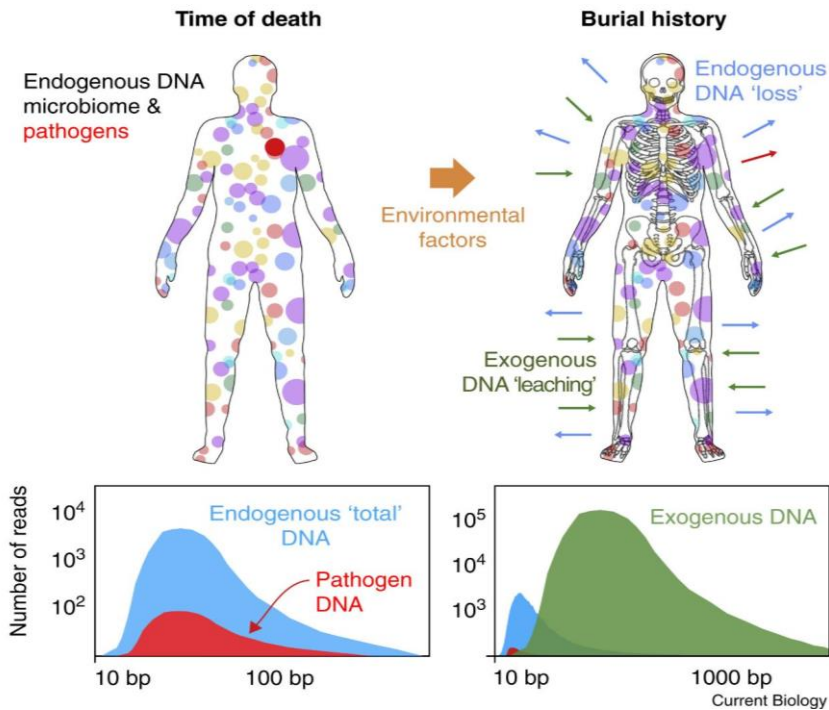




**Figure 6.** The upper two plots are histograms of the read lengths. The lower two plots are the empirical cumulative frequency of C->T and G->A misincorporations, normalized by the first 70 positions. Extracted from [190].

When the double strand of DNA is fragmented, single strands, or pieces of the strand without the complementary strand, typically remain at the ends. Single chains' nitrogenous bases are more vulnerable to harm. The most frequent and distinctive process of aDNA is hydrolytic deamination, which results in the conversion of cytosines (and to a lesser extent, guanines, and amines) into uracils. Given their resemblance to uracil, cytosines are more sensitive to hydrolytic deamination than other nucleotide bases. These genomic damages are corrected in living cells, but once a cell is dead, they start to accumulate. Uracils are switched out for thymines during PCR amplification, resulting in substitutions of the type C-T and G-A (Figure 6), which is also considered as another characteristic of authentic aDNA [191].

In aDNA studies it is very important to differentiate endogenous from exogenous DNA, although this is not a simple task. We refer as endogenous DNA to the DNA of the samples being studied, which includes the microbiome and DNA of the deceased creature (Figure 7). On the other hand, exogenous DNA would be all the environmental DNA from species outside the remains. Endogenous DNA tends to disappear in the environment during diagenesis while exogenous DNA colonizes the leftovers. Moreover, exogenous DNA can infiltrate and become contaminant even during the extraction of the remains and the subsequent laboratory processing (60–90% of endogenous DNA is thought to be lost during extraction, gene library construction, and purification) [192,193]. In the end, on average, endogenous DNA makes up only 0.5% of all sequenced DNA and about 10% to 15% in the study of endogenous bacterial DNA [189].



**Figure 7.** Environmental conditions contribute to DNA degradation and the remaining endogenous DNA is shortened but the majority is destroyed. Environmental (exogenous) DNA also overwhelms the original signal, with fragment-length distributions that both overlap and go beyond the sample's ancient DNA. Extracted from [189].

The success of aDNA investigations is also significantly influenced by the source of the DNA [99]. Although it has been feasible to extract from coprolites or hair, mummified bones and tissues are the most typical samples. Since the dentin of the most superficial layer of teeth is an excellent insulator from external conditions, many specialists claim that teeth are the remains in which more endogenous DNA may be detected. According to recent research, even the cementum layer of a tooth can provide more endogenous aDNA than the layer nearest to the dentin [95]. Another bone containing a significant amount of endogenous DNA is the temporal bear petrous plate, albeit its application to the metagenomics of ancient samples research has been found to be quite restricted [194].

### **7.2 Different hypotheses about the origin and spread of the treponematoses**

Historically, syphilis has been the cause of many different epidemics, beginning with the violent initial outbreak in Europe documented at the end of the 15th century. Today, it is largely acknowledged that this outbreak marked the start of the first known syphilis epidemic (Figure 8). The plague struck Charles VIII's army while they were advancing on Naples in 1495 [195]. Shortly after the battle, the army dispersed, and the mercenary-heavy soldiers went home and spread the plague over Europe [196,197].



**Figure 8.** Two of the earliest known representations of syphilis patients. The left picture represents a mercenary whose skin presents multiple chancres, woodcut by Albrecht Dürer, 1496 [198]. In the right picture, a physician examines the urine of a patient in the first European syphilis epidemic (Vienna in 1498) [199].

Ever since, the origins of this bacterium and its connection to the pathogens responsible for the other treponemal diseases have been the subject of debate, with three different hypotheses attempting to explain it: the Columbian hypothesis (also called post-Columbian), the pre-Columbian hypothesis and the Unitarian hypothesis. The Columbian hypothesis holds that syphilis originated in America and was introduced from there to Europe in 1493 [200]. The pre-Columbian hypothesis contends that syphilis or a closely related illness was previously prevalent in Europe, potentially as a result of prehistoric spread of the disease through African and Asian routes [201]. The third hypothesis, known as Hudson's Unitary hypothesis, which was put forth in 1963, claims that syphilis has plagued various human populations since ancient times [202–204]. Moreover, Hudson's theory affirms that venereal and non-venereal treponematoses are actually caused by the same pathogen, and the various illnesses are a result of various environmental factors [205]. Therefore, the usual clinical signs of yaws, especially in youngsters, would be caused by treponematoses that developed in locations with humid and warm climates during

the Paleolithic era; and the clinical signs of endemic syphilis are caused by the infection's ability to spread to drier regions. Afterwards, more people were able to attain sexual maturity without having previously been exposed to treponemes as a result of advancements in personal hygiene, which led to the development of a novel method of sexual transmission linked to venereal syphilis.

There was an additional theory related to the Unitary hypothesis proposed by Hackett *et al.* [206] which suggests that pinta was the first treponematosis from which yaws, bejel and, finally, syphilis originated by mutation. However, the Unitary hypothesis was progressively abandoned as advances in genetics and genomics studies were introduced and the new research revealed that *T. pallidum* subspecies are genetically different and had evolved in different ways [4,16,78]. Notably, cases of syphilis are mentioned in Medieval literature; however, because of its similarity to other diseases, it is frequently difficult to diagnose and is erroneously referred to as "venereal" or "hereditary" leprosy [207]. Although reports of pre-Columbian human remains bearing telltale signs of treponematoses have been made [208,209], they have not been confirmed by DNA yet [208–210]. Nevertheless, the geographical origin and timing of the genesis of TPA and TPE are still unknown; hence, it is reasonable to question their geographic spread and clinical symptoms as the primary means of categorization [78,211,212].

While the more general diversification of *T. pallidum* into subspecies is thought to have occurred in prehistoric times [203,213], a recent genomic analysis on present lineages of *T. pallidum* suggested a common ancestor of all current TPA strains in a median range between 1500-1900s, approximately [71,78,83,107]. However, because a sizable portion of the previous species diversity may have been lost over time [214], estimates of evolutionary rates derived from present-day genome isolates may be biased, and more studies including ancient genomes of *T. pallidum* are needed to shed light on the unresolved divergence periods, origins, and evolutionary histories of the corresponding subspecies [215–217].

### 7.3 Ancient *T. pallidum* genomes

In 1983, Russell Higuchi and colleagues succeeded in obtaining and sequencing DNA from bone fragments of the extinct quagga [218]. This is regarded as the initial result that gave rise to the paleogenomics discipline, despite the fact that earlier research with seeds had demonstrated that DNA could be maintained for hundreds of years [218,219]. Only a year later, a 3,400 base pair (bp) DNA fragment was extracted from dried mummy tissue by Pääbo *et al.* [220]. These discoveries, as well as others until the 2000s, were made feasible by the development and application of the PCR technology and Sanger sequencing.

But it was not until 1997, when Krings *et al.* [221] succeeded in sequencing a variable piece of a Neanderthal's mtDNA from skeletal bones, to shed light on the link of the humans with their closest ancestors. Nevertheless, ancient samples had substantial quantities of current DNA contamination, which tainted the conclusions because of the novelty and limits of these procedures. As a result, an attempt was made to develop a standard methodology of best practices [222] at this time in an effort to decrease contamination and increase the validity of the results.

Paleogenomics saw further advancements in the 2000s thanks to the revolution brought about by high-throughput sequencing (HTS) technologies. In fact, one of the major achievements in the field, the sequencing of the Neanderthal genome [223], was intimately tied to the development of these techniques. Although many discoveries have been achieved thus far, much more has to be learned.

Thanks to all these technological advances, it was possible to make great progress in the study of ancient bacterial pathogens. Originally, there were several reports of molecular detection of aDNA for *Mycobacterium tuberculosis* [224–228], *Mycobacterium leprae* [229–232], *Plasmodium* [233,234], *Yersinia pestis* [235,236] and *T. pallidum* [237], but some years later it was possible even to obtain whole genomes of different bacterial pathogens. By 2022, 151 ancient genomes

from different pathogens have been achieved, being *Yersinia pestis* the one for which the most genomes are available (Table 6).

**Table 6.** The total number of ancient genomes for each pathogen, acquired through manual search in 2022.

Pathogen	Number of genomes available	References
<i>Mycobacterium tuberculosis</i>	17	[217,238,239]
<i>Mycobacterium leprae</i>	46	[188,216,239–242]
<i>Yersinia pestis</i>	71	[188,243–249]
<i>Vibrio cholerae</i>	1	[250]
<i>Salmonella enterica</i>	8	[93,244,251,252]
<i>T. pallidum</i>	8	[92,98,253,254].

As the different infections caused by *T. pallidum* can produce bone lesions in the later stages of the three diseases, it facilitates the screening of bones for subsequent use in obtaining aDNA from this bacterium. However, the low degree of success of studies attempting to obtain *T. pallidum* aDNA caused great pessimism in the scientific community, and some studies even claimed its inaccessibility, such as a study that detected low levels of *T. pallidum in vivo* in infected rabbits under controlled laboratory conditions [255] or the negative result in the extraction of aDNA by PCR from bones with wounds compatible with syphilis [256]. This is mostly caused by the complete absence of an outer cell membrane in *T. pallidum* [257]. This characteristic, in combination with the low pathogenic burden of *Treponema* in late stages of infection, made it difficult to identify the organism in ancient remains [237,255,257], until the recent use of HTS techniques.

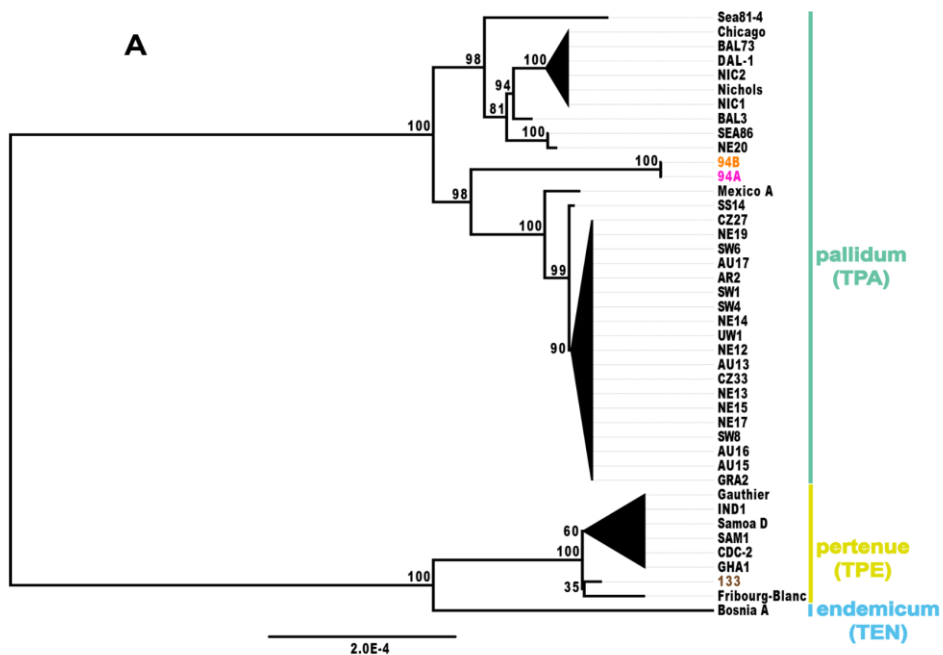
Due to all of these obstacles, it was not until 2018 when Schuenemann *et al.* [98] successfully sequenced the first three ancient *T. pallidum* genomes, two from TPA (94A and 94B) and one from TPE (133). The remains were from 19th century Mexico (Figure 9), which could not settle the question of the origins of syphilis but did demonstrate how *T. pallidum* ancient genomes might be obtained.



**Figure 9.** Examples for bone lesions for the three positive individuals for *T. pallidum* from which the aDNA extraction were done. Figure and information extracted from [98].

Additionally, Schuenemann *et al.* [98] emphasized how the many subspecies have coexisted in the same time and space, raising the possibility of events of recombination between them and highlighting the significance of molecular diagnosis in order to distinguish between the different *T. pallidum* subspecies (Figure 10).

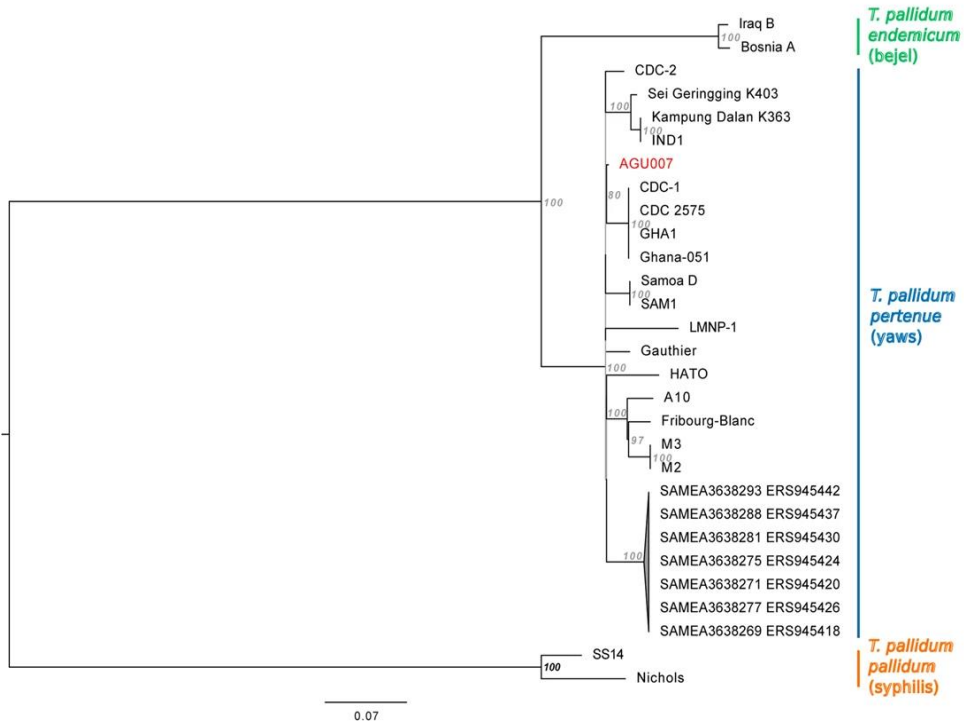




**Figure 10.** Maximum Likelihood tree for 39 contemporary strains and three ancient genomes obtained with bootstrap support. The new ancient genomes 94A (magenta) and 94B (orange) are basal to all SS14 clades (TPA), while strain 133 (brown) is placed with other TPE strains. The scale displays the average number of substitutions made at each location as calculated by the GTR+GAMMA. Figure and information extracted from [98].

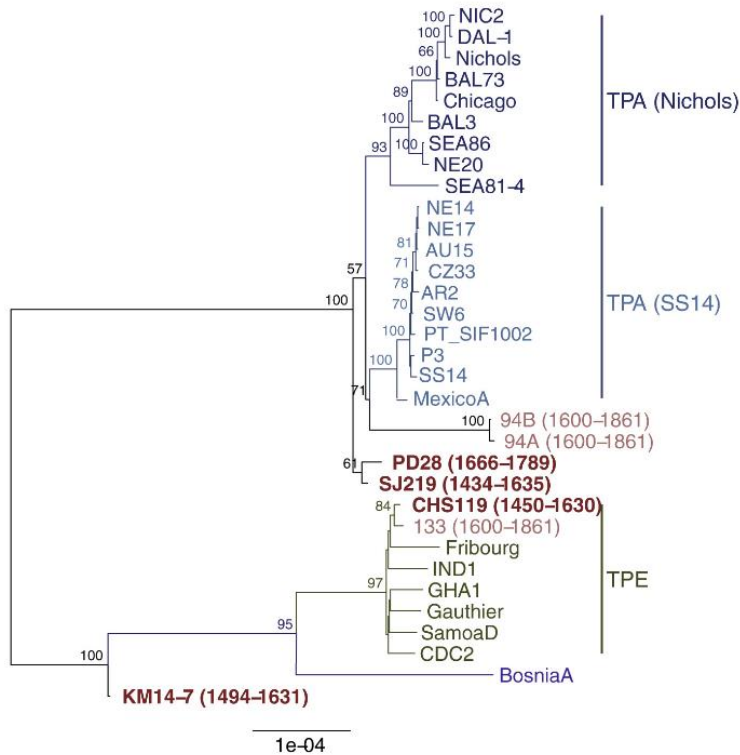
Two years later, two studies revolutionized the debate: the reconstruction of a TPE genome (AGU007) from the remains of a person from 15th century Lithuania by Giffin *et al.* [253] (Figure 11); and the reconstruction of four *T. pallidum* genomes dated from the 15th-17th centuries by Majander *et al.* [92]: two TPA genomes (PD28, SJ119) in Finland and Estonia, one TPE (CHS119) in Finland besides to a new possible TPA lineage (KM14-7) in the Netherlands (Figure 12).

a



**Figure 11.** Maximum Likelihood phylogenetic tree of 27 *T. pallidum* genomes, which was built using 1000 bootstrap replications after 13 homoplastic places and 54 recombination regions were removed. All sites with unclear bases were deleted. The tree was built from 929,012 bp, 1223 of which were polymorphic, and SNP calling was based on a 5-fold coverage. Branches that have less than 80 percent statistical support are shown in gray. Given a GTR + GAMMA substitution model, the scale indicates the mean substitutions per site. The AGU007 genome is highlighted in red. Figure and information extracted from [253].

## Introduction



**Figure 12.** Maximum-likelihood phylogenetic tree of *T. pallidum* strains. The tree is based on an alignment of 1,631 SNPs after exclusion of recombining and hypervariable genes as well as genomic positions with >25% missing data. The ancient genomes from this study are marked in red, whereas the ones from previous studies are marked in pink. Figure and information extracted from [92].

Regardless the opposition faced by the Columbus hypothesis for many years, the number of genomes that may date from before Columbus' journey has recently been increased, despite the dating range of the genomes being is very broad, by the discovery of syphilis-compatible bone wounds in Europe prior to European contact with America. This may be evidence that *T. pallidum* was present in Europe before it spread to other parts of the world. However, more studies are needed to confirm this assumption because of the wide range in the dating analysis. The discovery of TPE genomes in Europe [92,253] also disproved the conventional wisdom that yaws, and potentially also bejel, had always been confined to their current range as a result of environmental factors. Prior to genomic research, some theorized that *T.*

*pallidum* infections varied in phenotype because of variations in the infection's mode of transmission, environment, or patient phenotype. This was demoted, but these findings might offer justification for bringing it back or, at the very least, for revising our ideas about the distinctions and similarities between each [92].

Interestingly to all this debate about *T. pallidum* origins, Giffin *et al.* [253] proposed a new hypothesis: *T. pallidum* would have traveled to Europe via the slave trade from sub-Saharan Africa [253]. The fact that the highest genetic diversity of *T. pallidum* is found in Africa and that there were several sub-Saharan slaves residing in Europe as early as the 15th century would support this. They even went a step farther and speculated that TPE might have played a significant role in the outbreak. The decline in occurrences over the ensuing decades would correspond to the pattern of generation of herd immunity against TPE, which is prevalent in areas where it still affects today, they believe, because it was difficult to distinguish between the symptoms of yaws and syphilis.

However, despite the interesting collection of all the ancient genomes of *T. pallidum* described above in terms of knowing a little more about the history of this bacterium, its origin has not yet been elucidated with certainty. Likewise, in response to Majander *et al.* [92] and its contribution to the debate, Beale and Lukehart *et al.* [258] suggested that it could be wise to give priority to research on the evolution of *T. pallidum* in light just focus on the identification of the origin of the 16th century epidemic. They claimed that more genomes on persistent and endemic diseases were needed, and that a clearer understanding of the relationships between the various subspecies and infections they cause would help us better understand their evolutionary histories

— **OBJECTIVES** —



## Objectives

This PhD thesis is aimed at getting a deeper insight into the evolution and genomics of *T. pallidum* through the study of ancient and modern genomes of these bacteria and its applicability to the epidemiology of treponematosi. The specific objectives are the following:

- **Study the action of natural selection and recombination in the *T. pallidum* genome.** The greater amount of genomic information available will allow us to study the impact of different processes in the evolution of this bacterium. We will focus on recombination, because of its capacity to generate new genetic combinations almost instantaneously, and on natural selection, because it will inform us about the genes and proteins that have been subject to selection both adaptively and purifying in the different lineages.
- **Infer patterns of evolution and divergence through the study of ancient genomes of *T. pallidum*, in an attempt to elucidate the origin of this bacterium.** The discovery of ancient treponemal genomes provide an excellent opportunity to incorporate thorough data on evolutionary relationships and diversification into the present understanding of treponemal subspecies. Historical strains offer phylogenetic inference with depth and robustness that are not otherwise available. Moreover, no reliably pre-contact genetic evidence of any treponematosi has been recovered until now, from either the Americas or the Old World. Therefore, in the present thesis, ancient and pre-Columbian archaeological remains will be studied in order to try to obtain samples of *T. pallidum* from them that will allow us to elucidate the eternal debate on the origin of this bacterium.
- **Development of a MLST scheme for *T. pallidum* based on genomic data.** Given the current unavailability of a standardized culture system, it is important to have a good genetic typing method to study *T. pallidum*

samples. Therefore, one of the objectives of this thesis is to design new primers for PCR amplification of other loci to obtain new information to help improve the typing system by sequencing, increasing the power of discrimination between lineages and strains of TPA, especially in the SS14 lineage, and TPA against TPE and TEN. An important goal is to maximize genetic information (number of SNPs and STs) with minimal amplified genomic fragment sizes. Likewise, the efficiency of the designed primers will be tested, and the conditions of their use will be fine-tuned with a collection of *T. pallidum* samples



## — MATERIAL AND METHODS —



## Material and methods

It is important to emphasize that certain analyses and their corresponding methodologies will be presented in the Supplementary Material section, rather than in this section. These specific methodologies were carried out entirely by different collaborating researchers involved in various projects. The results obtained from these analyses play a crucial role for the interpretation of the findings of this thesis and drawing relevant conclusions for each project. Therefore, the employed methodologies are detailed in the Supplementary Material section, giving proper attribution to the authors responsible for the analyses. In cases where the methodology involved mutual collaboration rather than being solely performed by specific collaborators, the authorship of the analyses is shared between the designated collaborators and myself, and the methodology is provided in the present Material and Methods section or in the corresponding chapter.

### 1. Ancient samples recollection, archaeological site information and radiocarbon dating

The ancient samples recollection from the archeological site and the osteological exploration of the human remains to detect possible lesions caused by *T. pallidum* in the teeth and bones were performed by:

- **In Chapter 2:** Dr. Hanna Panagiotopoulou, Dr. Karolina Doa, Dr. Wiesław Bogdanowicz, all these three authors from the Polish Academy of Sciences in Wrocław (Poland), Dr. Paweł Dąbrowski, Dr. Maciej Oziembłowski and Dr. Joanna Grzelak, all three from Wrocław Medical University, in Wrocław (Poland). There are also people involved in initial genetic screening which revealed the presence of *Treponema* in one sample: Dr. Martyna Molak, Dr. Hanna Panagiotopoulou, Dr. Karolina

Doan and Dr. Wiesław Bogdanowicz from Polish Academy of Sciences, in Warsaw (Poland).

- **In Chapter 3:** Dr. Jose Filippini and Dr. Luis Pezo Lanfranco both from University of São Paulo in São Paulo (Brazil).

Details of the collection and archaeological site description of all samples obtained from ancient paleontological remains used in Chapters 2 and 3 of this dissertation are detailed in the Supplementary Notes 1 and 2. The ancient sample dating by radiocarbon was performed by Dr. Jose Filippini and Dr. Luis Pezo Lanfranco and is detailed in the Supplementary Note 4.

## **2. Ancient sample processing**

The ancient sample processing was performed in collaboration with Dr. Verena J. W. Schünemann's group, from the University of Basel (Switzerland).

### **2.1 Ancient remains sampling**

The ancient samples used in the present thesis were one upper-left premolar tooth used in chapter 2, and 99 bone samples in Chapter 3. To avoid possible human and environmental DNA contamination, the surfaces of the ancient samples employed were sanded off with a hand rotary tool. The samples were then washed with 5% sodium hypochlorite, molecular-grade water and 75% ethanol, UV-irradiated for 10 min 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 uses to minimize the chances of contamination.

### **2.2 DNA extraction and library preparation**

DNA extraction [94] and double-stranded Illumina library construction were performed according to an established protocol [259,260] in the clean room facilities dedicated to ancient DNA processing at the Institute of Evolutionary Medicine in Zurich [261]. For DNA extraction, 30-120 mg of bone powder were used per sample. The bone powder was obtained by drilling bone tissue using a

dental drill and dental drill bits. For different individuals, variable amounts of extracts were produced. During each extraction, one positive control (ancient cave bear bone powder sample) and one negative control were included for every ten samples. Positive extraction controls were carried along until the indexing of DNA libraries, and the negative controls were carried through all following experiments and sequenced.

Library pools were shotgun sequenced with an Illumina Nextseq platform using a NextSeq 500 Mid Output Kit (75 cycles paired-end) for the first pathogen screening.

Subsequently, to remove ancient DNA specific damage, the Uracil-DNA glycosylase (UDG) [262] was used to treat the additional libraries prepared from the same extraction.

### **2.3 Pathogen screening**

We did an initial screening of the candidate samples with the **MALT** program included in the **EAGER** pipeline [263] in Chapter 2 and by **Kraken2** [264] in Chapter 3, selecting the samples with more than 7 hits to *T. pallidum*. The samples selected were subjected to a target enrichment process and subsequently processed.

### **2.4 Damage profiles for aDNA authentication**

We used a damage profile obtained with the **DamageProfiler** tool [265] for aDNA authentication. We checked the misincorporation patterns and the damage at the end of the sequencing reads of the new ancient genomes obtained before the capture process. A pattern of cytosine-to-thymine misincorporation accumulated at the end of the reads is indicative of authentic ancient DNA in the sample.

### **2.5 Whole genome capture of the ancient samples**

An in-solution capture procedure was performed for the samples considered positive after investigating the shotgun sequencing data. A custom target enrichment kit (Arbor Biosciences) was used for the whole genome capture as in

Majander *et al.* [92]. For this purpose, 60 bp long RNA baits with a 4 bp tiling density and 99% identity were designed based on a selection of representative genomes (Nichols: CP004010.2, Fribourg Blanc: CP003902 and SS14: CP000805.1) in Chapter 2, plus the BosniaA genome (CP007548.1) in Chapter 3 from each *T. pallidum* subspecies or lineage. UDG-treated libraries were pooled in equimolar concentration, and 500 ng final pools were hybridized in 60°C for 48 hours following the manufacturer's instructions. 10 nM capture pools were sequenced using an Illumina NextSeq 500 High-throughput Kit (75 bp paired-end). Libraries from the same individual were merged and processed as paired-end sequencing reads.

Sequencing of the capture data was carried out on an Illumina NextSeq500 with  $2 \times 75 + 8 + 8$  cycles using the manufacturer's protocols for multiplex sequencing in University of Zurich for chapter 2 and in University of Vienna in Chapter 3.

### **3. Dataset selection, read processing and multiple genome alignment generation**

In each chapter, we generated distinct genomic datasets, as described in the specific materials and methods section of each chapter. These datasets primarily consisted of *T. pallidum* genomes previously published in various studies, with the exception of the newly obtained ancient genomes in Chapters 2 and 3.

To generate the datasets, we downloaded the raw data obtained from previous studies from the European Nucleotide Archive (ENA) [266] and the National Center for Biotechnology Information (NCBI) [267]. However, some genomes had only their assembly sequence and their raw data available. So, in Chapters 1 and 4, NGS-like reads were simulated based on the genome assemblies using the tool Genome2Reads (integrated in the **EAGER** pipeline). In chapters 2 and 3 we used a different approach and, instead of simulating reads, we used the assembly of the whole genomes.

The raw data of each new *T. pallidum* genome obtained and those downloaded from previous studies were processed and mapped against a reference, using different strategies also detailed in each chapter. Then, each genome sequence obtained was merged into a single multiple genome file for each dataset, which was aligned using different bioinformatics tools specified in each chapter, except for the datasets in chapters 2 and 3. In those two chapters, the multiple whole genomes file had to be corrected manually with **Aliview** 1.25 [268] after employing bioinformatics programs due to some sequences having a very large amount of missing data (especially the ancient genomes).

#### **4. Antibiotic resistance**

Two mutations in the 23S ribosomal RNA operons, A2058G and A2059G [34,107], were investigated to assess macrolide azithromycin resistance in the two new complete ancient genomes, W86 and ZH1540, obtained in chapters 2 and 3, respectively. 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 by checking the reads of the 23S ribosomal RNA operon with **Tablet** 1.21.02.08 [269] and subsequently with variant calling (with the same parameters detailed in each chapter).

#### **5. Recombination detection: phylogenetic incongruence method (PIM)**

To infer the presence of recombination in the whole genomes of *T. pallidum*, we used the Phylogenetic Incongruence Method (PIM) developed initially in our group [78,270,271] and completed and formalized in this thesis [108]. Putative recombination events are identified on a “per gene” basis with further verification of events spanning more than one gene as well as a detailed analysis of the intra-

genic portions actually involved in those events. In brief, the process involved the following steps (see below for more details):

1. A maximum likelihood (ML) tree was obtained for the multiple genome alignment using **IQ-TREE** 1.6.10 [272].
2. The number of SNPs for each gene extracted from the multiple genome alignment was calculated excluding genes with less than three SNPs.
3. The phylogenetic signal in each gene alignment for each of the remaining genes was evaluated by likelihood-mapping [273] in **IQ-TREE** 1.6.10, retaining only those genes that showed some phylogenetic signal (see below for details).
4. An ML tree was generated for each of the remaining genes using **IQ-TREE** 1.6.10.
5. For each remaining gene, we tested the phylogenetic congruence between trees using **IQ-TREE** 1.6.10, comparing the ML tree obtained from the gene alignment and the ML tree obtained from the whole genome alignment using two different methods: Shimodaira-Hasegawa [274] and Expected Likelihood Weights (ELW) [275].
6. The selected genes that displayed reciprocal incongruence were subsequently examined to assess and describe potential recombination events using **MEGAX** [276]. A gene had to have at least three congruent consecutive homoplastic SNPs that are shared by several groups (TPE, TEN, TPA-Nichols or TPA-SS14) resulting in a polyphyletic group on the reference tree. The consecutive homoplastic SNPs found in the gene alignment limit the boundaries of the recombination events.
7. Using a parsimony criterion on the distribution of alternative states of the homoplastic SNPs, the potential donor and recipient lineage/strains of each recombination event were inferred.



This method was applied to:

- **Chapter 1:** Three datasets generated after using different reference genomes with 75 *T. pallidum* genomes each: the Nichols-mapped dataset, SS14-mapped dataset, and the CDC2-mapped dataset.
- **Chapter 2:** A genome dataset generated and composed by 77 different *T. pallidum* genomes.
- **Chapter 3:** A genome dataset generated and composed by 99 different *T. pallidum* genomes.

### 5.1 Phylogenetic signal test (Likelihood mapping test)

The initial step consisted of an assessment of the phylogenetic information in each of the protein coding genes annotated in the reference genome (978 for the NIC-mapped dataset, 975 for the SS14-mapped dataset, and 1067 for the CDC2-mapped dataset) for chapter 1, and to the 1061 orthologous genes in the four reference genomes employed in chapters 2 and 3, using the likelihood mapping test in **IQ-TREE** 1.6.10 [277]. Prior to the test, individual sequences in each of these protein-coding genes were assigned to four groups, corresponding to the three different subspecies of *T. pallidum*, with TPA sequences further divided into Nichols and SS14 clades (TPE, TEN, TPA-NIC or TPA-SS14) [78]. For the test, we obtained 10,000 random quartets comprising one sequence from each group. For each of these draws, the likelihoods of the three possible unrooted trees for the four groups described above were compared. The genes that showed some phylogenetic signal, evaluated as likelihoods falling outside the central region in the LM triangle [273], were retained for the ensuing analyses. The *tp0897* gene, also known as *tprK*, was not included in chapter 1 in the recombination analyses because its hypervariable regions undergo intrastrain gene conversion and have been studied in detail elsewhere [79] but it was examined in detail in additional analyses, as explained below. In chapters 2 and 3, apart from the *tp0897* gene, two additional genes

(*tp0316*, *tp0317*), which contain repetitive regions and have previously been reported under gene conversion [2,79,146,175,278] were also discarded.

## 5.2 Phylogenetic congruence tests

A topology test was conducted as the second step of the PIM. For this, we constructed maximum-likelihood (ML) trees with **IQ-TREE** 1.6.10 for each of the genes showing some phylogenetic signal in the likelihood mapping analysis. We also obtained the ML tree of the whole genome alignment, which we used as the reference tree. We used the GTR+G+I as the evolutionary model in all the phylogenetic reconstructions. Next, we carried out topology tests for each gene, again with IQ-TREE, using two different methods: Shimodaira-Hasegawa (SH) [274] and Expected Likelihood Weights (ELW) [275] tests. Each topology test involves two comparisons. First, we compared the likelihood of each individual gene tree and the reference genome-wide data tree using the corresponding gene alignment. Secondly, we compared the same likelihoods using the whole genome alignment. A reciprocal incongruence was called when both tests rejected the topology not derived from the corresponding alignment (individual gene in the first comparison, the whole genome in the second). This procedure was performed for the three datasets. Genes for which at least one test rejected the reference tree topology with the gene alignment adopting a conservative approach ( $p < 0.2$ , weight value close to 0, for SH and ELW tests, respectively) and the whole genome alignment rejected the topology of the tree built using the gene alignment (reciprocal incongruence,  $p < 0.2$  and weight value close to 0) in at least one of them were selected and examined more closely in the next step.

## 5.3 Polyphyletic SNP distribution

The selected genes that showed reciprocal incongruence were further analyzed manually with **MEGAX** [276] in order to evaluate and define putative recombination events. To retain a gene as recombinant, we required at least 3 consecutive SNPs that were incongruent with the reference phylogeny but were

distributed congruently with each other, resulting in a polyphyletic group in the reference tree. Recombinant regions were delimited by the homoplastic SNPs detected in the gene alignment. Hence, the distance between the flanking SNPs represents the minimum size of the recombination event, but it might extend further into the non-variable positions upstream or downstream the homoplastic SNPs, given that the exact size of the recombining fragment cannot be estimated. The putative donor and recipient clade/strain of each recombination event were inferred applying a parsimony criterion to the distribution of alternative states of the homoplastic SNPs.

## 6. Recombination detection with alternative tools (chapter 1)

In chapter 1, we used **Gubbins** 2.2.0-1 [164,276] and **ClonalFrameML** 1.1 [166], two widely-used programs for the detection of recombination based on genome-wide data, as alternative tools to the PIM method. To reduce the computational load of these programs, they were run for a subset of the NIC-mapped dataset comprising 27 strains selected to obtain a balanced representation of the four lineages identified in the phylogenetic trees (TPE, TEN, TPA-Nichols, TPA-SS14).

## 7. Selection analyses

- **In Chapter 1**, selection analyses were conducted only with the NIC-mapped dataset. Manually, for each putative recombinant gene selected in our analyses, we extracted the SNPs, determined whether they were synonymous or non-synonymous, and computed the nonsynonymous-to-synonymous substitution ratio  $\omega = dN/dS$ . This ratio was computed for the entire gene as well as for the recombinant and non-recombinant regions separately. In addition, we also estimated  $\omega$  for all non-recombinant genes. A ratio above 1 is indicative of positive selection, while a ratio below 1 points to purifying selection. Additionally, for the non-recombinant genes, we used **CodeML** in the **PAML** 4.9 package [279,280] to estimate the total

number of synonymous and non-synonymous sites per region/gene, as well as SNPeff [279] to evaluate the number of synonymous and nonsynonymous changes in each gene. Additional analyses to detect positive selection were performed with the recombinant genes, the additional genes under positive selection detected with **CodeML** in**PAML**, plus the genes with a high number of SNPs. These additional analyses were performed with **HyPhy** 2.5.32 [281] using BUSTED [282], a gene-based method, and RELAX [283], a codon-based method.

- **In Chapter 2**, we tested for positive selection in a subset of 317 (out of 1,161) 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 the hypervariable regions and gene conversion signal present in these genes [79,133,175,284]. Then, to test if positive selection occurred along the different lineages in a phylogeny [285], we employed **HyPhy** 2.5.32 [286], using the aBSREL model (adaptive Branch-Site Random Effects Likelihood), which is an improved version of the commonly-used "branch-site" models [285,287]. We used default settings and the ML phylogenies of each gene. We assessed statistical significance using a Likelihood-ratio test (LRT).

## 8. Phylogeny reconstruction

In Chapters 1, 2, and 3, prior to the final phylogenetic analyses, certain genes identified as recombinant by PIM were excluded from the multiple genome alignment. Additionally, the well-known hypervariable *tp0897* gene was also excluded. However, in chapters 2 and 3, in order to maintain a phylogenetic assessment based strictly on vertical inheritance, two additional genes (*tp0316*, *tp0317*) were removed. These genes contain repetitive regions and have previously been reported to undergo gene conversion [2,74,79,146,278]. Notably, the *tp0317*

gene is embedded within the *tp0316* gene, and the coordinates for *tp0316* in the BosniaA reference genome span a longer region compared to the other reference genomes. Hence, *tp0316* and *tp0317* were removed based on the BosniaA reference genome coordinates for *tp0316*. All the phylogenetic trees were constructed using the GTR+G+I evolutionary model.

## **9. Mapping the TPE and TEN node-defining SNPs on the whole genome-based tree**

To analyze the genetic changes between the ancestral nodes of the TPE and TEN subspecies in the phylogeny in Chapter 2, an ancestral reconstruction of the multiple genome alignment and the reference phylogenetic tree was carried out with **TreeTime** 0.8.4 [288]. We used **IcyTree** [289] 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** 8.2.11 [290].

## **10. Molecular clock dating**

In Chapter 2, all the molecular clock dating analyses were conducted by Dr. Martyna Molak, affiliated with the Centre of New Technologies at the University of Warsaw and the Museum and Institute of Zoology at the Polish Academy of Sciences, both situated in Poland. On the other hand, in Chapter 3, Dr. Louis du Plessis, from the Department of Biosystems Science and Engineering at ETH and the Swiss Institute of Bioinformatics in Switzerland, performed the molecular clock dating analyses. For a comprehensive understanding of the methodology used in these analyses, please refer to the Supplementary Material in Chapter 2 (Supplementary Note 3) and Chapter 3 (Supplementary Note 6). However, to ensure clarity and readability, the results of these analyses are presented within their respective chapters.

# — CHAPTER 1 —

“Evolutionary processes in the emergence and recent spread  
of *T. pallidum*”





This chapter has been published as:

Pla-Díaz, M., Sánchez-Busó, L., Giacani, L., Šmajš, D., Bosshard, P. P., Bagheri, H. C., Schuenemann, J. V; Nielsen, K; Arora, N; & González-Candelas, F. (2022). Evolutionary processes in the emergence and recent spread of the syphilis agent, *Treponema pallidum*. *Molecular Biology and Evolution*, 39(1), msab318. Doi: <https://doi.org/10.1093/molbev/msab318>



## **Chapter 1: “Evolutionary processes in the emergence and recent spread of *T. pallidum*”**

### **1. Background**

Although genetic variation plays a central role in microbial evolution, some microorganisms have notably low levels of genetic variability, including some of the most virulent human pathogens, such as *Mycobacterium tuberculosis*, *Bacillus anthracis* or *Yersinia pestis* [291–293]. Interestingly, TPA displays strikingly low levels of sequence diversity, lower than that of other genetically monomorphic pathogens such as those mentioned above [291–293].

The increasing application of HTS technologies has generated a wealth of complete microbial genomes that, in turn, has allowed the comparisons required to characterize the extent of genetic variation in microbial pathogens, also including monomorphic ones. HST has also enabled the tracking of the rise and spread of antibiotic resistance [294], the detection of changes in pathogenicity and virulence [270,295], and outbreak investigations [270,296]. However, genomic information derived from mapping to a reference genome in HTS studies is highly dependent on the reference selected [103].

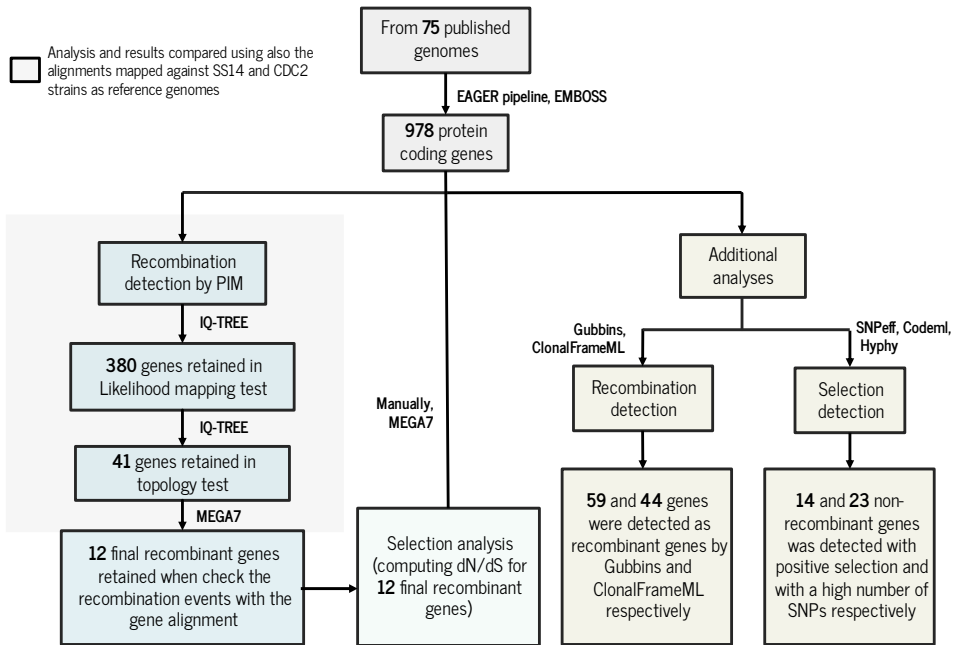
Together with other processes responsible for generating patterns of genetic diversity, it is imperative to examine recombination, which generates diversity [177,297] and natural selection, which shapes it by either maintaining or driving some alleles to fixation or extinction. Particularly the role of recombination is critical, because inferences on the evolutionary history of a species must take into account both vertical and non-vertical inheritance, assessing the contribution and consequences of each type of processes [291,298]. Since reconstructing the evolutionary history of an organism is useful in detecting loci under selection, these analyses would also be affected [299]. However, the detection of recombining loci

and those evolving under natural selection is a challenging task because natural selection may quickly remove allelic variants arising from recombination events, or drive them to fixation, which poses additional difficulties for accurately identifying recombinant loci [291].

Here, we have used a large data set of 75 genomes—all those available at the start of the analysis—across the three subspecies of *T. pallidum*, to explore how intersubspecies recombination and natural selection have shaped the current diversity patterns observed in TPA. To prevent the problems arising from using one single reference for mapping, we have used reference genomes from the major lineages of *T. pallidum*. Our findings provide insights into how lateral gene transfer is a source for change and adaptation in this pathogen, which is responsible for the re-emerging syphilis infections.

## **2. Material and methods**

In the following, an overview of the methods used for the analyses of recombination and selection events in *T. pallidum* genomes is given in Figure 13.



**Figure 13.** Analysis workflow for the study of recombination and selection in *T. pallidum* genomes for the NIC-mapped dataset. The same pipeline was applied to the SS14-mapped and CDC2-mapped datasets.

## 2.1 Dataset selection

We compiled a set of 75 *T. pallidum* genomes (67 TPA, 7 TPE and 1 TEN) from previous studies and public databases. The genomes were selected in order to obtain the best possible representation of the four groups identified in phylogenetic trees at the date (March 2019) when the analyses were started. Short read sequencing data was retrieved for 64 strains from three previous studies [78,79,106]. whole genome sequences for the remaining 11 strains were downloaded from GenBank (Table 7).

**Table 7.** Published samples and genomes used for all analyses.

Study or sample name	Bioproject or GenBank accession ID	Number of samples with at least 80% coverage of the Nichols genome covered by at least 3 reads
[78]	PRJNA313497	28
[79]	PRJNA322283	25
[106]	PRJNA305961	11
Nichols	CP004010.2	1
SS14	CP004011.1	1
Chicago	CP001752.1	1
MexicoA	CP003064.1	1
DAL-1	CP003115.1	1
Seattle81-4	CP003679.1	1
Fribourg	CP003902.1	1
SamoaD	CP002374.1	1
CDC2	CP002375.1	1
Gauthier	CP002376.1	1
BosniaA	CP007548.1	1

## 2.2 Read processing

Read processing was performed in collaboration with Dr. Kay Nieselt, from University of Tübingen (Germany).

To reconstruct the individual genomes from the raw short read data we applied **EAGER** [101], a pipeline including read pre-processing, mapping, deduplication, indel realignment, and variant identification. This pipeline has been applied in previous studies of TPA [92,78,300]. To generate the reads from genomes with no raw data available, the HTS-like reads based on genome assemblies were simulated using the tool **Genome2Reads** [101]. The individual steps are briefly described next. After adapter clipping, merging and quality trimming, the resulting reads for each sample were mapped to the Nichols genome (NC\_021490.2), SS14 (NC\_010741.1) and CDC-2 (NC\_016848.1) using **BWA-MEM** [301] with default

parameters. PCR duplicates were removed with **DeDUP** [101]. Coverage breadth of the reference genome and coverage depth were calculated using **QualiMap** 2.17 [302]. Indel realignments were performed using **GATK** 3.6 [303]. SNPs for the resulting mappings were called using **GATK UnifiedHaplotype**. Sequenced samples were required to cover at least 80% of the Nichols genome by at least 3 reads to be included in further analyses [78,79,106]. **MUSIAL** (<https://github.com/Integrative-Transcriptomics/MUSIAL>) was employed to obtain a multiple genome alignment (MSA) from the resulting VCF files.

In total, this yielded three different data sets, each comprising a total of 75 genomes, and each with a different reference genome (Nichols, SS14, or CDC-2). These are referred to as the NIC-mapped, SS14-mapped, and CDC-2-mapped data sets, respectively. In the following, gene positions and annotations, unless otherwise stated, refer to the Nichols strain. A coverage threshold of 3 and a minimum homozygous SNP allele frequency of 0.9 were required to assign a variant against the nucleotide in each reference genome.

### **2.3 Recombination detection of *tpr* genes**

Among the genes selected in the topology tests some were pertaining to the *tpr* family, which comprises groups of paralogous genes. Due to the repetitive nature of the DNA sequences of these genes, and the challenges of the mapping stage, a large proportion of sites had missing data. When possible, we examined these genes manually. In addition, to test for intrastrain recombination in the seven variable regions of the *tp0897* (*tprK*) gene [144], we generated a **BLAST** database with the 75 complete *T. pallidum* genomes described above and used the set of unique variable motifs found in the variable regions as query for **BLASTn** searches.

### 3. Results

#### 3.1 Reference-based alignments

We used **EAGER** [101] for read pre-processing, mapping, and variant identification of 75 *T. pallidum* samples, following Arora *et al.* [78], using three different references for the mapping step: the widely used TPA Nichols strain (accession CP004010.2), the TPA SS14 strain (accession CP004011.1), and the TPE CDC2 strain (accession CP002375.1). We then used **MUSIAL** (<https://github.com/Integrative-Transcriptomics/MUSIAL>) to compute the strain-specific SNPs and to generate three whole-genome alignment datasets. The SNP calling results for each sample and reference are listed in Supplementary Table 1. The resulting multiple sequence alignments spanned a total of 1,139,633 bp (NIC-mapped dataset), 1,139,569 bp (SS14-mapped dataset) and 1,139,744 bp (CDC2-mapped dataset), respectively.

#### 3.2 Recombination events in *T. pallidum*

To determine the effects of recombination as a force of genetic diversity and differentiation, we applied the **PIM** procedure to the three multiple alignments obtained from different genomes as reference for mapping. Firstly, we performed a likelihood-mapping test to ascertain which genes had some phylogenetic signal. For the NIC-mapped dataset, 380 out of the 978 genes showed a phylogenetic signal and were retained for the ensuing analyses (Supplementary Table 2). The remaining genes were discarded because for all the quartets considered the distribution of the corresponding likelihoods fell in the central zone of the triangle. Using the SS14-mapped and CDC2-mapped datasets, this step yielded 498 and 535 genes with some phylogenetic signal, respectively (Supplementary Tables 3-4).

Next, for each gene retained in the likelihood mapping analyses we tested the phylogenetic congruence between trees, comparing the tree obtained from the gene alignment and the tree obtained from the whole genome alignment using the SH



and ELW topology tests. For the NIC-mapped dataset, only 44 genes showed reciprocal incongruence. In contrast, 63 and 76 genes displayed reciprocal incongruence in the SS14-mapped and CDC2-mapped datasets, respectively. Overall, 90 genes displayed reciprocal incongruence in at least one of the three datasets while only 29 genes did so for all three datasets (Supplementary Table 5). Moreover, all the reciprocal tests, that is the whole genome alignment tested with the tree derived from it and with the trees derived from each gene alignment, yielded the same result for the three datasets: acceptance of the whole genome tree with probability or weight equal to 1 and rejection of the alternative tree with null probability or weight.

We checked for the presence of a minimum of three consecutive homoplastic SNPs contributing to the reciprocal incongruence results in these 90 genes. Only 12 genes had three or more consecutive homoplastic SNPs and the rest were not considered further (Table 8). It is worth noting that the recombination event observed in *tp0164* displayed reciprocal incongruence in the topology analyses of the NIC-mapped dataset only. However, the recombination event was detected in the SNP alignment for all three data sets.

Our analyses identified only one recombination event per gene (Table 8 and Supplementary Figures 1-12) except for genes *tp0136* and *tp0865*, for which seven and four events were detected, respectively. The average length of the recombinant region was around 470 bp (471.85, median = 391), with a minimum of 12 bp and a maximum of 1,900 bp. The average number of SNPs encompassed in these events was 15.38, with a minimum of 3 and a maximum of 45 SNPs. In total, the identified recombination events account for 294 out of the 927 SNPs (31.72%) found among the TPA strains analyzed here (Table 8).

**Table 8.** Recombination events in *T. pallidum* detected using PIM with the NIC-mapped dataset. For each recombination event (denoted as gene\_event) we report the start and end position of the event (coordinates according to the Nichols strain), its size in base pairs, the number of SNPs detected in the event, the donor (origin) cluster/strain, the recipient (receptor) cluster/strain (Supplementary Figures 1-12 for additional details), and the functional significance of the gene according to UniProt.

Gene_event	Start	End	Size (bp)	SNPs	Origin	Receptor	Function (UniProt)
<b>TP0136_1</b>	158092	158104	13	4	TPE	SW6	adhesin allowing binding to fibronectin
<b>TP0136_2</b>	158138	158149	12	4	Seattle 81-4	SW6	
<b>TP0136_3</b>	158149	158167	19	3	TPE	Seattle 81-4	
<b>TP0136_4</b>	158271	158336	66	6	TPE	Nichols clade	
<b>TP0136_5</b>	158346	158364	18	7	TEN	Nichols clade	
<b>TP0136_6</b>	158915	158976	62	3	TPE-TEN	Nichols clade	
<b>TP0136_7</b>	159312	159323	12	5	TPE-TEN	Nichols clade	
<b>TP0164</b>	187064	187177	113	4	TPE-TEN	NE20, SEA86 cluster	<i>troB</i> , iron/zinc/manganese ABC superfamily ATP binding cassette transporter, ABC protein
<b>TP0179</b>	198040	198428	391	9	TPE-TEN	Nichols clade	hypothetical protein
<b>TP0326</b>	347027	347956	929	32	TEN	SS14 clade excluding Mexico A	$\beta$ -barrel assembly machinery A ( <i>BamA</i> ) orthologue and rare outer membrane protein

## Chapter 1

<b>TP0462</b>	492772	493605	834	43	TPE-TEN	NE20, SEA86 cluster	hypothetical protein
<b>TP0488</b>	522981	523620	640	41	TEN	Mexico A	<i>Mcp2</i> , methyl-accepting chemotaxis protein
<b>TP0515</b>	555872	557771	1900	17	TPE-TEN	Nichols clade	<i>LptD</i> homolog
<b>TP0548</b>	593563	594215	653	45	TPE-TEN	Nichols clade	OMP- FadL family
<b>TP0558</b>	606171	606591	421	4	TPE-TEN	SS14 clade	NiCoT family nickel-cobalt transporter, high affinity
<b>TP0865_1</b>	945224	945542	319	13	TPE-TEN	NE20, SEA86 cluster	OMP- FadL family
<b>TP0865_2</b>	945224	945542	319	13	TPE-TEN	Seattle 81-4	as above
<b>TP0865_3</b>	945830	946298	469	18	TEN	NE20, SEA86 cluster	as above
<b>TP0865_4</b>	945830	946298	469	18	TEN	Seattle 81-4	as above
<b>TP0967</b>	1051257	1052302	1046	15	TPE-TEN	Seattle 81-4	<i>TolC</i> -homologue
<b>TP0968</b>	1052414	1053617	1204	22	TEN	Seattle 81-4	<i>TolC</i> -homologue

We found an additional recombination event in the *tp1031* (*tprL*) gene which was not among the 12 recombination events detected by **PIM** despite showing reciprocal incongruence in the topology tests in the SS14-mapped and CDC2-mapped data sets. Nevertheless, *tp1031* was classified as a gene with a high number of SNPs and also without signs of positive selection acting on it (see subsection below). A detailed analysis of this gene revealed 23 SNPs present only in SS14 clade strains (positions 745 to 872) (Figure 14). This high variation found in the

SS14 clade seems to be the result of a transfer from another *Treponema* subspecies or species not identified in the public databases.

```

!Domain=Data;
[
[
27777778 8888888888 8800012222 223]
[
2544557790 0111234457 7924480111 225]
[
7335453868 9148084831 2788979015 670]
#Nichols CGTGGCTCTC AAATGACCGA AACTAGAAT GGA
#SamoaD .A..... .TGTG.CCG. ..G
#CDC2 .AC..... .GTG.CCG. ..G
#Gauthier .A..... .GTG.CCG. ..G
#GHA1 .A..... .N... .GTG.CCG. ..G
#IND1 .A..... .GTG.CCG. ..G
#Fribourg .A..... .GTG.CCG. ..G
#BosniaA ..... .GTG.CCG. ..G
#BAL3 .....
#Chicago .....
#BAL73 .....
#NIC1 .....
#NIC2 .....
#Dallas .....
#Seattle81 .....
#NE20 .....C.....
#SEA86 NN....N... .C..... .N.....
#MexicoA GA.ACACTCN NNNNCGGTAG C.GTGG...N AC.
#SS14 GA.ACACTCN NNNNCGGTAG C.GTGG...N AC.
#NE14 GA.ACANNNN NNNNNNNNNN NNGTGG...N AC.
#SW6 GA.ACACTCA NNNNNGGTAG C.GTGG...N AC.
#CZ33 GA.NNNNNNN NNNNNNNNNN N.GTGG...N AC.
#AU15 GA.ACACTCA CGGACGGTAG C.GTGG...G AC.
#PT_SIF1002 GA.ACACTCA CGGACGGTAG C.GTGG...N AC.
#NE17 GA.ACACTCA CGGACGGTAG C.GTGG...N AC.
#P3 GA.ACACTCA CGGACGGTAG C.GTGG...G AC.
#AR2 GA.ACACTCA CGGACGGTAG C.GTGG...N AC.

```

**Figure 14.** Detail of the multiple alignment with the SNPs and strains involved in the additional recombination event in the *tp1031* (*tprL*) gene which was not among the 12 recombination events detected by PIM despite showing reciprocal incongruence in the topology tests in the SS14-mapped and CDC2-mapped data sets. The figure shows the 23 SNPs present only in SS14 clade strains (positions 745 to 872) corresponding to the possible recombination event.

Additionally, we checked for intragenomic recombination in *tp0897* (also known as *tprK*) (Table 9). The most similar genome fragments to those included in the variable regions of *tprK* correspond to coding and intergenic regions between *tp0126a* and *tp0138*. However, none of these putative donors for the variable regions in *tprK* was detected by PIM (Table 8).

**Table 9.** Summary of **BLASTN** searches of variable sequences (defined in [144]) in the *tp0897* (*tpkK*) gene (deposited in GenBank and reported in Supplementary Table 4 of [79,144]) using as query the 75 complete *T. pallidum* genome sequences analyzed in this study.

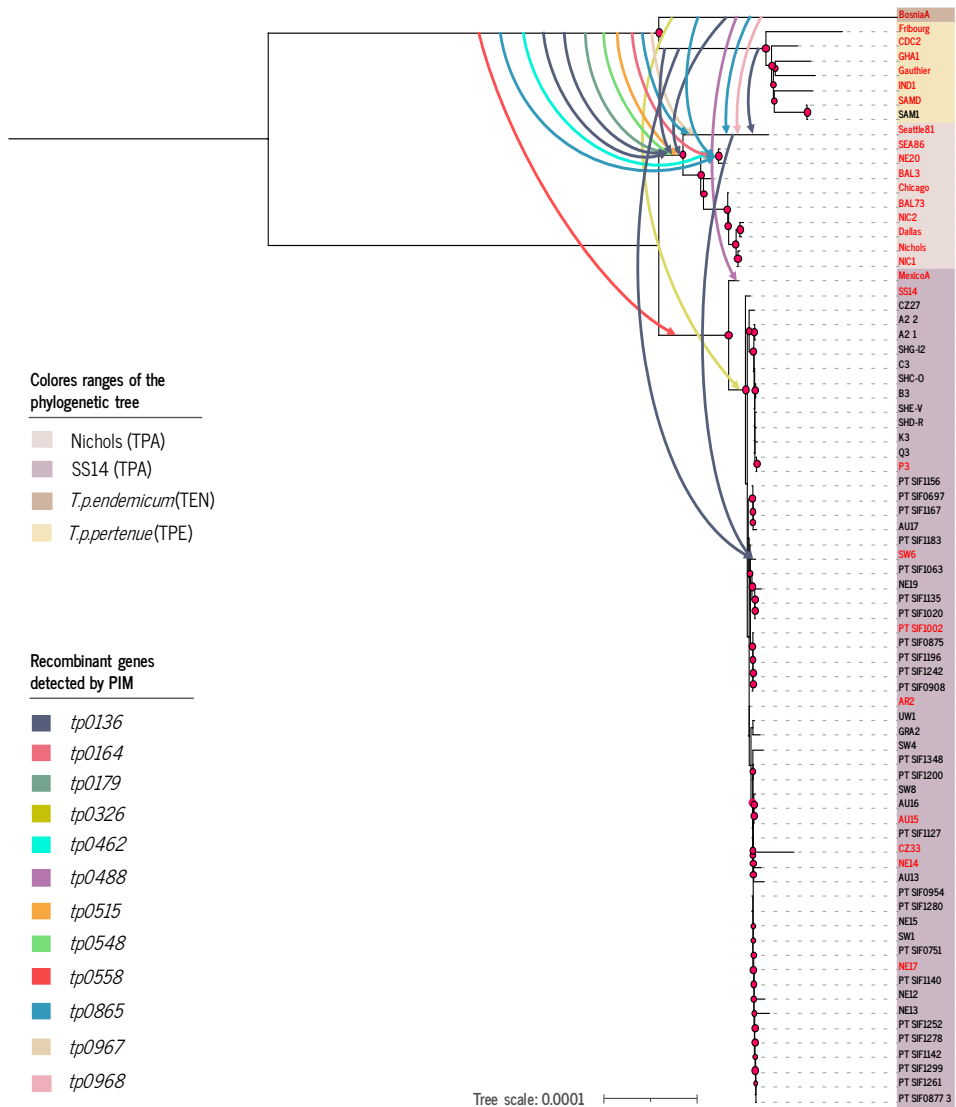
Variable region	Unique sequences	Matches
V1	31	<i>tp0126d, tp0126c, tp0129, tp0130</i>
V2	54	<i>tp0126c, tp0128, tp0129, tp0130</i>
V3	63	<i>tp0126c, IR-tp0127-tp0128, tp0128, tp0129, IR-tp0129-tp0130, IR-tp0130-tp0131</i>
V4	29	<i>tp0126a, IR-tp0126d-tp0126c, IR-tp0129-tp0130,</i>
V5	80	<i>tp0126a, IR-tp0126d-tp0126c, tp0126c, IR-tp0128d-tp0129, tp0129, IR-tp0129-tp0130, tp0130</i>
V6	155	<i>tp0126a, IR-tp0126d-tp0126c, IR-tp0127-tp0128, tp0128, tp0129, IR-tp0129-tp0130</i>
V7	102	<i>tp0126d, tp0126c, tp0127, tp0129, tp0130, IR-tp0136-tp0138</i>

### 3.3 Most recombination events have occurred between subspecies

Next, for each of the three datasets, we removed the 12 recombinant genes and *tp0897* from the multiple alignment, in order to build a non-recombinant genome phylogeny. As the three tree topologies were almost identical (Supplementary Figures 13-15), we selected the phylogenetic tree for the NIC-mapped dataset (Figure 15) to represent the 21 recombination events detailed in Supplementary Figures 1-12. The SS14-mapped and CDC2-mapped datasets were not considered for the remaining analyses. All but one recombination event corresponded to inter-subspecies transfers, from TPE/TEN to TPA. The only exception was the event *tp0136\_2* (Table 8), which corresponded to an intra-subspecies transfer (within TPA), specifically, to a transfer from the Nichols to the SS14 clade (Table 8, Supplementary Figures 1-12). These recombination events between the clades

further support a geographically close common history of the TPA and TPE lineages, which cannot be concluded from the geographical distribution of extant lineages [92].

Amongst the inter-subspecies transfers, 11 originated from the TPE/TEN clade: these events included the most frequent donor-recipient combination, which involved the entire Nichols clade as recipient (5 events). The assignment of an event to the TPE/TEN clade does not necessarily mean that the actual source of the event was an ancestor of TPE and TEN strains; it could have also been a more recent descendant of any of these two subspecies that share the same variants. Consequently, we cannot ascertain the exact phylogenetic location of the donor genome. Additionally, we detected six events originating from the TEN clade and three events originating from the TPE. Interestingly, strains in the Nichols clade were the most frequent recipient of external DNA, involving particular strains or clusters within (another 9 events) or the ancestor of the entire clade (in 7 events). Only one event involving *tp0558* had the whole SS14 clade as recipient, while there were two events in *tp0136* (one event originated in TPA), other involving *tp0488* and another involving *tp0326*, that had SW6, MexicoA and all SS14 clade but one (Mexico A) strains in this clade as recipients, respectively.



**Figure 15.** Maximum likelihood tree obtained from the NIC-based dataset without the 12 recombinant genes detected using PIM and the *tp0897* gene (resulting alignment length 1,117,857 bp). The different subspecies corresponding to yaws (TPE), bejel (TEN) and the Nichols and SS14 clades of syphilis (TPA) are indicated in the figure. Nodes with bootstrap support values larger than 70% are indicated by red circles.

The effect of recombination on the phylogenetic reconstruction of *T. pallidum* can be seen by comparing the topologies of two maximum likelihood trees: that obtained with all genes included in the alignment of the NIC-mapped dataset (1,139,633 bp), and that obtained after excluding the recombinant genes plus

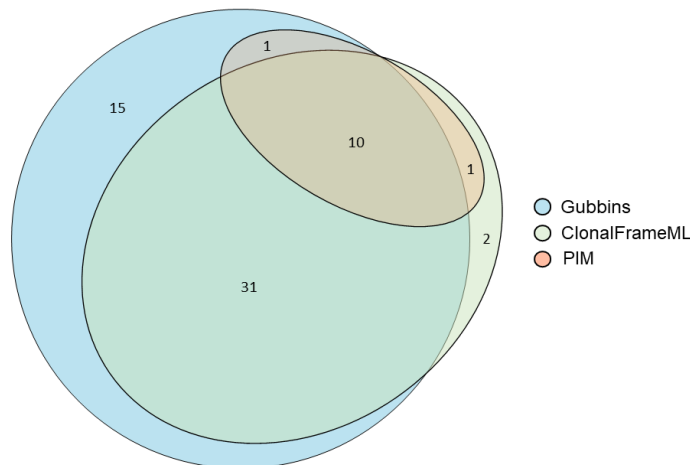
*tp0897* from the alignment (1,117,857 bp) (Supplementary Figure 15). As expected, the most remarkable differences between the two topologies involved those strains and clusters frequently implicated in recombination events. In the Nichols clade, Seattle 81-4 occupies a basal position in the non-recombinant tree whereas this position corresponds to NE20 and SEA86 in the whole genome-based tree. In fact, this difference in the topology is responsible for the doubling of recombination events inferred in locus *tp0865*, in which identical events had to be postulated for Seattle81-4 and for the NE20-SEA86 cluster. In *tp0865*, two different recombinant regions were detected which, based on the phylogeny derived from the non-recombinant core genome of *T. pallidum*, corresponded to four different transfers (Supplementary Figure 10). An alternative explanation, requiring only two recombination events, would involve separate transfers for each region from the source clade (TPE/TEN and TPE, respectively) to the Nichols clade, with a subsequent loss of the transferred fragments in the sister clade to NE20-SEA86 in the ML phylogenetic tree (Figure 15). These explanations rely on the assumption that the phylogenetic reconstruction after removal of recombinant regions is accurate. But there are other processes that can obscure phylogenetic reconstruction [304–306]. Thus, it is important to take into account that there might be alternative topologies providing a simpler explanation for the observation of four recombination events in *tp0865*.

The BAL3 strain also displays a dramatic topological change from a sister branch to the SS14 clade in the whole genome tree to a well-supported position within the Nichols clade in the non-recombinant tree. For the SS14 clade, there are numerous minor differences between the two trees. Nonetheless, the basal position of Mexico A is retained in both. Interestingly, the relationships of all strains from the yaws (TPE) and bejel (TEN) clades are consistent in both trees.



### 3.4 Recombination detection with alternative tools

The two alternative tools detected more recombination events than **PIM** (Supplementary Tables 6-7 and Supplementary Figures 16-17). **Gubbins** detected 58 potential recombination events spanning 64 genes whereas **ClonalFrameML** identified 92 potential events spanning 44 genes. Ten of the 12 recombinant genes detected by PIM (*tp0136*, *tp0179*, *tp0326*, *tp0462*, *tp0488*, *tp0515*, *tp0548*, *tp0865*, *tp0967*, *tp0968*) were also detected by **ClonalFrameML** and **Gubbins**. Additionally, **Gubbins** detected *tp0558* but did not detect *tp0164* whereas for **ClonalFrameML** the pattern was reversed (Figure 16). The average length of the recombinant regions detected by **Gubbins** and **ClonalFrameML** was 2,199 bp (range 33 – 12,566 bp) and 470 bp (range 4 – 3,508), respectively.



**Figure 16.** Venn diagram showing a summary of the number of recombinant genes detected by **PIM**, **Gubbins** and **ClonalFrameML**. 10 of the 12 genes detected using PIM were also detected by ClonalFrameML and Gubbins. ClonalFrameML and Gubbins detected 53 and 33 additional recombinant genes, respectively.

### 3.5 Selection analyses

The effects of selection on the 12 recombinant genes as identified by PIM were estimated by the non-synonymous to synonymous substitution ratios ( $\omega$ ). We found diverse patterns across the genes (Table 10).

**Table 10.** Recombinant genes with their synonymous and non-synonymous sites and changes and estimates of  $\omega = dN/dS$  for the recombinant and non-recombinant regions of each gene. (NC: non computable).

Gene	Recombinant regions					Recombinant regions					$\omega$	
	Size (nt)	Syn sites	Syn changes	Non syn sites	Non syn changes	Size (nt)	Syn sites	Syn changes	Non syn sites	Non syn changes	Non recombinant regions	Recombinant regions
<i>tp0136</i>	1274	432	18	801	84	214	64	6	128	25	2,52	2,08
<i>tp0164</i>	687	238	0	446	0	114	40	2	72	2	NC	0.56
<i>tp0179</i>	1495	547	1	947	1	389	137	2	245	7	0.58	1.96
<i>tp0326</i>	1632	567	2	1060	3	930	331	3	580	29	0.80	5.51
<i>tp0462</i>	345	119	2	219	8	834	271	2	545	41	2.17	10.19
<i>tp0488</i>	1898	719	0	1169	18	640	206	2	406	39	NC	9.89
<i>tp0515</i>	1076	386	1	688	0	1900	647	0	1239	17	0.00	NC
<i>tp0548</i>	652	211	7	423	27	653	204	5	420	40	1.92	3.89
<i>tp0558</i>	488	170	3	316	1	421	152	4	266	0	0.18	0.00
<i>tp0865</i>	652	211	3	430	10	788	277	14	497	17	1.64	0.68
<i>tp0967</i>	508	173	0	330	1	1046	371	6	666	9	NC	0.84
<i>tp0968</i>	419	152	1	264	1	1204	421	5	775	17	0.58	1.85

Both positive and purifying selection were observed in four genes, *tp0179*, *tp0865*, *tp0968* and *tp0326*, with clear differences between recombinant and non-recombinant regions of these two genes. In *tp0179* and *tp0968*, purifying selection was observed in the non-recombinant region ( $\omega=0.58$  in both genes) and positive selection in the recombinant region ( $\omega=1.96$  and  $\omega=1.85$ , respectively), with fewer changes in the non-recombinant compared to the recombinant regions (Table 10). In contrast, three non-recombinant and two recombinant regions were observed in *tp0865*, with indications for positive selection in the former ( $\omega = 1.64$ ) and purifying selection ( $\omega= 0.68$ ) in the latter. Interestingly, a more detailed analysis of  $\omega$  in the different subregions revealed some additional differences (Table 11). Values of  $\omega$  were similar in both recombinant regions, with estimates lower than 1 but not indicative of strong purifying selection (0.84 and 0.58, respectively). The three non-recombinant portions of this gene were not homogeneous: two of them, NR1 and NR3, were very constrained to changes, with only one synonymous mutation in NR3 and one non-synonymous change in NR1. The central, non-

recombinant portion of *tp0865* (NR2) accumulated more non-synonymous than synonymous changes, leading to  $\omega = 2.25$ .

**Table 11.** Analysis of selection in genes with more than one recombination event detected by PIM with the Nichols strain as reference with their synonymous and non-synonymous sites and changes and estimates of  $\omega$  ( $=dN/dS$ ) for the recombinant and non-recombinant regions of each gene. (NC=Non computable)

Gene	Region	Size	Syn. sites	Syn. changes	Non-syn. sites	Non syn. changes	$\omega$
<i>tp0136</i>	NR1	149	53	5	94	0	0.00
	R1	13	3	3	8	1	0.13
	NR2	33	12	3	21	0	0.00
	R2-3	40	11	3	26	4	0.56
	NR3	93	32	0	56	12	NC
	R4	66	22	0	43	6	NC
	NR4	9	0	2	2	5	NC
	R5	19	7	0	11	7	NC
	NR5	550	189	4	340	43	5.96
	R6	62	19	0	40	3	NC
	NR6	335	116	3	203	21	4.00
	R7	12	3	0	5	5	NC
	NR7	107	34	3	68	5	0.83
<i>tp0865</i>	NR1	257	85	0	169	1	NC
	R1	319	107	5	205	8	0.84
	NR2	287	93	2	186	9	2.25
	R2	469	170	9	292	9	0.58
	NR3	108	33	1	75	0	0.00

For *tp0326*, a weak signal of purifying selection was detected in the non-recombinant regions, with two synonymous substitutions and three non-synonymous substitutions detected. However, the recombinant portion of this gene seems to have evolved under strong positive selection ( $\omega = 5.51$ , with 29 non-synonymous changes and 3 synonymous ones).

Three genes (*tp0164*, *tp0558* and *tp0967*) showed negative or purifying selection ( $\omega < 1$ ) in both types of regions. For *tp0164* we only found SNPs within the recombinant region, associated with the recombination events (Table 10). In contrast, *tp0558* and *tp0967* showed some SNPs in both types of regions, but with

fewer non-synonymous than synonymous changes in the *tp0558* gene, and more non-synonymous changes than synonymous changes in the *tp0967* gene (Table 10). Positive selection in both regions was the dominant form of selection in four of the remaining recombinant loci (Table 10). The *tp0136* gene had similar  $\omega$  values ( $>1$ ) in the non-recombinant and recombinant regions indicating pervasive positive selection in this gene (Table 10 and Table 11). In *tp0462*, *tp0488* and *tp0548*, both recombinant and non-recombinant regions showed a strong signal of positive selection. Finally, *tp0515* also presented strong positive selection in the recombinant region, with no synonymous substitutions and 17 non-synonymous changes, but its non-recombinant region is highly constrained, with only one synonymous substitution found (Table 10).

In addition to examining the role of natural selection in the putative recombinant genes, we also investigated its role in the rest of the genome by **SNPeff** and **CodeML**. Among the 965 remaining genes (without considering the final 12 recombinant genes and the *tp0897* gene), only 14 yielded estimates of  $\omega > 1$  (Table 12), with estimates ranging from 1.11 to 2.61. Among them, only two had estimates of  $\omega > 2$  although, with 8 or fewer SNPs, the amount of variation contributed by these genes was quite low.

**Table 12.** Non-recombinant genes but with positive selection by dN/dS value detected using Codeml and SNPeff. For each gene the table shows the values of dN, dS,  $\omega$  ( $=dN/dS$ ), the expected and observed SNPs, the proportion of excess observed SNPs over the expected ones and their functional significance according to **Uniprot**, respectively.

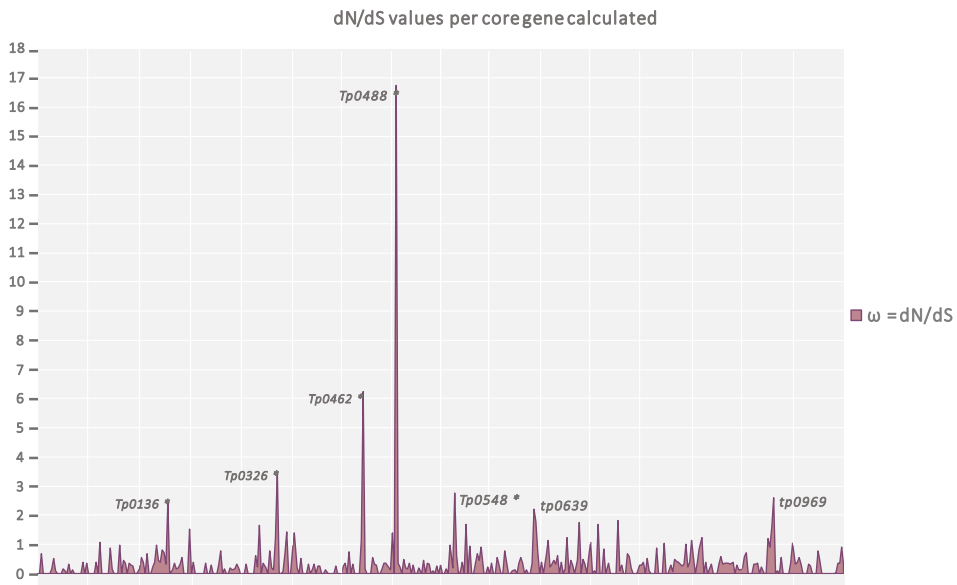
Gene	dN	dS	$\omega$	exp. SNPs	obs. SNPs	(O-E)/E	Functional significance
<i>tp0304</i>	0.008	0.005	1.67	6.04	21	2.48	Peptidase MA
<i>tp0346</i>	0.020	0.014	1.48	1.38	13	8.44	putative lipoprotein
<i>tp0369</i>	0.003	0.002	1.38	3.05	4	0.31	tetratricopeptide repeat containin protein
<i>tp0484</i>	0.003	0.002	1.40	3.91	5	0.28	FecR protein (regulation of iron dicitrate transport)
<i>tp0564</i>	0.003	0.002	1.69	4.01	5	0.25	signal peptide cleavable. predicted for secretion
<i>tp0639</i>	0.004	0.002	2.20	3.87	7	0.81	methyl-accepting chemotaxis protein
<i>tp0640</i>	0.004	0.002	1.80	3.63	6	0.65	methyl-accepting chemotaxis protein
<i>tp0687</i>	0.003	0.002	1.25	3.88	5	0.29	ATP-dependent DNA helicase RecG
<i>tp0705</i>	0.003	0.001	1.75	5.16	6	0.16	bifunctional membrane carboxypeptidase/penicillin-binding protein
<i>tp0729</i>	0.005	0.003	1.68	3.22	7	1.17	flagellar hook-length control protein FliK
<i>tp0746</i>	0.005	0.003	1.82	5.23	11	1.10	pyruvate. phosphate dikinase
<i>tp0859</i>	0.013	0.012	1.11	2.92	19	5.50	putative OMP
<i>tp0966</i>	0.009	0.008	1.19	3.16	14	3.43	putative OMP
<i>tp0969</i>	0.006	0.002	2.61	3.16	8	1.53	putative OMP

Although some of the genes identified as recombinant and with signs of positive selection had a large number of SNPs, especially for genomes with such a low genetic variability as those of *T. pallidum*, not all the genes with a relatively high proportion of SNPs have been under the action of adaptive selection. We detected 23 genes (Table 13) in which the number of observed SNPs exceeded twice the number of expected ones, but whose estimates for  $\omega$  were  $<1$ .

**Table 13.** Non-recombinant genes with no positive selection but with significantly more SNPs than expected. For each gene, the table shows the estimates of dN, dS,  $\omega$  ( $=dN/dS$ ), expected and observed SNPs, and the proportion of excess observed SNPs over the expected ones.

Genes	Protein function	dN	dS	$\omega$	exp.	obs.	(O-E)/E
<i>tp0110</i>	hypothetical protein	0.004	0.014	0.269	3.41	11	2.22
<i>tp0117</i>	Tpr protein C, TprC	0.098	0.180	0.548	0.27	16	59.13
<i>tp0131</i>	Tpr protein D, TprD	0.045	0.119	0.374	0.24	8	32.00
<i>tp0133</i>	hypothetical protein	0.009	0.011	0.802	2.40	12	4.00
<i>tp0134</i>	putative outer membrane protein	0.006	0.009	0.702	2.19	8	2.65
<i>tp0313</i>	Tpr protein E, TprE	0.006	0.008	0.784	2.25	8	2.56
<i>tp0433</i>	Signal peptide predicted for secretion	0.010	0.014	0.731	2.47	14	4.68
<i>tp0464</i>	tRNA (guanine-N(7))-methyltransferase, TrmB	0.004	0.021	0.173	1.45	6	3.12
<i>tp0483</i>	signal peptide for secretion	0.006	0.055	0.112	2.19	22	9.05
<i>tp0577</i>	putative membrane protein	0.005	0.006	0.914	3.63	10	1.75
<i>tp0617</i>	hypothetical protein	0.005	0.061	0.083	0.55	6	9.91
<i>tp0618</i>	putative lipoprotein	0.007	0.050	0.140	0.76	7	8.25
<i>tp0619</i>	hypothetical protein	0	0.063	0	0.44	4	8.14
<i>tp0620</i>	Tpr protein I	0.018	0.052	0.342	1.51	21	12.87
<i>tp0691</i>	segregation and condensation protein, ScpA	0.007	0.024	0.285	1.45	8	4.5
<i>tp0733</i>	OMP	0.029	0.034	0.853	1.31	20	14.44
<i>tp0856</i>	OMP for lipid transport (FadL family)	0.005	0.020	0.227	2.33	11	3.72
<i>tp0858</i>	OMP for lipid transport (FadL family)	0.048	0.122	0.395	2.41	79	31.74
<i>tp0861</i>	glutamine--fructose-6-phosphate_transaminase (isomerizing), GlmS	0.005	0.010	0.552	3.67	12	2.27
<i>tp0896</i>	hypothetical protein	0.028	0.076	0.369	0.29	6	19.71
<i>tp0898</i>	exodeoxyribonuclease V beta subunit, RecB	0.004	0.010	0.354	7.25	19	1.62
<i>tp1030</i>	hypothetical protein	0	0.125	0	0.31	4	12.01
<i>tp1031</i>	Tpr protein L, TprL	0.003	0.064	0.041	3.04	29	8.54

Globally, there were only 8 genes in the *T. pallidum* genome with estimates for  $\omega > 2$ . Six of those corresponded to genes involved in recombination events (*tp0136*, *tp0326*, *tp0462*, *tp0488*, *tp0515* and *tp0548*), whereas *tp0639* and *tp0969* had the lowest  $\omega$  values among the genes in this group and were not involved in recombination. The distribution of  $\omega$  values along the *T. pallidum* genome is shown in Figure 17 (detailed results of  $\omega$  values per gene are available in Supplementary Table 8).



**Figure 17.** Distribution of  $\omega$  values along the *T. pallidum* genome (detailed results of  $\omega$  value per gene are in Supplementary Table 11). Genes with  $\omega > 2$  are indicated. This parameter cannot be estimated for *tp0515* because it does not present synonymous substitutions. Genes marked with \* were detected as recombinant by PIM.

In addition to the computation of dN/dS, we also checked for the action of natural selection on *T. pallidum* genes using **Hyphy**. We analyzed the 12 recombinant genes, the 14 non-recombinant genes but with  $\omega$  values  $> 1$  and the 23 non-recombinant genes without signal of natural selection but with a large number of SNPs. Of these 49 genes, 5 genes could not be tested due to the presence of indels in their sequences. The BUSTED test implemented in Hyphy provided evidence for positive selection for 18 genes, comprising nine recombinant and nine non-

recombinant genes ( $p < 0.05$ ) (Table 14, Supplementary Table 9). Among the non-recombinant genes, one was found to display evidence of positive selection (Table 12), and eight had a high number of SNPs (Table 13). The two genes (*tp0639* and *tp0969*) found with  $\omega > 2$  (Table 12) were not detected by Hyphy to be under positive selection. Hyphy detected no genes as evolving under relaxed selection using the RELAX test.



**Table 14.** Genes tested with Hyphy for positive selection (Busted test) with their corresponding P-values. Red: recombinant genes with positive selection (Table 8). Green: Non-recombinant genes with  $\omega > 1$  evaluated by codeml+SNPeff (Table 12). Light blue: non-recombinant genes with excess of SNPs (Table 13). The genes with a significant p-value are underlined.

Gene	P-value	Gene	P-value
<i>tp0110</i>	1.00	<i>tp0618</i>	1.00
<i>tp0133</i>	<u>0.00</u>	<i>tp0620</i>	<u>0.00</u>
<i>*tp0136</i>	<u>0.00</u>	<i>*tp0639</i>	0.97
<i>tp0164</i>	0.98	<i>tp0640</i>	1.00
<i>tp0179</i>	1.00	<i>tp0687</i>	0.79
<i>tp0304</i>	0.71	<i>tp0691</i>	1.00
<i>tp0313</i>	<u>0.00</u>	<i>tp0705</i>	0.24
<i>*tp0326</i>	<u>0.00</u>	<i>tp0729</i>	0.09
<i>tp0346</i>	0.57	<i>tp0733</i>	0.06
<i>tp0369</i>	0.95	<i>tp0746</i>	0.27
<i>tp0433</i>	<u>0.00</u>	<i>tp0856</i>	<u>0.01</u>
<i>*tp0462</i>	<u>0.00</u>	<i>tp0858</i>	<u>0.02</u>
<i>tp0464</i>	1.00	<i>tp0859</i>	<u>0.03</u>
<i>tp0483</i>	<u>0.00</u>	<i>tp0861</i>	1.00
<i>tp0484</i>	0.98	<i>tp0865</i>	<u>0.00</u>
<i>*tp0488</i>	<u>0.00</u>	<i>tp0896</i>	0.45
<i>*tp0515</i>	<u>0.01</u>	<i>tp0898</i>	1.00
<i>*tp0548</i>	<u>0.00</u>	<i>tp0966</i>	0.16
<i>tp0558</i>	1.00	<i>tp0967</i>	<u>0.00</u>
<i>tp0564</i>	0.97	<i>tp0968</i>	<u>0.00</u>
<i>tp0577</i>	1.00	<i>*tp0969</i>	0.82
<i>tp0617</i>	0.50	<i>tp1031</i>	<u>0.00</u>

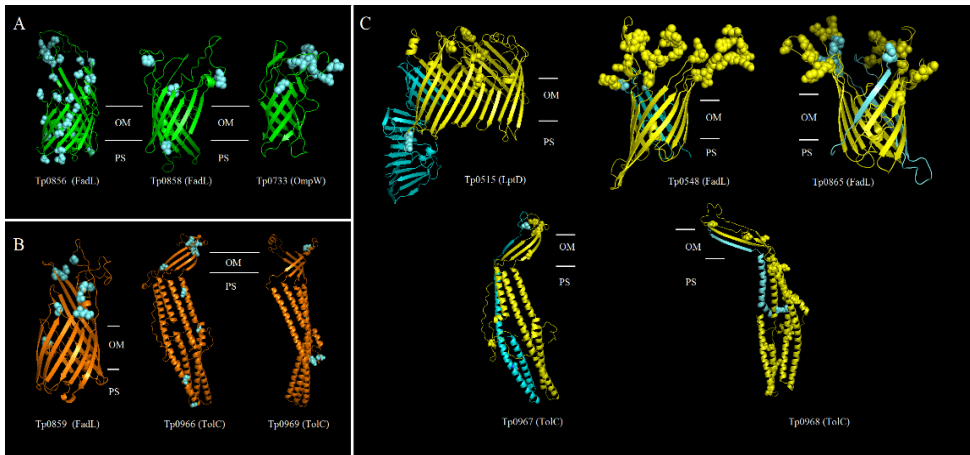
### **3.6 Gene implications for the selection of vaccine candidates and the design of a broadly protective syphilis vaccine**

During infection, in fact, *T. pallidum* clearance from early lesions is mediated by phagocytosis of opsonized treponemes by infiltrating macrophages [307,308]. Hence, *T. pallidum* surface-exposed antigens are regarded as the most promising vaccine candidates. Several of the genes analyzed were shown here to be highly variable among strains, whether or not variability is recombination-driven, encode putative (*tp0515*, *tp0548*, *tp0733*, *tp0865*, *Tp0856*, *tp0858*, *tp0859*, *tp0966*, *tp0967*, *tp0968*, *tp0969*, and *tp1031*) or *bona fide* (*tp0117*, *tp0131*, *tp0136*, *tp0326*, and *tp0620*) surface-exposed OMPs.

Most of the putative OMPs were never tested to evaluate their protective potential in pre-clinical immunization/challenge trials, mainly because of their recent identification as putative surface-exposed molecules. Upon assessing their protective ability, however, some of these antigens could become components of a multi-subunit vaccine, as current and past vaccine development endeavors clearly suggest that a single antigen will not be sufficient to elicit sterilizing immunity to the syphilis agent. Particular interest revolves around the Tp0117 (TprC) protein antigen. Past immunization experiments conducted with a conserved NH<sub>2</sub>-terminal fragment of this protein (AA 37-273) [309], also shared by the Tp0131 (TprD), Tp0316 (TprF), and Tp0620 (TprI) proteins, have shown high levels of protection upon infectious challenge in the rabbit model. The protein structure for *tprC* was hypothesized in the past by Centurion-Lara *et al.* [131], who predicted a beta-barrel structure with 11 external loops, a structure recently supported by an immunotopological analysis of the *tpr* ortholog of the major outer sheath protein (Msp) of *T. denticola* [310]. Three of these loops, known to be conserved among *T. pallidum* subspecies and strains, are encompassed in the AA 37-273 fragment. The COOH-terminal of *tprC* may also be protective upon infectious challenge of immunized animals, but the high variability detected in this region (Table 13), corresponding

to the predicted L9, L10, and L11 external loops could significantly increase the complexity of the antigen design to achieve a broadly protective vaccine.

In general, of the proteins that still await evaluation as protective antigens belong to two families of paralogs, which include orthologs to the FadL protein of *E. coli* (*tp0548*, *tp0856*, *tp0858*, *tp0859*, and *tp0865*), involved in long-chain fatty acid transport across the outer membrane, and orthologs to the outer membrane channel TolC (*tp0966*, *tp0967*, *tp0968*, and *tp0969*). Additionally, Tp0733 appears to be an OmpW [311] ortholog, and Tp0515 has significant homology with LptD, a protein that guides the assembly of lipopolysaccharide (LPS) at the surface of *E. coli*, and that in *T. pallidum* is likely involved in envelope biogenesis, as the syphilis agent lacks LPS. Interestingly, acceptable structural models for these *T. pallidum* proteins generated by **SWISS-MODEL** and/or **Pyre2** (Figure 18) revealed that a significant proportion of the variable residues are predicted to be surface-exposed. It is possible that such variability might have the dual purpose of affecting the substrate specificity of these proteins, and of contributing to the surface antigenic diversity of the syphilis spirochete as well as of the other subspecies, which could likely contribute to immune evasion during infection, and the ability of these pathogens to infect repeatedly.



**Figure 18.** Structural models. Models are visualized using pdb files generated by SWISS-MODEL (<https://swissmodel.expasy.org/>) or Phyre2 (<http://www.sbg.bio.ic.ac.uk/~phyre2/html/page.cgi?id=index>) the PyMol software. Variable residues mutated are pictured as spheres. (A) Structural models of proteins encoded by genes with elevated number of SNPs (Table 13). (B) Structural models of proteins encoded by genes shown to undergo selection (Table 12). (C) Structural models for proteins encoded by recombinant genes. The yellow color encompasses the region between areas in which recombination events were detected (Table 8). OM=outer membrane. PS=periplasmic space.

## 4. Discussion

In this study, we have examined a large set of nearly complete *T. pallidum* genomes to comprehensively evaluate recombination and selection in a pathogen displaying high clonality. We have followed a rigorous approach by analyzing the results using three different genome references for mapping, thus avoiding problems arising in differential variant calls [103] and three different methods for the detection of recombination, **PIM** developed in our group, **Gubbins** [164], and **ClonalFrameML** [166]. Using **PIM** we were able to identify 12 recombinant genes comprising 19 recombinant regions within them and involving at least 21 different recombination events. The estimates for the detection of recombination by **Gubbins** and **ClonalFrameML** were considerably larger. These differences are most likely because Gubbins evaluates the density of SNPs while concurrently constructing a phylogeny based on the putative point mutations outside of these regions. **ClonalFrameML** uses maximum likelihood inference to simultaneously

detect recombination and account for it in phylogenetic reconstruction. In contrast, **PIM** requires loci to fulfill three criteria to be classified as recombinant. First, the gene is required to have enough phylogenetic signal, which necessitates not only polymorphisms but specifically those that allow for phylogenetic resolution. Secondly, with PIM we check for phylogenetic congruence of the individual gene tree topologies with a reference genome tree. Finally, the last requirement is the presence of a minimum number of homoplastic SNPs in the putative recombination events. As a result, **PIM** is more conservative in the identification of recombination events and regions than Gubbins and ClonalFrame.

Few studies have focused on the identification of recombination in *T. pallidum* subspecies in detail because of the monomorphic nature of the organism, and because no mechanisms for recombination have been identified to date. Nonetheless, these studies have reported examples of recombination in *T. pallidum* and hypothesized that these events may have played a significant role in the evolution of this species [73,77–78,86,92,98,107,126,133,134,175,195,312,313]. Although the methods used to detect recombination are different from our workflow, most of the results previously reported were confirmed by our findings. For instance, recombination in *tp0136*, *tp0326*, *tp0462*, *tp0488*, *tp0548*, and *tp0865*, has also been detected in other studies [23,73,77–78,86,92,98,126,132–134,175,195,284,312,313], but we have identified additional genes (*tp0164*, *tp0179*, *tp0515*, *tp0558*, *tp0967* and *tp0968*) not previously reported as recombinant, probably as a result of the systematic analysis of 75 genomes. On the contrary, loci *tp0117/tp0131*, *tp0119*, *tp0317*, *tp0621*, *tp0856*, and *tp0858*, and the spacers of rRNA operons, which were detected in previous studies [73,74,77,172,173] were not detected as recombinant in our analyses because of the large number of missing positions resulting from mapping of short reads with stringent conditions applied to paralogous/duplicated genes.

Remarkably, we observed only one recombination event (tp0136\_2) between TPA strains, from one strain of the Nichols clade (Seattle 81-4) to the SS14 clade (SW6). All the other recombination events detected correspond to inter-subspecies transfers (TPE/TEN to TPA). Currently, it is not possible to discern whether this pattern is due to the difficulties in detecting intra-subspecific transfers, in light of the low levels of genetic variation in this species, or to the result of mechanisms that favor the transfer and incorporation of foreign material from a different subspecies. This might be the case for those genes in which only the recombinant regions have evolved under positive selection, such as *tp0179*, and *tp0326* and *tp0968*, but not all the recombinant regions do so [79,144,314].

For the two *tpr* genes that could be analyzed, a donor for the variable regions could not be identified. None of the 19 recombinant regions (Table 8) was identified as donor for the variable regions in the *tp0897* gene (Table 9). Variation in this gene accrues much more rapidly than in any other portion of the *T. pallidum* genome [79,144,315]. Hence, it is likely that gene conversion, the mechanism generating variation in *tp0897* is not the mechanism involved in the recombination events in the rest of the genome. In this regard, it is relevant to remark that all but one recombination event corresponds to inter-subspecies transfers, likely occurring much earlier than the continuously generated variability in *tp0897*. We have also identified a putative recombination event in *tp1031* involving a donor sequence of unknown origin, but very likely from the *Treponema* genus, as revealed by the identity in the constant positions (110/128) in the recombinant region (Supplementary Figure 14). The contribution of other donors to the genetic variation within and between TPA lineages remains to be explored.

Inter-subspecies transfers are particularly striking given that TPE and TEN are, to date, geographically restricted to specific regions. Furthermore, no evidence for co-infection has been found in humans so far. One possibility is that recombination occurred in a host other than the human species, for example in non-human

primates. However, the observation of at least one recombination event within the TPA clades Nichols and SS14 suggests that co-infection in humans is possible. These two clades are estimated to have shared their most recent common ancestor in the 18<sup>th</sup> century [149], and all strains so far identified were found in humans. Interestingly, in a recent investigation of medieval skeletons, Majander *et al.* [92] revealed the existence of a previously unknown *T. pallidum* lineage present in Europe probably until medieval times or even later. The historical geographical distribution of *T. pallidum* subspecies is not known but we cannot discard their possible coexistence thus providing ecological and temporal opportunities for coinfection and recombination in human hosts.

The absence of recent recombination events in *T. pallidum* has important implications for the design and use of MLST schemes in these subspecies. If there is only marginal or no recombination, then it is possible to use the genes involved in recombination events in an MLST scheme because most alleles will result from new mutations. The genes *tp0136* and *tp0548*, detected as recombinant in this work, are included in the recently proposed schemes for TPA [55,313] and TPE [60]. The former scheme also includes genes *tp0462* and *tp0865* whereas *tp0326* is incorporated in the scheme for TPE. Clearly, more research on the ecology and natural history of current and past *T. pallidum* strains is necessary to answer these questions.

We have observed a close relationship between recombination and selection. All the recombinant genes identified here display strong signals of either positive or purifying selection. Of the seven genes with evidence of positive selection, six had  $\omega$  values above 2. Only two of the 14 positively selected, non-recombinant genes in the rest of the genome had also  $\omega$  larger than 2, although **Hyphy** did not identify them to evolve under positive selection. Notably, a strong purifying selection is apparently acting on the recombinant portion of *tp0558* (Table 10), similar to the

non-recombinant portion of *tp0164* and many other genes in the *T. pallidum* genome.

The genes for which we found evidence of recombination as well as selection are functionally important. Most of these genes encode proteins that reside at the host-pathogen interface (Table 8, Table 12 and Table 13). This result is congruent with the results obtained in previous studies [78,173,182] in that the variation present in these genes maintained by the selective forces that act on them contributes significantly to the evolution of *T. pallidum*. Although lack of a culture system for *T. pallidum* has prevented experimental confirmation of most inferred protein functions for this species, these genes have been suggested as potentially involved in virulence, with an important role in the defense of the pathogen against the host and the evasion of the immune system [173]. These findings suggest that human hosts' selective pressures drive the diversity of TPA integral outer membrane proteins. The genes coding for OMPs of *T. pallidum* detected under positive selection are pivotal in ensuring and maintaining pathogen fitness and pathogenicity for humans and have important implications for the selection of vaccine candidates and the design of a broadly protective syphilis vaccine (Figure 18).

Our study shows the critical role that recombination and selection play in generating diversity in those genes most critical in the host-pathogen interactions [316]. These processes are known to play a key role in the emergence and adaptive evolution of many pathogens. Recent analyses of whole genome sequences of monomorphic bacteria have also revealed that these processes have been important in the initial stages of speciation, usually along with adaptation to a new niche, as it is the case of *Mycobacterium tuberculosis* [178], or *Vibrio cholerae* [317-318]. These cases are examples of the clonal expansion model from a panmictic pool [319], in which adaptation to a new niche, resulting in a successful spread, might be mediated by the introduction of variation through recombination followed by



the action of natural selection which contributes to the maintenance of the changes and the fixation of the carrier alleles in recombinant genes. Our results indicate that this might have been also the case in the early evolution of TPA and its more recent epidemic spread.

Our results point to a significant role of recombination and selection in the evolution and emergence of the syphilis agent as a human pathogen. However, several questions remain to be answered, including, which molecular processes are responsible for recombination, what is the relevance of intraspecies recombination, and, if this is an important phenomenon driving the evolution of this pathogen, how to improve the current methods available for its detection. Additional questions concern the frequency of co-infections, where they occur, and which subspecies/lineages are usually involved. The constantly increasing availability of whole genome sequences along with advances in the *in vitro* culturing and genetic manipulation [68] of *T. pallidum* will likely help in shedding light on all these pivotal questions.



## — CHAPTER 2 —

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



This chapter has been published as a preprint:

Akgül G., Pla-Diaz, M., Molak, M., du Plessis, L., Panagiotopoulou, H., Doan, K., Bogdanowicz, W., Dąbrowski, P., Oziębłowski, M., Grzelak, J., Arora, N., González-Candelas, F., Majander, K.; Schuenemann, J. V. (2023). Inferring patterns of recombination and divergence with ancient and modern treponemal genomes. *BioRxiv*, 2023-02. doi: <https://doi.org/10.1101/2023.02.08.526988>



## **Chapter 2: “Inferring patterns of recombination and divergence with ancient and modern treponemal genomes”**

### **1. Background**

Treponemes have occasionally been observed to adopt different transmission routes and clinical etiology than expected from the causative agent [11,12]. Whether these events necessitate genetic exchanges between the subspecies or represent responses to contemporary environments is yet unknown. However, they demonstrate the versatile adaptive behavior of treponemal strains. The increasing number of draft genomes available has yielded new findings on the roles of recombination [73,108,175,320] and natural selection [173,183]. 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

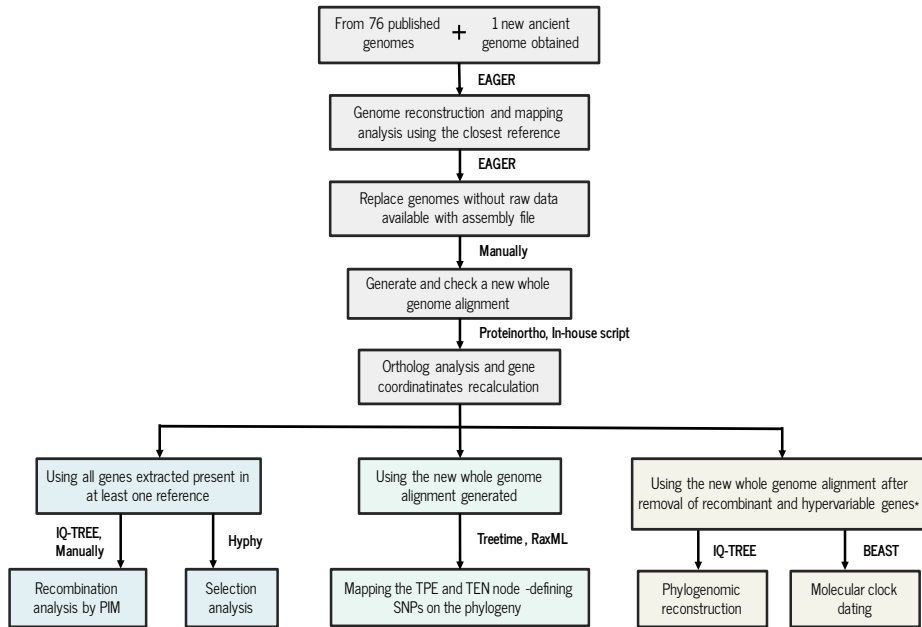
therefore of paramount importance to achieving accuracy of genomic and phylogenetic analyses.

In this chapter, we present a new 17th century genome “W86” of a syphilis-causing strain, sequenced at 35X coverage. We have analyzed the W86 genome together with a set of 76 modern and ancient genomes from the three *T. pallidum* subspecies. As an optimized mapping approach, we employ four different reference genomes and map each genome to its closest reference [103]. 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 the 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 this bacteria.



## 2. Material and methods

A summary of the workflow used in the genomic and phylogenomic analysis of the 77 *T. pallidum* genomes is shown in Figure 19.



**Figure 19.** 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.

### 2.1 Sample recollection

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.

### 2.2 Dataset selection

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 historical ancient draft genomes selected were those with a minimum of 5X coverage among those published at the start of this study in early 2021 [92,98,253,254]. 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** [101] analyzing C to T deamination at the terminal base of the DNA fragment.

### 2.3 Read processing

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 1.92.55 [101], as also done in previous studies [78,92,101]. After processing the de-multiplexed sequencing reads, sample sequencing quality was analyzed with **FastQC** 0.11.5 [321]. Following processing by **AdapterRemoval** 2.2.1a [322], the mapping was carried out using **CircularMapper** 1.0 [101], with default BWA (-l 32, -n 0.04, q 37) parameters [323] 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** [324] was applied to remove duplicate reads and **DamageProfiler** 0.3.12 [265] was utilized to estimate the DNA damage parameters [325]. Indel realignments were performed using **GATK** 3.6 [303] and SNPs for the resulting mappings were called using **GATK UnifiedGenotyper** with the following parameters: -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** 0.95 [326].

### 2.4 Multiple genome alignment generation

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 genome [103]. 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** 1.6.10 [277], 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. The details of each genome employed are details in Supplementary Table 10.

An orthology analysis was carried out with **Proteinortho** 6.0b [327] 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 1-2).

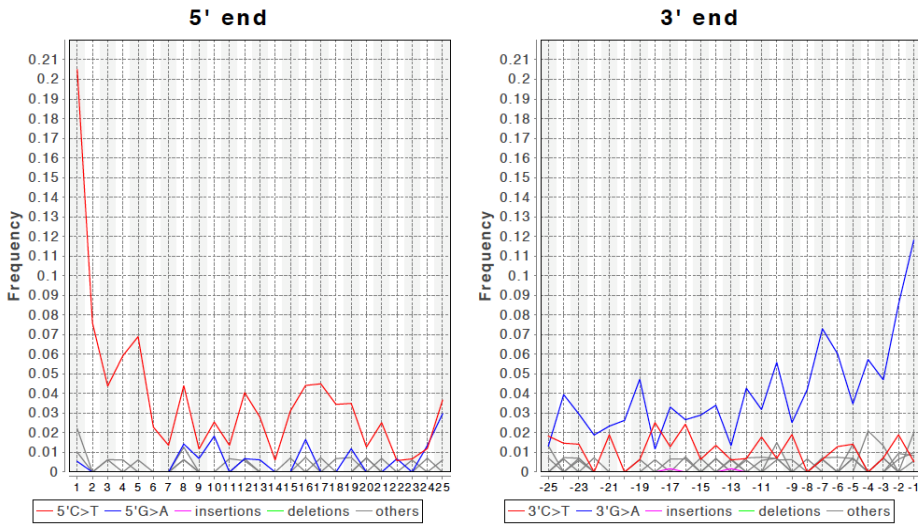
### **3. Results**

#### **3.1 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 [328]. 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 1-2.

#### **3.2 Pathogen screening for the new historical sample**

The W86 sample was subjected to a screening procedure using direct shotgun sequencing [78,92,329]. 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 [330,331] (Figure 20).



**Figure 20.** Damage profile obtained by MapDamage program 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.

Following these screening and authentication steps, genome-wide enrichment for *T. pallidum* DNA [78,92,329] and high throughput sequencing were conducted. The resulting 87.8 million raw reads were merged and duplicate reads were removed.

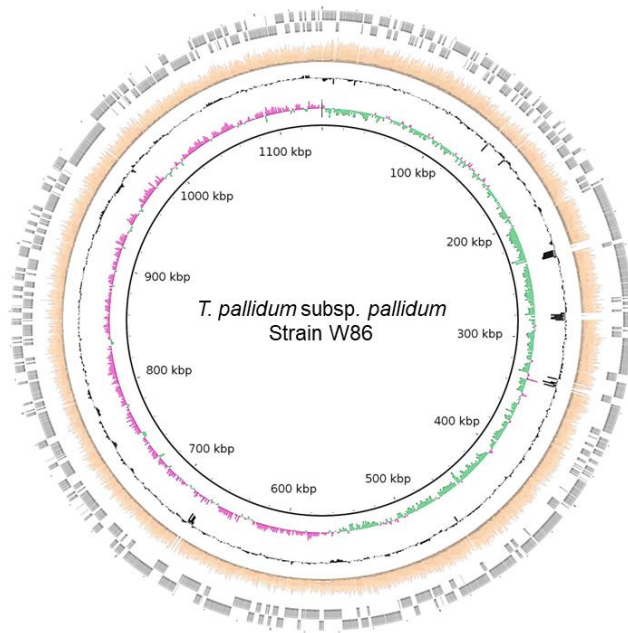
### 3.3 New ancient genome reconstruction and multiple reference-based genome alignment

As it is known that SNP calling is dependent on the choice of reference genome used for mapping [103], 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 BosniaA 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 (Supplementary Figure 18). 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 reference bases covered with a minimum of 3X and a median depth coverage of 35X (Supplementary Table 10 and Figure 21). Variant calling resulted in the identification of 163 SNPs.



**Figure 21.** 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 (grey) shows the direction of protein-coding regions according to the annotation of the SS14 reference genome (CP000805.1): forward, outermost circle.

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 with 4,822 SNPs detected.

To identify the location of genes in this multiple genome alignment, we conducted an orthology inference analysis with the four reference genomes. This analysis revealed a total of 1,161 genes, 24 of which were not shared across all four references (Supplementary Table 11). The genomic coordinates for the individual genes in the multiple genome alignment are detailed in Supplementary Table 12.

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.

### **3.3 Results of the PIM procedure**

After selecting 317 genes with more than 3 SNPs present in at least one reference genome (Supplementary Table 13), a likelihood mapping [273] (LM) test was performed to ascertain which genes had some phylogenetic signal. From the 317 genes, there were 53 genes with too many 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 14) and the remaining genes were discarded. Next, we tested the phylogenetic congruence between trees, comparing the tree obtained from each retained gene and the provisional reference tree obtained from the whole 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 15).

### **3.4 Recombinant events detected**

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 15 and Supplementary Figures 19-36). Six of these recombinant genes were novel compared to previously published results [78,92,108]. The average length of the recombinant regions was 469 bp, with a minimum length of 5 bp and a maximum of 2,097 bp. In addition to the identification of recombination, we also classified 11 of 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 15.** 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. 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. 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
<i>tp0131</i>	1	152624	153390	766	333	Gauthier, LMNP-1, CDC_2575 → Nichols, Chicago, BAL3, BAL73, NIC2	Modern
<i>tp0136</i>	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

## Chapter 2

<i>tp0164</i>	1*	187135	187320	185	5	TEN/TPE → Seattle 81-4, CW86, CW59, NE20, 94A, 94B	Modern
<i>tp0179</i>	1*	198183	198571	388	9	TEN/TPE→ Nichols-lineage clade, W86, PD28	Ancient
<i>tp0326</i>	1	346656	348753	2097	59	TEN → SS14-lineage clade, excluding MexicoA and syphilis ancient genomes	Modern
<i>tp0346</i>	1*	373266	373282	16	3	TEN/TPE → 94A, 94B	Ancient
<i>tp0462</i>	1	493382	494407	1025	55	TEN/TPE → CW86, Seattle86, NE20, CW59	Modern
<i>tp0488</i>	1*	524108	524906	798	54	TPE/TEN → MexicoA, W86	Ancient
<i>tp0515</i>	1	557164	559063	1899	24	TEN/TPE → Nichols-lineage clade	Ancient
<i>tp0548</i>	1*	596164	596828	664	57	TEN/TPE → Nichols-lineage clade, W86	Ancient
<i>tp0558</i>	1	607533	607953	420	4	TEN/TPE → SS14-lineage clade and syphilis ancient genomes	Ancient
<i>tp0621</i>	1	676649	676897	248	137	TEN/TPE → Seattle 81-4	Modern
	2	677030	677215	185	125	TEN/TPE→ Seattle 81-4	Modern
<i>tp0859</i>	1	939420	939442	22	11	External sources →TPE	Ancient
<i>tp0865</i>	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
<i>tp0896</i>	1*	977297	977302	5	4	All primate strains → NL14, CW59	Modern
<i>tp0967</i>	1	1052520	1053803	1283	22	TEN/TPE → Seattle 81-4	Modern
<i>tp0968</i>	1	1053915	1055118	1203	33	TEN → Seattle 81-4	Modern
<i>tp1031</i>	1	1128223	1128350	127	19	External sources →SS14-lineage clade	Modern

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 15, Supplementary Figures 19-36). Interestingly, the sequences corresponding to the putative recombination regions detected in two genes, *tp0859* and *tp1031* (Table 15), 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 15).

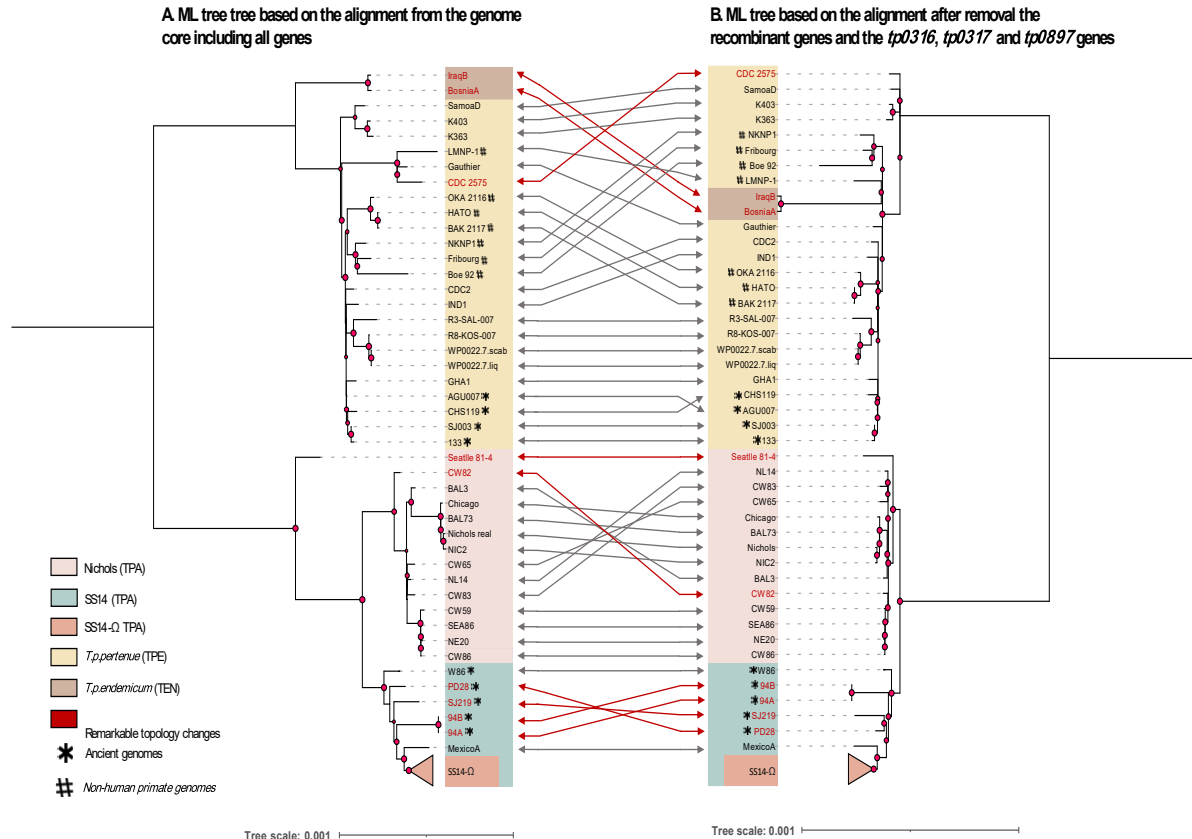
Notably, inclusion of all ancient genomes sequenced to date in our study allowed us to explore their role in recombination and their impact on current patterns of *T. pallidum*'s genetic diversity. Overall, we found eight recombination regions and events involving ancient genomes, four of these including W86, the new genome sequenced. One of these genes, *tp0488*, was found in previous studies to have an unusual sequence in the strain MexicoA [23,108,133], 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. 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.

### 3.5 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). The resulting alignment encompassed 1,107,881 bp with 3,221 SNPs. 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 22.

## Chapter 2



**Figure 22.** 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. Bootstrap support values higher than 70% are indicated by red circles, with circle size proportional to bootstrap support percentage.

Based on our findings, and building on previous phylogenetic classifications and nomenclature of the SS14-lineage [78,107], we defined the SS14- $\Omega$  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 and widespread use of antibiotics [78,107]. 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 22). 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 22).

By contrast, we found several topological incongruences between the two maximum likelihood trees obtained (Figure 22). 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, changed after removal of the aforementioned 21 genes, placing it at a basal position within the Nichols lineage. A similar topological change occurred 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 was included in a subclade with strains LMNP-1 and Gauthier in the whole genome-based tree but occupied 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 involved the TEN genomes. In this phylogeny, both TEN genomes were nested within the TPE clade with high support and did not form a separate sister clade, which was the case in the phylogeny derived from the multiple genome alignment.

### 3.6 Different possible scenarios for the detected recombination events

It is important to note that, although the TPE and TPA ancient genomes are included in the corresponding clades as possible donors or recipients in some recombination events, they often lack data for the SNPs defining recombination events (see Supplementary Figures 19-36). Moreover, the identification of TPE/TEN as putative donors of a genomic transfer does not necessarily mean that the source of the recombination event was their MRCA. As TPE and TEN share identical SNPs, it is not possible to ascertain the exact phylogenetic location of the donor sample(s) [108].

Supplementary Figures 19-36 show, for each recombinant locus, the inferred recombination events and the SNP alignments that support them. The recombination events were mapped onto the maximum likelihood tree obtained from the whole genome alignment (Figure 22) because this tree generally provided a better framework to represent all the genomic transfers as fewer events were needed to explain the recombinant regions detected. Nevertheless, we are aware that the phylogeny of these strains is not fully resolved and that the global relationships might have changed throughout evolution as different recombination events occurred.

We detected 15 genes with a single recombinant region or event, and three genes with more than one region: *tp0136*, with 7 recombinant regions, and *tp0621* and *tp0865*, with two recombinant regions each (Table 15). However, based on the phylogenetic reconstruction used in these representations, more than one genomic transfer are necessary to explain the recombinant regions detected in seven genes (*tp0136*, *tp0164*, *tp0179*, *tp0488*, *tp0548*, *tp0865* and *tp0896*).

The different transfer events that might explain the recombining regions detected in those seven genes are detailed next, with the exception of *tp0865*, whose two recombinant regions were explained in detail in Chapter 1 and Pla-Díaz *et al.* [108].

We will consider only most parsimonious alternatives, although others, involving more transfer events, might also explain the observed distribution of variants in these loci.

Two different transfers might explain the recombination events detected in *tp0164*, both with a donor in TPE/TEN. One event had some modern TPA strains in the Nichols clade as recipients and the syphilis ancient genomes 94A and 94B were receptors of the other transfer event. All the alternative explanations would involve one single transfer from the same source to the common ancestor of all ancient and modern TPA strains and several independent reversals.

A similar explanation applies to genes *tp0179* and *tp0548*, with three and two different transfers, respectively, as the most parsimonious scenario for the recombinant regions detected in those genes. For both loci, the Nichols clade of TPA is detected as the putative receptor of a transfer event from TPE/TEN, in addition to the ancient genomes W86 and PD28, for *tp0179*, and just W86, for *tp0548*. This is also valid for locus *tp0488*, in which two different transfers from TPE/TEN to MexicoA, a modern strain in the SS14 clade of TPA, and to W86, an ancient genome, might also explain the recombination regions detected in the recipient genomes.

Two pairs of two different transfers might explain the recombinant region detected in *tp0896* where, interestingly, all primate genomes from the dataset employed in this study are putative donors and two Nichols strains, NL14 and CW59, are the recipients. The alternative transfers arise from the polyphyletic positions of the primate genomes and of the two receiving strains. An alternative phylogeny, with monophyletic clustering of both donors and recipients would certainly be more parsimonious, as only one single transfer would explain the observed pattern of diversity in this gene.



Seven different recombinant regions were detected in the *tp0136* gene. Two different transfers are needed to explain recombination events 2 and 6 (Table 15). Furthermore, a reversion in combination with the action of the natural selection maintaining the variation in the CW82 strain not detected as involved in the recombination events is the most plausible explanation for recombinant region 6. An analogous situation may explain the recombinant regions 3 and 4 of this gene, where again a reversion occurred for the strains CW82 for recombinant region 4, and also in the three primate strains: HATO, OKA-2116 and LMNP-1 for recombinant region 3 maintained by the action of natural selection seems to be the most parsimonious explanation.

Additionally, in the seven genes where a single genomic transfer cannot explain the recombinant regions, they would correspond to homoplasy events, produced independently for some strains without forming monophyletic groups. However, an alternative possibility would be that the variation detected as recombinant regions correspond to reversion events, causing the paraphyletic strain groups observed in the recombinant regions. Moreover, other possible interpretations might also explain the recombination reported in those genes, similarly to the external transfers occurring to the other clades or subspecies as in *tp0859* and *tp1031* genes. Finally, it is also possible that yet undiscovered *T. pallidum* lineages could be the source of this variation, as it has been described for the different ancient human populations [332,333].

### **3.7 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 whole genome-based tree with the multiple genome alignment using TreeTime.

We focused the analysis on the polymorphisms specific to either the TPE or TEN clades, 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 16 (see also Supplementary Tables 16-17). 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 (version 8.2.11), as detailed in Supplementary Table 18.

**Table 16.** 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 positions in the multiple genome alignment are provided. The total number 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
<i>tp0131</i>	152453	154252	372	4	3
<i>tp0136</i>	158086	159573	176	46	21
<i>tp0164</i>	186980	187780	6	0	0
<i>tp0179</i>	197925	199808	11	2	0
<i>tp0316</i>	332459	336618	53	2	3
<i>tp0317</i>	334348	336618	52	1	0
<i>tp0326</i>	346280	348793	60	13	2
<i>tp0346</i>	373237	373938	30	1	0
<i>tp0462</i>	493325	494605	69	0	0
<i>tp0488</i>	523745	526282	79	19	3
<i>tp0515</i>	556216	559191	29	3	0
<i>tp0548</i>	595868	595884	91	28	7
<i>tp0558</i>	607120	608028	8	1	0
<i>tp0621</i>	675919	678195	283	1	0
<i>tp0859</i>	938709	940211	47	26	2
<i>tp0865</i>	946375	947820	67	20	2
<i>tp0896</i>	977201	977353	12	0	1
<i>tp0897</i>	977335	978924	134	6	2
<i>tp0967</i>	1052318	1053883	23	1	0
<i>tp0968</i>	1053880	1055502	35	10	1
<i>tp0131</i>	1126974	1129023	58	2	0

From the total 2,325 SNPs present between TEN and TPE subspecies, 1,695 SNPs (Table 16) 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 16) 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 17).

The marginal probabilities for the ancestral nodes of TPE and TEN subspecies calculated with RaxML (Supplementary Table 18) 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.

### 3.8 Selection analysis

To study the effects of positive selection, from the total of 1,161 genes (Supplementary Table 12), 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 19), all of which have a high number of SNPs in the strains comprising our dataset (Supplementary Table 13). 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* (*tpr1*), 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 [175] as having a modular genetic structure in certain *T. pallidum* 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 [133,175], 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 [284], 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* (*tprIJ*) in all the TPE and TEN strains analyzed.

In addition, other loci (*tp0856*, *tp0856a*, and *tp0858*) also showed a modular structure in previous analyses [175], apart from the putative recombinant gene *tp0136* and *tp0620* (*tprI*), both previously described. All 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

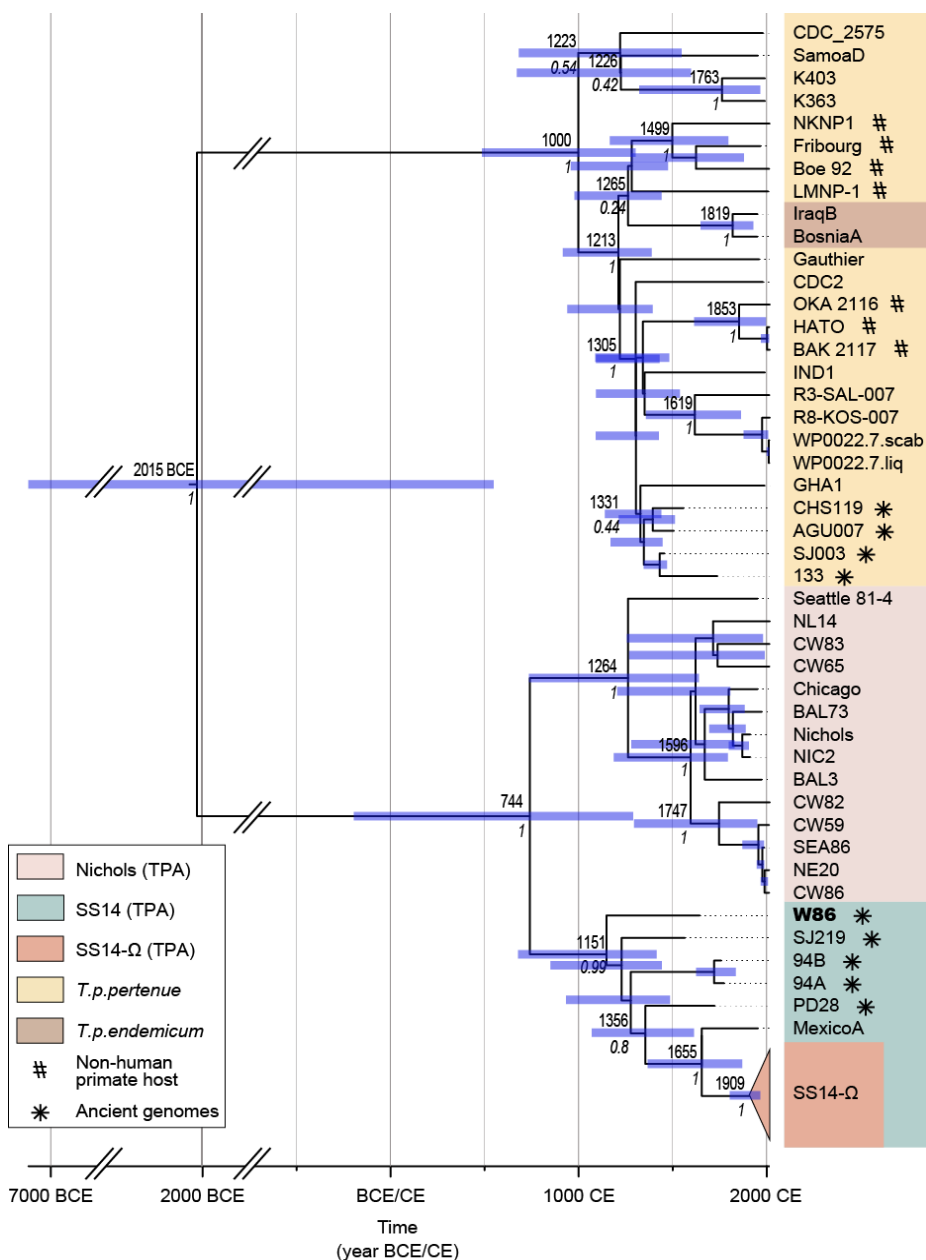
associated with the branch leading to the MRCA of the historical genomes. Specifically, only the branches leading to the OKA2116 and HATO strains for the *tp0134* gene and the branches leading to the WP00227liq and WP00227scab strains for the *tp0833* gene exhibited signal of positive selection.

As we have observed a close relationship between recombination and selection, we also examined the functional roles of the proteins encoded by the recombinant and additional genes found to be under positive selection, according to Uniprot and a literature search (detailed in Supplementary Table 20). Despite some gene proteins having an unknown function, most 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.

### **3.9 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 (Figure 23). 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 17. 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 [92]. Consistent with previous studies [83,92], we dated the tMRCA of all TPA lineages between 190 BCE and 1290 CE (median 745 CE), 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- $\Omega$  to 1805 - 1965 CE (median 1910 CE). The median evolutionary rate was estimated to  $6.77 \times 10^{-5}$  substitutions/SNP-site/year (95% HPD interval:  $3.78 - 9.95 \times 10^{-5}$ ), which corresponds to  $1.91 \times 10^{-7}$  substitutions/site/year (95% HPD

interval:  $1.07 - 2.81 \times 10^{-7}$ ) for the full *Treponema* genome with hypervariable and recombining genes excluded.



**Figure 23.** Maximum clade credibility (MCC) tree of the dataset consisting of 66 modern and 9 historical genomes estimated in BEAST [149] 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 37.



**Table 17.** Time to the most recent common ancestor and 95% HPD interval (Highest Posterior Density, HPD) estimated by molecular clock dating for the major clades.

Sample/Clade	Posterior Date Estimate	
	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

#### 4. Discussion

In this study, we report the sequencing and subsequent analysis of a new, high coverage (35X) ancient syphilis genome, W86, integrated in a 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 better 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 [103,334,335]. 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, specially for ancient genomes, was clearly demonstrated according to the different mapping results obtained for the new ancient genome W86 (Supplementary Figure 18), 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 [73,86,108,133,146,173,175,182,183,284] 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 [78,92,108,270], 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 [78,92,108,270] but detected as

recombinant in other studies [73,77,129,175,314] 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 [92,98,253] are integrated and analyzed together for the first time, with a further inclusion of several genomes sampled from non-human primates. Nevertheless, despite detecting five new recombinant genes not previously detected by PIM but reported in other studies [73,77,129,175,314], 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, as it is often the case of *T. pallidum* 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 [92,98,253]. 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 [92,203]. 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) detailed in Table 15.

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 15, Supplementary Figures 19-36), 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. Furthermore, most of the genes detected as recombinant or under positive selection pressure 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 20).

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 22). 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 important 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 [80,336]. As some of the latest findings on historical yaws genomes suggest [253,254], 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 that 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.

## — CHAPTER 3 —

“Redefining the treponemal history through pre-Columbian  
ancient genomes from Brazil”





## **Chapter 3: “Redefining the treponemal history through pre-Columbian ancient genomes from Brazil”**

### **1. Background**

As similar manifestations can appear in syphilis, yaws and bejel [33,337,338], the diagnostic distinction of the treponematoses at the subspecies level may be unreliable, especially in developing countries suffering from limited analytic tools and facilities [3,4,339]. Diagnoses from historical cases are even more difficult to make: although treponematoses often leave lesions and other pathological alterations in bones, securely identifying a treponemal infection based on these pathologies is challenging. Furthermore, the characteristic lesions appear in only approximately 5-30% of all advanced treponematoses, [340,341], which has probably resulted in underestimating past prevalence of treponematoses. Attempts have been made to identify types of treponemal infections from paleopathological bone lesions, but these assignments remain ambiguous at best, and require DNA evidence to confirm the diagnosis to the subspecies level (see Section 1.7 of the present thesis).

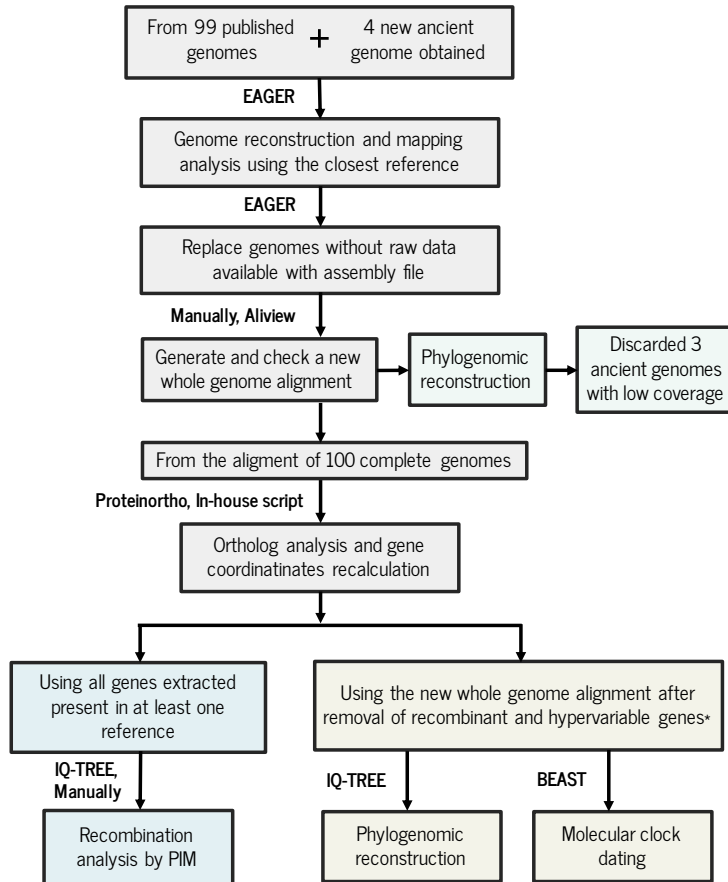
Treponemal outbreaks are likely to spread rapidly and aggressively in a naïve population. The European epidemic at the end of the 15th century could represent one such event, rising in the wake of the transatlantic contact. The later known historical outbreaks on both continents, though occasionally severe, have remained locally limited [342–344]. Furthermore, according to the observations in the historical medical reports, syphilis appears to have assumed a less virulent form in the course of a few decades after the initial outbreak [211,345], which would have likely benefited its spread as a sexually transmitted disease and functioned as an adjustment to a new host population. Evolutionary adaptations responsible for the virulence and the many manifestations of treponemal infections are still poorly understood. The recombination detected both in ancient and modern genomes

[11,12,73,108,346] suggests that co-infections in the host pool are an important source of acquired genetic variation in treponemes, although it is unclear to what extent recombinant genes are responsible for the characteristic symptoms associated with the three subspecies. Furthermore, it is challenging to show which treponemal agents concurrently afflicted past human populations and thus had the opportunity to conduct the horizontal gene exchange through coinfection. As treponemes appear to possess the ability to adjust to various environments [347], and have apparently occupied geographical regions outside their present distributions, only unequivocally pre-Columbian treponemal DNA evidence can shed light on the origins of syphilis, and possibly unravel entirely unforeseen aspects of all treponemes' evolutionary history.

In this chapter, we present four reconstructed genomes of *T. pallidum* from a nearly 2'000 year old archaeological burial site in the Laguna region of Santa Catarina on the Brazilian coast, South America. A comprehensive set of 99 samples was chosen from the Jabuticabeira II sambaqui (burial shellmound), where previous osteological analyses had revealed several infection-related pathologies, including four cases strongly suggestive of treponemal infections [68]. With confidently dated stratigraphy and vast bioarchaeological studies, the site was considered optimal to test the hypotheses of pre-Columbian treponematoses in the ancient coastal Brazilian populations [68]. The genomes reconstructed in this study constitute the first firmly established pre-Columbian evidence of treponemal disease in the New World. Exceptionally well preserved DNA is recovered from one of the samples, yielding over 30-fold coverage genome that falls phylogenetically basal to the modern diversity of the bejel-causing subspecies, *T. pallidum* subsp. *endemicum*. This genome also enables in-depth analyses on divergence time estimates and evolutionary dynamics between the *T. pallidum* subspecies, especially illuminating for the so-far poorly studied bejel clade.

## 2. Material and methods

A summary of the workflow used in the genomic and phylogenomic analysis of the 99 *T. pallidum* genomes is shown in Figure 24.



**Figure 24.** Analysis workflow for the genomic and phylogenomic analysis of the 99 previously published *T. pallidum* genomes and one four historical genomes obtained in this study. The hypervariable genes indicated by \* are *tp0897* and *tp0316*.

### 2.1 Dataset selection

We assembled a genomic dataset comprising 98 publicly available *T. pallidum* genomes (8 TEN, 30 TPE and 60 TPA) from previously published studies (including 8 ancient genomes), and the newly generated ZH1540 genome (Supplementary Table 21). The genomes represent the genetic variation of the three

known subspecies of *T. pallidum* (TPA, TPE and TEN) available by December 2022, and were selected with a focus on TEN and TPE, because of their proximity to the new ancient genome classified as TEN.

We selected all 8 publicly available TEN genomes, all of which have more than 99.4% genome coverage, with the exception of C77 (81.4%) [2]. We selected 30 TPE genomes. To represent each lineage/sublineage, we selected at least one genome, preferring the ones with the highest sequencing depth and genome coverage. All included TPE genomes have more than 95.3% genome coverage, except the four ancient TPE genomes: SJN003, AGU007, 133, and CHS119, displaying 97.4%, 92.7%, 57% and 62% genome coverage, respectively. Furthermore, 60 TPA genomes from the major lineages and sublineages described in previous studies were included. All of these genomes had more than 90% coverage, except the four ancient genomes, PD28, W86, SJ219 and 94B, all of which have the genome coverage of 30% or more. All the genomes in the dataset are separated from each other by at least 5 SNPs. The TPA strain Seattle-81 was excluded from the final dataset, due to the mutations likely accumulated during extensive passaging in rabbits that can cause ambiguous placement in phylogenies [78,92,108].

The raw data and/or assembly files for each genome in our dataset were downloaded from the public databases: European Nucleotide Archive (ENA)[266] and National Center for Biotechnology Information (NCBI) [348]. Accession numbers are given in Supplementary Table 21.

### **2. 2 Read processing and multiple reference-based genome alignment generation**

To reconstruct the individual genomes from the raw data, we carried out raw read quality control and preprocessing, removing duplicates, variant calling and filtering. After processing the de-multiplexed sequencing reads, sample sequencing

quality was analyzed with **FastQC** 0.11.9 [321], filtering reads with a QC value <25. Following processing by **Cutadapt** 4.1 [322], in order to reduce the reference bias, and improve the posterior phylogenetic inference and assignment [103], the genome reference selection for mapping each sample was determined according to the results from the original manuscript where the genomes were published (see Supplementary Table 21). The mapping was carried out by **BWA mem** [323] using parameters: -k 19, -r 2.5. Four reference genomes were used; the well-studied TEN and TPE genomes BosniaA (NZ\_CP007548.1) and CDC2 (NC\_016848.1), as well as the Nichols (NC\_021490.2) and SS14 (NC\_010741.1) genomes, representing the two main lineages within TPA. However, for the new ancient samples obtained here, genomes for each sample were reconstructed by mapping to three high-quality reference genomes, representing the three *T. pallidum* subspecies (CDC2, BosniaA and Nichols).

**CleanSam**, from **Picard Toolkit** 2.18.29 (<http://broadinstitute.github.io/picard>), was used to clean the provided SAM/BAM files. Duplicate reads were removed using **MarkDuplicates**, from **Picard toolkit** 2.18.29 [324]. **AddOrReplaceReadGroups**, from **Picard Toolkit** 2.18.29, was used to assign all the reads in a file to a single new read-group before using **mapDamage** 2.2.0-86-g81d0aca [265] to estimate the DNA damage parameters and rescale quality scores of likely damaged positions in the reads (using parameter: --rescale).

After generating a text pileup output for the BAM files with the **mpileup** tool from **Samtools** 1.7 [349], SNPs were called using **VarScan** 2.4.3 [350] (using parameters: -p-value 0.01, -min-reads2 1, -min-coverage 1, -min-freq-for-hom, 0.4 -min-var-freq 0.05, -output-vcf 1). Next, a SNP filtering was also carried out with **VarScan**, using for the modern samples parameters: -p-value 0.01, -min-reads2, 5 -min-coverage 10, -min-avg-qual 30 -min-freq-for-hom 0.4, -min-var-freq 0.9, -output-vcf 1; and modifying some parameters for the ancient samples because of their lower read coverage and quality: -p-value 0.01 -min-reads2 3, -min-coverage

5, -min-avg-qual 30, -min-freq-for-hom 0.4, -min-var-freq 0.9 -output-vcf 1. Additionally, all positions with less than 3 mapped reads were masked with **Genomecov** from **Bedtools** 2.26.0 [351] for modern and ancient samples. All steps of genome generation were visualised and manually confirmed with **Tablet** 1.21.02.08[269], checking each SNP one by one and discarding the possible spurious SNPs from the new ancient genome ZH1540. The resulting final sequences were obtained by maskfasta from **Bedtools** 2.26.0.

Additionally, we used tested sequencing and posterior analysis methodologies [2,36] to obtain higher coverage and more reliable modern *T. pallidum* genomes. Where possible, assembly files were obtained rather than raw data (Supplementary Table 21). A multiple reference-based genome alignment for all sequences was generated in MAFFT v7.467 [352]. However, due to the use of different genomic references, regions with low coverage for some genomes, corresponding mostly to *tpr* and *arp* genes, were reviewed and manually aligned with Aliview version 1.25 [268].

**Proteinortho** 6.0b [327] was used to conduct an orthology analysis in order to find orthologous genes in the four reference genomes used [327]. Each gene present in at least one of the four reference genomes had its genomic coordinates determined based on its location in the final merged alignment.

To verify the accuracy of the final multiple genome alignment, and that no protein-coding gene was inadvertently truncated, the protein translations for every gene present in at least one reference genome were compared to the original gff3 files of each of the four references (Supplementary Table 21). The reconstructed ZH1540 genome and its main features were represented graphically using **BRIG** 0.95 [326]. We also represented with BRIG the positions of 60 candidate genes associated with virulence and outlined in previous studies [92,98].

### 3. Results

#### 3.1 Geographical origins and osteological analyses of samples

Altogether, 99 specimens from Jabuticabeira II, both with and without pathologies, were incorporated in the analyses of this study, 37 of which were considered preliminarily positive for treponemal DNA after initial screening (Supplementary Table 22). Four bone samples, from four different individuals, yielded sufficient genomic data for comprehensive analyses.



**Figure 25.** A map showing the location of Jabuticabeira II excavation in the South Coast of Santa Catarina state, Brazil, and the samples ZH1390, ZH1540, ZH1541 and ZH1557 for which genomes were reconstructed.

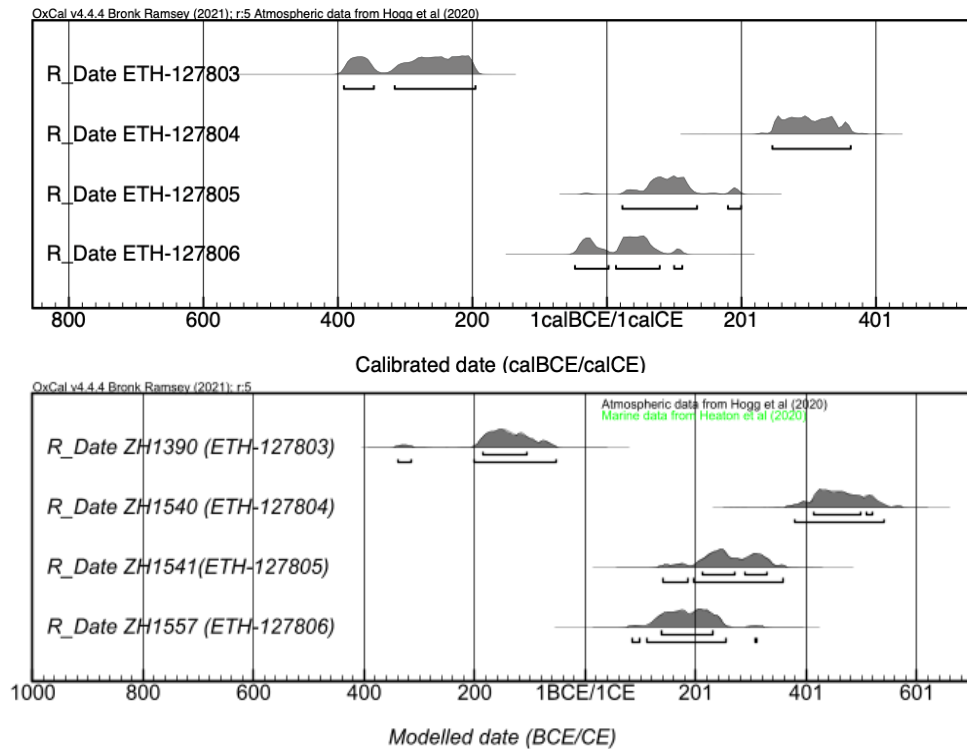
Sample ZH1390 (Table 18 and Figure 25) represented a tibia fragment showing periostitis. Sample ZH1540 came from a set of commingled bones of an incomplete skeleton, namely from a fibula with pathological lesions (Table 18 and Figure 25). Samples ZH1541 and sample ZH1557 originated in long bones without any identified pathologies (Table 18 and Figure 25).

**Table 18. Summary of the sample information and central statistics.** ID:s and molecular sexing used for the four individuals representing the samples yielding the reconstructed genomes. Statistical data for the DNA content in the samples, including the number of raw reads, reads after duplicate removal, average coverage and genomic coverage from 1 to 5 fold per sample, and the final number of SNPs covered for each.

Ind. ID	Arch. ID	Mol. Sex from SG-data	Raw Reads	Mapped Reads (Post-duplicate Removal)	AVG Cov.	% Gen. Cov. 1×	% Gen. Cov. 2×	% Gen. Cov. 3×	% Gen. Cov. 5×	SNPs
ZH1540	FS9-L3-T2	XX	19,661,672	567,158	33.60	99.71	99.65	99.57	99.38	125
ZH1390	41A-L2.05-E4	?	43,980,864	83,348	2.10	9.22	9	3.63	2.31	272
ZH1541	FS3B-L3-T4	XX	99,335,748	122,086	2.67	18.45	18.32	8.66	5.74	215
ZH1557	2B-L6-E3	?	88,186,424	179,285	3.92	19.41	19.25	9.74	6.70	316

All samples were radiocarbon-dated and tested for the marine reservoir effect (Figure 26, Supplementary Table 23 and Supplementary Note 4).





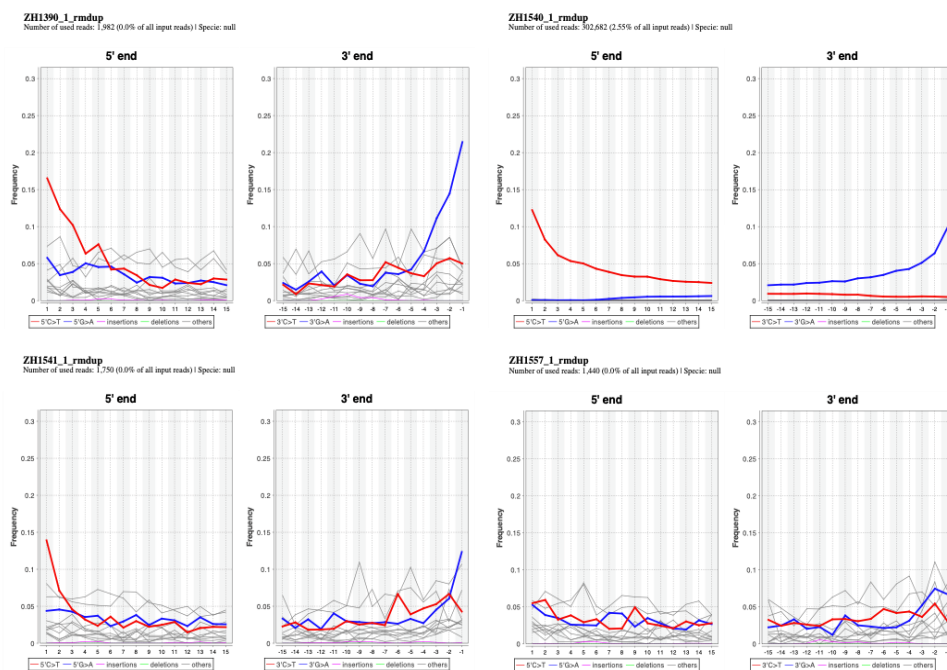
**Figure 26.** Pre- and post-reservoir effect correction (cal) dates of samples used for ancient genome reconstruction. For pre-corrected calibrated curves, ETH-127803 corresponds to the sample ZH1390, ETH-127804 to sample ZH1540, ETH-127805 to ZH1541 and ETH-127805 to ZH1557, respectively. For more information on the modeled dates for marine reservoir effects, see Supplementary Table 23.

### 3. 2 Preliminary pathogen screening and authenticity estimation of ancient DNA

In the initial screening from shotgun sequencing data, 37 out of 99 samples showed between 7 and 133 hits to *Treponema* family taxa in the Kraken database, and were included in the subsequent target enrichment process (Supplementary Table 22). Of these samples, 9 had over 5,000 reads mapping to three *T. pallidum* references (BosniaA, CDC2, Nichols) post-capture, and were thus considered positive for treponemal infection (Supplementary Table 24). For these positive samples, three additional double-stranded libraries were produced for a second round of genome-wide enrichment[92,260]. For detailed methodology, see Methods: Sample

processing. After the additional enrichment, the data from all libraries and both rounds of capture were combined. Four samples, namely ZH1390, ZH1540, ZH1541 and ZH1557, had reads covering 9.2-99.4% of the BosniaA reference genome at 1X, with an average coverage between 2X and 33X (Table 18). These four samples were considered having the most potential for whole-genome reconstruction and downstream analysis.

The authenticity of ancient DNA was confirmed by examining the deamination of bases at the ends of reads: 21%, 10%, 12%, and 7% at the 5' ends and 17% , 12%, 14%, and 6% at the 3' ends for the ZH1390, ZH1540, ZH1541 and ZH1557 samples, respectively (Figure 27).



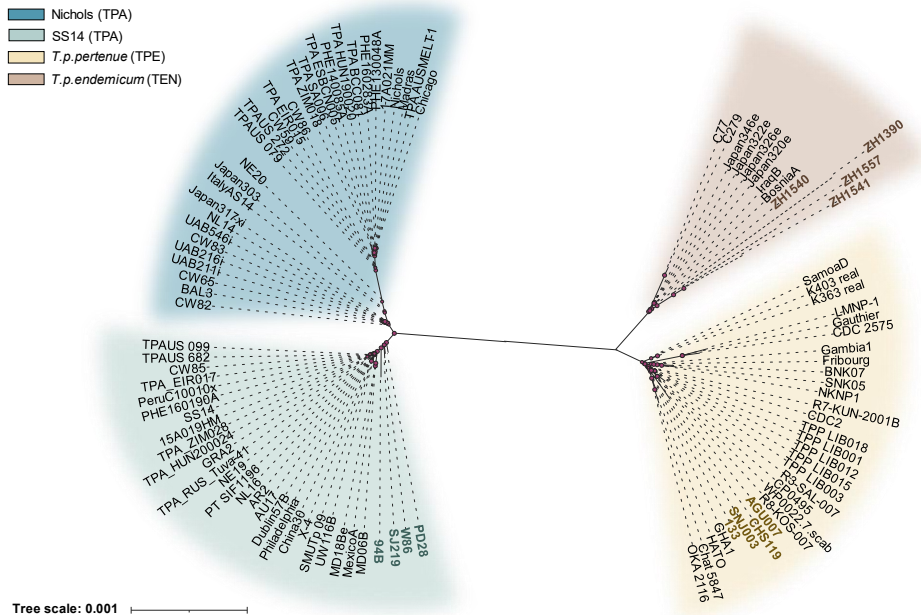
**Figure 27.** Damage profiles for aDNA authentication. Damage profiles obtained with DamageProfiler tool [265] showing the misincorporation patterns induced by age, for each sample that yielded a reconstructed genome. A pattern of cytosine-to-thymine base misincorporation accumulating at the end of the reads is indicative of authentic ancient DNA in the sample.

The samples had an average fragment length ranging from 64 bp to 74 bp [330,353,354] (Table 18). Data from shotgun sequencing was also used for molecular sexing (Table 18). Samples ZH1540 and ZH1541 were consistent with the XX female karyotype, whereas ZH1390 and ZH1557 could not be assigned due to insufficient data. Nonetheless, based on osteological analysis of near-complete skeletons they are most probably males (see Supplementary Note 4).

### 3.3 Genome reconstruction

After high-throughput Illumina sequencing of the enriched DNA from the four selected samples, the resulting 20–100 million raw reads were merged sample-wise and duplicate reads were removed (Table 18). Genomes were reconstructed by mapping each sample to three representative high-quality reference genomes of *T. pallidum* subspecies: CDC2 for TPE, BosniaA for TEN, and Nichols for TPA. For more information. We filtered positions based on read coverage, variant allele frequency, p-value, and base quality, and obtained three different consensus sequences for each sample, each with a different number of covered bases, as well as single nucleotide polymorphisms (SNPs). The samples ZH1390, ZH1541, and ZH1557 had sufficient data to attempt a genome reconstruction and were determined to have the most SNPs in common with the TEN reference, but they were excluded from downstream analyses due to the limited coverage attained for each of them, which made the obtained SNPs less reliable. The sample ZH1540, however, yielded a remarkable 33-fold genomic coverage and was selected for subsequent in-depth analyses.

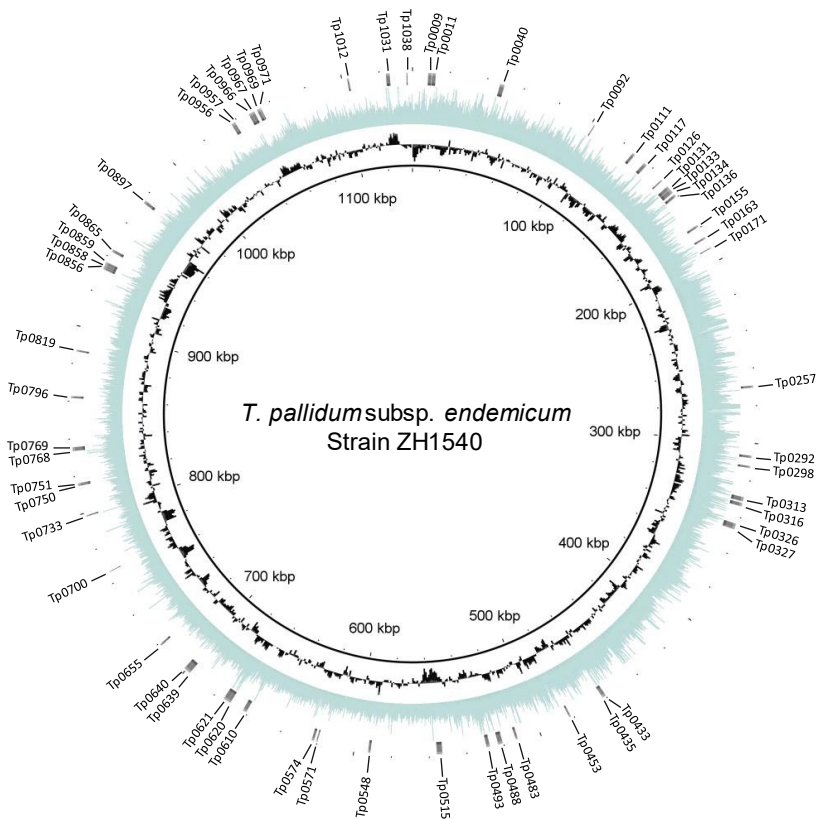
The number of SNPs obtained for each sequence (Table 18) and the subsequent phylogenetic analyses consistently supported a placement of all four samples within the TEN clade (Figure 28). Although the consensus sequences from three samples, ZH1390, ZH1541 and ZH1557, were found to be assigned for TPE (Figure 28), read coverage was below the threshold required for subsequent analyses.



**Figure 28.** A network tree visualization of the modern and ancient *T. pallidum* strains. Brown denotes for TEN clade, yellow for TPE clade, light blue for the SS14 strains of TPA clade and dark blue for the Nichols strains for the TPA clade. The four ancient genomes from this study are marked in bold font.

The final sequence obtained for the ZH1540 sample resulted in 99.38 % coverage with respect to the TEN reference genome (BosniaA), a minimum coverage depth of 5X and a median depth of 34X (Table 18 and Figure 29).

Variant calling resulted in the identification of 123 SNPs, each of which was checked individually (Supplementary Note 5). The 123 SNPs identified for ZH1540 are detailed in Supplementary Table 25. Of the available modern references, the new ancient TEN genome has a difference of 123 SNPs to BosniaA and IraqB samples. However, the number of differing SNPs is much higher compared to the four Japanese TEN genomes (205 SNPs) and the Cuban TEN genomes (504 SNPs).



**Figure 29.** Circular plot of the ZH1540 genome obtained by BRIG 0.95 [326]. Circles indicate, from inside outwards: genomic position, GC content (black) and coverage (blue). The outer rim (grey) shows a set of 60 candidate genes associated with virulence and outlined in previous studies [92,98].

### 3.4 Multiple reference-based genome alignment

The new ancient genome ZH1540 was analyzed together with an additional 98 publicly available genomes, including 8 modern TEN strains, 30 TPE strains (including 9 genomes from primates and 4 ancient genomes), 30 Nichols- and 30 SS14-lineage TPA strains (including 4 ancient genomes) (Supplementary Table 21). Assembly files were available for 33 of these 98 genomes, which were downloaded directly from the public databases European Nucleotide Archive (ENA) and National Center for Biotechnology Information (NCBI). For the remaining 65 genomes, we mapped the raw sequencing data to the closest of four

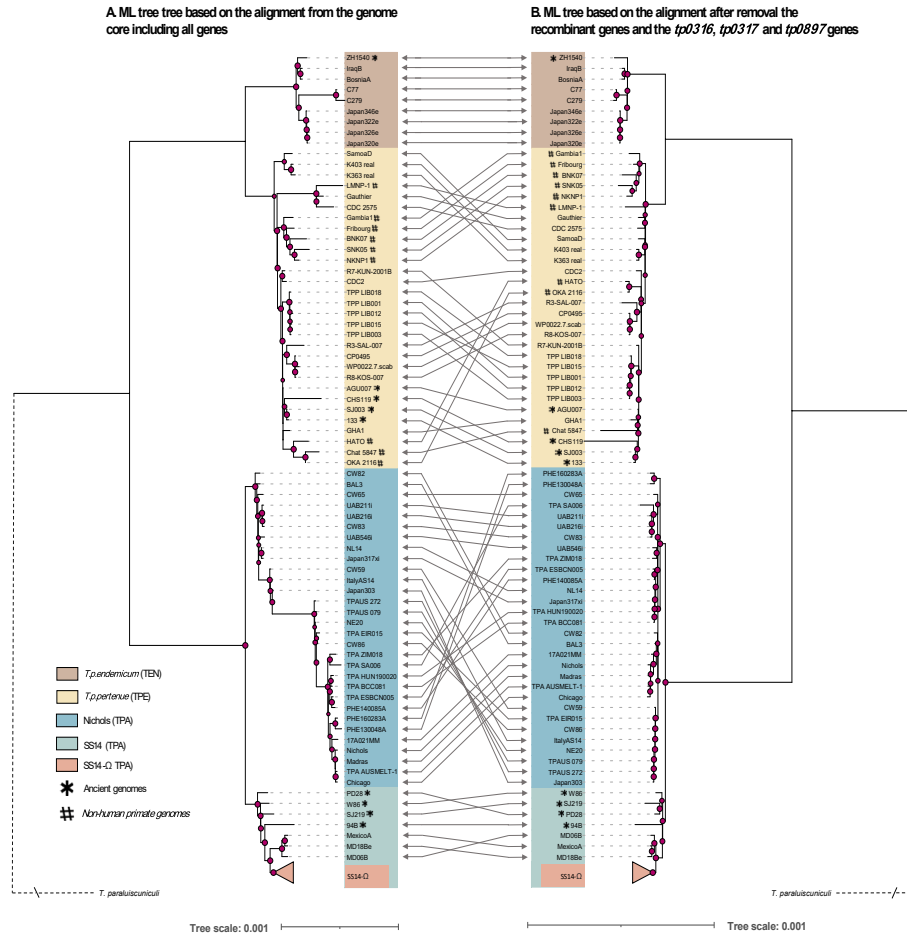
representative reference genomes (CDC2, BosniaA, Nichols, and SS14), to obtain new assembly files. The genome reference selected for each sample was based on the subspecies and/or lineage classification of each sample from the original publications (Supplementary Table 21). A multiple reference-based genome alignment of 98 sequences from several sources was generated according to the previously published methodology [346]. The resulting alignment spanned a total of 1,141,812 nucleotides with 6149 SNPs detected. 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 1,161 genes, 24 of which were not shared across all four references. The genomic coordinates for the individual 1,161 genes in the multiple genome alignment are detailed in Supplementary Table 26.

### **3.5 Phylogenetic analysis and genetic recombination**

A reliable phylogenetic reconstruction required the removal of genomic regions that are not strictly vertically inherited, such as recombinant regions or loci with intra- or intergenic conversion. To this end, we conducted a recombination analysis, applying the phylogenetic incongruence method [108] to the 98-genome dataset generated for this study. We detected 34 recombinant regions across 27 genes, encompassing a total of 957 SNPs (15.56 % of the total SNPs) (Supplementary Tables 27-30). Due to the exclusion of the highly passaged Seattle-81 strain, three of the previously detected recombinant genes were not detected here, while 11 detected genes were novel compared to the previously published results. The average length of the recombinant regions was 368 bp, with a minimum length of 4 bp and a maximum of 2,209 bp. Interestingly, all the recombination events detected here correspond to inter-subspecies transfers with the exception of an intra-subspecies recombination event found in the *tp0117* gene and three additional genes where the putative donors are unidentified external sources.

Subsequently, in order to construct a strictly vertical-inheritance alignment we removed the 27 recombinant genes detected here along with three genes (*tp0316*, *tp0317*, and *tp0897*) that are known to be hypervariable and/or subject to gene conversion[175,284] from the initial alignment (See: Data availability, Methods: Statistical analyses; Phylogenetic analysis). The final recombination-free alignment spanned 1,103,436 bp with 3,718 SNPs. Maximum-likelihood trees were built using both multiple genome alignments (Figure 30 and Supplementary Figures 3-6). In Figure 30, the topologies of the two maximum-likelihood trees with and without the recombinant/hypervariable loci are compared.

# Chapter 3





**Figure 30. (Located in previous page)** Comparison of the topologies of maximum likelihood trees. **A)** ML tree topology with all genes included in the multiple genome alignment. **B)** ML tree topology 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 color-coded according to the legend. Bootstrap support values higher than 70% are indicated by red circles, with circle size proportional to bootstrap support percentage. Subclade SS14-Ω, which includes all SS14-lineage TPA strains except for the ancient TPAs and MD06B, MD18Be and MexicoA genomes, is shown collapsed. Link to the online figure with a higher resolution: <https://drive.google.com/file/d/1uX4Z6yvK15Ot5YYxBFzs9FVKdth-3iMh/view?usp=sharing>

The elimination of non-vertically inherited genes has a significant impact on the reconstruction of the *T. pallidum*'s phylogeny (Figure 30 and Supplementary Figures 38-40). However, the only topological change in the TEN clade is related to the new ancient genome ZH1540, which, after removal of the recombinant and the other three additional genes, occupies a more basal position to the other modern genomes of this subspecies (Supplementary Figures 38-40). Additionally, gene mutations (A2058G and A2059G) related to macrolide antibiotic resistance were assessed. Four geographically clustered modern TEN strains from Japan were shown to possess the A2048G mutation, whereas none of them was present in the new ancient genome ZH1540 (see Supplementary Note 5).

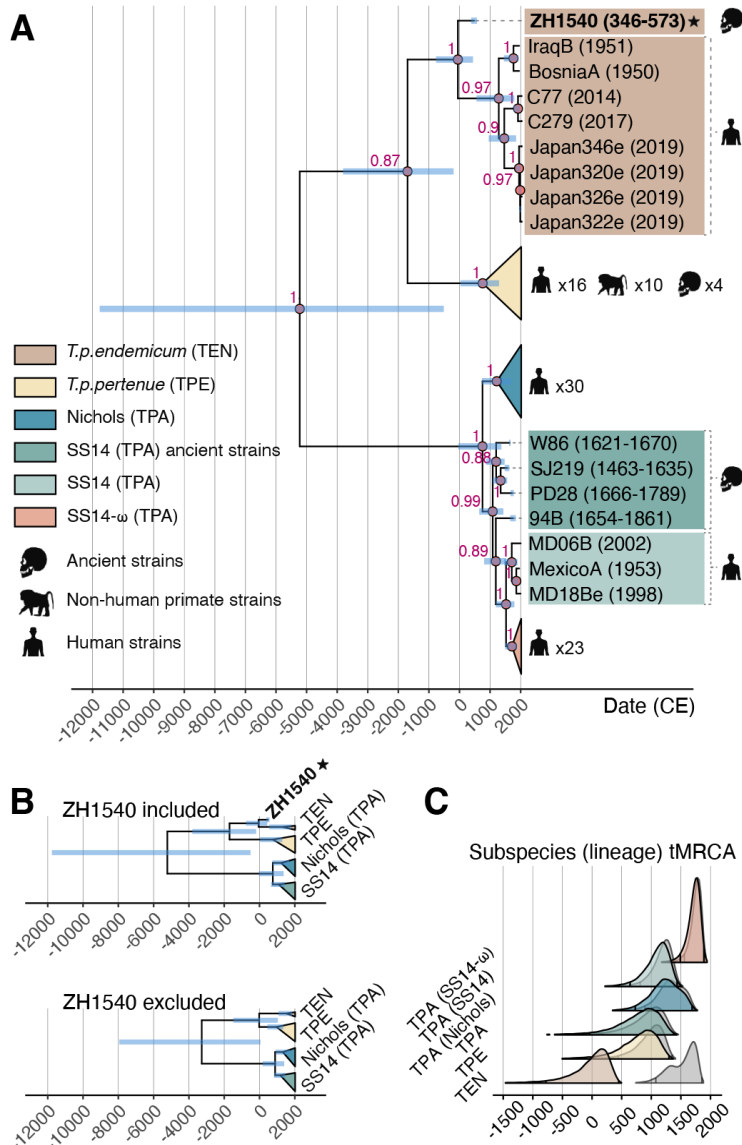
### 3.5 Molecular clock dating

Molecular clock dating was performed on the same dataset as above, with 27 recombinant genes, *tp0316*, *tp0317*, and *tp0897* removed. In the estimated time-calibrated phylogeny, all three subspecies (TEN, TPE and TPA), as well as the SS14 and Nichols lineages of TPA, received high support for forming monophyletic clades (posterior probability >0.97; Figure 31 and Table 19).

**Table 19.** Divergence dates. Times of the most recent common ancestors (tMRCAs, in years CE) and monophyly statistics for the *T. pallidum* subspecies and major lineages within TPA, as estimated by molecular clock dating.

	Including ZH1540 (n=99)			Excluding ZH1540 (n=98)		
	Median	(95% HPD)	Pr(monophyletic)	Median	(95% HPD)	Pr(monophyletic)
<b>TPA (SS14-Ω)</b>	1737.5	(1490.91 - 1888.38)	0.98	1769.07	(1559.16 - 1904.91)	0.98
<b>TPA (SS14)</b>	1127.06	(641.83 - 1436.09)	0.97	1208.11	(812.43 - 1471.93)	0.97
<b>TPA (Nichols)</b>	1237.65	(729.93 - 1689.21)	0.98	1334.36	(894.51 - 1725.94)	0.98
<b>TPA</b>	844.36	(-41.67 - 1375.99)	0.97	975.18	(166.79 - 1402.34)	0.97
<b>TPE</b>	835.12	(27.62 - 1299.16)	0.98	1015.88	(439.89 - 1351.22)	0.97
<b>TEN</b>	47.2	(-779.82 - 449.48)	0.99	1586.82	(1077.29 - 1855.49)	0.99

As in the maximum-likelihood phylogeny, the new ancient genome, ZH1540, occupies a basal position within the TEN clade, with all modern TEN strains forming a monophyletic subclade (posterior probability 0.96; Figure 31).



**Figure 31. Molecular clock dating results.** **A**) Maximum clade credibility (MCC) tree of previously published ancient and modern genomes, and the new ancient genome ZH1540 from this study ( $n = 99$ ). Blue bars indicate the 95% HPD intervals of node ages and red text the posterior probability that a clade is monophyletic (only shown for nodes with posterior probability  $>0.8$ ). **B**) Simplified collapsed MCC tree of the analysis with ZH1540 included (top) and excluded (bottom). **C**) Posterior densities of the times of the most recent ancestors (tMRCAs) of the *T. pallidum* subspecies and major lineages as estimated by the molecular clock dating with ZH1540 included (coloured densities, corresponding to panel B top) and excluded (grey densities, corresponding to panel B bottom). Vertical lines indicate the upper and lower limits of the 95% HPD intervals.

The majority of SS14 strains fall within the previously defined SS14- $\Omega$  subclade [78,92], which also receives high posterior support. Consistent with the molecular clock dating results previously reported [92,346], we find that all historical TPA strains fall basal to all modern SS14 strains and together these form a well-supported, monophyletic clade (posterior probability 0.97). We therefore consider the historical strains to fall within the SS14 clade.

The age of ZH1540, which is parameterised by the radiocarbon dating results, acts as a constraint on the time of the most recent common ancestor (tMRCA) of the TEN clade. The inclusion of one far older sample results in earlier divergence times, with wider credible intervals, for all major clades in the tree (Figure 31 and Supplementary Figures 41-42). This effect is most pronounced for TEN, where the 95% highest posterior density (HPD) interval of the tMRCA stretches from 780 BCE to 449 CE (236-1845 CE for the subclade comprising only modern TEN strains), but is limited to 1077-1855 CE when excluding ZH1540 (Table 19, Supplementary Figures 41-42). For all other major lineages, the effect is more moderate, and while the lower limit of the 95% HPD interval can be several hundred years older when including ZH1540 (around 400 years in the case of TPE), the upper limit is never much more than half a century older (Figure 31, Table 19, and Supplementary Figures 41-42).

While the median estimates of lineage divergence times are older than those reported previously (TEN, 47 CE; TPE, 835 CE; TPA, 844 CE; Nichols, 1238 CE; SS14, 1127 CE; SS14- $\Omega$ , 1738 CE; Table 19), the 95% HPD intervals largely overlap with estimates reported elsewhere [253,346](Table 19). The two exceptions are TEN and SS14- $\Omega$ , which we estimate to have a possibly much older origin than previously thought, regardless of whether ZH1540 is included. This is likely due to the more diverse dataset used here, which more accurately represents the full genetic diversity of the SS14- $\Omega$  lineage. Similarly, the overall *T. pallidum*

tMRCA is estimated to be much older when including ZH1540, and also much older than previously estimated [92,253,346].

We caution that although we performed a relaxed-clock analysis, we did not explicitly model lineage-specific or time-dependent substitution rates. Both phenomena could explain the older age of the TPA lineages estimated here compared to previous studies, while time-dependent rates could also push the subspecies and overall *T. pallidum* tMRCAs even further into the past.

#### **4. Discussion**

Here, we present the first ancient DNA evidence of a pre-Columbian New World treponematoses. Many previous hypotheses on the origin of syphilis have relied on paleopathological evidence suggesting early treponemal infections among the native, prehistoric populations in the Americas [68,344,355]. This study presents unequivocal evidence of New World treponematoses by presenting a reconstructed, high-coverage *T. pallidum* genome retrieved from nearly 2,000-year-old Brazilian indigenous human remains, along with three low-coverage genomes from the same spatiotemporal context. Surprisingly, these genomes are remarkably similar to those of the causative agent of modern-day bejel, *T. pallidum endemicum*. Even though the emergence of sexually transmitted syphilis still remains unaccounted for, the genomes presented here contribute to the understanding of treponemal evolution and estimating the divergences of the subspecies in this bacterial family.

With syphilis as the central point of interest, previous hypotheses have rarely focused on the history of the endemic treponematoses with any significant depth [355,356]. Likewise, bejel is particularly under-represented in the modern diversity of genomes, with only eight sequenced genomes available to date [2,36,284,357]. Contrary to the findings from recent historical periods that showcase yaws and syphilis both in the Old and New World contexts [92,98,253,254], genetic evidence of a bejel-causing (TEN) agent has not been recovered from any of the studied ancient remains so far. Analysis of the remarkably early TEN-like genomes in this

study provides the first evidence of treponemal infections in the New World, establishing ancient DNA as an invaluable tool for investigating prehistoric treponemal pathogens. The presence of this unexpected subspecies in the coastal region of South-America puts the possible endemicity of the non-venereal treponematoses in a new perspective, and gives a tangible starting point to form new hypotheses on their spread across the globe. Furthermore, a high-coverage TEN genome in this study represents the first ancient point of comparison to the small set of available modern genomes of this relatively understudied subspecies.

The prehistoric distribution and spread of the endemic treponematoses have mainly been assessed in the Unitarian hypothesis that treats all treponematoses as one and the same, very old and global disease, and considers its varying manifestations as responses to environmental and cultural factors. The hypothesis was contested by genetic and genomic studies showing a clear distinction among the bacteria causing syphilis, yaws and bejel, with sequences clustering in reciprocally monophyletic clades[195,203]. However, findings in this study support some aspects of the unitarian view: particularly the identification of an ancient, TEN-like agent, found far from the disease's modern-day geographical niche in a humid Brazilian coastal region attests to treponemes' ability to adapt to various climates and geographic locations. Distinction of these infections based on clinical manifestations, mode of transmission, and environmental habitat, is in itself problematic. For example, recently isolated DNA samples from presumed syphilis infections from patients in Cuba and Japan were found to pertain to yaws and bejel-causing bacteria, respectively, challenging the geographical categorizations and providing observations on sexual or congenital transmission of genetically yaws- and bejel-type treponematoses [11,12,22].

Finding an endemic treponematosis in prehistoric Brazil restages many of the existing hypotheses on the spread of *T. pallidum*, worldwide. Paleopathology has provided ample examples of treponemal diseases from the prehistoric Americas [68,344], as well as some, though notably rare, European cases prior to the late 15th

century [355,358–360]. These analyses have, however, lacked the sufficient sensitivity to diagnose the diseases to the subspecies level, as is demonstrated by the recent ancient DNA studies that showcase a large, previously unresolved diversity of treponemes in historical Europe [92,98,253,254]. Perhaps one of the most challenging questions to address is whether the three *T. pallidum* subspecies are responsible for inherently different clinical manifestations, and thus cause distinct infections, or whether the observed differences are mostly dependent on environment and cultural practices. Overall, very little can be reliably presumed of the treponemes' adaptations and prevalence in the past.

Apart from inflicting bone damage, it is unknown whether the newly found, genetically TEN-like ancient pathogen resembled bejel in its clinical manifestations, and which adaptations allowed it to thrive in the host population of ancient Brazil. Despite its considerable age, one ancient genome alone cannot securely date the divergence of each subspecies. It also leaves the possible presence of multiple treponemal diseases in the New World open to questioning, proving neither the so-called pre-Columbian hypothesis, that postulates a global distribution of syphilis and its origin outside the New World, nor the contrary Columbian hypothesis, which associates the emergence of syphilis with Columbus' return from the Americas. However, the recombination from TPE/TEN strains to TPA strains, shown in both ancient and modern strains, suggests that at least one of the endemic forms remained in the geographical proximity and within a common host pool with TPA after their initial divergence. Future findings may further localize extinct ancestral forms of treponemal infections and reveal the patterns and routes of their spread across the human host population. Recovering a high-quality genome from a prehistoric source consolidates the use of ancient DNA techniques for unraveling the origins of syphilis, and may in time help to establish an entirely novel, more informed hypothesis on the events leading to the spread of the *T. pallidum* subspecies across the world.

Bejel is mainly found in arid African, Middle-Eastern and Mediterranean regions today, making it an unlikely candidate when examining a potential South American treponematoses from a coastal context. It is considered one of the neglected tropical diseases, and the genomic data from modern strains remains extremely sparse, rendering the current genetic diversity underexplored.

Paleopathological findings, although tentative, point to a global prehistoric existence of bejel, with the earliest proposed cases in the Old World Sudan, dating as early as 15,000 years ago, and the North American Nevada up to 9,000 years ago [203]. At the focus point of this study, the southern coast of Brazil, the earliest potential cases of treponematoses reportedly date to around 6,300 BP [68]. Also, some of the Jabuticabeira II individuals carried pathologic lesions that indicated infections of treponemal origin [68]. Although these findings were not diagnosed as bejel and concerned different individuals than the genetically confirmed cases in this study, it can be reasoned that they possibly represent the same bacterial biovar in the host population. These findings suggest that bejel was possibly more widespread in the past, and not necessarily associated with the same environmental habitats than today.

Our genomic investigation, together with the radiocarbon datings of both human remains and stratigraphy, places the TEN-like treponematoses in South America long before the European contact in the 15th century, and even predates the viking expeditions to the North American coast, firmly attesting to the presence of bejel-like treponemal infections in the pre-contact New World. Phylogenetically, this prehistoric form belongs indisputably to the TEN clade, basal to all of its modern strains and distinctly apart from the yaws and syphilis groups. Among the few modern TEN strains available for comparison, those found in Europe and the Middle-East appear relatively close to the new ancient genome, whereas recently published Cuban and Japanese bejel strain sequences are part of geographically structured and more distant clades. Overall, the TEN genomes are highly similar to each other across the clade, which may indicate a slow evolution of the lineage as



a whole, at least until recently. Regardless of the improved genomic representation of the modern TEN genomes and the newly reconstructed pre-Columbian genome in this study, a larger representation of this lineage would be needed to draw robust conclusions about the evolution and diversification of the subspecies.

With the ancient genomes presented here, we push back the dates for the oldest reconstructed ancient *T. pallidum* strains by more than a thousand years. The recovery of an ancient treponemal genome with substantial amounts of relatively high-quality DNA (over 10-fold coverage) has only been achieved on a few occasions and in more recent historical contexts [92,253,254,346].

A high-coverage genome is a powerful tool for exploring a pathogen's adaptations across time, since it allows for a detailed analysis at the level of genes and single nucleotide polymorphisms. The short fragment size of ancient DNA material, however, makes it challenging to construct ancient genomes *de-novo*, including the potentially evolutionarily crucial hypervariable genes. So far, the reconstructed ancient *T. pallidum* genomes have largely resembled modern strains and possessed very few detectable genetic changes that could illuminate the ancestral conditions of treponemes. Only one subspecies-intermediate lineage, basal to both yaws and bejel according to a SNP based analysis, has been recovered from a historical context in the Netherlands [92] and offered a glimpse of the diversity that may have existed in the past. Unfortunately, the low whole-genome coverage ruled out more in-depth analyses and molecular dating, leaving this extinct pathogen's genetic characteristics largely unaccounted for.

Recombination analyses are conducted for the *T. pallidum* genomes in this study as a way of assessing the processes of subspecies divergence and gene selection. For instance, the evolutionary relationship of the TEN and TPE clades appears to become more ambiguous when putative recombinant loci are excluded [346]. This effect likely reflects the evolutionary adaptations of the different subspecies, and showcases the importance of recombination in their diversification. For example,

the genes in the *tpv* family, most of which are recombinant [73,108,175], are often considered as strong virulence factor candidates [143,182] due to their role in constructing cell membrane structures that improve the treponemes' ability to avoid host immunity [341]. Here, 34 recombination events are detected. In most of these events, the syphilis-related TPA strains appear to be recipients of genetic material, whereas the lineages responsible for the endemic treponematoses are shown as donors. Eight of the events show the TPE lineage as the likeliest donor, exclusively. Likewise, six events point only to a TEN donor. In the majority of events, altogether 15 of them, no clear distinction can be made of which lineage, TPE or TEN, has functioned as a donor. These events could in fact predate the divergence of the two lineages, and horizontal gene transfer may have happened from their common ancestor to the TPA group, at a time when co-infections among the *T. pallidum* strains were common. The new ancient genome from Brazil is involved in the recombination events as part of the TEN donors. According to this evidence, the bejel-causing agents, roughly in their current form, have coexisted with the early forms of today's syphilis-causing strains, and the endemic- and syphilis-like treponemes have exchanged genetic material since at least two millennia; likely more.

The aforementioned hypotheses are in large part based on the divergence times estimated via molecular clock dating for the different branches of the *T. pallidum* phylogenetic tree. The calibration of this method is based on the known ages of the utilized genomes, making securely radiocarbon-dated ancient genomes indispensable to the analysis. With reliable radiocarbon-dates assigned to our new high-coverage ancient genome, we gain an unprecedented, prehistoric calibration point for estimating the subspecies-level divergence dates. Employing this ancient lineage in the molecular clock dating allows us to conclude that all three subspecies had already diverged from each other before Columbus' voyages. Specifically, it places the most recent common ancestor of *T. pallidum* between 12,006 and 545 BCE, the emergence of the TEN clade between 780 BCE and 449 CE, TPE between

28 and 1299 CE and TPA between 42 BCE and 1376 CE. These divergence dates are much older than previous estimates conducted with only modern genomes, as well as analyses including only previously available historical genomes (all dating from the early modern period). Including ZH1540 in the analysis constrains the time of the most recent common ancestor (tMRCA) of the TEN clade, forcing it to be at least as old as ZH1540, which pulls all other lineage tMRCA further back in time and widens the credible intervals. The effect is well-known and has been shown, for example, with *Y. pestis*, where including ancient genomes from the first pandemic shifted divergence dates of other lineages within the global phylogeny by more than a thousand years[361]. This well illustrates how calibration points close to the root of the tree can have an outsized contribution to the divergence times compared to the more recent representatives, and highlights the importance of including older samples in molecular clock analyses.

Defining a nearly two-thousand-year-old bejel-like treponematoses at the genomic level provides a novel perspective on the endemic treponematoses' prehistoric distribution. If the spread of treponemes was facilitated by the early human dispersals, a common ancestor of all modern *T. pallidum* strains may have originated in Africa or Asia, and been transported to the American continent by its initial settlers, some 15,000-23,000 years ago [203,362]. Indeed, the divergence of the subspecies could also have happened even earlier than this: although the tMRCA for the individual clades are here estimated to be later, these are only the lower bounds of the possible age of these events. Should prehistoric evidence of yaws or syphilis emerge, it still has ample opportunity to alter the views on history of treponemes, and influence the dates that mark the rise of each disease and its spread across the world.

Even with an exceptionally well-preserved pre-Columbian genome now available, many unresolved questions remain. Only clearly ancestral forms of treponemes, should they be discovered, can answer whether the American *T. pallidum* arose endemically, perhaps due to a zoonotic event, or accompanied humans from the

colonization of the continent. The events leading to the emergence of sexually transmitted syphilis are also still enigmatic. If the emergence of syphilis *de facto* dates around the contact period, it may prove extremely challenging to find solid evidence anchoring its origin either to the New or the Old World. For this we need more pre-contact *Treponema pallidum* genomes from each continent.

The discovery of bejel, a disease with a largely unknown past, from an ancient context informs us of the continuous interconnections between the subspecies and the timescale in which the functionalities currently separating the strains developed. Together with equally successful specimens, this high-coverage representative of an evidently old treponemal strain can help us understand, which evolutionary pressures or environmental opportunities lead to the rise of a sexually transmitted treponematosi. Overall, the findings in this study underline the importance of horizontal gene exchange related to virulence and transmission routes as an essential factor in the past epidemics, and stress the role of flexible adaptiveness in the global success of the *Treponema* family.

## — CHAPTER 4 —

“Development and evaluation of a new multilocus sequence  
typing (MLST) scheme for *T. pallidum*”



## **Chapter 4: “Development and evaluation of a new multilocus sequence typing (MLST) scheme for *T. pallidum*”**

### **1. Background**

With the increasing incidence of treponemal diseases in recent decades and the presence of endemic yaws and bejel in various regions worldwide, there is a pressing need to enhance our epidemiological understanding of these diseases incorporating molecular biology methods. Cultivating these bacteria in the laboratory is challenging, and genome sequencing from clinical samples is often costly and not always successful. The existing typing schemes, as outlined in Section 1.2.2 of this thesis, rely on a limited number of loci and face significant technical difficulties when applied to clinical samples. This emphasizes the necessity for a novel approach.

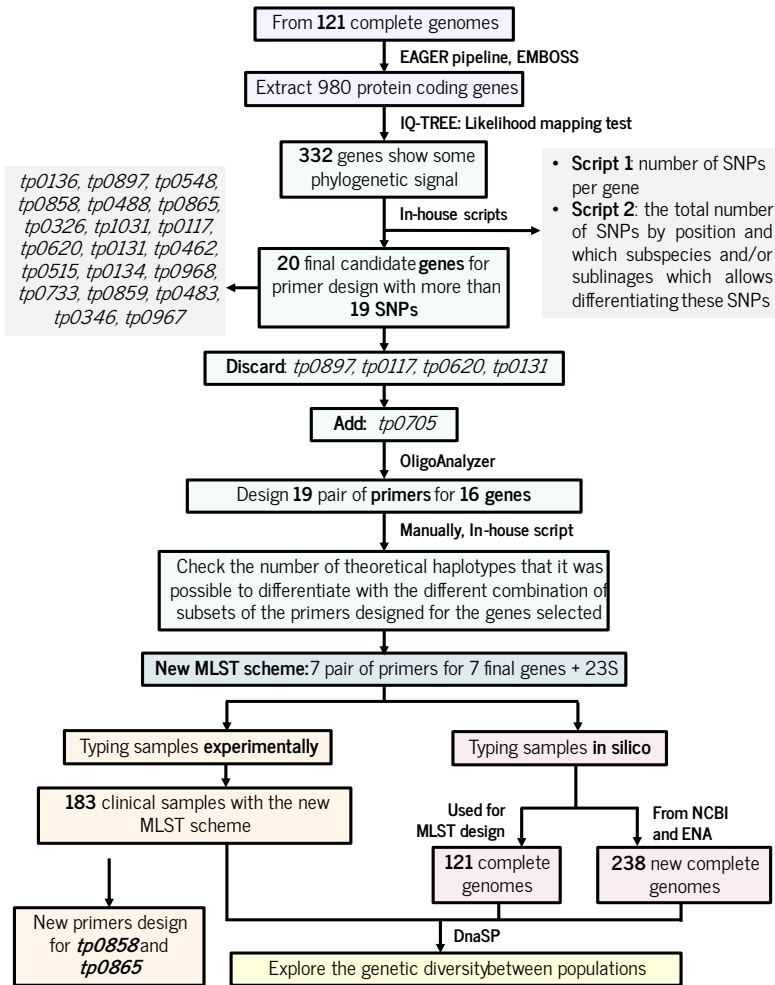
Moreover, these typing systems were not originally designed incorporating information from representative genomes of all three subspecies and their respective variations. To address these limitations and improve our understanding of treponemal diseases and their molecular epidemiology, a new approach is required. By adopting a molecular biology approach that integrates whole genome data and representative genomes from all three subspecies, our epidemiological understanding of treponematoses can be significantly enhanced. This novel approach offers the potential to overcome the challenges associated with cultivating these bacteria and the limitations of existing typing schemes when applied to clinical samples. It will enable a comprehensive analysis of genetic variations and their implications for treponemal diseases.

In this work, we have used the information from 121 complete *T. pallidum* genomes to design a new multilocus sequence typing (MLST) scheme that allows differentiating between the three subspecies of *T. pallidum* and, also, within lineages of TPA. This new typing scheme has been fine-tuned and tested experimentally with a collection of *T. pallidum* samples from different regions worldwide and for the three subspecies. Our proposed scheme will make it possible to identify genetic diversity and transmission patterns for the three diseases, as well as to establish a global network for surveillance and investigation of possible outbreaks as it will be accessible in a public database.

### **2. Material and methods**

A summary of the workflow used to design a new MLST scheme for *T. pallidum* and fine-tuning it experimentally and *in silico* with a representative panel of *T. pallidum* samples is shown in Figure 32.





**Figure 32.** General workflow for the design of a new MLST scheme and testing it in a complete panel of samples experimentally and *in silico*.

## 2.1 Genomic dataset generation for the new MLST scheme design

We gathered a complete genomic dataset with 121 *T. pallidum* genomes (Supplementary Table 31). We used 101 genomes from TPA (14 from Nichols-lineage and 87 from SS14-lineage), 17 from TPE and 3 TEN from previous studies and public databases by December of 2018. From these genomes, 31 genomes were unpublished, of which 28 have not been published at present (2023), and were

obtained through a collaboration with Dr. Natasha Arora, from the University of Zurich (Switzerland).

To reconstruct the individual genomes from the raw short reads data, we applied the **EAGER** pipeline [101]. The read processing was performed in collaboration with Dr. Kay Nieselt, from University of Tübingen (Germany). To generate reads from genomes with no raw data available, HTS-like reads based on genome assemblies were simulated using the tool **Genome2Reads** [101]. After adapter clipping, merging and quality trimming, the resulting reads for each sample were mapped to the Nichols genome (NC\_021490.2), using **BWA-MEM** [301] with default parameters. PCR duplicates were removed with **DeDUP** [101]. **QualiMap** 2.17 was used to calculate the coverage breadth of the reference genome and coverage depth [302]. SNP calling was performed using **GATK UnifiedHaplotype** 3.6 [303]. Sequenced samples were required to cover at least 80% of the Nichols genome with at least 3 reads to be included in further analyses [78,79,106]. The assignment of a variant from a reference nucleotide needed a coverage threshold of 3 and a minimum homozygous SNP allele frequency of 0.9. **MUSIAL** (<https://github.com/Integrative-Transcriptomics/MUSIAL>) was applied to obtain a multiple genome alignment (MSA) from the resulting VCF files.

### 2.2. Design of the new MLST scheme

We identified candidate loci for the new typing system using the information from 121 whole *T. pallidum* genomes. First, we performed an assessment of the phylogenetic information using the likelihood mapping test as implemented in **IQ-TREE** (see section 5.1 of Material and Methods) for each of the protein-coding genes, following the annotation of the Nichols reference genome. The genes that showed some phylogenetic signal, evaluated as likelihoods falling outside the central region in the LM triangle [273], were retained for the ensuing analyses. Using two in house scripts (Supplementary Files 4 and 5), we checked the total number of SNPs per gene, and which subspecies and/or sublineages could be

differentiated by each of these SNPs. We selected those genes with the highest level of variation and power of discrimination to be used for designing primers.

The primers were designed with the following criteria:

1. Primers had to be in conserved regions.
2. The 3' end had to correspond to a second position of the codon of a coding sequence.
3. Primer length of around 20 bp.
4. Amplicon size of 650 bp, approximately.
5. 45%-60% of GC content.
6. Direct and reverse primers should not be complementary.
7. Minimum dG value of -6 kcal/mol.

After the primer design, we checked the number of haplotypes that it was possible to differentiate with the different combination of primers designed for the genes selected. The number of haplotypes was checked using an in-house script (Supplementary File 6) in order to discard haplotypes that could not be considered due to missing data in the sequences employed. Finally, we selected the set of primers with the highest level of resolution for a MLST scheme according to the expected variability in the set of 121 genomes. The selected seven pairs of primers were tested to check if they could be amplified correctly by PCR, using a serial dilution of the Nichols' sample plus a negative control.

### **2.3. Fine-tuning of the new MLST scheme**

#### **2.3.1 Sample collection and DNA extraction**

We collected 183 samples of *T. pallidum* to be typed with the new MLST scheme. 179 samples (Table 20) corresponded to clinical samples obtained in collaboration with Dr. Pablo Hernández Bel (Hospital General de València, València, Spain), Prof. David Šmajš's group (Masaryk University, Brno, Czech Republic), Prof. Lorenzo Giacani's group (University of Washington, Washington D.C., USA), Dr.

Takuya Kawahata (Osaka Institute of Public Health, Osaka, Japan), and Dr. Allan Pillay (Centers for Disease Control and Prevention, Atlanta, USA). We also included 4 historical samples (Table 21) from a collaboration with Prof. Steven Norris's group (McGovern Medical School, Houston, USA), which were obtained by culture in rabbits.

**Table 20.** Clinical samples collected to be typed by the new MLST scheme. The table shows the *T. pallidum* subspecies in which the samples were classified, who provided them and who performed the DNA extraction and typing of the samples using the new MLST scheme.

Subspecies	Number of samples	Source	DNA extraction	MLST tests
TPA	76	Hospital General de València	Our research group	Our research group
	21	David Šmajš (Masaryk University)	David Šmajš (Masaryk University)	David Šmajš (Masaryk University)
	36	Lorenzo Giacani (University of Washington)	Lorenzo Giacani (University of Washington)	Lorenzo Giacani (University of Washington)
TPE	16	David Šmajš (Masaryk Univ.)	David Šmajš (Masaryk Univ.)	Our research group
	13	Lorenzo Giacani (Univ. Washington)	Lorenzo Giacani (Univ. Washington)	Lorenzo Giacani (Univ. Washington)
	8	Allan Pillay (Centers for Disease Control and Prevention)	Allan Pillay (Centers for Disease Control and Prevention)	Allan Pillay (Centers for Disease Control and Prevention)
TEN	5	Takuya Kawahata (Osaka Inst. Public Health)	Takuya Kawahata (Osaka Inst. Public Health)	Takuya Kawahata (Osaka Inst. Public Health)
	4	David Šmajš (Masaryk Univ.)	David Šmajš (Masaryk Univ.)	David Šmajš (Masaryk Univ.)

**Table 21.** Historical samples from cultures in rabbit collected to be typed with the new MLST scheme. The table shows the sample's name, the *T. pallidum* subspecies in which each sample was classified, the source, and who performed the DNA extraction and typing of the samples using the new MLST scheme.

Sample Name	Subspecies	Source and DNA extraction	Laboratory which tested the MLST in the samples
4202 (Gauthier)	TPE	Steven Norris (McGovern Medical School)	Our research group
5217 (Samoa D)	TPE		
4920 (Samoa F)	TPE		
4301 (Nichols)	TPA		

The DNA extraction of the samples was performed by each research group that provided the samples (Table 20) with the exception of the samples provided by the Hospital General de València, which were extracted by our research group using Nuclisens® easyMAG® (bioMérieux).

### 2.3.2 PCR and DNA sequencing

Seven candidate loci (*tp0136*, *tp0326*, *tp0548*, *tp0705*, *tp0858*, *tp0865* and *tp1031*) plus the 23S rRNA gene were amplified by PCR with a touchdown protocol.

The total volume of a PCR reaction was 25 µL, with the following composition: 3 µL of DNA, 1.5 µL of dNTP mix from TaKaRa Ex Taq® DNA Polymerase 250 Units kit, 2.5 µL Mg<sup>2+</sup> plus buffer and 0.05 µL Taq polymerase, 0.025 3 µL of each primer, and 0.75 µL of DMSO (3%) to increase the specificity and yield of the PCR reaction.

The protocol for the touchdown PCR consisted of 40 amplification cycles. In the first 10 cycles, we used a melting temperature 4°C higher than the optimal melting temperature in order to obtain more specificity. For the remaining 30 cycles, we used the optimum melting temperature for each primer calculated in the primer design.

PCR products were purified using a NucleoFast 96 PCR Plate, 96-well ultrafiltration plate for PCR clean up (Macherey-Nagel, Cultek) and sequenced using the ABI 3730XL Capillary Electrophoresis Sequencing System (Sequencing Service of the University of Valencia, Spain). Sequence analyses were performed using the **Staden** package v4.11.2-r [363] and merged in a gene alignment for each locus to check each sequence and obtain its corresponding allele and the different ST's per sample manually by **Aliview** 1.25 [268].

For the 23S rRNA gene, nucleotides corresponding to positions 2058 and 2059 were checked for A→G mutations indicative of macrolide resistance [152,364].

#### **2.4 *In silico* application of the new MLST scheme**

We downloaded all assembly files available by September of 2022 from the two main public databases, the European Nucleotide Archive (ENA) and the National Center for Biotechnology Information (NCBI), in order to type them *in silico* using the new MLST scheme. In total, we obtained 238 different assembly files, 226 from TPA, 5 from TPE and 7 from TEN. Details of all these genomes are provided in Supplementary Table 31.

All the genomes were consolidated into a single fasta file, which was then aligned using **MAFFT** 7.467 to generate a comprehensive whole genome alignment. From this alignment, the flanking regions of the newly designed primers for each locus in the novel MLST scheme were extracted, resulting in eight distinct fasta files, one for each locus. To determine the allelic profile for each sequenced sample, **CD-HIT** 4.7 [365] was employed to identify similar sequences and generate clusters. The results were subsequently verified manually using **Aliview** 1.25 [268] to ensure accuracy. In detail, the obtained sequences from the samples were compared with the available sequences. Through this comparison, sequence differences, such as single nucleotide polymorphisms (SNPs) or insertion/deletion events (indels), were identified, enabling the assignment of different alleles at each locus. Once an allele

was assigned, it contributed to the determination of the sequence type (ST) of the organism. The combination of alleles at multiple loci, in a specific order, yielded the ST.

### **2.5 Genetic diversity and population divergence in *T. pallidum***

We compared the genetic diversity (D) of STs between the different subspecies/sublineages of *T. pallidum* using the known expression  $D = 1 - \sum p_i^2$  [366] where  $p$  is the frequency of each ST for each subspecies/sublineages of *T. pallidum*.

Additionally, to analyze the geographical distribution of the different STs, we examined the genetic diversity within and differentiation between populations using **DnaSP** 6.12.03 [367], considering as population the countries and continents of origin of the samples. Due to the small sample size available for the TPE and TEN subspecies, we performed these analyses only for TPA. Additionally, we classified the TPA samples according to their continent of origin (four different continents), to analyze the genetic diversity at continental level. However, we excluded Oceania as a continent to analyze its genetic divergence, because it only had two samples. Next, at both the continental and country level, we estimated the average number of nucleotide differences (k) within and between populations. Moreover, we also estimated the nucleotide diversity ( $\pi$ ) to examine the degree of polymorphism within populations, and between populations using the Jukes-Cantor model ( $D_{xy}(JC)$ ) of nucleotide substitution.

## **3. Results**

### **3.1 Reference-based alignment for the design of a new MLST scheme**

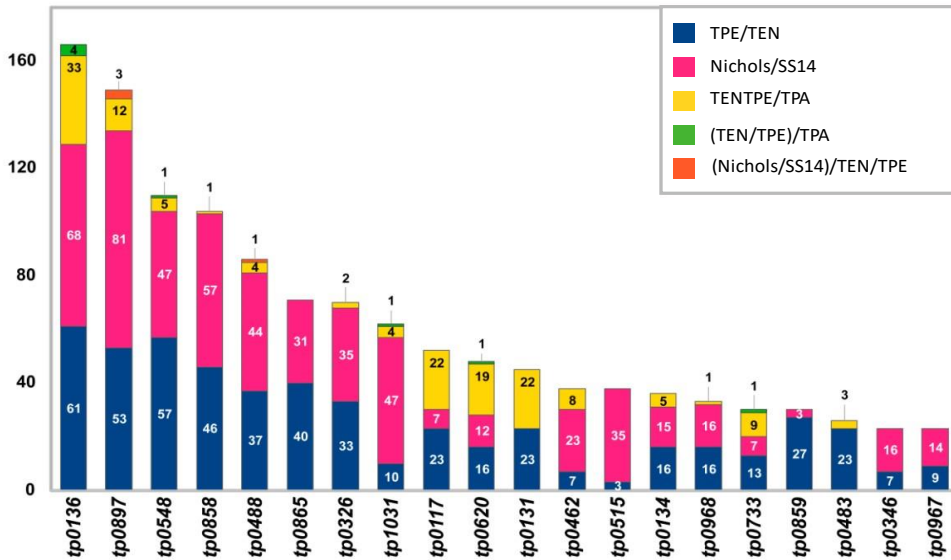
The resulting multiple sequence alignment from the 121 whole genomes spanned a total of 1,139,633 bp and had 3,465 SNPs.

### **3. 2 Selection of loci for primer design**

From the 978 protein-coding genes extracted according to Nichols genome as reference, a likelihood mapping test was performed selecting 332 genes that showed some phylogenetic signal (Supplementary Table 32). It was not possible to perform this test for 198 genes due to the large number of undetermined positions in the corresponding multiple alignments, and those genes were discarded from the subsequent analyses.

Using an in-house script (Supplementary File 4), we checked the number of SNPs per gene (Supplementary Table 33). Moreover, to know which genes could provide optimal resolution for the new MLST scheme, we checked which subspecies and/or sublineages could be differentiated by each SNP using another in-house script (Supplementary File 5). We finally selected 20 candidate genes (Figure 33) with more than 29 SNPs each, because they had the most variation and power of discrimination between the three different subspecies to be used for primers design (Supplementary Table 34).





**Figure 33.** Graph summarizing the discriminatory power at the different levels used to select 20 candidate genes for primer design. For each gene with a different color, it is indicated of all the SNPs detected in that gene, how many allow to differentiate between each of the different subspecies and/or sublineages of *T. pallidum*. In dark blue are shown the SNPs that allow differentiation between TPE and TEN. In fuchsia, the changes that allow differentiation between Nichols and SS14. In yellow, SNPs that differentiate TPE and TEN from TPA (but do not differentiate TPE from TEN). In green, SNPs that differentiate TPE from TEN and in turn from TPA. Orange shows SNPs that differentiate Nichols from SS14 and in turn from TPE and TEN (but do not differentiate TPE from TEN)..

Next, we designed 19 different pairs of primers for 15 of the 20 previously selected genes and we discarded 5 genes because it was not possible to design primers for them. The genes discarded were four *tpr* genes (*tp0897*, *tp0117*, *tp0620*, *tp0131*) and *tp0733*, a gene encoding a putative member of the OmpW family of porins/virulence factors. Although its phylogenetic signal could not be checked by a likelihood mapping test due to the large amount of missing data for some of its sequences, the *tp0705* gene was included as a candidate gene for primer design in this step, because it is part of the panel of the only MLST scheme available for *T. pallidum* designed for TPA [55].

After designing the primers, we performed an analysis to determine the potential number of haplotypes that could be differentiated using different combinations of

the primers developed for the 16 selected genes (Supplementary File 6 and Supplementary Table 35). These theoretical haplotypes represent unique sequences (alleles) assigned with integer numbers. The combination of alleles at each locus, referred to as the "allelic profile," defines the sequence type (ST) of a strain. Based on the analysis of haplotypes that could be differentiated by various combinations of the designed primer set, we selected a set of seven primers that exhibited the highest discriminatory power among the four main subspecies/sublineages of *T. pallidum*. These primers were specifically designed for genes *tp0136*, *tp0326*, *tp0548*, *tp0705*, *tp0858*, *tp0865*, and *tp1031* (Table 22). This primer set not only demonstrated the highest resolution for an MLST scheme, considering the expected variability in the set of 121 genomes (Supplementary Table 31), but also consistently amplified a serial dilution of the Nichols sample (Supplementary Figure 43). In Supplementary Table 36, the theoretical number of STs calculated based on information from the 121 whole genomes is presented in the "strain" column. These STs were differentiated by the final selected combination of genes: 12, 16, 11, and 3 STs for Nichols, SS14, TPE, and TEN, respectively.

Additionally, we included the 23S rRNA gene previously used in [55] in the new typing MLST scheme proposed, to check for A→G mutations indicative of macrolide resistance.

**Table 22.** Final primers designed for the new *T. pallidum* MLST scheme. The primers for the 23S rRNA gene are the same primers used in [55] to complete the typing MLST scheme and to know if the samples are resistant to tetracycline or not. The amplicon size is the final size obtained after Sanger sequencing.

Genes	Primer sequence	Melting temperature	Amplicon size (pb)
<i>tp0136</i>	5'- AGCGACGGGTGCTATCACTA -3'	58.75°C	566
	5'- TTACTCGCGGTTCCAGGAGC -3'		
<i>tp0326</i>	5'- CATTTCGTTTCGCTCCGACAC -3'	55.5°C	545
	5'- TACCGTGAACGACAACACAA -3'		
<i>tp0548</i>	5'- ATGATATCGTGTTTCGGTGCG -3'	55.8°C	506
	5'- ACAGAAGGTGTGAGACGCAT -3'		
<i>tp0705</i>	5'- ACCGACCATATCCAGTACAC -3'	57.85°C	545
	5'- TCTTCTCTCACACACGTTGC -3'		
<i>tp0858</i>	5'- AAGTGTGGTTGCTGCAAGGA -3'	57.55°C	445
	5'- ATTCGGCCGAGCAGTATCG -3'		
<i>tp0865</i>	5'- GGCAATCGCTTCCTCATAGT -3'	59°C	640
	5'- GGCATCAGTGTGGGAACCAA -3'		
<i>tp1031</i>	5'- TTGCTGAGCATGCAGTGGAA -3'	57.85°C	398
	5'- CACGTGGTACTGCATTGCCT -3'		
23S	5'- GTACCGCAAACCGACACAG -3'	59°C	592
	5'- AGTCAAACCGCCACCTAC -3'		

Nevertheless, we were reported some problems with the amplification of loci *tp0865* and *tp0858* for some samples. Although it seems that the regions used for primer design were conserved for this loci in the 121 genomes used for designing the new MLST scheme, this might not be the case when additional samples were analyzed. Hence, we designed alternative primers.

For locus *tp0858*, a new set of primers flanking the initial primer set were designed (Table 23). The new reverse primer is not within the *tp0858* gene; it is in an intergenic flanking region of the 3' gene end. The new forward primer is within the *tp0858* gene, and it is flanking the 5' region of the original forward primer. The new primers were designed following the same conditions to all primers for the other genes.

**Table 23.** New alternative primers designed for loci *tp0858* and *tp0865*.

Genes	Primer sequence		Melting temperature	Amplicon size (pb)
<i>tp0858</i>	F_2	5'- ACCGTAAGGTCTCGGACAA -3'	57,6 °C	544 pb
	R_2	5'- GTGCCCTGCTGAAGAATGCG -3'		
<i>tp0865</i>	F_2	5'- CACGCCCCGTATAAAGAACA -3'	55 °C	-
	F_3	5'- GCAACCGCCGAGGGTGTCTT -3'	62,9 °C	-
	R_2	5'- CCACCAGGAGATAGGGGAAC -3'	56,8 °C	-

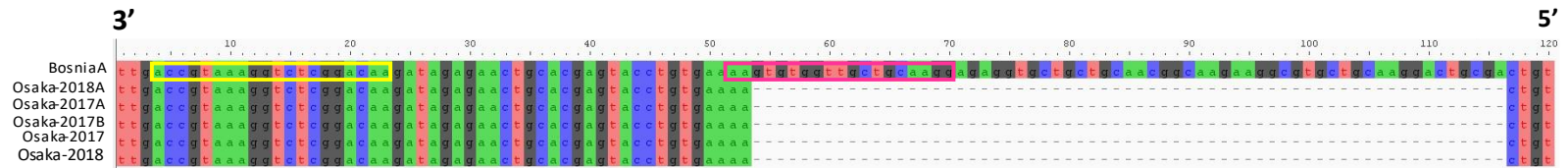
For locus *tp0865*, it was not possible to design new primers flanking the previous ones because, as the product size of the initial primers for the locus *tp0865* was 645 bp, flanking primers would have a too large amplicon size to be amplified and sequenced successfully by a single PCR. For this reason, we designed three new primers for the *tp0865* gene to test them in combination with the previous ones. The different combinations of the primers were selected according to the best melting temperature and amplicon size (Table 24).

**Table 24.** Primer combinations used to test the new primers designed for the gene *tp0865* with the old ones.

Gene	Combination	Primers	Primer sequence	Melting temperature	Amplicon size (pb)
<i>tp0865</i>	1	F_2	5'-CACGCCCCGTATAAAGAACA-3'	56.5 °C	836
		R_1	5'-GGCATCAGTGTGGGAACCAA-3'		
	2	F_1	5'-GGCAATCGCTTCCTCATAGT-3'	55.9 °C	957
		R_2	5'-CCACCAGGAGATAGGGGAAC-3'		
	3	F_3	5'-GCAACCGCCGAGGGTGTCTT-3'	60.4 °C	903
		R_1	5'-GGCATCAGTGTGGGAACCAA-3'		

We tested the new primers for the *tp0858* gene in five Japan TEN strains (Osaka-2017A, Osaka-2017B, Kyoto-2017, Osaka-2018, Osaka-2018B), and it was revealed that the cause of the absence of amplification in the old *tp0858* primer (Figure 34) set was a deletion of 63 bases in the 5' end (the old forward primer matched only the first two bases).

## Chapter 4



**Figure 34.** The deletion found in the original *tp0858* primer for the five TEN samples from Japan. In yellow is highlighted the new forward primer designed for the *tp0858* gene and in pink the old one.

Moreover, we also tested all the different combinations possible with the original and new primers designed for the locus *tp0865* (Table 24) using Nichols strain DNA as a positive sample and a negative control, and the amplification was observed in all sets. However, all the combinations of primers were tested for the same Japan TEN samples used to test the new *tp0858* set of primers (Osaka-2017A, Osaka-2017B, Kyoto-2017, Osaka-2018, Osaka-2018B) and the amplification was not successful.

Following the analysis of these results, it is crucial to determine the definitive primers for each locus in the new MLST scheme. Based on the findings, it is recommended to utilize the primers specified in Table 22. However, in cases of failed amplification for the *tp0858* and *tp0865* loci, the alternative primers provided in Table 23 and Table 24 should be employed for the problematic samples. Therefore, when calculating the sequence type (ST) for a particular sample, it is important to consider the sequence flanked by the primers specific to the *tp0858* and *tp0865* loci, as outlined in Table 22, rather than the sequence flanked by the alternative primers provided in Table 23 and Table 24. Specifically, the gene coordinates used to obtain the allele for those genes are 759 bp to 1204 bp for *tp0858* and 141 bp to 781 bp for *tp0865*.

### **3. 3 Allelic profiles identified with the new MLST scheme**

Most MLST schemes include 7 housekeeping genes, but schemes with as few as 5 and as many as 10 genes have also been developed [368]. To keep the number of typing loci minimal but with the highest level of resolution for a MLST scheme according to the expected variability in the set of 121 genomes, we propose seven loci *tp0136*, *tp0326*, *tp0548*, *tp0705*, *tp0858*, *tp0865*, and *tp1031* plus the 23S rRNA gene as a new typing scheme of *T. pallidum*.

The MLST scheme is based on the different sequences present in *T. pallidum* samples, assigned as distinct alleles and, for each isolate, the alleles at each of the seven loci define the allelic profile or sequence type (ST) (e.g., 1.3.1.1.2.2.1).

We conducted experimental analysis on 183 clinical samples using the new MLST scheme. In addition, we performed *in silico* analysis on another 238 samples using whole genome sequences obtained from public databases. This analysis was complemented with the examination of 121 whole genomes that were utilized for the design of the new MLST scheme (Supplementary Table 37). With the 542 samples typed with the new MLST, we obtained different alleles per gene, as detailed in Table 25. The gene with the largest number of alleles was the *tp0548* gene, whereas the gene with the fewest alleles was the *tp0705* gene. The sequences of the representative alleles obtained in this study for each gene can be found in the Supplementary Files 7-13.

**Table 25.** Number of alleles obtained for each gene of the new MLST scheme designed for the total of samples analyzed experimentally and *in silico*. The alleles obtained from the *in silico* analysis with missing data are also detailed in the table.

	N	<i>tp0136</i>	<i>tp0326</i>	<i>tp0548</i>	<i>tp0705</i>	<i>tp0858</i>	<i>tp0865</i>	<i>tp1031</i>
<b>Experimental</b>	183	6	18	30	5	17	10	6
<i>In silico</i>	358	15	19	29	2	20	22	3
<b>All alleles obtained</b>	542	21	37	59	7	37	32	9
<b>Alleles with missing data from <i>in silico</i> analysis</b>	-	9	10	10	1	9	4	1

We are currently working on the configuration of the new MLST scheme for *T. pallidum* in PubMLST, to ensure its availability to the scientific and clinical communities. When it is ready, all the sequences of the alleles identified for each gene as well as their corresponding ST for each sample will be uploaded to PubMLST. This will generate a database with all the allelic profiles of each ST and



the different STs identified for *T. pallidum* where other researchers will be able to deposit their sequences and obtain directly there the allelic profile of each one of them as well as their ST. It is important to note that we have decided to establish as a requirement for depositing allelic sequences in the database that sequences deposited as sequences representing new alleles must be complete sequences without indeterminate positions. For this reason, 44 sequences representing different alleles for the seven genes (Supplementary files 7-13) will not be uploaded to PubMLST (Table 25). This is because, although we identified different SNPs in these sequences that have allowed us to determine a different allele for these genes, some positions in these sequences contain missing data (Ns) so they do not meet the requirement to be uploaded to PubMLST.

### **3. 4 Sequence types (STs) identified among all typed samples**

For the 542 samples analyzed experimentally and *in silico*, we were able to determine the ST for 386 samples (Supplementary Table 38), 82 from experiments and 304 from *in silico* genomes (Table 26). Out of the total of 542 samples tested (Table 26), experimental testing revealed macrolide resistance in 93 samples, while *in silico* analysis detected resistance in 101 samples. Conversely, experimental testing showed no macrolide resistance in 58 samples, and *in silico* analysis confirmed the absence of resistance in 141 samples. However, among the 321 samples with assigned ST, the allelic profile of the 23S gene was experimentally determined for 79 samples (57 macrolide-resistant and 22 macrolide-sensitive) and 242 *in silico* samples (101 macrolide-resistant and 141 macrolide-sensitive).

**Table 26.** Total number of samples for which the ST was obtained, complete or partial, experimentally or *in silico*. It is also specified for all these samples whether the allelic profile of the 23S gene can be obtained and whether the sample was resistant or sensitive to macrolides.

	N	STs Completed	Partial	23S (R)	23S (S)	STs + 23S (R)	STs + 23S (S)
<b>Experimental</b>	184	82	102	93	58	57	22
<i>In silico</i>	358	304	54	101	141	101	141
<b>All</b>	542	386	156	194	199	158	163

We also analyzed how many STs can be differentiated per sample and for each subspecies/lineage of *T. pallidum*, as well as how many samples of each of them were resistant to macrolides or not (Table 27).

**Table 27.** Number of complete STs obtained according to the subspecies/lineage of the samples analyzed and the number of samples for which the allelic profile obtained is partial or incomplete. The table also shows the total number of different STs obtained and the genetic diversity for each subspecies/lineage.

Subspecies/lineage	N	23S (R)	23S (S)	STs Completed	STs + 23S	Partial	Different STs	Genetic diversity
<b>TPA-SS14</b>	311	166	46	229	188	82	37	0.84
<b>TPA-Nichols</b>	127	12	102	121	110	6	35	0.87
<b>TPA*</b>	20	12	1	-	-	20	-	-
<b>TPE</b>	66	0	38	25	14	41	18	0.91
<b>TEN</b>	19	4	12	11	9	8	7	0.83
<b>All</b>	543	194	199	386	325	157	97	-

\*Samples unable to be assigned to a specific TPA sublineage

We obtained 37 different STs for the TPA-SS14 lineage (Table 27 and Figure 35), with ST3 as the most abundant one, with 65 samples. For TPA-Nichols, we obtained 35 different STs (Table 27 and Figure 36), with ST83 as the most frequent with 34 samples. Moreover, we obtained 18 different STs for TPE (Table 27 and

Figure 37) and 7 for TEN (Table 27 and Figure 38), with ST68 being the most abundant STs with 6 samples for TPE, and ST37 and ST44 for TEN, both with 2 samples. According to these results (Table 27), the SS14 sublineage was the sublineage for which more samples could be typed (229/311) and for which more different STs could be differentiated (37), followed by Nichols (121/127 samples typed and 35 different STs). For TPE and TEN, in addition to a much smaller number of samples being collected, fewer samples were typed (25 and 11, respectively) and, therefore, the number of typed STs was lower (18 and 7, respectively).

The ST diversity (D) among the different subspecies of *T. pallidum* was calculated with the frequencies of the different STs obtained for each of them (Table 27). Among the different subspecies of *T. pallidum*, TPE exhibited the highest diversity (D=0.91), followed by Nichols (D=0.87), SS14 (D=0.84), and finally TEN (D=0.83). This indicates that, despite the different number of samples available and tested for the different subspecies/sublineages, the inferred levels of ST diversity are very similar for all of them.

Samples showing macrolide resistance were detected in all the subspecies/sublineages (Table 27), with the exception of TPE, for which we were able to test 38 samples, all sensitive to macrolides. SS14 was the sublineage with the largest number of resistant samples (166/212), compared to Nichols and TEN, with 12/114 and 4/16 resistant samples, respectively.

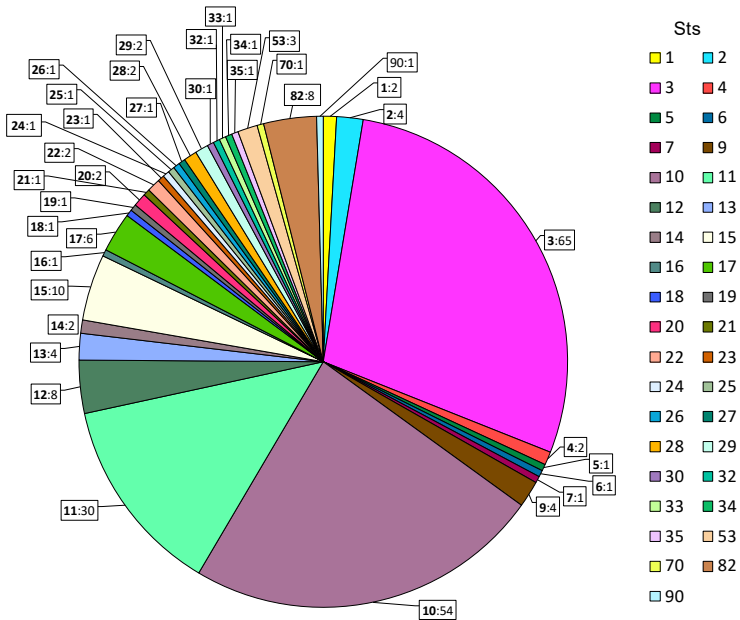


Figure 35. Distribution of STs detected for TPA-SS14.

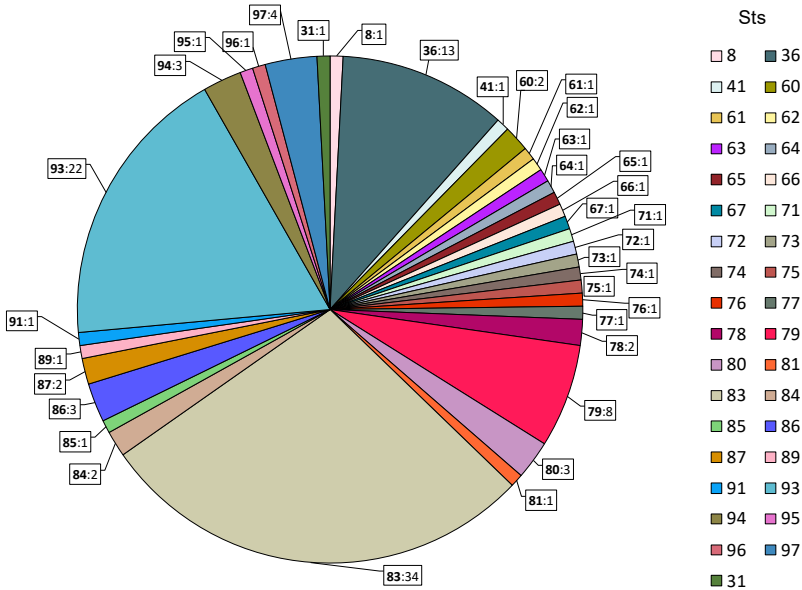


Figure 36. Distribution of STs detected for TPA-Nichols.

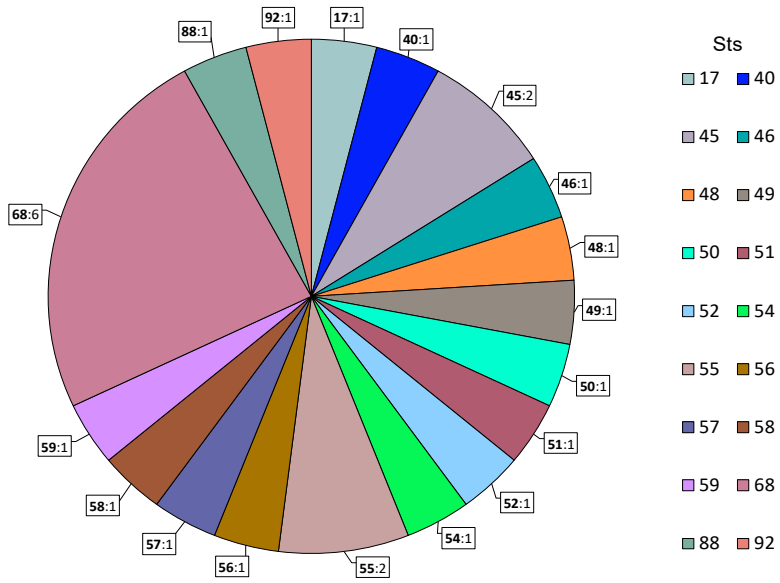


Figure 37. Distribution of STs detected for TPE.

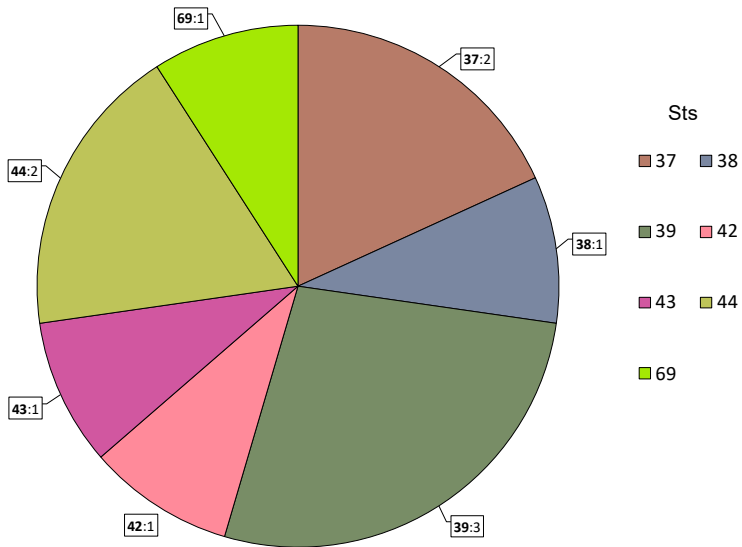


Figure 38. Distribution of STs detected for TEN

In this study, we included 21 repeated samples, which means that, despite having different identification, they originated from the same DNA isolate (Table 28). These repeated samples were analyzed to identify any sequence differences that might result in different sequence types (STs) despite their common origin. The gene sequences for these repeated samples were obtained using different methodologies, as indicated in Table 28. The genome sequences of samples Sea81\_4\_1 and Sea81\_4\_2, obtained through resequencing the original DNA sample isolated in 1951 [369] and subjected to different culture passages in rabbits, are still unpublished (Table 28).

## Chapter 4

**Table 28.** Details of the 21 repeated samples included. Each strain in column "Strain 1" corresponds to the strain in column "Strain 2" of the same row. The allelic profile for each gene is provided in the follow order: *tp0136*, *tp0326*, *tp0548*, *tp0705*, *tp0548*, *tp0858*, *tp0865* and *tp1031* and is specified for each sample, highlighting in red the different alleles between the samples. Moreover, it is also specified the ST obtained for each sample and how many SNPs differed between the repetitive samples compared. For samples highlighted in blue, their gene sequences were obtained experimentally, whereas for samples highlighted in green, the gene sequences were extracted from their original assembly file obtained by de novo assembly. The remaining samples are highlighted in orange because their gene sequences were obtained, by processing the reads obtained in the different studies specified, by mapping using Nichols as genome reference.

Strain 1	Raw reads origin	Allelic profile	ST	Strain 2	Raw reads origin	Allelic profile	ST	N° of SNPs
<b>C279_1</b>	[2]	3,14,34,4,3,5,4	44	<b>C279_2</b>	-	3,14,34,4,3,5,4	44	0
<b>CDC2_1</b>	[113]	4, <b>4</b> ,6,3, <b>13</b> ,8,4	47	<b>CDC2_2</b>	-	4, <b>6</b> , <b>4</b> ,2, <b>4</b> ,4,4	48	37
<b>SS14_1</b>	[370]	1,1, <b>59</b> ,2,1,1,1	82	<b>SS14_2</b>	-	1,1, <b>12</b> ,2,1,1,1	11	5
<b>GHA1</b>	[78]	4, <b>16</b> ,23,2,15, <b>10</b> ,4	54	<b>Ghana_051</b>	[150]	4, <b>15</b> ,23,2,15, <b>4</b> ,4	45	8
<b>CDC-1</b>	[78]	4,32, <b>54</b> ,2, <b>33</b> ,4,4	58	<b>CDC2575</b>	[150]	4,32, <b>23</b> ,2, <b>15</b> ,4,4	45	8
<b>GRA2</b>	[78]	-,1, <b>18</b> ,2,-,-,1	NA	<b>Grady</b>	[73]	1,1, <b>12</b> ,2,1,1,1	51	3
<b>IND1</b>	[78]	4, <b>13</b> ,7,7, <b>12</b> ,7,4	52	<b>Kampung_Dalan_363</b>	[74]	4, <b>25</b> , <b>39</b> ,7, <b>27</b> , <b>19</b> ,4	56	20
<b>Sea81_4_1</b>	Unpublished	5,18,22,2,20, <b>3</b> ,2	60	<b>Seattle81</b>	[73,369]	5,18,22,2,20, <b>14</b> ,2	62	5
<b>Sea81_4_2</b>	Unpublished	5,18,22,2,20, <b>3</b> ,2	60					
<b>NIC1</b>	[78]	2,2,2,2,2,2	36	<b>Nichols_1/ Nichols_2*</b>	[370]/ this study	2,2,2,2,2,2	36	0
<b>NIC2</b>	[78]	2,2,2,2,2,2	36					

As can be seen in Table 28, of these 21 repeated samples, only three sets of samples (CZ279\_1 and CZ279\_2, all Nichols samples and Sea81\_4\_1 and Sea81\_4\_2) had coincident STs. In contrast, samples CDC2\_1/CDC2\_2, GHA1/Ghana\_051, CDC-1/CDC2575, IND1/Kampun\_Dalan\_363 and Sea81\_4\_1-Sea81\_4\_2/SEA81 were assigned to different STs despite being originally the same sample. We could not make the comparison of the STs for samples GRA2 and Grady, because the ST could not be assigned to GRA2 due to missing data in its genome sequence. Nevertheless, we were able to learn that the assignment of a different allele for the *tp0548* gene between both samples was caused by 3 SNPs.

For samples CDC2\_1 and CDC2\_2, we obtained the following allelic profiles: (4,**4,6,3,13,8**,4) and (4,**6,4,2,4,4**,4), respectively, that led to different STs (Table 28). At the gene level, the different alleles between them were caused by 37 SNPs, specifically, by 15 SNPs in *tp0326*, 10 SNPs in *tp0548*, and six SNPs in *tp0858* and *tp0865* each. For the IND1 and Kampun\_Dalan\_K363 samples, we obtained the allelic profiles (4,**13,7,7,12,7**,4) and (4,**25,39,7,27,19**,4) that led to different STs for each one (Table 28). The different alleles between them were caused by 20 SNPs, one SNP in *tp0326*, four SNPs in *tp0548*, eight SNPs in *tp0858*, and seven SNPs in *tp0865*.

We obtained the following allelic profiles for samples GHA1 and Ghana\_051: (4,**16**,23,2,15,**10**,4) and (4,**15**,23,2,15,**4**,4), respectively (Table 28), which originated different STs between both. The different alleles were caused at the gene level by eight SNPs, one SNP in *tp0326*, and seven SNPs in *tp0865*. We also obtained different STs for samples CDC-1 and CDC2575, with the following allelic profiles: (4,32,**54**,2,**33**,4,4) and (4,32,**23**,2,**15**,4,4) that led to different STs for each (Table 28). The different alleles between these two samples were caused by eight SNPs, two SNPs in *tp0548* and six SNPs in *tp0858*.

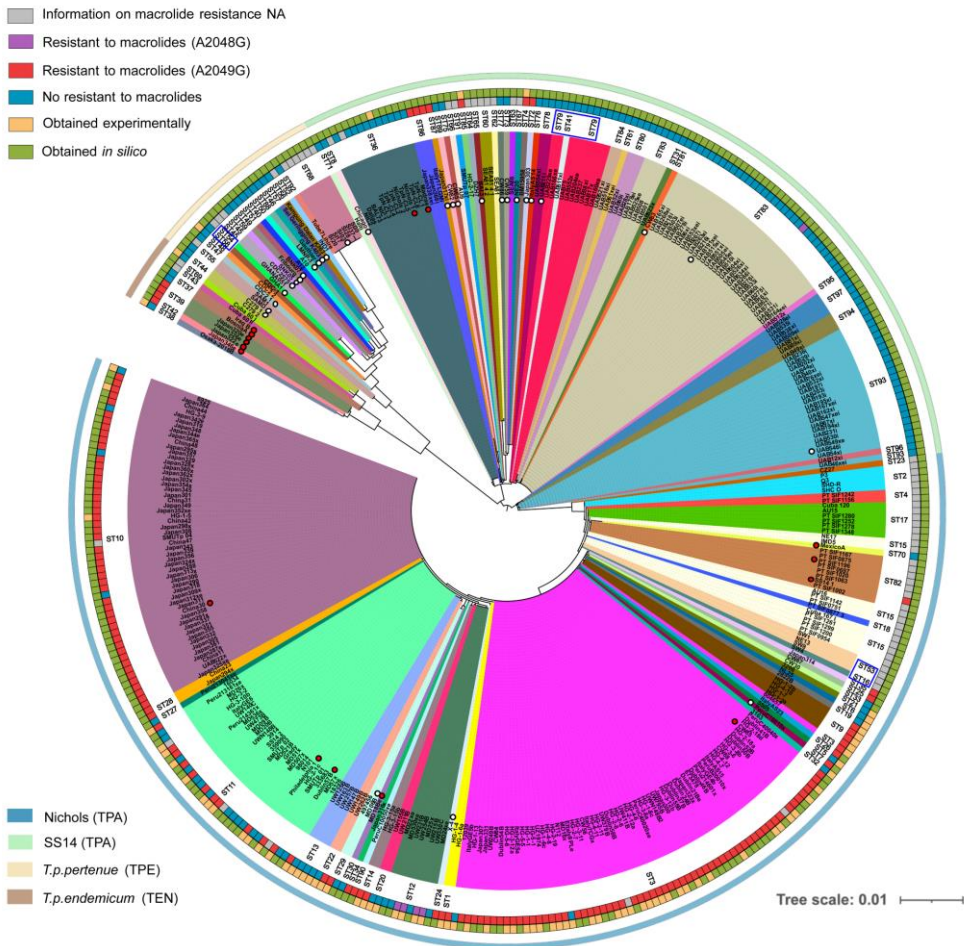
For samples Sea81\_4\_1/Sea81\_4\_2 and Seattle81, which have allelic profiles (5,18,22,2,20,**3**,2) and (5,18,22,2,20,**14**,2), we also obtained different STs (Table 28). The different alleles between these two samples were caused by 3 SNPs



in the *tp0865* gene. Finally, the different allelic profiles for samples SS14\_1 (1,1,**59**,2,1,1,1) and SS14\_2 (1,1,**12**,2,1,1,1) that led to different STs for each one (Table 26), were caused by 5 SNPS in *tp0548*.

### **3. 5 Phylogenetic analysis**

A maximum likelihood tree was constructed from the concatenated data of the seven loci included in the new MLST scheme (Figure 39).



**Figure 39.** Phylogenetic tree of the different STs assigned *in silico* or experimentally and the macrolide resistance profile per sample. In blue are highlighted the ST which form a paraphyletic clade. Samples marked with a blank circle are those in common with phylogeny obtained by whole genomes in Chapter 3 (Figure 40). Samples marked with a circle are the samples also present in the WG-phylogeny (Figure 40) from Chapter 3. In addition, from samples with a circle, those samples with a red circle are the samples compared versus the whole genomes phylogeny from Chapter 3 (Figure 40) in the main text. Samples marked with a Link to the online figure with a higher resolution: <https://drive.google.com/file/d/1dgM0E7aE5pz9IyGihLDbM1oDGvSdqYav/view?usp=sharing>

The different STs assigned *in silico* or experimentally and the macrolide resistance profile obtained with the 23S gene per sample is also shown in Figure 39. In the

---

phylogenetic tree, the four major clades corresponding to each of the *T. pallidum* subspecies and sublineages (TEN, TPE, SS14 and Nichols) could be distinguished clearly.

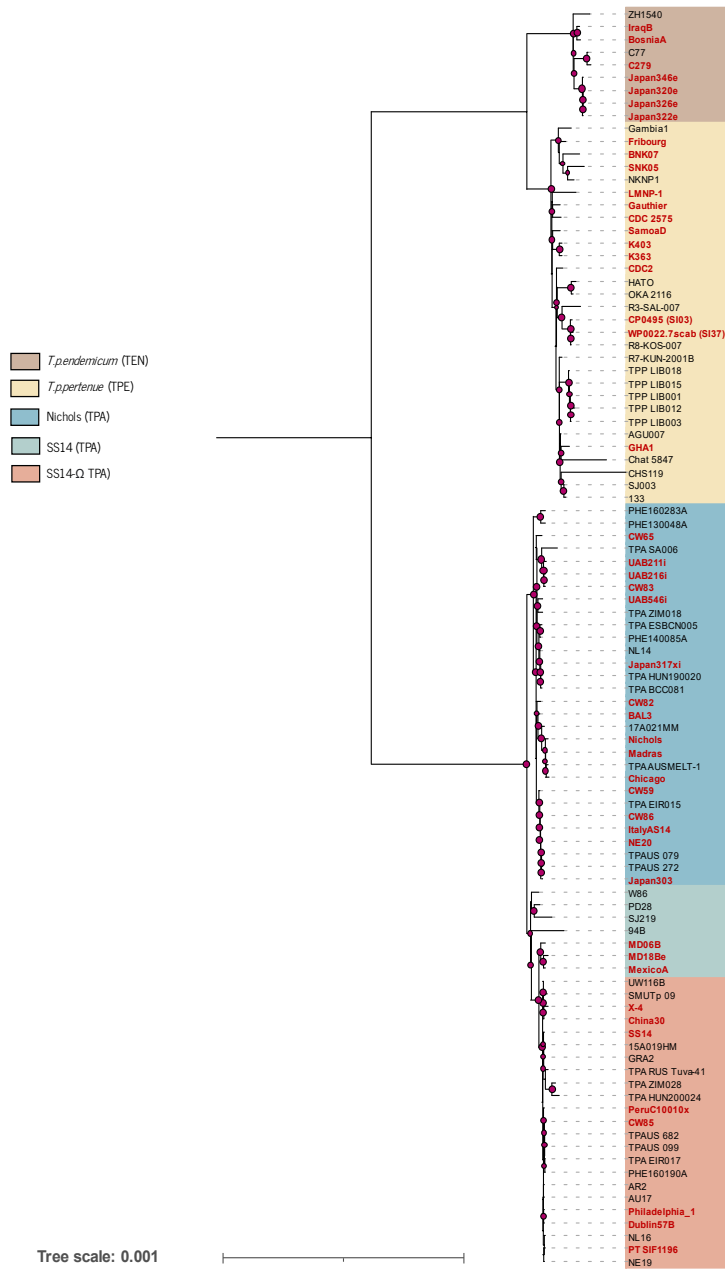
Every ST identified corresponded to a monophyletic clade, except in three cases where a paraphyletic clade was formed (highlighted in blue in the phylogeny from Figure 39). This was the case for the clade formed by ST79 and ST41. According to the phylogeny obtained (Figure 39), the only strain of ST41 (UAB511xi) was identical to the other ST79 strains. However, in the concatenate gene sequences of the UAB511xi strain, there is a deletion in the *tp0136* gene that results in a different allele, which differentiates this sample as a new ST (ST101) and not as ST79.

The same situation occurs for ST54 and ST58, corresponding to samples GHA1 and CDC-1, respectively. According to the phylogeny obtained (Figure 39) both strains appeared to be identical. However, after reviewing the sequences of these two strains, some deletions in the *tp0326*, *tp0548*, *tp0858* and *tp0865* genes were identified, which originate different alleles in those genes between these two samples as well as a distinct STs for each. These deletions are not properly reflected in the ML phylogeny. The last case where a paraphyletic group is formed by ST53 and ST16. According to the phylogeny obtained (Figure 39), it seems more likely that the SW4 sample, which is the only sample in ST16, is in fact identical to the samples that conform ST53. However, after reviewing the sequence of the concatenated data of the SW4 sample, we verified the presence of two SNPs in the *tp0548* gene, which corresponded to a deletion in the other sequences of ST53. This means that ST16 had a different allele for the *tp0548* gene that differentiated it from ST53.

Finally, according to the phylogeny obtained, sample SI37, which belongs to ST68, seems to form an independent clade from the other five samples of this ST. However, after reviewing the concatenated sequences of all ST68 samples, we found that the differentiation was due to two SNPs found in the *tp0865* gene of the SI37 sample, which corresponded to undetermined positions (Ns) in the other five

ST68 sequences. Therefore, sample SI37 cannot be considered different from the other ST68 strains.

We also compared the resolution obtained with the phylogeny built using the concatenated seven genes of the new MLST scheme (Figure 39) with that obtained using whole genomes from Chapter 3 (Figure 40).



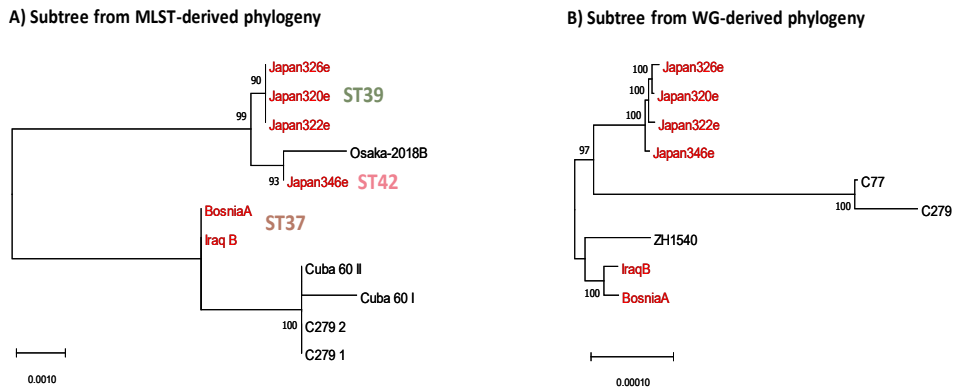
**Figure 40.** ML tree obtained after excluding *tp0897*, *tp0316*, *tp0317*, and recombinant genes from the multiple genome alignment generated in Chapter 3. Samples in common with the phylogeny obtained from the concatenated genes sequences of the MLST scheme are highlighted in red (Figure 39). Link to the online figure with a higher resolution: [https://drive.google.com/file/d/1B402D\\_r7Gk5ZSlipacsXqTcPFBL04rz/view?usp=sharing](https://drive.google.com/file/d/1B402D_r7Gk5ZSlipacsXqTcPFBL04rz/view?usp=sharing)

This phylogeny (Figure 40) was chosen for the comparison because it includes a very complete selection of genomes, representative of the variation in *T. pallidum*. To facilitate the ensuing description, we will refer to MLST-derived phylogeny (Figure 39) as the one obtained by concatenating the seven genes of the new MLST scheme, and to WG-derived phylogeny (Figure 40) as the one obtained from whole genomes in Chapter 3 of the present thesis.

For the comparison, we only considered from the 47 samples present in both phylogenies, the 16 samples that were part of STs with more than one sample from MLST-derived phylogeny (highlighted with a red circle in Figure 39), with the exception of sample Japan346e, because of its interesting genomic relationship with other three samples of interest from TEN clade (Japan326e, Japan322e and Japan320e). We were specifically interested in comparing whether the samples were also placed in WG-derived phylogeny (Figure 40) as monophyletic groups or not. STs formed by repeated samples, as detailed previously (Table 28), were not included in this comparison, except for a repeated sample from ST82 (SS14\_1) because, unlike the other STs with repeated samples that had only one or two samples, this one had eight samples. Additionally, to further enhance the comparison, we extracted the subtrees containing the samples or sublineages being compared and employed **MEGAX** [276] for visualization, as illustrated in the figures below.

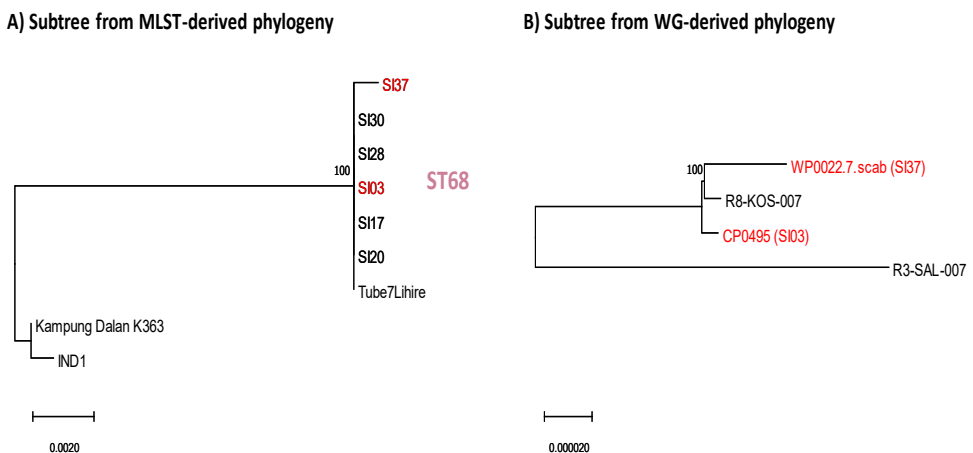
We started by comparing the STs of TEN. In the MLST-derived phylogeny (Figure 39), ST39 includes three of the four Japanese TEN strains whereas in the WG-derived phylogeny (Figure 40) these three strains (Japan326e, Japan322e and Japan320e) are grouped in a paraphyletic clade that includes sample Japan346e, which is assigned to ST42 (Figure 41). Interestingly, the classification of these four samples into two different STs is attributed to the presence of 2 single nucleotide polymorphisms (SNPs) between them, both present in two loci included in the MLST scheme, *tp0326* and *tp0865*.

Moreover, the two strains from ST37 (IraqB and BosniaA), form a monophyletic clade both in MLST-derived phylogeny and WG-derived phylogeny (Figure 41).



**Figure 41.** Enhanced view of TEN lineage subtrees extracted from A) MLST-derived phylogeny and B) WG-derived phylogeny, offering a detailed analysis of their genetic relationships. The highlighted samples in red indicate shared samples between the two phylogenies, characterized by multiple samples forming their respective STs (Figure 40), allowing for their comparison.

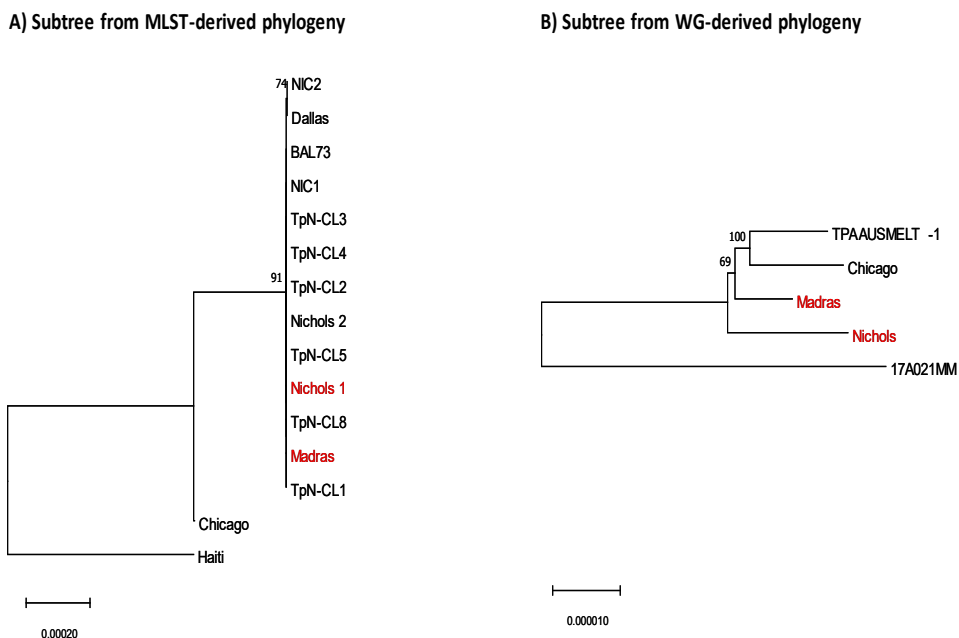
Next, we compared the STs found for TPE. The only ST from TPE formed by more than one sample is ST68, including strains SI03 and SI37, and which is monophyletic in both phylogenies (Figure 42).



**Figure 42.** Enhanced view of TPE subtrees extracted from A) the MLST-derived phylogeny and B) the WG-derived phylogeny, offering a detailed analysis of their genetic relationships. The highlighted samples in red indicate common samples in the

two phylogenies, characterized by multiple samples forming their respective STs (Figure 39), allowing for their comparison.

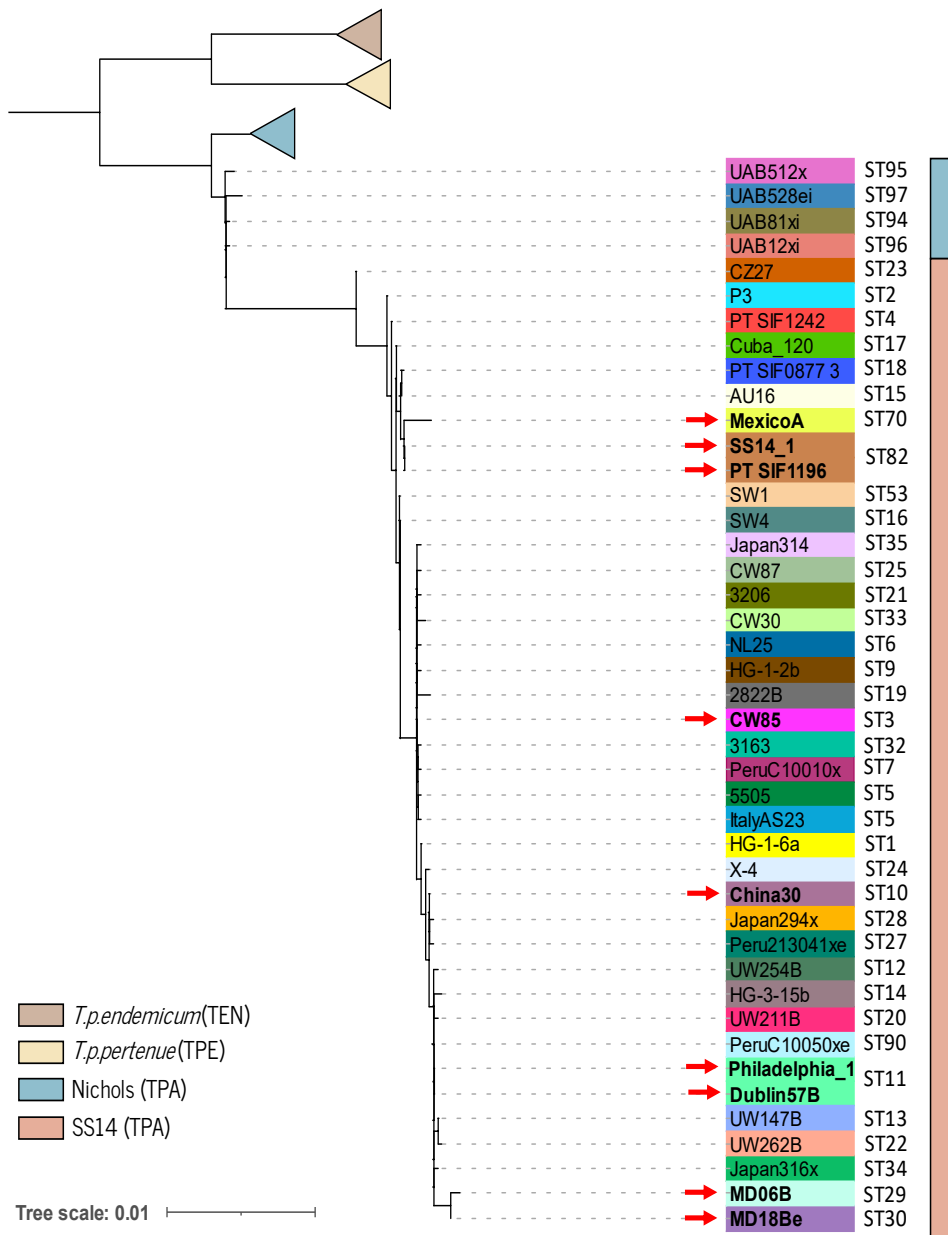
Regarding the Nichols sublineage of TPA, of the 13 strains included in ST36 in MLST-derived phylogeny (Figure 39), the only two samples included in WG-derived phylogeny (Figure 40), Madras and Nichols, belong to a monophyletic group (Figure 43).



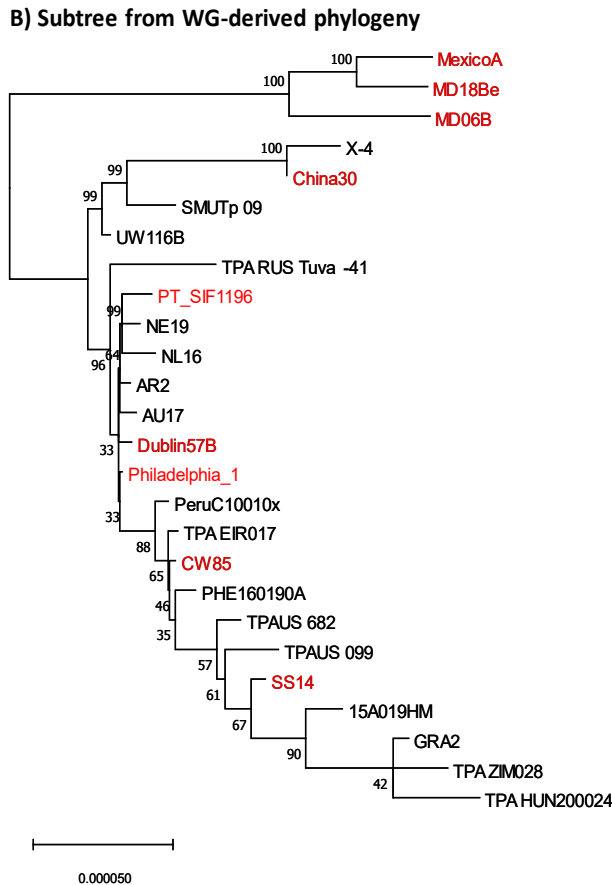
**Figure 43.** Enhanced view of a specific TPA-Nichols sublineage subtrees extracted from A) MLST-derived phylogeny and B) WG-derived phylogeny, offering a detailed analysis of their genetic relationships. The highlighted samples in red indicate common samples in the two phylogenies, characterized by multiple samples forming their respective STs (Figure 39), allowing for their comparison.

With respect to the sublineage SS14 of TPA, in MLST-derived phylogeny (Figure 39 and Figure 44), ST82 had eight strains of which only two are included in WG-derived phylogeny, SS14\_1 and PTSIF1196, and both are placed as two monophyletic clades in the phylogeny B (Figure 40 and Figure 45).





**Figure 44.** MLST-derived phylogeny with the TEN, TPE, and TPA-Nichols lineages condensed to highlight the TPA-SS14 sublineage. Furthermore, to enhance the resolution of the phylogeny, only one strain has been retained for each sequence type (ST), except in cases where the ST have multiple strains shared with the whole-genome-derived phylogeny. The highlighted samples with a red arrow indicate common samples in this phylogeny and the WG-derived one, characterized by multiple samples forming their respective STs (Figure 40), allowing for a comparison.



**Figure 45.** Enhanced view of a TPA-SS14 sublineage subtree extracted from the WG-derived phylogeny, offering a detailed analysis of their genetic relationships. The highlighted samples in red indicate common samples in this and MLST-derived phylogeny, characterized by multiple samples forming their respective STs (Figure 39), allowing for their comparison.

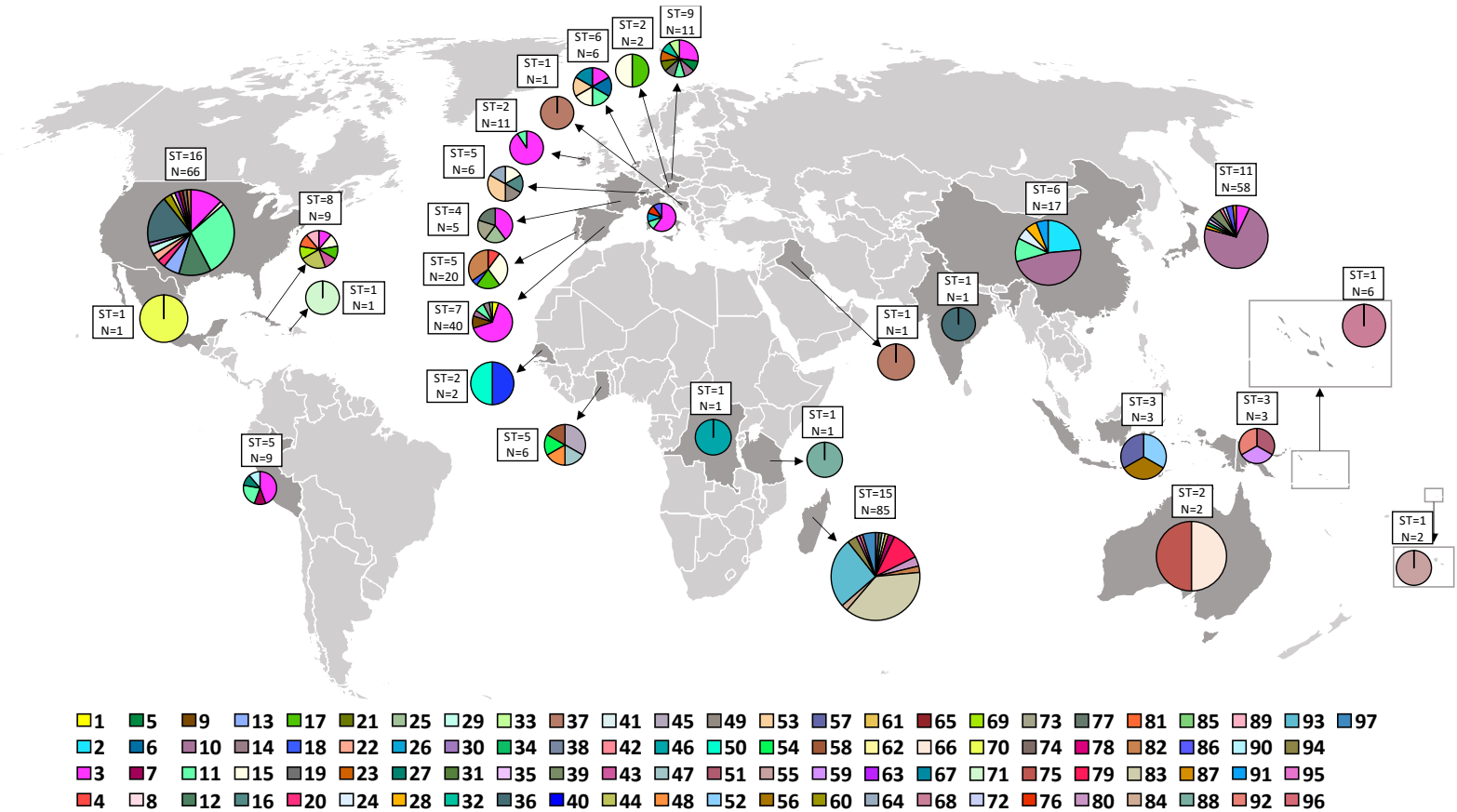
The strain Mexico A, from ST70, was also placed forming a monophyletic group in WG-derived phylogeny (Figure 40 and Figure 45) together with strain MD18Be. However, MD18Be is classified as ST29 in the MLST-derived phylogeny (Figure 39 and Figure 44), which would form a polyphyletic group with ST70 in MLST-derived phylogeny (Figure 39 and Figure 44).

Lastly, ST3, ST10, and ST11 are represented as monophyletic clades in the MLST-derived phylogeny (Figure 39 and Figure 44). Within these STs, one sample from

ST3 (CW85), other from ST10 (China30), and another two from ST11 (Dublin57B and Philadelphia), are shared in the WG-derived phylogeny (Figure 40 and Figure 45) forming distinct monophyletic clusters.

### **3. 6 Population genetic structure**

The distribution of STs identified by country for all samples typed with the new MLST scheme for *T. pallidum* is shown in Figure 46. Madagascar is the country with the largest number of studied samples (n=85) followed by the USA (n=66). However, the country with the largest number of different STs is the USA, with 18 different STs, being ST11 the most abundant one, with 19 samples. After the USA, Madagascar is the next country with the largest number of different STs (15), being ST93 the most abundant ST with 22 samples.



**Figure 46.** Distribution of STs identified by country for all samples typed with the new MLST scheme for *T. pallidum*. For each country, the number of samples and the number of different STs obtained is indicated. The legend shows the different colors for each identified ST.

We assessed the potential relationship between the identified sequence types (STs) and the geographical origin of the samples. Populations were defined based on the countries and continents from which the samples originated. Notably, we focused our analyses solely on the TPA subspecies due to the limited availability of samples for TPE and TEN. We excluded Oceania as a continent from the analysis due to the presence of only two samples.

One continent, America, had a remarkably larger within-group nucleotide diversity ( $k=12.81$ ,  $\pi = 0.00419$ ) than the rest, which had very similar values (Table 29 and Table 30). The average numbers of nucleotide differences between continents were larger than the average number of nucleotide differences within continents, with the notable exception of the American continent, whose within-group value was larger than the net differentiation with Europe ( $k=9.20$ ,  $D_{xy}(JC) = 0.00039$ ) and Asia ( $k=10.41$ ,  $D_{xy}(JC) = 0.00049$ ). The lowest intracontinental diversity was found in Africa ( $k=4.58$ ,  $\pi = 0.00166$ ). Africa and Asia showed the highest differentiation ( $k=37.16$ ,  $D_{xy}(JC) = 0.00902$ ), while Europe and Asia exhibited the lowest value ( $k=7.29$ ,  $D_{xy}(JC) = 0.0003$ ).

At the country level (Table 31 and Table 32), Ireland and Portugal had the lowest within-group nucleotide diversity, with ( $k=0.18$ ,  $\pi = 0.00103$ ) and ( $k=0.52$ ,  $\pi = 0.00014$ ), respectively. On the other hand, Italy and the USA showed the highest values, with ( $k=19.16$ ,  $\pi = 0.00258$ ) and ( $k=17.59$ ,  $\pi = 0.00582$ ), respectively. The mean number of nucleotide differences between countries was generally higher than the mean number of nucleotide differences within countries for Madagascar, Portugal, Italy, and Ireland. However, there were exceptions. For example, Ireland ( $k=0.18$ ,  $D_{xy}(JC)=0.00042$ ), Italy ( $k=19.16$ ,  $D_{xy}(JC)=0.00023$ ), and Czech Republic ( $k=3.86$ ,  $D_{xy}(JC)=0.00021$ ) showed a lower mean number of nucleotide differences between continents compared to within their respective countries. Additionally, Spain ( $k=3.29$ ,  $\pi=0.00092$ ) and Czech Republic ( $k=3.86$ ,  $\pi=0.00194$ ) demonstrated a higher mean number of nucleotide differences within countries

when compared to Ireland. However, when comparing countries, the mean number of nucleotide differences between them was generally lower than the mean number of nucleotide differences within each country. This pattern held true for the USA ( $k=17.59$ ,  $\pi=0.00582$ ), with the exception of comparisons involving Madagascar ( $k=9.43$ ,  $D_{xy}(JC)=0.00569$ ) and China ( $k=7.154$ ,  $\pi=0.00074$ ). Similarly, for China, the mean number of nucleotide differences between countries was lower, except for comparisons with Italy ( $k=19.16$ ,  $D_{xy}(JC)=0.00026$ ), Madagascar ( $k=9.43$ ,  $D_{xy}(JC)=0.00917$ ), and the USA ( $k=17.59$ ,  $D_{xy}(JC)=0.0007$ ).

**Table 29.** Average number of nucleotide differences ( $k$ ) in a population and between populations at continental level. The table is colored with the next color codes: green for the lowest values, yellow for intermediate values and purple for the highest values. The number of samples ( $N$ ) from each continent is also indicated.


N	Continents	Europe	Africa	Asia	America	Color legend
110	Europe	4.66				
85	Africa	24.91	4.58			
34	Asia	7.29	37.16	5,12		
82	America	9.20	19.20	10.41	12.81	

**Table 30.** Average number of nucleotide substitutions per site within the four continents considered ( $\pi$ ), and and between all of them with Jukes and Cantor, Dxy(JC) on the left side of the matrix, and Standard deviation of Dxy(JC) on the left right of the matrix. The table is colored with the next color codes: green for the lowest values, yellow for intermediate values and purple for the highest values. The number of samples (N) from each continent is also indicated.

N	Continents	Europe	Africa	Asia	America	Color legend
110	Europe	<b>0.00187</b>	0.0008	0.00048	0.00048	
85	Africa	0.00589	<b>0.00166</b>	0.00134	0.00084	
34	Asia	0.0003	0.00902	<b>0.00175</b>	0.00064	
82	America	0.00039	0.00407	0.00049	<b>0.00419</b>	

Chapter 4

**Table 31.** Average number of nucleotide differences (k) in a population and between populations at country level. The table is colored with the next color codes: green for the lowest values, yellow for intermediate values and purple for the highest values.

N	Countries	Spain	USA	Madagascar	Portugal	Japan	Italy	Czech Republic	China	Ireland	Color legend
40	Spain	3.29									
66	USA	13.49	17.59								
85	Madagascar	40.90	27.34	9.43							
20	Portugal	5.77	13.47	36.26	0.52						
53	Japan	6.66	14.39	39.54	8.71	7.99					
10	Italy	13.20	16.56	36.45	14.38	15.63	19.16				
11	Czech Republic	3.33	11.63	34.02	5.48	6.19	12.00	3.86			
17	China	6.29	14.09	37.13	6.19	7.30	14.89	5.69	7.154		
11	Ireland	1.76	12.84	41.52	4.46	5.59	12.14	2.04	5.27	0.18	



Chapter 4

**Table 32.** Average number of nucleotide substitutions per site within the nine countries considered ( $\pi$ ), and and between all of them with Jukes and Cantor, Dxy(JC) on the left side of the matrix, and Standard deviation of Dxy(JC) on the left right of the matrix. The table is colored with the next color codes: green for the lowest values, yellow for intermediate values and purple for the highest values.

N	Countries	Spain	USA	Madagascar	Portugal	Japan	Italy	Czech Republic	China	Ireland	Color legend
40	Spain	<b>0.00092</b>	0.00097	0.00148	0.00044	0.00075	0.00247	0.00045	0.00094	0.00046	
66	USA	0.00098	<b>0.00582</b>	0.00114	0.00092	0.00103	0.00157	0.00089	0.00109	0.00157	
85	Madagascar	0.01091	0.00569	<b>0.00295</b>	0.00152	0.00159	0.00289	0.0017	0.0019	0.00228	
20	Portugal	0.00108	0.00142	0.00985	<b>0.00014</b>	0.00067	0.00228	0.0004	0.00088	0.00035	
53	Japan	0.00028	0.00074	0.00983	0.00127	<b>0.00224</b>	0.00258	0.00061	0.00106	0.00074	
10	Italy	0.00018	-0.00008	0.00686	0.00101	0.00023	<b>0.00258</b>	0.00194	0.00221	0.0029	
11	Czech Republic	-0.00002	0.00076	0.00873	0.00091	0.00021	0.00014	<b>0.00194</b>	0.00074	0.00039	
17	China	0.00031	0.0007	0.00917	0.00065	-0.00002	0.00026	0.00022	<b>0.00074</b>	0.00103	
11	Ireland	0	-0.00008	0.01154	0.00114	0.00042	0.00031	0.00001	0.00045	<b>0.00103</b>	

## 4. Discussion

In recent years, the incidence of treponematosi s has markedly increased worldwide and has become a serious global health problem. Moreover, currently there is no universally adopted culture method for *T. pallidum* that can be conveniently used to obtain comprehensive genomes of the bacterium. Although the number of complete *T. pallidum* genomes available has increased due to the introduction of an enrichment technique and WGS directly from clinical samples since 2016 [2,21,36,70–77], obtaining whole genome sequences of *T. pallidum* is still a very time-consuming and costly process. For all these reasons, quick and economical typing procedures that may be used in any laboratory with access to Sanger sequencing resources are required. These techniques would reveal details on the frequency of strain types and variations over time, whether specific strain types are related to particular patient populations, and trends among macrolide resistant strains. Moreover, the availability of more *T. pallidum* whole genomes presents a chance to discover additional genes useful for typing.

Here, we analyzed 121 whole or draft genome sequences of *T. pallidum* to design an innovative and practical molecular typing scheme. Given that *T. pallidum* cannot yet be cultured routinely, and the low DNA content in clinical samples, our objective was to minimize the number of typing loci while adhering to the standardized gene count for an MLST scheme [45–47]. Additionally, since *T. pallidum* is considered a monomorphic bacterium for which each SNP can provide valuable information, we focused on loci with the highest SNP densities to ensure adequate variability for capturing the overall diversity of species. Furthermore, we aimed at maximizing the resolution of the new MLST scheme, enabling differentiation within each lineage and sublineage of this bacterium while distinguishing between all *T. pallidum* subspecies.

Our findings point to seven variable genes (*tp0136*, *tp0326*, *tp0548*, *tp0705*, *tp0858*, *tp0865* and *tp1031*) as the most suitable for typing. These genes—along with the

23S rRNA genes—are proposed here as a new molecular typing scheme for *T. pallidum*. All these loci have been previously recognized as including recombinant regions involving different *T. pallidum* subspecies [73,78,108,175]. These partially recombining regions are very helpful in increasing the level of discrimination for the selected gene fragments included in the new MLST scheme. Incorporating these genes, which exhibit significant heterozygosity, will facilitate forthcoming population genetics analyses based on MLST haplotypes. This inclusion is valuable as these genes can provide essential insights into genetic variations, enabling enhanced discrimination among various strains or subtypes.

The proposed *T. pallidum* MLST scheme shares five loci (*tp0136*, *tp0326*, *tp0548*, *tp0705* and *tp0865* with previous typing schemes available for the different subspecies of *T. pallidum* [5,55,59–61] (see Table 3). Hence, this novel MLST approach can be regarded as a refined and unified version of previous typing schemes, now applicable to all subspecies of *T. pallidum*, after the incorporation of *tp0858* and *tp1031*. Consequently, it would be highly valuable to assess the enhanced resolution achieved by the new scheme proposed in this study when compared to all previous typing schemes.

The efficiency of amplification in typing schemes is influenced by various factors, such as the type of material, time interval between sample collection and DNA isolation, DNA extraction technique, length of amplification product, and amplification protocol. To enhance resolution capability, we opted for relatively shorter loci compared to previous typing schemes. As a result, we were able to amplify samples using a single PCR instead of nested PCRs, which significantly reduced time and cost as well as improved the efficiency. These improvements led to a moderate overall rate of 45% (82/184) for fully typed samples in experimental settings. In contrast, we obtained a higher efficiency (82%,151/184) for the amplification of the 23S gene compared to that obtained for the 7 genes of the new MLST scheme. These rates are comparable to the amplification efficiency observed

in previous *T. pallidum* typing studies, which ranged from 20% to 93% with a median of 54% [35,61,62,313,371–384]. However, a limitation of comparing these studies is that they employed different criteria for sample collection and investigation of *T. pallidum*, which could introduce heterogeneity in the efficiency obtained for each study.

It is important to note that a significant portion of the samples used in this study consisted of remnants from previous investigations, resulting in limited quantities of DNA with compromised quality. These factors may have significantly impacted the efficiency achieved in our analyses, particularly for the 7 genes included in the new MLST scheme. In many cases, we were able to obtain the sequence of the 23S gene because it was tested in the first place, but not the sequence of some of the genes in the MLST scheme due to insufficient sample or poor DNA quality. As a result, the efficiency of obtaining the 23S gene sequence was higher compared to all the genes of the MLST scheme. Therefore, to obtain a more reliable assessment of efficiency, it is crucial to conduct further testing using recently collected samples of higher quality.

We encountered some issues with the amplification efficiency of primers targeting the *tp0858* and *tp0865* genes in TEN samples. Consequently, new primers were specifically designed for these genes. The new primers for the *tp0858* gene were successfully tested on a set of TEN samples, leading to the identification of a deletion in the forward primer that hindered the amplification of these samples (Figure 34). On the contrary, the newly designed primers for the *tp0865* gene, while amplifying properly a Nichols sample used as a positive control, did not amplify the target region in a subset of TEN samples that previously exhibited amplification problems. To ascertain the effectiveness of the entire set of new and old primers for the *tp0865* gene, additional testing is required. This entails utilizing more clinical samples from the TEN subspecies, as well as investigating the gene sequence in more whole genomes, in order to determine the cause of primer ineffectiveness

because at present, no specific reason for this ineffectiveness has been identified. Based on the results mentioned above, we recommend utilizing the primers listed in Table 22, with the exception of samples exhibiting amplification issues for the *tp0858* and *tp0865* genes. In such cases, we advise employing the alternative primers outlined in Table 23 and Table 24.

In our study, we applied the new typing scheme to analyze a total of 542 samples, both experimentally and *in silico*. The results demonstrated a high level of discrimination, as we identified 97 different allelic profiles in 386 out of the 542 samples. However, we encountered challenges in assigning STs to 157 samples. These challenges arose primarily from unsuccessful gene amplification in the experimentally typed samples and a significant amount of missing data in the gene sequences of the *in silico*-typed samples. Among the 97 STs identified, 37 belonged to the SS14-sublineage, 35 to the Nichols-sublineage, 18 to TPE, and 7 to TEN. Nevertheless, values of genetic diversity from the frequency distributions of the different STs within subspecies/sublineages did not show large differences (ranging from 0.91 to 0.83, Table 27).

Consistent with prior research, our findings reveal a significantly high proportion of macrolide-resistant strains (49.4% of samples) among those tested for resistance (393 out of 542 samples) [36,83,107]. As a result, second-line therapy with azithromycin is no longer advised for individuals who have a penicillin allergy or bleeding problems [385]. This underscores the necessity for further epidemiological investigations aimed at monitoring and characterizing the escalating dissemination of macrolide resistance in *T. pallidum*. Among the samples tested for resistance (393/542), a remarkable 86% of the resistant strains were identified as belonging to the SS14-sublineage (166/393). Additionally, we detected 12 resistant strains from the Nichols-sublineage and 4 from TEN, while 12 additional resistant strains were attributed to TPA, without precise sublineage determination. The majority of the resistant samples exhibited the A2058G mutation, whereas only 6 resistance samples (6/194) displayed the A2059G

mutation. Among these six samples, five originated from the SS14-sublineage, and one from the Nichols-sublineage. Notably, we encountered challenges in determining the macrolide sensitivity or resistance of 116 samples through *in silico* analysis, as well as 33 samples through experimental means. This was primarily due to substantial missing data in the 23S gene sequences during *in silico* analysis and difficulties in amplifying this gene for these particular samples, likely attributed to low DNA quality and/or quantity. Overall, these findings underscore the ongoing challenges posed by macrolide resistance in *T. pallidum* and emphasize the need for comprehensive epidemiological investigations and genomic studies to inform public health efforts and combat the escalating dissemination of resistance.

Furthermore, the observed clustering patterns among various subspecies and lineages of *T. pallidum* confirm their clear differentiation in the phylogeny generated using the concatenated gene sequences from the new MLST scheme (Figure 32). This distinction remains consistent when compared to the phylogenetic relationships obtained from genome-wide data (see Supplementary Figure 39), indicating congruence with the vertical relationships of all *T. pallidum* strains. The MLST-derived phylogeny is topologically congruent with that derived from whole genomes, enabling differentiation between the three subspecies and the two major sublineages of *T. pallidum* (TPE, TEN, TPA-Nichols, and TPA-SS14), as well as finer distinctions within them. However, a comprehensive phylogeny of whole genomes that encompasses all samples typed by the new MLST scheme in this project is currently unavailable and such a phylogeny would be necessary for a thorough comparison of the resolution provided by MLST versus whole genomes. Nonetheless, considering the number of distinct sequence types (97) obtained and comparing it to WG-derived phylogeny (Supplementary Figure 39), it can be inferred that the resolution offered by MLST is nearly equivalent to that of whole genomes.

This study included twenty-one duplicate samples, each obtained by different methods (as shown in Table 28). Interestingly, only seven of these samples yielded the same sequence type (ST), despite originating from the same DNA sample but collected through different means. These findings emphasize the significance of sequence quality in epidemiological studies, as errors in sequencing can lead to incorrect assignment of sequence types. The results presented in Table 28 demonstrate that Sanger sequencing is the most reliable methodology in terms of results. Although *de novo* assembly for genomes with high coverage can yield similar outcomes, they still exhibit ambiguous positions, even when combined with Sanger sequencing to resolve complex genomic regions. Consequently, in the absence of a standardized culture system for *T. pallidum* that guarantees obtaining closed genomes with high coverage, MLST schemes remain a dependable and essential tool for *T. pallidum* epidemiology.

The assessment of genetic diversity and population structure of a pathogen holds great significance in disease surveillance as it offers valuable insights into the dissemination and transmission routes of the pathogen. To this end, we explored the genetic variations within and between populations, taking into consideration the geographic origin of the samples. We observed genetic variations within and between populations based on geographic origin, highlighting the existence of distinct population structures. This suggests that *T. pallidum* strains may have diverged over time in different geographical regions, indicating potential localized transmission patterns. Furthermore, our analysis revealed a consistent pattern of a greater number of nucleotide differences between groups compared to within subpopulations at both the continent and country levels. This observation suggests a clear geographic clustering of *T. pallidum* strains. This implies that certain genetic variants are more prevalent in specific regions, indicating possible localized transmission networks or restricted movement of the bacteria across geographic boundaries.

Interestingly, we found a higher genetic variation in Africa and Asia compared to other continents, which could imply more diverse *T. pallidum* populations in these areas. This could be indicative of higher transmission rates, increased population mixing, or longer-standing endemicity in these regions, potentially influencing the spread and persistence of the infection. Furthermore, the consistently higher number of nucleotide differences observed in Madagascar compared to other countries suggests a unique genetic profile and potential epidemiological connectivity to multiple regions. This finding underscores the importance of considering specific regional factors and transmission dynamics when analyzing the spread of *T. pallidum*. Moreover, all these results emphasize the need for further research to investigate transmission routes, conduct comparative analyses, and perform longitudinal studies to enhance our understanding of *T. pallidum* dynamics and inform targeted control measures.

To sum up, we have used information from 121 *T. pallidum* genomes to design a new multi-locus scheme for typing (MLST) that differs from the previous ones and allows differentiating between the three subspecies of *T. pallidum* and within lineages of TPA with a higher resolution than pre-existing approaches. This scheme is based on the sequencing of seven loci, and the additional analysis of the 23S rDNA genes to determine macrolide resistance/sensitivity. Furthermore, this refined scheme has undergone experimental fine-tuning and testing using a collection of *T. pallidum* samples.

Remarkably, all targets can be amplified simultaneously using the same amplification protocol, eliminating the need for a nested amplification step. This approach significantly broadens the scope of treponemal samples that can be typed, even in cases where the sample amount is scarce. Moreover, we are currently working with the PubMLST development team to make this new MLST scheme public in their database. Then, the sequences of the representative alleles of each identified ST, which are complete and do not have indeterminate positions, will also be made public in PubMLST.



We hearten the scientific community and public health authorities to embrace the utilization of this new MLST scheme, which can be readily implemented in standard laboratories, offering ease and speed in its application. This powerful tool is expected to bring forth new possibilities in the field of epidemiology, enabling longitudinal studies of *T. pallidum* allelic profiles across different locations. Additionally, it will enhance the tracking of the infections and facilitate the identification of associations between specific strains and particular patient groups. These advancements will greatly contribute to our understanding of *T. pallidum* epidemiology, ultimately leading to improved public health interventions as well as to establish a global network for molecular surveillance as the results will be in a publicly accessible database.



— **DISCUSSION** —



## Discussion

In recent years, there has been a significant rise in the application of high throughput sequencing (HTS) technology for genomic and epidemiological research on *T. pallidum*. When this doctoral thesis was started, only around 75 complete *T. pallidum* genomes were available for analysis [78,79]. However, the field has undergone a remarkable transformation, and the number of accessible genomes is now over 1,400, representing a substantial increase in genomic resources for studying this bacterium [2,21,36,70–77]. This exponential growth in available *T. pallidum* genomes provides researchers with unprecedented opportunities to explore its genetic diversity, evolutionary patterns, and population dynamics. Therefore, this doctoral thesis aims to shed light on these aspects by examining both ancient and contemporary genomes of *T. pallidum* through four distinct studies.

In Chapter 1, our study focused on a comprehensive collection of 75 contemporary *T. pallidum* genomes to assess recombination and selection in this highly clonal pathogen. To accomplish this objective, we developed a novel pipeline, called PIM, specifically designed for accurate recombination detection in *T. pallidum* genomes. Our findings demonstrated that PIM outperformed Gubbins and ClonalFrameML, two widely utilized tools for recombination detection, in terms of accuracy. Furthermore, PIM has been effectively used in other studies to detect recombination in different bacteria such as *Lactococcus garvieae* [386] and *Leptospira* [387]. Despite this, Gubbins and ClonalFrameML continue to be extensively employed in epidemiological studies involving large datasets comprising hundreds or even thousands of *T. pallidum* genomes [36,71,83,107]. The rationale behind this preference lies in the fact that those programs offer a simpler and more automated approach suitable for handling vast genomic datasets necessary for epidemiology analysis, despite potentially overestimating recombination events in *T. pallidum* when compared to PIM. However, ensuring accuracy in detecting recombination is

essential for maintaining the integrity of the genome alignment used in phylogenetic reconstruction and subsequent analyses.

Removing genes or regions from the alignment can lead to a significant loss of variability, potentially influencing the obtained results and their interpretation, particularly in the case of *T. pallidum*, a pathogen recognized for its clonality. Moreover, in contrast to other longitudinal genomic studies primarily focused on epidemiology [36,71,83,107], the research conducted in this doctoral thesis has not been aimed at generating new genomes of this bacterium. Instead, we utilized the wealth of existing genomes accessible in public databases for a comprehensive analysis of the genomics and evolution of *T. pallidum*. Consequently, the tool used for recombination detection in Chapters 1-3 has consistently been PIM. Given the technical limitations of utilizing PIM for extensive datasets, we conducted a meticulous selection process for the genomes and dataset composition in each project. The goal was to exclude unnecessary genomes and prioritize those that would offer the most representative and detailed understanding of the variation and diversity within *T. pallidum*, based on the genomes available at the start of each project. This approach aimed to ensure the highest quality and relevance of the selected genomes, allowing for a comprehensive analysis of the pathogen.

The recombination analyses conducted in Chapters 1-3 revealed variations in the number of recombinant genes detected among the different datasets. In Chapter 1, 12 recombinant genes were identified from the analysis of 75 genomes. Chapter 2 yielded 18 recombinant genes from 77 genomes, while in Chapter 3 we detected 27 recombinant genes from a dataset of 99 genomes. This discrepancy in the number of identified recombinant genes can be attributed to the specific number and composition of genomes used in each project. Each chapter of the thesis started at different times, Chapter 1 of the thesis commenced in early 2019, followed by Chapter 2 in early 2021, and finally Chapter 3 in late 2022. Despite efforts to select genomes that encompassed the major subspecies and sublineages of *T. pallidum*

and provided optimal variability, the limited number of available genomes at each time, particularly for TPE and TEN genomes in Chapter 1 and 2, imposed a limitation that might have impacted the detection of recombination events.

In addition to detecting varying quantities of recombinant genes in each chapter, we also observed a distinct tree topology in Chapter 2 compared to the topologies obtained in the other chapters when recombinant genes were excluded from the whole genome alignment used for phylogenetic reconstruction. Specifically, our analysis revealed that in the reference phylogeny obtained in Chapter 2, the TEN clade, which was consistently identified as a distinct, sister clade of TPE, was unexpectedly found to be nested within the TPE clade, still with high bootstrap support. This incongruence in the tree topologies could be attributed to several factors. Firstly, the composition of the genomes included in each dataset varied, directly impacting the analyzed diversity and variation in each project. The number of genomes alone does not significantly influence this outcome, but rather the extent of genetic diversity among them and their contribution to overall variation and resolution representative of the species (in terms of genome-wide differences). Additionally, the removal of recombinant genes in each dataset, which affects the variability retained for phylogenetic reconstruction, can contribute to the different topologies observed.

The detection of recombinant genes differs depending on the genomes included in the analyzed dataset, and thus, the variation removed also varies. For instance, in the whole genome alignment analyzed in Chapter 1, we removed 486 out of 2,625 SNPs (18.5%). In Chapter 2, 1,601 out of 4,822 SNPs (33.2%) were removed, and in Chapter 3, we removed 2,431 out of 6,149 SNPs (39.5%). These significant differences in the amount of variation eliminated from each dataset, combined with the overall increase in total variation analyzed, likely contribute to the topological incongruence observed in Chapter 2. Despite the observed topological incongruence in Chapter 2, it is essential to emphasize that all relevant clades in the

phylogenies obtained in the present thesis, including the one showing the nesting of the TEN clade within the TPE clade in Chapter 2, are supported by robust bootstrap values. This indicates that the depicted branching patterns and relationships in the phylogenetic trees have strong data support and can be considered reliable in the context of the dataset and analytical methods employed. Therefore, while the results suggest that recombination has played a significant role in the diversification of these subspecies, further analyses are required to understand the specific factors contributing to the observed nesting of the TEN clade within the TPE clade in the Chapter 2 phylogeny.

We have also observed a close relationship between recombination and selection as evidenced by the strong signals of positive selection detected in all the identified recombinant genes in both Chapters 1 and 2. Additionally, we analyzed the functional roles of the proteins encoded by these recombinant genes, as well as other genes under positive selection detected. Although some proteins had unknown functions, the majority of the proteins identified in both studies appear to have crucial roles in the pathogen's defense against the host immune system and are potentially implicated in virulence.

Recent investigations with whole genome sequences of different bacteria have provided valuable insights into the significance of recombination during the initial stages of speciation. Notably, bacteria such as *Mycobacterium tuberculosis* [316] and *Vibrio cholerae* [317] serve as examples where recombination has played a vital role in adapting to new ecological niches and achieving successful expansion. These cases align with the clonal expansion model from a diverse population [318], wherein the introduction of genetic variation through recombination, coupled with natural selection, facilitates the preservation of changes and the fixation of advantageous alleles in recombinant genes. Our findings suggest that a similar mechanism may have played a role in the early evolution of *T. pallidum*, leading to its subsequent epidemic spread. Moreover, extensive research has demonstrated the



importance of recombination in shaping the population structure and genetic diversity of various bacteria, including *Neisseria meningitidis* [388,389], *Streptococcus pneumoniae* [177,390], *Escherichia coli* [391,392] and *Salmonella enterica* [393,394]. These studies highlight the significant contribution of recombination and selection in driving gene diversity, promoting evolutionary and adaptive processes in the context of host-pathogen interactions. Therefore, we can confidently conclude that recombination and selection are key factors in the evolutionary dynamics of *T. pallidum*, facilitating its adaptation and survival.

In our investigations about recombination in *T. pallidum*, we made an intriguing observation: all the recombination events identified primarily occurred between different subspecies rather than within the same subspecies. The underlying factors contributing to this pattern remain uncertain and warrant further investigation. One possibility is that the species exhibits low levels of genetic variation, which can pose challenges for detecting recombination events within subspecies/sublineages using the PIM method. Another hypothesis is that recombination between different subspecies facilitates the exchange of genetic material, thereby enhancing the bacterium's adaptability and survival in diverse environments. Given that distinct subspecies face varying selective pressures, recombination events between them can result in the acquisition of advantageous traits, such as increased virulence or resistance to host immune responses. Moreover, limited opportunities for recombination within the same subspecies may be influenced by factors like geographic isolation, host specificity, or other ecological factors. These elements can restrict the intermixing and interaction of strains within a subspecies, reducing the likelihood of recombination occurring within this particular group. However, due to the current lack of understanding regarding the specific molecular processes responsible for recombination in *T. pallidum*, further research is necessary to elucidate the underlying mechanisms driving these observed patterns. Gaining insights into the molecular mechanisms and factors influencing recombination will

greatly contribute to our understanding of the evolutionary dynamics and adaptive processes of this pathogen.

The absence of recent recombination events in *T. pallidum* has implications for incorporating the recombinant genes into MLST schemes. The reason is that the observed alleles in these genes are more likely to arise from new mutations rather than from genetic exchanges between different strains. Consequently, the allelic variations identified through MLST analysis provide a reliable representation of the genetic diversity and evolutionary relationships among *T. pallidum* strains. This makes them well-suited for constructing phylogenetic trees and conducting population genetic studies. During the development of the new MLST scheme presented in Chapter 4, we were able to select recombinant genes that not only offered high resolution for the new scheme but also performed well in practical applications.

Furthermore, the phylogenetic tree topology in Chapter 4 demonstrated consistency with all other trees obtained in this thesis (excluding the phylogenetic tree of Chapter 2 discussed previously) and with those derived from previous studies. This further confirms the suitability of the chosen genes for the new MLST scheme. Moreover, the newly developed MLST scheme for *T. pallidum* offers valuable insights into the epidemiology of this bacterium by enabling tracking the transmission and propagation dynamics among its distinct subspecies. However, due to the lack of a standardized culturing method for *T. pallidum* and the challenges associated with obtaining genomes from this bacterium, an additional, noteworthy application of this MLST scheme arises, which was not explicitly mentioned in the chapter. This application pertains to the selection of appropriate genomes for sequencing, thus facilitating evolutionary studies. By employing the MLST scheme, researchers can effectively identify and classify *T. pallidum* strains into their respective subspecies and sublineages. This ensures that the genomes selected for analysis represent a diverse range of strains from each subspecies.

Consequently, this additional use of the proposed MLST scheme contributes to the enhancement of evolutionary studies by enabling researchers to choose genomes that accurately represent the genetic diversity and evolutionary relationships among *T. pallidum* subspecies.

Thanks to the remarkable advancements in sequencing technologies, obtaining ancient genomes of *T. pallidum* is now a reality, overturning the notion that it was an impossible endeavor just a few years ago. In this dissertation, we have successfully obtained two ancient high-coverage whole genomes of *T. pallidum*. The first genome obtained in Chapter 2 (W86) was classified as TPA and originates from Poland, dating back to the 17th century. The second genome obtained in Chapter 3 (ZH1540) was classified as TEN and is derived from human remains in Brazil, dating back approximately 2,000 years. It represents the first available pre-Columbian *T. pallidum* genome from the Americas.

By including and contextualizing both new ancient genomes in two diverse datasets of *T. pallidum* genomes, chosen to capture the wide range of variability in this bacterium, we were able to uncover several additional novel recombinant genes. The identification of the strains involved in each recombination event provides insights into the potential occurrence of recombination between TPE/TEN and TPA strains in the Old World. This suggests the coexistence and circulation of these subspecies in the same region, as evidenced by the presence of ancient genomes from both TPE/TEN and TPA lineages in the observed recombination events. Furthermore, the divergence dates are much older than previous estimates based on modern genomes alone. The inclusion of these ancient genomes in Bayesian molecular clock dating, extends the credible ranges, enhancing the accuracy and precision of evolutionary timeline estimations. This highlights the importance of including older samples in molecular clock analyses.

Other studies examining the inclusion of aDNA from increasingly older time periods have yielded valuable insights [188]. For instance, Bos *et al.* [395]

investigated the impact of including aDNA from three different *Y. pestis* datasets, progressively spanning older time periods. Their findings revealed that the inclusion of older sequences pushed back the estimated ages of events and significantly reduced dating uncertainty. Additionally, Duchêne *et al.* [396] conducted an analysis using aDNA from three different bacteria from previous studies: *M. leprae* [216], *M. tuberculosis* [217], and *Y. pestis* [246]. This analysis confirmed the existence of the rate phenomenon and highlighted the importance of aDNA in reconciling discrepancies in rate estimates across different time periods. The inclusion of aDNA in bacterial genomics has proven to be an invaluable tool in addressing incongruity and gaining deeper insights into evolutionary rates. By bridging the gap between species and lineage comparisons, aDNA contributes to a better understanding of bacterial evolutionary processes and helps refine our knowledge of their timelines [188].

Interestingly, the dating methods used for the two bone samples were different. Sample ZH1540 was dated using the radiocarbon technique, whereas sample W86 relied on archaeological context for dating. However, it is important to note that radiocarbon dating becomes less reliable as we approach more recent centuries. Initially developed for longer geological time spans, it is based on isotopic average values. Furthermore, there is a carbon anomaly called the Maunder Minimum, a period of low solar activity that occurred from approximately 1645 to 1715, which is known to have had effects on radiocarbon dating [397]. During this time, there was a decrease in the number of sunspots and solar flares, which are indicators of solar activity. The decrease in solar activity during the Maunder Minimum resulted in a reduction of cosmic ray intensity reaching the Earth's atmosphere [398]. With fewer cosmic rays interacting with nitrogen in the atmosphere, the production rate of radiocarbon (C-14) was likely lower compared to periods of higher solar activity. As a result, organic materials from the Maunder Minimum period may appear older than they actually are when dated using radiocarbon dating [397,399]. It is worth noting that the specific effects of the Maunder Minimum on radiocarbon dating can

vary depending on factors such as the geographic location, sample type, and calibration methods used [397]. Consequently, if there is strong evidence suggesting that a sample originates from a period after 1500, radiocarbon dating is seldom conducted. In such cases, young graves often provide more accurate dating through the presence of dates, grave goods, or other items, surpassing the precision offered by natural science methods.

In addition to *T. pallidum*, several other bacterial species have also yielded ancient genomes, as is detailed in section 7.3 of the Introduction. Among these bacteria, *Yersinia pestis* has the largest number of available ancient genomes (n=71), followed by *M. leprae* (n=46) and *M. tuberculosis* (n=17). In contrast, there are only 8 available genomes of *S. enterica*, 1 of *V. cholerae* and 8 of *T. pallidum* (plus the two new ancient genomes obtained in this dissertation). While the number of ancient genomes is important, what truly matters is the knowledge they can provide about the bacteria. By studying these ancient genomes and comparing them with modern strains, it has been possible to confirm or challenge hypotheses regarding the etiological agents of certain epidemics or the spread of these diseases in human populations [92,93,98,188,217,238–250,252,254].

Nevertheless, it is important to note that *T. pallidum* is the only bacterium among these species that cannot be cultured using a standardized system. This poses a challenge not only for studying the genomes of modern strains but also for ancient ones. Most of the known protein functions of *T. pallidum* are hypothetical or inferred through sequence comparisons with proteins of known function. This makes it challenging to establish changes in pathogenicity or virulence in this bacteria using ancient genomes, as it has been possible for other bacteria [93,188,238–250,252]. However, the analysis of ancient *T. pallidum* genomes has shown that there have been no major changes in its genome, and virulence genes seem to have persisted for centuries, despite variations in their prevalence across

different populations and time periods, as it was also observed in the case of *M. leprae* [188,216,239–242].

Despite the valuable insights provided by ancient genomes, many questions remain unanswered. We can confirm the presence of bejel in the New World, but the origin and emergence of syphilis are still unknown. However, the utilization of molecular clock dating, along with the identification of putative recombination events among contemporary TPA lineages, where these lineages exhibit shared genomic regions from a common TPE/TEN donor, suggest that both TPA and TPE/TEN lineages coexisted in the Old World before the arrival of Columbus. The diversity and wide geographic distribution of contemporary lineages in early modern Europe support this idea. Therefore, obtaining more pre-contact *T. pallidum* genomes from different continents is crucial to establish the origin of syphilis in the New World. Additionally, further investigation into the evolutionary processes underlying the obtained genomic sequences is necessary to better understand the processes of expansion and divergence of these three subspecies.

Valiente-Mullor *et al.* [103] showed that employing a single reference genome for aligning HTS reads in microbial genomics can lead to errors impacting downstream analyses, including SNP detection and phylogenetic inference, particularly when analyzing datasets with genetically diverse isolates. Concretely, this study examined the impact of reference choice using short read sequence data from five high genomic variable bacteria: *Klebsiella pneumoniae*, *Legionella pneumophila*, *Neisseria gonorrhoeae*, *Pseudomonas aeruginosa*, and *Serratia marcescens*.

Building upon this knowledge, in Chapter 1, we conducted a comprehensive comparison by using three different genome references (Nichols, CDC2, and SS14) for mapping and analyzing recombinant data. Interestingly, only a limited number of studies have addressed the potential bias associated with using a single *T. pallidum* genome reference, including the research discussed in Chapter 1 of this thesis [83,92,108]. Consistent with previous exploratory studies, our findings

indicated that the choice of genome reference did not significantly affect the results or impact the conclusions. The discrepancy between our results and those of Valiente-Mullor *et al.* [103] may be attributed to the clonality of *T. pallidum*. Given that the divergence between the three subspecies is 0.03% [2], it is not surprising that using a reference from a different *T. pallidum* subspecies would not lead to significant erroneous variation in SNP calling, unlike more genetically variable species such as the bacteria explored in Valiente-Mullor *et al.* [103]. However, it is important to note that the number of genomes included in Chapter 1 and previous studies for each *T. pallidum* subspecies was unequal or limited, mainly due to the scarcity of available genomes of this bacterium at that time, particularly for TPE and TEN.

With the increasing availability of *T. pallidum* genomes, in Chapter 2 we devised a new methodological approach to mitigate the bias introduced by reference choice and ensure result robustness without the need for repetitive analyses. In that chapter, we carefully selected one reference genome from each *T. pallidum* subspecies (TPE and TEN) and one from each major TPA lineage (Nichols and SS14). Based on the strain classification obtained from previous studies, we mapped the reads of each strain against the corresponding closest reference genome. However, for the newly obtained ancient genome (W86), we mapped the reads against all four possible genome references of *T. pallidum* to assess the consistency of results across different reference choices. The results of this chapter clearly demonstrated the significance of selecting the closest reference when analyzing ancient genomes, as exemplified by the diverse mapping outcomes obtained for the new ancient genome W86 (Supplementary Figure 18). While the identification of subspecies remained consistent across all four mapping results for the W86 genome, the phylogenetic placement of this ancient genome varied depending on the reference genome used for mapping its reads.

Our innovative approach not only improved genome coverage, particularly important when working with sparse ancient DNA reads, but also reduced reference bias, enhanced the accuracy of phylogenetic inference and assignment. Additionally, it ensured result consistency regardless of the reference chosen, eliminating the need for laborious comparisons with multiple reference genomes in subsequent analyses. These results highlight that the use of any of the four *T. pallidum* references does not appear to affect the classification of strains within their respective subspecies or subsequent genomic analyses. However, it can influence the positioning of the genome in the phylogenetic tree, particularly for ancient genomes with lower coverage compared to modern genomes, in addition to improving SNP calling, which could have a major impact on subsequent recombination and selection inferences [103].

It is important to note that, in Chapter 2, we focused on the detailed examination of the effects of mapping for the specific ancient genome (W86) we generated, while for other genomes in the dataset we ensured the use of the closest reference genome without evaluating the impact of using a different one. In Chapter 3, we followed the same methodological approach as in Chapter 2, using the closest genomic reference for mapping and evaluating the best genome reference choice for the new ancient genomes obtained. In contrast, in Chapter 4, we exclusively utilized the Nichols reference strain to obtain the genomic dataset for the design of the new MLST scheme, considering the project objectives and the time when it was started.

Despite the significant findings obtained, we are currently working on expanding the results presented in this doctoral thesis. Our aim is to further investigate and quantify the true impact of genomic reference choice on the reconstruction of *T. pallidum* genomes in a comprehensive dataset, striving to represent each subspecies or sublineage of this bacterium as accurately as possible, besides exploring its effects on phylogenomic and evolutionary analyses of this bacterium.



## Discussion

---

Overall, the present doctoral thesis makes significant contributions to our understanding of the evolution, genomics, and epidemiology of *T. pallidum*. The incorporation of ancient genomes, the innovative mapping approach, and the development of a new MLST scheme collectively contribute to the advancement of research in this field. The findings and insights presented in this thesis lay the foundation for further investigations and provide valuable resources for addressing the challenges posed by the different treponematoses.



— **CONCLUSIONS** —



## Conclusions

1. A novel pipeline, called PIM, was developed for accurate recombination detection in *T. pallidum* genomes and outperformed other widely used tools in terms of accuracy.
2. We have also explored the role of natural selection in this clonal pathogen. We found that recombination and selection drive *T. pallidum* evolution and contribute to gene diversity in host-pathogen interactions and adaptive pathogen evolution.
3. Remarkably, the recombinant genes detected also showed strong signals of positive or purifying selection, emphasizing their functional significance in host-pathogen interactions, virulence, and immune evasion. However, there are still unanswered questions regarding the molecular processes underlying recombination and the frequency and specific sites of coinfections that are likely essential for recombination to take place.
4. The absence of recent recombination events in *T. pallidum* has implications for the design and use of MLST schemes. Genes involved in recombination events can potentially be used in MLST schemes as most alleles result from new mutations.
5. In Chapter 1, we conducted a comparison of three different genome references and found that the choice of reference did not significantly impact the results. However, the limited number of available genomes, especially for certain subspecies, hindered a more comprehensive analysis.
6. In Chapter 2, a new mapping approach was developed to improve genome coverage, reduce reference bias, and enhance the accuracy of phylogenetic inference and assignment. This eliminated the need for comparisons with multiple reference genomes, streamlining subsequent analyses. The application of this new approach revealed that the choice of reference influenced the phylogenetic placement of ancient genomes but did not affect the classification of strains within subspecies.

7. We used 121 *T. pallidum* genomes to develop a novel MLST scheme incorporating seven variable genes and the 23S rRNA genes. This scheme significantly enhances the ability to distinguish between different strains and can be effectively applied to all *T. pallidum* subspecies. The amplification efficiency of the MLST scheme was moderate, and improvements were made to reduce time and costs. However, further testing with higher-quality samples is recommended.
8. Notably, our typing scheme successfully identified genetic diversity and demonstrated the presence of macrolide resistance across all *T. pallidum* subspecies and sublineages, with a particularly high prevalence observed in the SS14 sublineage. Moreover, our investigation into the genetic diversity and population structure revealed variations within and between populations based on geographic origin. These findings suggest the existence of localized transmission patterns and the potential influence of regional factors in the spread of *T. pallidum*.
9. We successfully acquired two complete high-coverage genomes of ancient *T. pallidum*. Through the analysis of these new ancient genomes, along with comprehensive datasets of whole genomes, we gained insights into the adaptive nature of treponemal agents, showcased the potential of ancient DNA in unraveling significant events in pathogen evolution and emergence, and opened up new avenues for investigating the historical spread of treponematoses.

— **REFERENCES** —





---

## References

1. Miao RM, Fieldsteel AH. Genetic relationship between *Treponema pallidum* and *Treponema pertenuae*, two noncultivable human pathogens. *Journal of Bacteriology*. 1980. pp. 427–429. doi:10.1128/jb.141.1.427-429.1980
2. Vrbová E, Noda AA, Grillová L, Rodríguez I, Forsyth A, Oppelt J, et al. 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*. 2022;16: e0009900.
3. Radolf JD, Deka RK, Anand A, Šmajš D, Norgard MV, Frank Yang X. *Treponema pallidum*, the syphilis spirochete: making a living as a stealth pathogen. *Nature Publishing Group*. 2016;14: 744–759.
4. Giacani L, Lukehart SA. The endemic treponematoses. *Clin Microbiol Rev*. 2014;27: 89–115.
5. Noda AA, Grillová L, Lienhard R, Blanco O, Rodríguez I, Šmajš D. Bejel in Cuba: molecular identification of *Treponema pallidum* subsp. *endemicum* in patients diagnosed with venereal syphilis. *Clin Microbiol Infect*. 2018;24: 1210.e1–1210.e5.
6. Grange PA, Allix-Beguec C, Chanal J, Benhaddou N, Gerhardt P, Morini JP, et al. Molecular subtyping of *Treponema pallidum* in Paris, France. *Sex Transm Dis*. 2013. doi:10.1097/OLQ.0000000000000006
7. Smith JL, Pesetsky BR. The current status of *Treponema cuniculi*. Review of the literature. *Br J Vener Dis*. 1967;43: 117–127.
8. DiGiacomo RF, Lukehart SA, Talburt CD, Baker-Zander SA, Condon J, Brown CW. Clinical course and treatment of venereal spirochaetosis in New Zealand white rabbits. *Br J Vener Dis*. 1984;60: 214–218.
9. Šmajš D, Zobaníková M, Strouhal M, Čejková D, Dugan-Rocha S, Pospíšilová P, et al. Complete Genome Sequence of *Treponema paraluis-cuniculi*, Strain Cuniculi A: The Loss of Infectivity to Humans Is Associated with Genome Decay. *PLoS ONE*. 2011. p. e20415. doi:10.1371/journal.pone.0020415
10. Marks, M, Lebari, D, Solomon, A. W, Higgins, S. P. Yaws. *Int J STD AIDS*. 2015;26: 696-703.

## References

---

11. Shinohara K, Furubayashi K, Kojima Y, Mori H, Komano J, Kawahata T. 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*. 2022;28: 444–450.
12. Goh, B. T. Syphilis in adults. Goh, B. T. (2005). Syphilis in adults. *Sex. Transm. Infect.* 81(6), 448-452.. 2005;81: 448-452.
13. World Health Organization. Global progress report on HIV, viral hepatitis and sexually transmitted infections, 2021. 2021. Available from: <https://www.who.int/publications/i/item/9789240027077>
14. Stockholm: ECDC (European Centre for Disease Prevention and Control). Syphilis. In: ECDC. Annual epidemiological report for 2019. 2022. Available from: <https://www.ecdc.europa.eu/en/publications-data/syphilis-annual-epidemiological-report-2019>
15. Adhikari EH. Syphilis in Pregnancy. *Obstet Gynecol*. 2020;135: 1121–1135.
16. Mitjà O, Marks M, Konan DJP, Ayelo G, Gonzalez-Beiras C, Boua B, et al. Global epidemiology of yaws: a systematic review. *Lancet Glob Health*. 2015;3: e324–31.
17. Stamm LV. Syphilis: antibiotic treatment and resistance. *Epidemiology and Infection*. 2015. pp. 1567–1574. doi:10.1017/s0950268814002830
18. Dofitas BL, Kalim SP, Toledo CB, Richardus JH. Yaws in the Philippines: first reported cases since the 1970s. *Infect Dis Poverty*. 2020;9: 1.
19. Mitjà O, Godornes C, Houinei W, Kapa A, Paru R, Abel H, et al. Re-emergence of yaws after single mass azithromycin treatment followed by targeted treatment: a longitudinal study. *Lancet*. 2018;391: 1599–1607.
20. Ayã AG, Barogui YT, Wadagni AC, et al. Resurgence of yaws in Benin: Four confirmed cases in the district of Z, Southern Benin. *Journal of Public Health*. 2019;11: 201-208.
21. Timothy JWS, Beale MA, Rogers E, Zaizay Z, Halliday KE, Mulbah T, et al. Epidemiologic and Genomic Reidentification of Yaws, Liberia. *Emerg Infect Dis*. 2021;27: 1123–1132.
22. Kawahata T, Kojima Y, Furubayashi K, Shinohara K, Shimizu T, Komano J, et al. Bejel, a Nonvenereal Treponematoses, among Men Who Have Sex with Men, Japan. *Emerg Infect Dis*. 2019;25: 1581–1583.

## References

---

23. Mikalová L, Strouhal M, Oppelt J, Grange PA, Janier M, Benhaddou N, et al. 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. 2017. doi:10.1371/journal.pntd.0005434
24. Fanella S, Kadkhoda K, Shuel M, Tsang R. Local transmission of imported endemic syphilis, Canada, 2011. Emerg Infect Dis. 2012;18: 1002–1004.
25. Mitjà O, Šmajš D, Bassat Q. Advances in the diagnosis of endemic treponematoses: yaws, bejel, and pinta. PLoS Negl Trop Dis. 2013;7: e2283.
26. Stamm LV. Pinta: Latin America’s Forgotten Disease? Am J Trop Med Hyg. 2015;93: 901–903.
27. Rompalo AM, Lawlor J, Seaman P, Quinn TC, Zenilman JM, Hook EW 3rd. Modification of syphilitic genital ulcer manifestations by coexistent HIV infection. Sex Transm Dis. 2001;28: 448–454.
28. Wu MY, Gong HZ, Hu KR, Zheng H-Y, Wan X, Li J. Effect of syphilis infection on HIV acquisition: a systematic review and meta-analysis. Sex Transm Infect. 2021;97: 525–533.
29. Zetola NM, Klausner JD. Syphilis and HIV infection: an update. Clin Infect Dis. 2007;44: 1222–1228.
30. Workowski KA. Centers for Disease Control and Prevention Sexually Transmitted Diseases Treatment Guidelines. Clinical Infectious Diseases. 2015. pp. S759–S762. doi:10.1093/cid/civ771
31. Little JW. Syphilis: An update. Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology. 2005. pp. 3–9. doi:10.1016/j.tripleo.2005.03.006
32. Kurban M, Abbas O. Comment on: “Sexually acquired syphilis: Laboratory diagnosis, management, and prevention.” Journal of the American Academy of Dermatology. 2020. p. e241.
33. Román GC, Román LN. Occurrence of congenital, cardiovascular, visceral, neurologic, and neuro-ophthalmologic complications in late yaws: a theme for future research. Rev Infect Dis. 1986;8: 760–770.
34. Beale MA, Noguera-Julian M, Godornes C, Casadellà M, González-Beiras C, Parera M, et al. Yaws re-emergence and bacterial drug resistance selection after mass administration of azithromycin: a genomic epidemiology investigation. The Lancet Microbe. 2020;1: e263–e271.

## References

---

35. Morando N, Vrbová E, Melgar A, Rabinovich RD, Šmajš D, Pando MA. High frequency of Nichols-like strains and increased levels of macrolide resistance in *Treponema pallidum* in clinical samples from Buenos Aires, Argentina. *Scientific Reports*. 2022. doi:10.1038/s41598-022-20410-5
36. Lieberman NAP, Lin MJ, Xie H, Shrestha L, Nguyen T, Huang M-L, et al. *Treponema pallidum* genome sequencing from six continents reveals variability in vaccine candidate genes and dominance of Nichols clade strains in Madagascar. *PLoS Negl Trop Dis*. 2021;15: e0010063.
37. Kenyon C. Prevalence of macrolide resistance in *Treponema pallidum* is associated with macrolide consumption. *J Med Microbiol*. 2019;68: 119–123.
38. Marra CM, Colina AP, Godornes C, Tantaló LC, Puray M, Centurion-Lara A, et al. Antibiotic selection may contribute to increases in macrolide-resistant *Treponema pallidum*. *J Infect Dis*. 2006;194: 1771–1773.
39. Grillova L, Petrošova H, Mikalova L, Strnadel R, Dastychova E, Kuklova I, et al. Molecular typing of *Treponema pallidum* in the czech republic during 2011 to 2013: Increased prevalence of identified genotypes and of isolates with macrolide resistance. *J Clin Microbiol*. 2014;52. doi:10.1128/JCM.01292-14
40. Luo Y, Xie Y, Xiao Y. Laboratory Diagnostic Tools for Syphilis: Current Status and Future Prospects. *Front Cell Infect Microbiol*. 2020;10: 574806.
41. World Health Organization. The Use of Rapid Syphilis Tests. World Health Organization; 2007.
42. Brischetto A, Gassiep I, Whiley D, Norton R. Retrospective Review of *Treponema pallidum* PCR and Serology Results: Are Both Tests Necessary? *J Clin Microbiol*. 2018;56: e01782–17.
43. Hay PE, Clarke JR, Strugnell RA, Taylor-Robinson D, Goldmeier D. Use of the polymerase chain reaction to detect DNA sequences specific to pathogenic treponemes in cerebrospinal fluid. *FEMS Microbiol Lett*. 1990;56: 233–238.
44. Zhou C, Zhang X, Zhang W, Duan J, Zhao F. PCR detection for syphilis diagnosis: Status and prospects. *J Clin Lab Anal*. 2019;33: e22890.
45. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res*. 2018;3: 124.

## References

---

46. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A*. 1998;95: 3140–3145.
47. Maiden MCJ. Multilocus sequence typing of bacteria. *Annu Rev Microbiol*. 2006;60: 561–588.
48. Maiden MCJ, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat Rev Microbiol*. 2013;11: 728–736.
49. Enright MC, Day NP, Davies CE, Peacock SJ, Spratt BG. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus*. *J Clin Microbiol*. 2000;38: 1008–1015.
50. Dingle KE, Colles FM, Wareing DR, Ure R, Fox AJ, Bolton FE, et al. Multilocus sequence typing system for *Campylobacter jejuni*. *J Clin Microbiol*. 2001;39: 14–23.
51. Enright MC, Spratt BG. A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. *Microbiology*. 1998;144 ( Pt 11): 3049–3060.
52. Pillay A, Liu H, Chen CY, Holloway B, Sturm WA, Steiner B, et al. Molecular Subtyping of *Treponema pallidum* subspecies *pallidum*. *Sexually Transmitted Diseases*. 1998. pp. 408–414. doi:10.1097/00007435-199809000-00004
53. Katz KA, Pillay A, Ahrens K, Kohn RP, Hermanstynne K, Bernstein KT, et al. Molecular Epidemiology of Syphilis—San Francisco, 2004–2007. *Sex. Transm. Dis*. 2010; 660–663. doi:10.1097/olq.0b013e3181e1a77a
54. Mikalová L, Pospíšilová P, Woznicová V, Kuklová I, Zákoucká H, Šmajš D. Comparison of CDC and sequence-based molecular typing of syphilis treponemes: tpr and arp loci are variable in multiple samples from the same patient. *BMC Microbiol*. 2013;13: 178.
55. Grillová L, Bawa T, Mikalová L, Gayet-Ageron A, Nieselt K, Strouhal M, et al. Molecular characterization of *Treponema pallidum* subsp. *pallidum* in Switzerland and France with a new multilocus sequence typing scheme. *PLoS One*. 2018;13: e0200773.
56. Grillova L, Jolley K, Šmajš D, Picardeau M. A public database for the new MLST scheme for *Treponema pallidum* subsp. *pallidum* : surveillance and epidemiology of the causative agent of syphilis. *PeerJ*. 2019;6: e6182.

## References

---

57. Flasarová M, Smajs D, Matejková P, Woznicová V, Heroldová-Dvoráková M, Votava M. Molecular detection and subtyping of *Treponema pallidum* subsp. *pallidum* in clinical specimens. *Epidemiol Mikrobiol Imunol*. 2006;55: 105–111.
58. Woznicová V, Smajs D, Wechsler D, Matějková P, Flasarová M. Detection of *Treponema pallidum* subsp. *pallidum* from skin lesions, serum, and cerebrospinal fluid in an infant with congenital syphilis after clindamycin treatment of the mother during pregnancy. *J Clin Microbiol*. 2007;45: 659–661.
59. Marra CM, Sahi SK, Tantalo LC, Godornes C, Reid T, Behets F, et al. Enhanced molecular typing of *Treponema pallidum*: geographical distribution of strain types and association with neurosyphilis. *J Infect Dis*. 2010;202: 1380–1388.
60. Godornes C, Giacani L, Barry AE, Mitja O, Lukehart SA. Development of a Multilocus Sequence Typing (MLST) scheme for *Treponema pallidum* subsp. *pertenue*: Application to yaws in Lihir Island, Papua New Guinea. Norris SJ, editor. *PLoS Negl Trop Dis*. 2017;11: e0006113.
61. Chuma IS, Roos C, Atickem A, Bohm T, Anthony Collins D, Grillová L, et al. Strain diversity of *Treponema pallidum* subsp. *pertenue* suggests rare interspecies transmission in African nonhuman primates. *Sci Rep*. 2019;9: 14243.
62. Katz SS, Chi K-H, Nachamkin E, Danavall D, Taleo F, Kool JL, et al. Molecular strain typing of the yaws pathogen, *Treponema pallidum* subspecies *pertenue*. Kalendar R, editor. *PLoS One*. 2018;13: e0203632.
63. Turner TB, Hollander DH, Organization WH, Others. Biology of the treponematoses. World Health Organization; 1957. 1957. Available from: <https://apps.who.int/iris/handle/10665/41677>
64. Lukehart SA, Marra CM. Isolation and Laboratory Maintenance of *Treponema pallidum*. *Curr Protoc Microbiol*. 2007;7: 12A.1.1–12A.1.18.
65. Edmondson DG, Hu B, Norris SJ. Long-Term In Vitro Culture of the Syphilis Spirochete *Treponema pallidum* subsp. *pallidum*. *MBio*. 2018;9: e01153–18.
66. Edmondson DG, Delay BD, Kowis LE, Norris SJ. Parameters affecting continuous in vitro culture of *Treponema pallidum* strains. *MBio*. 2021;12: 1–21.

## References

---

67. Edmondson DG, Norris SJ. In Vitro Cultivation of the Syphilis Spirochete *Treponema pallidum*. *Curr Protoc*. 2021;1: e44.
68. Romeis E, Tantalò L, Lieberman N, Phung Q, Greninger A, Giacani L. Genetic engineering of *Treponema pallidum* subsp. *pallidum*, the Syphilis Spirochete. *PLoS Pathog*. 2021;17: e1009612.
69. Fraser CM, Norris SJ, Weinstock GM, White O, Sutton GG, Dodson R, et al. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science*. 1998. doi:10.1126/science.281.5375.375
70. Nishiki S, Lee K, Kanai M, Nakayama S-I, Ohnishi M. Phylogenetic and genetic characterization of *Treponema pallidum* strains from syphilis patients in Japan by whole-genome sequence analysis from global perspectives. *Sci Rep*. 2021;11: 3154.
71. Taouk ML, Taiaroa G, Pasricha S, Herman S, Chow EPF, Azzatto F, et al. Characterisation of *Treponema pallidum* lineages within the contemporary syphilis outbreak in Australia: a genomic epidemiological analysis. *The Lancet Microbe*. 2022. pp. e417–e426. doi:10.1016/s2666-5247(22)00035-0
72. Mubemba B, Gogarten JF, Schuenemann VJ, Dux A, Lang A, Nowak K, et al. Geographically structured genomic diversity of non-human primate-infecting subsp. *Microb Genom*. 2020;6. doi:10.1099/mgen.0.000463
73. Grillová L, Oppelt J, Mikalová L, Nováková M, Giacani L, Niesnerová A, et al. Directly Sequenced Genomes of Contemporary Strains of Syphilis Reveal Recombination-Driven Diversity in Genes Encoding Predicted Surface-Exposed Antigens. *Front Microbiol*. 2019;10: 1691.
74. Strouhal M, Mikalová L, Haviernik J, Knauf S, Bruisten S, Noordhoek GT, et al. Complete genome sequences of two strains of *Treponema pallidum* subsp. *pertenue* from Indonesia: Modular structure of several treponemal genes. *PLoS Negl Trop Dis*. 2018;12: e0006867.
75. Mediannikov O, Fenollar F, Davoust B, Amanzougaghene N, Lepidi H, Arzouni J-P, et al. Epidemic of venereal treponematosi in wild monkeys: a paradigm for syphilis origin. *New Microbes New Infect*. 2020;35: 100670.
76. Liu D, Tong M-L, Liu L-L, Lin L-R, Zhang H-L, Yang T-C. Characterisation of the novel clinical isolate X-4 containing a new sequence-type. *Sex Transm Infect*. 2021;97: 120–125.

## References

---

77. Marks M, Fookes M, Wagner J, Butcher R, Ghinai R, Sokana O, et al. Diagnostics for Yaws Eradication: Insights From Direct Next-Generation Sequencing of Cutaneous Strains of *Treponema pallidum*. *Clin Infect Dis*. 2018;66: 818–824.
78. Arora N, Schuenemann VJ, Jäger G, Peltzer A, Seitz A, Herbig A, et al. Origin of modern syphilis and emergence of a pandemic *Treponema pallidum* cluster. *Nat Microbiol*. 2016;2: 16245.
79. Pinto M, Borges V, Antelo M, Pinheiro M, Nunes A, Azevedo J, et al. Genome-scale analysis of the non-cultivable *Treponema pallidum* reveals extensive within-patient genetic variation. *Nature Microbiology*. 2016. doi:10.1038/nmicrobiol.2016.190
80. Knauf S, Gogarten JF, Schuenemann VJ, De Nys HM, Dux A, Strouhal M, et al. Nonhuman primates across sub-Saharan Africa are infected with the yaws bacterium *Treponema pallidum* subsp. *pertenue*. *Emerg Microbes Infect*. 2018;7: 1–4.
81. Grillová L, Giacani L, Mikalová L, Strouhal M, Strnadl R, Marra C, et al. Sequencing of *Treponema pallidum* subsp. *pallidum* from isolate UZ1974 using Anti-Treponemal Antibodies Enrichment: First complete whole genome sequence obtained directly from human clinical material. *PLoS One*. 2018;13: e0202619.
82. Thurlow CM, Joseph SJ, Ganova-Raeva L, Katz SS, Pereira L, Chen C, et al. Selective whole genome amplification as a tool to enrich specimens with low *Treponema pallidum* genomic DNA copies for whole genome sequencing. *bioRxiv*. 2022. p. 2021.07.09.451864. doi:10.1101/2021.07.09.451864
83. Beale MA, Marks M, Cole MJ, Lee M-K, Pitt R, Ruis C, et al. Global phylogeny of *Treponema pallidum* lineages reveals recent expansion and spread of contemporary syphilis. *Nat Microbiol*. 2021;6: 1549–1560.
84. Chen W, Šmajš D, Hu Y, Ke W, Pospíšilová P, Hawley KL, et al. Analysis of *Treponema pallidum* Strains From China Using Improved Methods for Whole-Genome Sequencing From Primary Syphilis Chancres. *J Infect Dis*. 2021;223: 848–853.
85. Matejková P, Strouhal M, Smajš D, Norris SJ, Palzkill T, Petrosino JF, et al. Complete genome sequence of *Treponema pallidum* ssp. *pallidum* strain SS14 determined with oligonucleotide arrays. *BMC Microbiol*. 2008;8: 76.



## References

---

86. Staudová B, Strouhal M, Zobaníková M, Cejková D, Fulton LL, Chen L, et al. 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*. 2014;8: e3261.
87. Reuter S, Ellington MJ, Cartwright EJP, Köser CU, Török ME, Gouliouris T, et al. Rapid bacterial whole-genome sequencing to enhance diagnostic and public health microbiology. *JAMA Intern Med*. 2013;173: 1397–1404.
88. Petersen LM, Martin IW, Moschetti WE, Kershaw CM, Tsongalis GJ. Third-Generation Sequencing in the Clinical Laboratory: Exploring the Advantages and Challenges of Nanopore Sequencing. *J Clin Microbiol*. 2019;58. doi:10.1128/JCM.01315-19
89. Fulton TL, Shapiro B. Setting Up an Ancient DNA Laboratory. In: Shapiro B, Barlow A, Heintzman PD, Hofreiter M, Paijmans JLA, Soares AER, editors. *Ancient DNA: Methods and Protocols*. New York, NY: Springer New York; 2019. pp. 1–13.
90. Latorre SM, Lang PLM, Burbano HA, Gutaker RM. Isolation, Library Preparation, and Bioinformatic Analysis of Historical and Ancient Plant DNA. *Curr Protoc Plant Biol*. 2020;5: e20121.
91. Stone AC, Ozga AT. Chapter 8 - Ancient DNA in the Study of Ancient Disease. In: Buikstra JE, editor. *Ortner's Identification of Pathological Conditions in Human Skeletal Remains (Third Edition)*. San Diego: Academic Press; 2019. pp. 183–210.
92. Majander K, Pfrengle S, Kocher A, Neukamm J, du Plessis L, Pla-Díaz M, et al. Ancient Bacterial Genomes Reveal a High Diversity of *Treponema pallidum* Strains in Early Modern Europe. *Curr Biol*. 2020;30: 3788–3803.e10.
93. Vågane ÅJ, Herbig A, Campana MG, Robles García NM, Warinner C, Sabin S, et al. *Salmonella enterica* genomes from victims of a major sixteenth-century epidemic in Mexico. *Nature ecology & evolution*. 2018;2: 520–528.
94. Dabney J, Knapp M, Glocke I, Gansauge M-T, Weihmann A, Nickel B, et al. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc Natl Acad Sci U S A*. 2013;110: 15758–15763.
95. Lindqvist C, Rajora OP. Paleogenomics: Genome-Scale Analysis of Ancient DNA. *Springer*. 2019; 323-360.

## References

---

96. Wales N, Carøe C, Sandoval-Velasco M, Gamba C, Barnett R, Samaniego JA, et al. New insights on single-stranded versus double-stranded DNA library preparation for ancient DNA. *Biotechniques*. 2015;59: 368–371.
97. Bennett EA, Massilani D, Lizzo G, Daligault J, Geigl E-M, Grange T. Library construction for ancient genomics: single strand or double strand? *Biotechniques*. 2014;56: 289–300.
98. Schuenemann VJ, Kumar Lankapalli A, Barquera R, Nelson EA, Iraíz Hernández D, Acuña Alonzo V, et al. Historic *Treponema pallidum* genomes from Colonial Mexico retrieved from archaeological remains. Norris SJ, editor. *PLoS Negl Trop Dis*. 2018;12: e0006447.
99. Orlando L, Allaby R, Skoglund P, Der Sarkissian C, Stockhammer PW, Ávila-Arcos MC, et al. Ancient DNA analysis. *Nature Reviews Methods Primers*. 2021;1: 1–26.
100. Horner DS, Pavesi G, Castrignanò T, De Meo PD, Liuni S, Sammeth M, et al. Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Brief Bioinform*. 2010;11: 181–197.
101. Peltzer A, Jäger G, Herbig A, Seitz A, Kniep C, Krause J, et al. EAGER: efficient ancient genome reconstruction. *Genome Biol*. 2016;17: 60.
102. Sohn J-I, Nam J-W. The present and future of de novo whole-genome assembly. *Brief Bioinform*. 2018;19: 23–40.
103. Valiente-Mullor C, Beamud B, Ansari I, Francés-Cuesta C, García-González N, Mejía L, et al. One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads. *PLoS Comput Biol*. 2021;17: e1008678.
104. Liao X, Li M, Zou Y, Wu F-X, Yi-Pan, Wang J. Current challenges and solutions of de novo assembly. *Quantitative Biology*. 2019;7: 90–109.
105. Edmondson DG, De Lay BD, Hanson BM, Kowis LE, Norris SJ. Clonal isolates of *Treponema pallidum* subsp. *pallidum* Nichols provide evidence for the occurrence of microevolution during experimental rabbit infection and in vitro culture. *PLoS One*. 2023;18: e0281187.
106. Sun J, Meng Z, Wu K, Liu B, Zhang S, Liu Y, et al. Tracing the origin of *Treponema pallidum* in China using next-generation sequencing. *Oncotarget*. 2016;7: 42904–42918.

## References

---

107. Beale MA, Marks M, Sahi SK, Tantaló LC, Nori AV, French P, et al. Genomic epidemiology of syphilis reveals independent emergence of macrolide resistance across multiple circulating lineages. *Nat Commun.* 2019;10: 1–9.
108. Pla-Díaz M, Sánchez-Busó L, Giacani L, Šmajš D, Bosshard PP, Bagheri HC, et al. Evolutionary Processes in the Emergence and Recent Spread of the Syphilis Agent, *Treponema pallidum*. *Mol Biol Evol.* 2022;39. doi:10.1093/molbev/msab318
109. Paciência FMD, Rushmore J, Chuma IS, Lipende IF, Caillaud D, Knauf S, et al. Mating avoidance in female olive baboons (*Papio anubis*) infected by *Treponema pallidum*. *Sci Adv.* 2019;5: eaaw9724.
110. Nováková M, Najt D, Mikalová L, Kostková M, Vrbová E, Strouhal M, et al. First report of hare treponematosis seroprevalence of European brown hares (*Lepus europaeus*) in the Czech Republic: seroprevalence negatively correlates with altitude of sampling areas. *BMC Vet Res.* 2019;15: 350.
111. Hisgen L, Abel L, Hallmaier-Wacker LK, Lueert S, Siebert U, Faehndrich M, et al. High syphilis seropositivity in European brown hares (*Lepus europaeus*), Lower Saxony, Germany. *Transbound Emerg Dis.* 2020. doi:10.1111/tbed.13551
112. Mikalová L, Strouhal M, Čejková D, Zobaníková M, Pospíšilová P, Norris SJ, et al. Genome analysis of *Treponema pallidum* subsp. *pallidum* and subsp. *pertenue* strains: most of the genetic differences are localized in six regions. *PLoS One.* 2010;5: e15713.
113. Čejková D, Zobaníková M, Chen L, Pospíšilová P, Strouhal M, Qin X, et al. Whole Genome Sequences of Three *Treponema pallidum* ssp. *pertenue* Strains: Yaws and Syphilis Treponemes Differ in Less than 0.2% of the Genome Sequence. Lukehart S, editor. *PLoS Negl Trop Dis.* 2012;6: e1471.
114. Zobaníková M, Strouhal M, Mikalová L, Čejková D, Ambrožová L, Pospíšilová P, et al. Whole Genome Sequence of the *Treponema* Fribourg-Blanc: Unspecified Simian Isolate Is Highly Similar to the Yaws Subspecies. Picardeau M, editor. *PLoS Negl Trop Dis.* 2013;7: e2172.
115. Jaiswal AK, Tiwari S, Jamal SB, De Castro Oliveira L, Alves LG, Azevedo V, et al. The pan-genome of *Treponema pallidum* reveals differences in genome plasticity between subspecies related to venereal and non-venereal syphilis. *BMC Genomics.* 2020;21: 1–16.
116. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol.* 2008;11: 472–477.

## References

---

117. Salzberg, S. L. Next-generation genome annotation: we still struggle to get it right. *Genome Biol.* 2019;20: 1-3.
118. Knauf S, Batamuzi EK, Mlengeya T, Kilewo M, Lejora IAV, Nordhoff M, et al. *Treponema* infection associated with genital ulceration in wild baboons. *Vet Pathol.* 2012;49: 292–303.
119. Deng X, Phillippy AM, Li Z, Salzberg SL, Zhang W. Probing the pan-genome of *Listeria monocytogenes*: new insights into intraspecific niche expansion and genomic diversification. *BMC Genomics.* 2010;11: 500.
120. Poulsen BE, Yang R, Clatworthy AE, White T, Osmulski SJ, Li L, et al. Defining the core essential genome of *Pseudomonas aeruginosa*. *Proc Natl Acad Sci U S A.* 2019;116: 10072–10080.
121. Bosi E, Monk JM, Aziz RK, Fondi M, Nizet V, Palsson BØ. Comparative genome-scale modelling of *Staphylococcus aureus* strains identifies strain-specific metabolic capabilities linked to pathogenicity. *Proc Natl Acad Sci U S A.* 2016;113: E3801–9.
122. Rousset F, Cabezas-Caballero J, Piastra-Facon F, Fernández-Rodríguez J, Clermont O, Denamur E, et al. The impact of genetic diversity on gene essentiality within the *Escherichia coli* species. *Nat Microbiol.* 2021;6: 301–312.
123. Weinstock GM, Hardham JM, McLeod MP, Sodergren EJ, Norris SJ. The genome of *Treponema pallidum*: new light on the agent of syphilis. *FEMS Microbiol Rev.* 1998;22: 323–332.
124. Hooper SD, Berg OG. On the nature of gene innovation: duplication patterns in microbial genomes. *Mol Biol Evol.* 2003;20: 945–954.
125. Bratlie MS, Johansen J, Sherman BT, Huang DW, Lempicki RA, Drabløf F. Gene duplications in prokaryotes can be associated with environmental adaptation. *BMC Genomics.* 2010;11: 588.
126. Gray RR, Mulligan CJ, Molini BJ, Sun ES, Giacani L, Godornes C, et al. Molecular Evolution of the *tprC*, *D*, *I*, *K*, *G*, and *J* Genes in the Pathogenic Genus *Treponema*. *Mol Biol Evol.* 2006;23: 2220–2233.
127. Smajs D, McKevitt M, Howell JK, Norris SJ, Cai W-W, Palzkill T, et al. Transcriptome of *Treponema pallidum*: gene expression profile during experimental rabbit infection. *J Bacteriol.* 2005;187: 1866–1874.

## References

---

128. Liu H, Rodes B, George R, Steiner B. Molecular characterization and analysis of a gene encoding the acidic repeat protein (Arp) of *Treponema pallidum*. J Med Microbiol. 2007;56: 715–721.
129. Kumar S, Caimano MJ, Anand A, Dey A, Hawley KL, LeDoyt M, et al. Sequence variation of rare outer membrane protein  $\beta$ -barrel domains in clinical strains provides insights into the evolution of *Treponema pallidum* subsp. *pallidum*, the syphilis spirochete. Mbio. 2018; 9. Doi: 10.1128/mBio.01006-18
130. Haynes AM, Fernandez M, Romeis E, Mitjà O, Konda KA, Vargas SK, et al. Transcriptional and immunological analysis of the putative outer membrane protein and vaccine candidate TprL of *Treponema pallidum*. PLoS Negl Trop Dis. 2021;15: e0008812.
131. Centurion-Lara A, Giacani L, Godornes C, Molini BJ, Brinck Reid T, Lukehart SA. Fine analysis of genetic diversity of the tpr gene family among treponemal species, subspecies and strains. PLoS Negl Trop Dis. 2013;7: e2222.
132. Centurion-Lara, A., Sun, E. S., Barrett, L. K., Castro, C., Lukehart, S. A., & Van Voorhis, W. C. Multiple alleles of *Treponema pallidum* repeat gene D in *Treponema pallidum* isolates. J. Bacteriol. 2000;182: 2332-2335.
133. Pětrošová H, Zabaníková M, Čejková D, Mikalová L, Pospíšilová P, Strouhal M, et al. Whole Genome Sequence of *Treponema pallidum* ssp. *pallidum*, Strain Mexico A, Suggests Recombination between Yaws and Syphilis Strains. PLoS Negl Trop Dis. 2012. doi:10.1371/journal.pntd.0001832
134. Čejková, D., Zabaníková, M., Pospíšilová, P., Strouhal, M., Mikalova, L., Weinstock, G. M., & Šmajš, D. Structure of rrn operons in pathogenic non-cultivable treponemes: sequence but not genomic position of intergenic spacers correlates with classification of *Treponema pallidum* and *Treponema paraluis-cuniculi* strains. J. Clin. Microbiol. 2013;62: 196.
135. de Vries MC, Siezen RJ, Wijman JGE, Zhao Y, Kleerebezem M, de Vos WM, et al. Comparative and functional analysis of the rRNA-operons and their tRNA gene complement in different lactic acid bacteria. Syst Appl Microbiol. 2006;29: 358–367.

## References

---

136. Mikalová L, Janečková K, Nováková M, Strouhal M, Čejková D, Harper KN, et al. 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 *Treponema pallidum* subsp. *pertenue* strains. PLOS ONE. 2020. p. e0230926. doi:10.1371/journal.pone.0230926
137. Liu D, Tong M-L, Lin Y, Liu L-L, Lin L-R, Yang T-C. Insights into the genetic variation profile of *tprK* in *Treponema pallidum* during the development of natural human syphilis infection. PLoS Negl Trop Dis. 2019;13: e0007621.
138. Addetia A, Lin MJ, Phung Q, Xie H, Huang M-L, Ciccarese G, et al. Estimation of Full-Length *TprK* Diversity in *Treponema pallidum* subsp. MBio. 2020;11. doi:10.1128/mBio.02726-20
139. Giacani, L., Molini, B., Godornes, C., Barrett, L., Van Voorhis, W., Centurion-Lara, A., & Lukehart, S. A. Quantitative analysis of *tpr* gene expression in *Treponema pallidum* isolates: differences among isolates and correlation with T-cell responsiveness in experimental syphilis. Infection and immunity. 2007. 75: 104-112.
140. Centurion-Lara A, Godornes C, Castro C, Van Voorhis WC, Lukehart SA. The *tprK* gene is heterogeneous among *Treponema pallidum* strains and has multiple alleles. Infect Immun. 2000;68: 824–831.
141. LaFond RE, Centurion-Lara A, Godornes C, Rompalo AM, Van Voorhis WC, Lukehart SA. Sequence diversity of *Treponema pallidum* subsp. *pallidum tprK* in human syphilis lesions and rabbit-propagated isolates. J Bacteriol. 2003;185: 6262–6268.
142. Vink C, Rudenko G, Seifert HS. Microbial antigenic variation mediated by homologous DNA recombination. FEMS Microbiol Rev. 2012;36: 917–948.
143. Centurion-Lara A, Giacani L, Godornes C, Molini BJ, Brinck Reid T, Lukehart SA. Fine analysis of genetic diversity of the *tpr* gene family among treponemal species, subspecies and strains. PLoS Negl Trop Dis. 2013;7: e2222.
144. Centurion-Lara A, LaFond RE, Hevner K, Godornes C, Molini BJ, Van Voorhis WC, et al. Gene conversion: a mechanism for generation of heterogeneity in the *tprK* gene of *Treponema pallidum* during infection. Mol Microbiol. 2004;52: 1579–1596.

## References

---

145. Giacani L, Brandt SL, Puray-Chavez M, Reid TB, Godornes C, Molini BJ, et al. Comparative investigation of the genomic regions involved in antigenic variation of the TprK antigen among treponemal species, subspecies, and strains. *J Bacteriol.* 2012;194: 4208–4225.
146. Molini, B., Fernandez, M. C., Godornes, C., Vorobieva, A., Lukehart, S. A., & Giacani, L. B-cell epitope mapping of TprC and TprD variants of *treponema pallidum* subspecies informs vaccine development for human treponematoses. *Frontiers in Immunology.* 2022; 1326.
147. Arenas M, Araujo NM, Branco C, Castelhana N, Castro-Nallar E, Pérez-Losada M. Mutation and recombination in pathogen evolution: Relevance, methods and controversies. *Infect Genet Evol.* 2018;63: 295–306.
148. Cooper, T. F. Recombination speeds adaptation by reducing competition between beneficial mutations in populations of *Escherichia coli*. *PLoS biology.* 2007; 9:e225.
149. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol.* 2007;7: 214.
150. Strouhal M, Mikalová L, Havlíčková P, Tenti P, Čejková D, Rychlík I, et al. Complete genome sequences of two strains of *Treponema pallidum* subsp. *pertenue* from Ghana, Africa: Identical genome sequences in samples isolated more than 7 years apart. *PLoS Negl Trop Dis.* 2017;11: e0005894.
151. Stamm LV, Bergen HL. The sequence-variable, single-copy *tprK* gene of *Treponema pallidum* Nichols strain UNC and Street strain 14 encodes heterogeneous TprK proteins. *Infect Immun.* 2000;68: 6482–6486.
152. Matejkova P, Flasarova M, Zakoucka H, Borek M, Kremenova S, Arenberger P, et al. Macrolide treatment failure in a case of secondary syphilis: a novel A2059G mutation in the 23S rRNA gene of *Treponema pallidum* subsp. *pallidum*. *J Med Microbiol.* 2009;58: 832–836.
153. John LN, Beiras CG, Houinei W, Medappa M, Sabok M, Kolmau R, et al. Trial of Three Rounds of Mass Azithromycin Administration for Yaws Eradication. *N Engl J Med.* 2022;386: 47–56.
154. Stott CM, Bobay LM. Impact of homologous recombination on core genome phylogenies. *BMC Genomics.* 2020;21: 829.
155. Posada D. Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Mol Biol Evol.* 2002;19: 708–717.

## References

---

156. Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, et al. Full-Length Human Immunodeficiency Virus Type 1 Genomes from Subtype C-Infected Seroconverters in India, with Evidence of Intersubtype Recombination. *Journal of Virology*. 1999. pp. 152–160. doi:10.1128/jvi.73.1.152-160.1999
157. Martin D, Rybicki E. RDP: detection of recombination amongst aligned sequences. *Bioinformatics*. 2000. pp. 562–563. doi:10.1093/bioinformatics/16.6.562
158. Milne I, Wright F, Rowe G, Marshall DF, Husmeier D, McGuire G. TOPALi: software for automatic identification of recombinant sequences within DNA multiple alignments. *Bioinformatics*. 2004. pp. 1806–1807. doi:10.1093/bioinformatics/bth155
159. Jakobsen IB, Easteal S. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Bioinformatics*. 1996. pp. 291–295. doi:10.1093/bioinformatics/12.4.291
160. Worobey M. A novel approach to detecting and measuring recombination: new insights into evolution in viruses, bacteria, and mitochondria. *Mol Biol Evol*. 2001;18: 1425–1434.
161. Holmes EC, Worobey M, Rambaut A. Phylogenetic evidence for recombination in dengue virus. *Mol Biol Evol*. 1999;16: 405–409.
162. Grassly NC, Holmes EC. A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol Biol Evol*. 1997;14: 239–247.
163. Salminen MO, Carr JK, Burke DS, McCutchan FE. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. *AIDS Res Hum Retroviruses*. 1995;11: 1423–1425.
164. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res*. 2015;43: e15.
165. Didelot X, Falush D. Inference of bacterial microevolution using multilocus sequence data. *Genetics*. 2007;175: 1251–1266.
166. Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol*. 2015;11: e1004041.



## References

---

167. Yang C, Pei X, Wu Y, Yan L, Yan Y, Song Y, et al. Recent mixing of *Vibrio parahaemolyticus* populations. *ISME J.* 2019;13: 2578–2588.
168. Zukancic A, Khan MA, Gurmen SJ, Gliniecki QM, Moritz-Kinkade DL, Maddox CW, et al. Staphylococcal Protein A Is a Hot Spot for Recombination and Horizontal Gene Transfer in *Staphylococcus pseudintermedius*. *mSphere.* 2020;5. doi:10.1128/mSphere.00666-20
169. Chaguza C, Andam CP, Harris SR, Cornick JE, Yang M, Bricio-Moreno L, et al. Recombination in *Streptococcus pneumoniae* Lineages Increase with Carriage Duration and Size of the Polysaccharide Capsule. *MBio.* 2016;7. doi:10.1128/mBio.01053-16
170. Lewis JM, Mphasa M, Banda R, Beale MA, Heinz E, Mallewa J, et al. Colonization dynamics of extended-spectrum beta-lactamase-producing Enterobacterales in the gut of Malawian adults. *Nat Microbiol.* 2022;7: 1593–1604.
171. Anand, Arvind, Amit Luthra, Star Dunham-Ems, Melissa J. Caimano, Carson Karanian, Morgan LeDoyt, Adriana R. Cruz, Juan C. Salazar, and Justin D. Radolf. *J. bacteriology.* 2012: 2321-2333.
172. Brinkman MB, McGill MA, Pettersson J, Rogers A, Matejková P, Smajs D, et al. A novel *Treponema pallidum* antigen, TP0136, is an outer membrane protein that binds human fibronectin. *Infect Immun.* 2008;76: 1848–1857.
173. Centurion-Lara, A., Castro, C., Barrett, L., Cameron, C., Mostowfi, M., Van Voorhis, W. C., & Lukehart, S. A. *Treponema pallidum* major sheath protein homologue Tpr K is a target of opsonic antibody and the protective immune response. *The Journal of experimental medicine.* 1999;189:647-656.
174. Marks M, Fookes M, Wagner J, Butcher R, Ghinai R, Sokana O, et al. Diagnostics for Yaws Eradication: Insights From Direct Next-Generation Sequencing of Cutaneous Strains of *Treponema pallidum*. *Clin Infect Dis.* 2018;66: 818–824.
175. Radolf, J. D., & Kumar, S. *The Treponema pallidum* outer membrane. *Spirochete Biology: The Post Genomic Era.* 2018. pp 1-38.
176. Grillová L, Oppelt J, Mikalová L, Nováková M, Giacani L, Niesnerová A, et al. Directly Sequenced Genomes of Contemporary Strains of Syphilis Reveal Recombination-Driven Diversity in Genes Encoding Predicted Surface-Exposed Antigens. *Front Microbiol.* 2019;10: 1691.

## References

---

177. Lefébure T, Stanhope MJ. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biol.* 2007;8: R71.
178. Chiner-Oms Á, López MG, Moreno-Molina M, Furió V, Comas I. Gene evolutionary trajectories in *Mycobacterium tuberculosis* reveal temporal signs of selection. *Proc Natl Acad Sci U S A.* 2022;119: e2113600119.
179. Pepperell CS, Casto AM, Kitchen A, Granka JM, Cornejo OE, Holmes EC, et al. The role of selection in shaping diversity of natural *M. tuberculosis* populations. *PLoS Pathog.* 2013;9: e1003543.
180. Cui Y, Schmid BV, Cao H, Dai X, Du Z, Ryan Easterday W, et al. Evolutionary selection of biofilm-mediated extended phenotypes in *Yersinia pestis* in response to a fluctuating environment. *Nat Commun.* 2020;11: 281.
181. Zhang J. Positive selection, not negative selection, in the pseudogenization of *rcaA* in *Yersinia pestis*. *Proceedings of the National Academy of Sciences of the United States of America.* 2008. p. E69; author reply E70.
182. Maděránková D, Mikalová L, Strouhal M, Vadják Š, Kuklová I, Pospíšilová P, et al. 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.* 2019;13: e0007463.
183. Giacani L, Chattopadhyay S, Centurion-Lara A, Jeffrey BM, Le HT, Molini BJ, et al. Footprint of positive selection in *Treponema pallidum* subsp. *pallidum* genome sequences suggests adaptive microevolution of the syphilis pathogen. *PLoS Negl Trop Dis.* 2012;6: e1698.
184. Zhang Z, Wang J, Wang J, Wang J, Li Y. Estimate of the sequenced proportion of the global prokaryotic genome. *Microbiome.* 2020;8: 134.
185. Pearce ME, Alikhan N-F, Dallman TJ, Zhou Z, Grant K, Maiden MCJ. Comparative analysis of core genome MLST and SNP typing within a European *Salmonella* serovar *Enteritidis* outbreak. *Int J Food Microbiol.* 2018;274: 1–11.
186. Witteveen S, Hendrickx APA, de Haan A, Notermans DW, Landman F, van Santen-Verheuvél MG, et al. Genetic Characteristics of Methicillin-Resistant *Staphylococcus argenteus* Isolates Collected in the Dutch National MRSA Surveillance from 2008 to 2021. *Microbiol Spectr.* 2022;10: e0103522.

## References

---

187. Gona F, Comandatore F, Battaglia S, Piazza A, Trovato A, Lorenzin G, et al. Comparison of core-genome MLST, coreSNP and PFGE methods for *Klebsiella pneumoniae* cluster analysis. *Microbial Genomics*. 2020. doi:10.1099/mgen.0.000347
188. Arning N, Wilson DJ. The past, present and future of ancient bacterial DNA. *Microb Genom*. 2020;6. doi:10.1099/mgen.0.000384
189. Duchêne S, Ho SYW, Carmichael AG, Holmes EC, Poinar H. The Recovery, Interpretation and Use of Ancient Pathogen Genomes. *Curr Biol*. 2020;30: R1215–R1231.
190. Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*. 2013. pp. 1682–1684. doi:10.1093/bioinformatics/btt193
191. Hofreiter M, Jaenicke V, Serre D, von Haeseler A, Pääbo S. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res*. 2001;29: 4793–4799.
192. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol*. 2014;12: 87.
193. Pääbo S, Poinar H, Serre D, Jaenicke-Després V, Hebler J, Rohland N, et al. Genetic Analyses from Ancient DNA. *Annual Review of Genetics*. 2004. pp. 645–679. doi:10.1146/annurev.genet.37.110801.143214
194. Margaryan A, Hansen HB, Rasmussen S, Sikora M, Moiseyev V, Khoklov A, et al. Ancient pathogen DNA in human teeth and petrous bones. *Ecol Evol*. 2018;8: 3534–3542.
195. Harper KN, Ocampo PS, Steiner BM, George RW, Silverman MS, Bolotin S, et al. On the origin of the treponematoses: a phylogenetic approach. *PLoS Negl Trop Dis*. 2008;2: e148.
196. Oviedo GF de, de Oviedo GF. Sumario de la Natural Historia de las Indias. 2010. doi:10.31819/9783964560407
197. Pardo-Castello V. A History of Syphilis. *Archives of Dermatology*. 1963. p. 408. doi:10.1001/archderm.1963.01590150124030
198. Tampa M, Sarbu I, Matei C, Benea V, Georgescu SR. Brief history of syphilis. *J Med Life*. 2014;7: 4–10.

## References

---

199. Medieval DNA suggests Columbus didn't trigger syphilis epidemic in Europe. *Science*. 2020. Available from: <https://www.science.org/content/article/medieval-dna-suggests-columbus-didn-t-trigger-syphilis-epidemic-europe>
200. Maatouk I, Moutran R. History of syphilis: between poetry and medicine. *J Sex Med*. 2014;11: 307–310.
201. Meyer C, Jung C, Kohl T, Poenicke A, Poppe A, Alt KW. Syphilis 2001--a palaeopathological reappraisal. *Homo*. 2002;53: 39–58.
202. Lopez B, Lopez-Garcia JM, Costilla S, Garcia-Vazquez E, Dopico E, Pardiñas AF. Treponemal disease in the Old World? Integrated palaeopathological assessment of a 9th–11th century skeleton from north-central Spain. *Anthropological Science*. 2017. pp. 101–114. doi:10.1537/ase.170515
203. de Melo FL, de Mello JCM, Fraga AM, Nunes K, Eggers S. Syphilis at the crossroad of phylogenetics and paleopathology. *PLoS Negl Trop Dis*. 2010;4: e575.
204. Jedidi H, Laverdeur C, Depierreux-Lahaye F, Beckers A. A brief history of syphilis. The disease through the art and the artist]. *Rev Med Liege*. 2018;73: 363–369.
205. Hudson EH. Treponematoses in perspective. *Bull World Health Organ*. 1965;32: 735–748.
206. Hackett CJ. On the origin of the human treponematoses (pinta, yaws, endemic syphilis and venereal syphilis). *Bull World Health Organ*. 1963. Available: from <https://www.ncbi.nlm.nih.gov/pubmed/14043755>
207. Filippini J, Pezo-Lanfranco L, Eggers S. Estudio regional sistemático de treponematoses en conchales (sambaquis) precolombinos de Brasil. *Chungará (Arica)*. 2019;51: 403–425.
208. Henneberg M, Henneberg RJ. Treponematoses in an ancient Greek colony of Metaponto, southern Italy, 580–250 BCE. The origin of syphilis in Europe, before or after. 1994.
209. Blondiaux J, Bagousse AA. A treponematoses dated from the Late Roman Empire in Normandy, France. *L'Origine de la syphilis en Europe: Avant ou après*. 1994.

## References

---

210. Crane-Kramer GMM. The paleoepidemiological examination of treponemal infection and leprosy in medieval populations from northern Europe. University of Calgary; 2000.
211. Knell RJ. Syphilis in renaissance Europe: rapid evolution of an introduced sexually transmitted disease? *Proc Biol Sci.* 2004;271 Suppl 4: S174–6.
212. Mulligan CJ, Norris SJ, Lukehart SA. Molecular studies in *Treponema pallidum* evolution: toward clarity? *PLoS neglected tropical diseases.* 2008. p. e184.
213. Rothschild BM. History of syphilis. *Clin Infect Dis.* 2005;40: 1454–1463.
214. Spyrou MA, Bos KI, Herbig A, Krause J. Ancient pathogen genomics as an emerging tool for infectious disease research. *Nat Rev Genet.* 2019;20: 323–340.
215. Bos KI, Schuenemann VJ, Golding GB, Burbano HA, Waglechner N, Coombes BK, et al. A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature.* 2011;478: 506–510.
216. Schuenemann VJ, Singh P, Mendum TA, Krause-Kyora B, Jäger G, Bos KI, et al. Genome-wide comparison of medieval and modern *Mycobacterium leprae*. *Science.* 2013;341: 179–183.
217. Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, et al. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature.* 2014;514: 494–497.
218. Higuchi R, Bowman B, Freiburger M, Ryder OA, Wilson AC. DNA sequences from the quagga, an extinct member of the horse family. *Nature.* 1984;312: 282–284.
219. Sivori E, Nakayama F, Cigliano E. Germination of achira seed (*Canna* sp.) approximately 550 years old. *Nature.* 1968;219: 1269–1270.
220. Pääbo S. Molecular cloning of Ancient Egyptian mummy DNA. *Nature.* 1985;314: 644–645.
221. Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M, Pääbo S. Neandertal DNA Sequences and the Origin of Modern Humans. *Cell.* 1997. pp. 19–30. doi:10.1016/s0092-8674(00)80310-4
222. Cooper A, Poinar HN. Ancient DNA: Do It Right or Not at All. *Science.* 2000. pp. 1139–1139. doi:10.1126/science.289.5482.1139b

## References

---

223. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. *Science*. 2010;328: 710–722.
224. Braun M, Cook DC, Pfeiffer S. DNA from *Mycobacterium tuberculosis* Complex Identified in North American, Pre-Columbian Human Skeletal Remains. *Journal of Archaeological Science*. 1998. pp. 271–277. doi:10.1006/jasc.1997.0240
225. Donoghue HD, Spigelman M, Zias J, Gernaey-Child AM, Minnikin DE. *Mycobacterium tuberculosis* complex DNA in calcified pleura from remains 1400 years old. *Lett Appl Microbiol*. 1998;27: 265–269.
226. Faerman M, Jankauskas R, Gorski A, Bercovier H. Prevalence of human tuberculosis in a medieval population of Lithuania studied by ancient DNA analysis. *Anc Biomol*. 1997.
227. Fletcher HA, Donoghue HD, Holton J, Pap I, Spigelman M. Widespread occurrence of *Mycobacterium tuberculosis* DNA from 18th-19th century Hungarians. *Am J Phys Anthropol*. 2003;120: 144–152.
228. Haas CJ, Zink A, Molnár E, Szeimies U, Reischl U, Marcsik A, et al. Molecular evidence for different stages of tuberculosis in ancient bone samples from Hungary. *Am J Phys Anthropol*. 2000;113: 293–304.
229. Haas CJ, Zink A, Pálfi G, Szeimies U, Nerlich AG. Detection of leprosy in ancient human skeletal remains by molecular identification of *Mycobacterium leprae*. *Am J Clin Pathol*. 2000;114: 428–436.
230. Montiel R, García C, Cañadas MP, Isidro A, Guijo JM, Malgosa A. DNA sequences of *Mycobacterium leprae* recovered from ancient bones. *FEMS Microbiol Lett*. 2003;226: 413–414.
231. Rafi A, Spigelman M, Stanford J, Lemma E, Donoghue H, Zias J. DNA of *Mycobacterium leprae* detected by PCR in ancient bone. *Int J Osteoarchaeol*. 1994;4: 287–290.
232. Taylor GM, Widdison S, Brown IN, Young D, Molleson T. A Mediaeval Case of *Lepromatous Leprosy* from 13–14th Century Orkney, Scotland. *J Archaeol Sci*. 2000;27: 1133–1138.
233. Sallares R, Gomzi S. Biomolecular archaeology of malaria. *Anc Biomol*. 2001.
234. Taylor GM, Rutland P, Molleson T. A sensitive polymerase chain reaction method for the detection of *Plasmodium* species DNA in ancient human remains. *Anc Biomol*. 1997.

## References

---

235. Drancourt M, Aboudharam G, Signoli M, Dutour O, Raoult D. Detection of 400-year-old *Yersinia pestis* DNA in human dental pulp: an approach to the diagnosis of ancient septicemia. *Proc Natl Acad Sci U S A*. 1998;95: 12637–12640.
236. Raoult D, Aboudharam G, Crubézy E, Larrouy G, Ludes B, Drancourt M. Molecular identification by “suicide PCR” of *Yersinia pestis* as the agent of medieval black death. *Proc Natl Acad Sci U S A*. 2000;97: 12800–12803.
237. Kolman CJ, Centurion-Lara A, Lukehart SA, Owsley DW, Tuross N. Identification of *Treponema pallidum* subspecies pallidum in a 200-year-old skeletal specimen. *J Infect Dis*. 1999;180: 2060–2063.
238. Hershkovitz I, Donoghue HD, Minnikin DE, Besra GS, Lee OY-C, Gernaey AM, et al. Detection and Molecular Characterization of 9000-Year-Old *Mycobacterium tuberculosis* from a Neolithic Settlement in the Eastern Mediterranean. *PLoS One*. 2008;3: e3426.
239. Donoghue HD. Tuberculosis and leprosy associated with historical human population movements in Europe and beyond - an overview based on mycobacterial ancient DNA. *Ann Hum Biol*. 2019;46: 120–128.
240. Köhler K, Marcsik A, Zádori P, Biro G, Szeniczey T, Fábíán S, et al. Possible cases of leprosy from the Late Copper Age (3780-3650 cal BC) in Hungary. *PLoS One*. 2017;12: e0185966.
241. Monot M, Honoré N, Garnier T, Araoz R, Coppée J-Y, Lacroix C, et al. On the origin of leprosy. *Science*. 2005;308: 1040–1042.
242. Pfrengle S, Neukamm J, Guellil M, Keller M, Molak M, Avanzi C, et al. *Mycobacterium leprae* diversity and population dynamics in medieval Europe from novel ancient genomes. *BMC Biol*. 2021;19: 220.
243. Susat J, Lübke H, Immel A, Brinker U, Macāne A, Meadows J, et al. A 5,000-year-old hunter-gatherer already plagued by *Yersinia pestis*. *Cell Rep*. 2021;35: 109278.
244. Neumann GU, Skourtanioti E, Burri M, Nelson EA, Michel M, Hiss AN, et al. Ancient *Yersinia pestis* and *Salmonella enterica* genomes from Bronze Age Crete. *Curr Biol*. 2022;32: 3641–3649.e8.
245. Rasmussen S, Allentoft ME, Nielsen K, Orlando L, Sikora M, Sjögren K-G, et al. Early divergent strains of *Yersinia pestis* in Eurasia 5,000 years ago. *Cell*. 2015;163: 571–582.

## References

---

246. Wagner DM, Klunk J, Harbeck M, Devault A, Waglechner N, Sahl JW, et al. *Yersinia pestis* and the Plague of Justinian 541–543 AD: a genomic analysis. *Lancet Infect Dis.* 2014;14: 319–326.
247. Mordechai L, Eisenberg M, Newfield TP, Izdebski A, Kay JE, Poinar H. The Justinianic Plague: An inconsequential pandemic? *Proc Natl Acad Sci U S A.* 2019;116: 25546–25554.
248. Drancourt M, Signoli M, Dang LV, Bizot B, Roux V, Tzortzis S, et al. *Yersinia pestis* Orientalis in remains of ancient plague patients. *Emerg Infect Dis.* 2007;13: 332–333.
249. Spyrou MA, Tukhbatova RI, Feldman M, Drath J, Kacki S, Beltrán de Heredia J, et al. Historical *Y. pestis* Genomes Reveal the European Black Death as the Source of Ancient and Modern Plague Pandemics. *Cell Host Microbe.* 2016;19: 874–881.
250. Devault AM, Golding GB, Waglechner N, Enk JM, Kuch M, Tien JH, et al. Second-pandemic strain of *Vibrio cholerae* from the Philadelphia cholera outbreak of 1849. *N Engl J Med.* 2014;370: 334–340.
251. Zhou Z, Lundstrøm I, Tran-Dien A, Duchêne S, Alikhan N-F, Sergeant MJ, et al. Pan-genome Analysis of Ancient and Modern *Salmonella enterica* Demonstrates Genomic Stability of the Invasive Para C Lineage for Millennia. *Curr Biol.* 2018;28: 2420–2428.e10.
252. de-Dios T, Carrión P, Olalde I, Llovera Nadal L, Lizano E, Pàmies D, et al. *Salmonella enterica* from a soldier from the 1652 siege of Barcelona (Spain) supports historical transatlantic epidemic contacts. *iScience.* 2021;24: 103021.
253. Giffin K, Lankapalli AK, Sabin S, Spyrou MA, Posth C, Kozakaité J, et al. A treponemal genome from an historic plague victim supports a recent emergence of yaws and its presence in 15 century Europe. *Sci Rep.* 2020;10: 9499.
254. Barquera R, Lamnidis TC, Lankapalli AK, Kocher A, Hernández-Zaragoza DI, Nelson EA, et al. Origin and Health Status of First-Generation Africans from Early Colonial Mexico. *Curr Biol.* 2020;30: 2078–2091.e11.
255. von Hunnius TE, Yang D, Eng B, Wayne JS, Saunders SR. Digging deeper into the limits of ancient DNA research on syphilis. *J Archaeol Sci.* 2007;34: 2091–2100.



## References

---

256. Bouwman AS, Brown TA. The limits of biomolecular palaeopathology: ancient DNA cannot be used to study venereal syphilis. *Journal of Archaeological Science*. 2005. pp. 703–713. doi:10.1016/j.jas.2004.11.014
257. Cruz AR, Pillay A, Zuluaga AV, Ramirez LG, Duque JE, Aristizabal GE, et al. Secondary Syphilis in Cali, Colombia: New Concepts in Disease Pathogenesis. Lukehart S, editor. *PLoS Negl Trop Dis*. 2010;4: e690.
258. Beale MA, Lukehart SA. Archaeogenetics: What Can Ancient Genomes Tell Us about the Origin of Syphilis? *Current biology: CB*. 2020. pp. R1092–R1095.
259. Kircher M, Sawyer S, Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res*. 2012;40: e3.
260. Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc*. 2010;2010: db.prot5448.
261. Cooper A. Ancient DNA: Do It Right or Not at All. *Science*. 2000. p. 1139b–1139. doi:10.1126/science.289.5482.1139b
262. Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res*. 2010;38: e87.
263. Herbig A, Maixner F, Bos KI, Zink A, Krause J, Huson DH. MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman. *bioRxiv*. 2016. p. 050559. doi:10.1101/050559
264. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol*. 2019;20: 257.
265. Neukamm J, Peltzer A, Nieselt K. DamageProfiler: Fast damage pattern calculation for ancient DNA. *Bioinformatics*. 2021. doi:10.1093/bioinformatics/btab190
266. Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, Bates K, et al. The EMBL Nucleotide Sequence Database. *Nucleic Acids Res*. 2005;33: D29–33.
267. Sayers EW, Beck J, Bolton EE, Bourexis D, Brister JR, Canese K, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2021;49: D10–D17.

## References

---

268. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*. 2014;30: 3276–3278.
269. Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, et al. Using Tablet for visual exploration of second-generation sequencing data. *Brief Bioinform*. 2013;14: 193–202.
270. Sánchez-Busó L, Comas I, Jorques G, González-Candelas F. Recombination drives genome evolution in outbreak-related *Legionella pneumophila* isolates. *Nat Genet*. 2014;46: 1205–1211.
271. Beamud B, Bracho MA, González-Candelas F. Characterization of New Recombinant Forms of HIV-1 From the Comunitat Valenciana (Spain) by Phylogenetic Incongruence. *Front Microbiol*. 2019;10: 1006.
272. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 2015;32: 268–274.
273. Strimmer K, von Haeseler A. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proc Natl Acad Sci U S A*. 1997;94: 6815–6819.
274. Shimodaira H, Hasegawa M. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Mol Biol Evol*. 1999;16: 1114–1114.
275. Strimmer K, Rambaut A. Inferring confidence sets of possibly misspecified gene trees. *Proc Biol Sci*. 2002;269: 137–142.
276. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol*. 2018;35: 1547–1549.
277. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 2015;32: 268–274.
278. Giacani, L., Lukehart, S., & Centurion-Lara, A. Length of guanosine homopolymeric repeats modulates promoter activity of subfamily II tpr genes of *Treponema pallidum* ssp. *pallidum*. *FEMS Immunology & Medical Microbiology*. 2007;51: 289-301.

## References

---

279. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*. 2012;6: 80–92.
280. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007;24: 1586–1591.
281. Pond SLK, Frost SDW, Muse SV. HyPhy: hypothesis testing using phylogenies. *Bioinformatics*. 2005;21: 676–679.
282. Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, et al. Gene-wide identification of episodic selection. *Mol Biol Evol*. 2015;32: 1365–1371.
283. Wertheim JO, Murrell B, Smith MD, Kosakovsky Pond SL, Scheffler K. RELAX: Detecting Relaxed Selection in a Phylogenetic Framework. *Mol Biol Evol*. 2014;32: 820–832.
284. Giacani, L., Godornes, C., Puray-Chavez, M., Guerra-Giraldez, C., Tompa, M., Lukehart, S. A., & Centurion-Lara, A. TP0262 is a modulator of promoter activity of tpr subfamily II genes of *Treponema pallidum* ssp. *pallidum*. *Molecular microbiology*. 2009;72:1087-1099.
285. Yang Z, dos Reis M. Statistical Properties of the Branch-Site Test of Positive Selection. *Molecular Biology and Evolution*. 2011. pp. 1217–1228. doi:10.1093/molbev/msq303
286. Pond SLK, Muse SV. HyPhy: Hypothesis Testing Using Phylogenies. *Statistical Methods in Molecular Evolution*. New York: Springer-Verlag; 2005. pp. 125–181.
287. Lu A, Guindon S. Performance of standard and stochastic branch-site models for detecting positive selection among coding sequences. *Mol Biol Evol*. 2014;31: 484–495.
288. Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol*. 2018;4: vex042.
289. Vaughan TG. IcyTree: rapid browser-based visualization for phylogenetic trees and networks. *Bioinformatics*. 2017;33: 2392–2394.
290. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014. pp. 1312–1313. doi:10.1093/bioinformatics/btu033

## References

---

291. Didelot X, Maiden MCJ. Impact of recombination on bacterial evolution. *Trends Microbiol.* 2010;18: 315–322.
292. Achtman M. Insights from genomic comparisons of genetically monomorphic bacterial pathogens. *Philos Trans R Soc Lond B Biol Sci.* 2012;367: 860–867.
293. Gagneux S. Ecology and evolution of *Mycobacterium tuberculosis*. *Nat Rev Microbiol.* 2018;16: 202–213.
294. Davies J, Davies D. Origins and evolution of antibiotic resistance. *Microbiol Mol Biol Rev.* 2010;74: 417–433.
295. Liu X, Gutacker MM, Musser JM, Fu Y-X. Evidence for recombination in *Mycobacterium tuberculosis*. *J Bacteriol.* 2006;188: 8169–8177.
296. Francés-Cuesta C, Sánchez-Hellín V, Gomila B, González-Candelas F. Is there a widespread clone of *Serratia marcescens* producing outbreaks worldwide? *J Hosp Infect.* 2021;108: 7–14.
297. Awadalla P. The evolutionary genomics of pathogen recombination. *Nat Rev Genet.* 2003;4: 50–60.
298. Anisimova M, Nielsen R, Yang Z. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics.* 2003;164: 1229–1236.
299. Joseph SJ, Didelot X, Gandhi K, Dean D, Read TD. Interplay of recombination and selection in the genomes of *Chlamydia trachomatis*. *Biol Direct.* 2011;6: 28.
300. Knauf, S., Gogarten, J. F., Schuenemann, V. J., De Nys, H. M., Duex, A., Strouhal, M. Et al. African nonhuman primates are infected with the yaws bacterium *Treponema pallidum* subsp. *pertenue*. 2007. *BioRxiv*, 135491.
301. Li C. A Burrows-Wheeler Transform Based Method for DNA Sequence Comparison. *Computational Biology and Bioinformatics.* 2014. p. 33. doi:10.11648/j.cbb.20140203.11
302. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics.* 2016;32: 292–294.

## References

---

303. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20: 1297–1303.
304. Brocchieri L. Phylogenetic inferences from molecular sequences: review and critique. *Theor Popul Biol.* 2001;59: 27–40.
305. Maddison WP, Knowles LL. Inferring phylogeny despite incomplete lineage sorting. *Syst Biol.* 2006;55: 21–30.
306. Degnan JH, DeGiorgio M, Bryant D, Rosenberg NA. Properties of consensus methods for inferring species trees from gene trees. *Syst Biol.* 2009;58: 35–54.
307. Baker-Zander SA, Lukehart SA. Macrophage-mediated killing of opsonized *Treponema pallidum*. *J Infect Dis.* 1992;165: 69–74.
308. Shaffer JM, Baker-Zander SA, Lukehart SA. Opsonization of *Treponema pallidum* is mediated by immunoglobulin G antibodies induced only by pathogenic treponemes. *Infect Immun.* 1993;61: 781–784.
309. Sun ES, Molini BJ, Barrett LK, Centurion-Lara A, Lukehart SA, Van Voorhis WC. Subfamily I *Treponema pallidum* repeat protein family: sequence variation and immunity. *Microbes Infect.* 2004;6: 725–737.
310. Godovikova V, Goetting-Minesky MP, Timm JC, Fenno JC. Immunotopological Analysis of the Major Surface Protein (Msp). *J Bacteriol.* 2019;201. doi:10.1128/JB.00528-18
311. Cox DL, Luthra A, Dunham-Ems S, Desrosiers DC, Salazar JC, Caimano MJ, et al. Surface immunolabeling and consensus computational framework to identify candidate rare outer membrane proteins of *Treponema pallidum*. *Infect Immun.* 2010;78: 5178–5194.
312. Tong M-L, Zhao Q, Liu L-L, Zhu X-Z, Gao K, Zhang H-L, et al. Whole genome sequence of the *Treponema pallidum* subsp. *pallidum* strain Amoy: An Asian isolate highly similar to SS14. *PLoS One.* 2017;12: e0182768.
313. Pospíšilová P, Grange PA, Grillová L, Mikalová L, Martinet P, Janier M, et al. Multi-locus sequence typing of *Treponema pallidum* subsp. *pallidum* present in clinical samples from France: Infecting treponemes are genetically diverse and belong to 18 allelic profiles. *PLoS One.* 2018;13: e0201068.

## References

---

314. Martin, I. E, Tsang, R. S, Sutherland K, Anderson B, Read R, Roy C, et al. Molecular typing of *Treponema pallidum* strains in western Canada: predominance of 14d subtypes. *Sexually transmitted diseases*. 2010; 544-548.
315. Romeis, E, Lieberman, N. A, Molini, B, Tantalo, L. C, Chung, B, Phung, Q, et al. *Treponema pallidum* subsp. *pallidum* with an Artificially impaired TprK antigenic variation system is attenuated in the Rabbit model of syphilis. *PLoS Pathogens*. 2023;19: e1011259.
316. Vos, M. Why do bacteria engage in homologous recombination?. *Trends microbiol*. 2009;17: 226-232.
317. Shapiro BJ, Jesse Shapiro B, Levade I, Kovacicova G, Taylor RK, Almagro-Moreno S. Origins of pandemic *Vibrio cholerae* from environmental gene pools. *Nature Microbiology*. 2017. doi:10.1038/nmicrobiol.2016.240
318. Pretzer, C., Druzhinina, I. S., Amaro, C., Benediktsdóttir, E., Hedenström, I., Hervio-Heath, D., et al. High genetic diversity of *Vibrio cholerae* in the European lake Neusiedler See is associated with intensive recombination in the reed habitat and the long-distance transfer of strains. *Environmental microbiology*. 2007; 19: 328-344.
319. Shapiro BJ. How clonal are bacteria over time? *Curr Opin Microbiol*. 2016;31: 116–123.
320. Vrbová E, Mikalová L, Grillová L, Pospíšilová P, Strnadel R, Dastychová E, et al. 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*. 2020;15: e0237949.
321. Wingett SW, Andrews S. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res*. 2018;7: 1338.
322. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res Notes*. 2016;9: 88.
323. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25: 1754–1760.
324. Broad Institute. Picard Toolkit. In: GitHub Repository [Internet]. 2019. Available: <https://broadinstitute.github.io/picard/>
325. Mitchell, David, Eske Willerslev, and Anders Hansen. Damage and repair of ancient DNA. *Mutat Res-Fund Mol M*. 2005;571: 265-276.

## References

---

326. Alikhan N-F, Petty NK, Ben Zakour NL, Beatson SA. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics*. 2011;12: 402.
327. Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics*. 2011;12: 124.
328. Pankiewicz A, Witkowski J. Dewocjonalia barokowe odkryte na cmentarzysku przy kościele św. Piotra i Pawła na Ostrowie Tumskim we Wrocławiu, *Wroclavia antiqua*. 2012;17: 1621–1670.
329. Hodges E, Rooks M, Xuan Z, Bhattacharjee A, Benjamin Gordon D, Brizuela L, et al. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat Protoc*. 2009;4: 960–974.
330. Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prüfer K, et al. Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci U S A*. 2007;104: 14616–14621.
331. Sawyer S, Krause J, Guschanski K, Savolainen V, Pääbo S. Temporal Patterns of Nucleotide Misincorporations and DNA Fragmentation in Ancient DNA. *PLoS ONE*. 2012. p. e34131. doi:10.1371/journal.pone.0034131
332. Skoglund P, Mathieson I. Ancient Genomics of Modern Humans: The First Decade. *Annu Rev Genomics Hum Genet*. 2018;19: 381–404.
333. Sikora M, Pitulko VV, Sousa VC, Allentoft ME, Vinner L, Rasmussen S, et al. The population history of northeastern Siberia since the Pleistocene. *Nature*. 2019;570: 182–188.
334. Kushwaha B, Kumar R, Agarwal S, Pandey M, Nagpure NS, Singh M, et al. Assembly and variation analyses of *Clarias batrachus* mitogenome retrieved from WGS data and its phylogenetic relationship with other catfishes. *Meta Gene*. 2015. pp. 105–114. doi:10.1016/j.mgene.2015.06.004
335. Yang X, Lee W-P, Ye K, Lee C. One reference genome is not enough. *Genome Biol*. 2019;20: 104.
336. Šmajš D, Strouhal M, Knauf S. Genetics of human and animal uncultivable treponemal pathogens. *Infect Genet Evol*. 2018;61: 92–107.
337. Kanan MW, Abbas M, Girgis HY. Late mutilating bejel in the nomadic Bedouins of Kuwait. *Dermatologica*. 1971;143: 277–287.

## References

---

338. Erdelyi RL, Molla AA. Burned-Out Endemic Syphilis (Bejel). *Plastic and Reconstructive Surgery*. 1984. pp. 589–600. doi:10.1097/00006534-198411000-00001
339. Baker BJ. Treponemal Infection. *The Routledge Handbook of Paleopathology*. doi:10.4324/9781003130994-18/treponemal-infection-brenda-baker
340. Ortner DJ. *Identification of Pathological Conditions in Human Skeletal Remains*. Academic Press; 2003.
341. Forrestel AK, Kovarik CL, Katz KA. Sexually acquired syphilis: Historical aspects, microbiology, epidemiology, and clinical manifestations. *J Am Acad Dermatol*. 2020;82: 1–14.
342. Morton RS. The sibbens of Scotland. *Med Hist*. 1967;11: 374–380.
343. Morton RS. Another look at the Morbus Gallicus. Postscript to the meeting of the Medical Society for the Study of Venereal Diseases, Geneva, May 26–28, 1967. *Br J Vener Dis*. 1968;44: 174–177.
344. Powell ML, Cook DC, Others. *The myth of syphilis: the natural history of treponematosi s in North America*. University Press of Florida; 2005.
345. Tognotti E. The rise and fall of syphilis in Renaissance Europe. *J Med Humanit*. 2009;30: 99–113.
346. Akgül G, Pla-Díaz M, Molak M, du Plessis L, Panagiotopoulou H, Doan K, et al. Inferring patterns of recombination and divergence with ancient and modern treponemal genomes. *bioRxiv*. 2023. p. 2023.02.08.526988. doi:10.1101/2023.02.08.526988
347. Porcella, S. F., & Schwan, T. G. *Borrelia burgdorferi* and *Treponema pallidum*: a comparison of functional genomics, environmental adaptations, and pathogenic mechanisms. *The Journal of clinical investigation*. 2001; 107: 651-656.
348. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res*. 2022;50: D20–D26.
349. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. *Gigascience*. 2021;10. doi:10.1093/gigascience/giab008



## References

---

350. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 2012;22: 568–576.
351. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics.* 2014;47: 11.12.1–34.
352. Nakamura T, Yamada KD, Tomii K, Katoh K. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics.* 2018;34: 2490–2492.
353. Adler CJ, Haak W, Donlon D, Cooper A. Survival and recovery of DNA from ancient teeth and bones. *J Archaeol Sci.* 2011;38: 956–964.
354. Dabney J, Meyer M, Pääbo S. Ancient DNA damage. *Cold Spring Harb Perspect Biol.* 2013;5. doi:10.1101/cshperspect.a012567
355. Harper KN, Zuckerman MK, Harper ML, Kingston JD, Armelagos GJ. The origin and antiquity of syphilis revisited: an appraisal of Old World pre-Columbian evidence for treponemal infection. *Am J Phys Anthropol.* 2011;146 Suppl 53: 99–133.
356. Baker BJ, Crane-Kramer G, Dee MW, Gregoricka LA, Henneberg M, Lee C, et al. Advancing the understanding of treponemal disease in the past and present. *Am J Phys Anthropol.* 2020;171 Suppl 70: 5–41.
357. Štaudová B, Strouhal M, Zobaníková M, Čejková D, Fulton LL, Chen L, et al. 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.* 2014;8: e3261.
358. Schwarz S, Skytte L, Rasmussen KL. Pre-Columbian treponemal infection in Denmark?- a paleopathological and archaeometric approach. *Heritage Science.* 2013;1: 1–12.
359. Rissech C, Roberts C, Tomás-Batlle X, Tomás-Gimeno X, Fuller B, Fernandez PL, et al. A Roman skeleton with possible treponematosis in the North-East of the Iberian peninsula: A morphological and radiological study. *Int J Osteoarchaeol.* 2013;23: 651–663.
360. Gaul JS, Grossschmidt K, Gusenbauer C, Kanz F. A probable case of congenital syphilis from pre-Columbian Austria. *Anthropol Anz.* 2015;72: 451–472.

## References

---

361. Eaton K, Featherstone L, Duchene S, Carmichael AG, Varlık N, Golding GB, et al. Plagued by a cryptic clock: insight and issues from the global phylogeny of *Yersinia pestis*. *Commun Biol*. 2023;6: 23.
362. Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, et al. POPULATION GENETICS. Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science*. 2015;349: aab3884.
363. Staden R, Judge DP, Bonfield JK. Managing Sequencing Projects in the GAP4 Environment. *Introduction to Bioinformatics*. pp. 327–344. doi:10.1385/1-59259-335-6:327
364. Molini BJ, Tantalo LC, Sahi SK, Rodriguez VI, Brandt SL, Fernandez MC, et al. Macrolide Resistance in *Treponema pallidum* Correlates With 23S rDNA Mutations in Recently Isolated Clinical Strains. *Sex Transm Dis*. 2016;43: 579–583.
365. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28: 3150–3152.
366. Nei M. F-statistics and analysis of gene diversity in subdivided populations. *Ann Hum Genet*. 1977;41: 225–233.
367. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, et al. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Mol Biol Evol*. 2017;34: 3299–3302.
368. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, et al. Multilocus sequence typing of total-genome-sequenced bacteria. *J Clin Microbiol*. 2012;50: 1355–1361.
369. Giacani L, Iverson-Cabral SL, King JCK, Molini BJ, Lukehart SA, Centurion-Lara A. Complete Genome Sequence of the *Treponema pallidum* subsp. *pallidum* Sea81-4 Strain. *Genome Announc*. 2014;2: 333–314.
370. Pětrošová H, Pospíšilová P, Strouhal M, Čejková D, Zobaníková M, Mikalová L, et al. Resequencing of *Treponema pallidum* ssp. *pallidum* strains Nichols and SS14: correction of sequencing errors resulted in increased separation of syphilis treponeme subclusters. *PLoS One*. 2013;8: e74319.
371. Gallo Vaulet L, Grillová L, Mikalová L, Casco R, Rodríguez Fermepin M, Pando MA, et al. Molecular typing of *Treponema pallidum* isolates from Buenos Aires, Argentina: Frequent Nichols-like isolates and low levels of macrolide resistance. *PLoS One*. 2017;12: e0172905.

## References

---

372. Mikalová L, Grillová L, Osbak K, Strouhal M, Kenyon C, Crucitti T, et al. Molecular Typing of Syphilis-Causing Strains Among Human Immunodeficiency Virus-Positive Patients in Antwerp, Belgium. *Sex Transm Dis.* 2017;44: 376–379.
373. Flores JA, Vargas SK, Leon SR, Perez DG, Ramos LB, Chow J, et al. *Treponema pallidum pallidum* Genotypes and Macrolide Resistance Status in Syphilitic Lesions among Patients at 2 Sexually Transmitted Infection Clinics in Lima, Peru. *Sex Transm Dis.* 2016;43: 465–466.
374. Wu H, Chang S-Y, Lee N-Y, Huang W-C, Wu B-R, Yang C-J, et al. Evaluation of macrolide resistance and enhanced molecular typing of *Treponema pallidum* in patients with syphilis in Taiwan: a prospective multicenter study. *J Clin Microbiol.* 2012;50: 2299–2304.
375. Read P, Tagg KA, Jeffreys N, Guy RJ, Gilbert GL, Donovan B. *Treponema pallidum* Strain Types and Association with Macrolide Resistance in Sydney, Australia: New TP0548 Gene Types Identified. *J Clin Microbiol.* 2016;54: 2172–2174.
376. Noda AA, Matos N, Blanco O, Rodríguez I, Stamm LV. First Report of the 23S rRNA Gene A2058G Point Mutation Associated With Macrolide Resistance in *Treponema pallidum* From Syphilis Patients in Cuba. *Sex Transm Dis.* 2016;43: 332–334.
377. Xiao Y, Liu S, Liu Z, Xie Y, Jiang C, Xu M, et al. Molecular Subtyping and Surveillance of Resistance Genes In *Treponema pallidum* DNA From Patients With Secondary and Latent Syphilis in Hunan, China. *Sex Transm Dis.* 2016;43: 310–316.
378. Zondag HCA, Bruisten SM, Vrbová E, Šmajš D. No bejel among Surinamese, Antillean and Dutch syphilis diagnosed patients in Amsterdam between 2006-2018 evidenced by multi-locus sequence typing of *Treponema pallidum* isolates. *PLoS One.* 2020;15: e0230288.
379. Vrbová E, Grillová L, Mikalová L, Pospíšilová P, Strnadel R, Dastychová E, et al. MLST typing of *Treponema pallidum* subsp. *pallidum* in the Czech Republic during 2004-2017: Clinical isolates belonged to 25 allelic profiles and harbored 8 novel allelic variants. *PLoS One.* 2019;14: e0217611.
380. Fernández-Naval C, Arando M, Espasa M, Antón A, Fernández-Huerta M, Silgado A, et al. Multilocus sequence typing of *Treponema pallidum* subsp. *pallidum* in Barcelona. *Future Microbiol.* 2021;16: 967–976.

## References

---

381. Sahi SK, Zahlan JM, Tantaló LC, Marra CM. A Comparison of *Treponema pallidum* Subspecies *pallidum* Molecular Typing Systems: Multilocus Sequence Typing vs. Enhanced Centers for Disease Control and Prevention Typing. *Sex Transm Dis.* 2021;48: 670–674.
382. Garcia LN, Morando N, Otero AV, Moroni S, Moscatelli GF, Gonzalez N, et al. Multilocus sequence typing of *Treponema pallidum pallidum* in children with acquired syphilis by nonsexual contact. *Future Microbiol.* 2022;17: 1295–1305.
383. Venter JME, Müller EE, Mahlangu MP, Kularatne RS. *Treponema pallidum* Macrolide Resistance and Molecular Epidemiology in Southern Africa, 2008 to 2018. *J Clin Microbiol.* 2021;59: e0238520.
384. Pillay A, Lee M-K, Slezak T, Katz SS, Sun Y, Chi K-H, et al. Increased Discrimination of *Treponema pallidum* Strains by Subtyping With a 4-Component System Incorporating a Mononucleotide Tandem Repeat in *rpsA*. *Sex Transm Dis.* 2019;46: e42–e45.
385. Janier M, Hegyi V, Dupin N, Unemo M, Tiplica GS, Potočnik M, et al. 2014 European guideline on the management of syphilis. *J Eur Acad Dermatol Venereol.* 2014;28: 1581–1593.
386. Francés-Cuesta C, Ansari I, Fernández-Garayzábal JF, Gibello A, González-Candelas F. Comparative genomics and evolutionary analysis of *Lactococcus garvieae* isolated from human endocarditis. *Microb Genom.* 2022;8. doi:10.1099/mgen.0.000771
387. Mejía L, Prado B, Cárdenas P, Trueba G, González-Candelas F. The impact of genetic recombination on pathogenic *Leptospira*. *Infect Genet Evol.* 2022;102: 105313.
388. Joseph B, Schwarz RF, Linke B, Blom J, Becker A, Claus H, et al. Virulence evolution of the human pathogen *Neisseria meningitidis* by recombination in the core and accessory genome. *PLoS One.* 2011;6: e18441.
389. Budroni S, Siena E, Dunning Hotopp JC, Seib KL, Serruto D, Nofroni C, et al. *Neisseria meningitidis* is structured in clades associated with restriction modification systems that modulate homologous recombination. *Proc Natl Acad Sci U S A.* 2011;108: 4494–4499.
390. Chaguza C, Cornick JE, Everett DB. Mechanisms and impact of genetic recombination in the evolution of *Streptococcus pneumoniae*. *Comput Struct Biotechnol J.* 2015;13: 241–247.

## References

---

391. Didelot X, Méric G, Falush D, Darling AE. Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. *BMC Genomics*. 2012;13: 256.
392. Patiño-Navarrete R, Rosinski-Chupin I, Cabanel N, Gauthier L, Takissian J, Madec J-Y, et al. Stepwise evolution and convergent recombination underlie the global dissemination of carbapenemase-producing *Escherichia coli*. *Genome Med*. 2020;12: 10.
393. Desai PT, Porwollik S, Long F, Cheng P, Wollam A, Bhonagiri-Palsikar V, et al. Evolutionary Genomics of *Salmonella enterica* Subspecies. *MBio*. 2013;4. doi:10.1128/mBio.00579-12
394. Park CJ, Andam CP. Distinct but Intertwined Evolutionary Histories of Multiple *Salmonella enterica* Subspecies. *mSystems*. 2020;5. doi:10.1128/mSystems.00515-19
395. Bos KI, Kühnert D, Herbig A, Esquivel-Gomez LR, Andrades Valtueña A, Barquera R, et al. Paleomicrobiology: Diagnosis and Evolution of Ancient Pathogens. *Annu Rev Microbiol*. 2019;73: 639–666.
396. Duchêne S, Geoghegan JL, Holmes EC, Ho SYW. Estimating evolutionary rates using time-structured data: a general comparison of phylogenetic methods. *Bioinformatics*. 2016;32: 3375–3379.
397. Usoskin IG, Arlt R, Asvestari E, Hawkins E, Käpylä M, Kovaltsov GA, et al. The Maunder minimum (1645–1715) was indeed a grand minimum: A reassessment of multiple datasets. *Astron Astrophys Suppl Ser*. 2015;581: A95.
398. Eddy JA. The maunder minimum. *Science*. 1976;192: 1189–1202.
399. Stuiver M, Quay PD. Changes in atmospheric carbon-14 attributed to a variable sun. *Science*. 1980;207: 11–19.
400. Montiel R, Solórzano E, Díaz N, Álvarez-Sandoval BA, González-Ruiz M, Cañadas MP, et al. Neonate human remains: a window of opportunity to the molecular study of ancient syphilis. *PLoS One*. 2012;7: e36371.
401. Ho SYW, Duchêne S. Molecular-clock methods for estimating evolutionary rates and timescales. *Molecular Ecology*. 2014. pp. 5947–5965. doi:10.1111/mec.12953
402. Bouckaert RR, Drummond AJ. bModelTest: Bayesian phylogenetic site model averaging and model comparison. *BMC Evol Biol*. 2017;17: 42.

## References

---

403. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology*. 2018. pp. 901–904. doi:10.1093/sysbio/syy032
404. Lima TA. Em busca dos frutos do mar os pescadores-coletores do litoral centro-sul do Brasil. *Revista Usp*. 1999. Available: <https://www.revistas.usp.br/revusp/article/download/29850/31736>
405. Gaspar MD, DeBlasis P, Fish SK, Fish PR. Sambaqui (Shell Mound) Societies of Coastal Brazil. In: Silverman H, Isbell WH, editors. *The Handbook of South American Archaeology*. New York, NY: Springer New York; 2008. pp. 319–335.
406. Fish SK, DeBlasis P, Gaspar MD. Eventos incrementais na construção de sambaquis, litoral sul do Estado de Santa Catarina. *Revista do Museu de*. 2000. Available: <https://www.revistas.usp.br/revmae/issue/download/8303/546#page=78>
407. Klokler DM. Food for body and soul: mortuary ritual in shell mounds (Laguna-Brazil). 2008. Tese (Doutorado)--University of Arizona, Tucson, Arizona. 2008.
408. Blasis PAD, Kneip A, Scheel-Ybert R. Sambaquis e paisagem: dinâmica natural e arqueologia regional no litoral do sul do Brasil. *Arqueología*. 2007. Available: <https://repositorio.usp.br/bitstreams/b60a80ea-7720-4f4a-ad58-bb5a9533f7a7>
409. Villagran, XS, Klokler, D, Nishida, P, Gaspar, M. D, & DeBlasis, P. Lecturas estratigráficas: Arquitectura funerária y depositación de residuos en el sambaquí Jabuticabeira II. *Lat. Am. Antiq.* 2010;21: 195-216.
410. Edwards HGM, Farwell DW, de Faria DLA, Monteiro AMF, Afonso MC, De Blasis P, et al. Raman spectroscopic study of 3000-year-old human skeletal remains from a sambaqui, Santa Catarina, Brazil. *J Raman Spectrosc.* 2001;32: 17–22.
411. Beck A, Pereira JBB. Variação do conteúdo cultural dos sambaquis: litoral de Santa Catarina. 1972. Available: <https://repositorio.usp.br/item/000722285>
412. Prous A, Fogaça E. Archaeology of the Pleistocene-Holocene boundary in Brazil. *Quat Int.* 1999;53-54: 21–41.
413. Barbosa PN. A coisa ficou preta: estudo do processo de formação da Terra Preta do sítio arqueológico Jabuticabeira II. Universidade de São Paulo. 2007. Available: <https://www.teses.usp.br/teses/disponiveis/71/71131/tde-28032008-085815/en.php>

## References

---

414. Okumura MMM, Eggers S. The people of Jabuticabeira II: reconstruction of the way of life in a Brazilian shellmound. *Homo*. 2005;55: 263–281.
415. Pezo-Lanfranco L, Filippini J, Di Giusto M, Petronilho C, Wesolowski V, DeBlasis P, et al. Child development, physiological stress and survival expectancy in prehistoric fisher-hunter-gatherers from the Jabuticabeira II shell mound, South Coast of Brazil. *PLoS One*. 2020;15: e0229684.
416. Rothschild BM, Rothschild C. Treponemal disease revisited: skeletal discriminators for yaws, bejel, and venereal syphilis. *Clin Infect Dis*. 1995;20: 1402–1408.
417. Toso A, Hallingstad E, McGrath K, Fossile T, Conlan C, Ferreira J, et al. Fishing intensification as response to Late Holocene socio-ecological instability in southeastern South America. *Sci Rep*. 2021;11: 23506.
418. Heaton TJ, Köhler P, Butzin M, Bard E, Reimer RW, Austin WEN, et al. Marine20—The Marine Radiocarbon Age Calibration Curve (0–55,000 cal BP). *Radiocarbon*. 2020;62: 779–820.
419. Hogg AG, Heaton TJ, Hua Q, Palmer JG, Turney CSM, Southon J, et al. SHCal20 Southern Hemisphere Calibration, 0–55,000 Years cal BP. *Radiocarbon*. 2020;62: 759–778.
420. Pezo-Lanfranco L, DeBlasis P, Eggers S. Weaning process and subadult diets in a monumental Brazilian shellmound. *Journal of Archaeological Science: Reports*. 2018;22: 452–469.
421. Cejkova D, Strouhal M, Smajs D. P1.008 Intrastrain Genetic Heterogeneity in *Treponema Pallidum* Ssp. *Pallidum*. *Sexually Transmitted Infections*. 2013. pp. A76.1–A76. doi:10.1136/sextrans-2013-051184.0229
422. Giacani L, Molini BJ, Kim EY, Godornes BC, Leader BT, Tantalo LC, et al. Antigenic variation in *Treponema pallidum*: TprK sequence diversity accumulates in response to immune pressure during experimental syphilis. *J Immunol*. 2010;184: 3822–3829.
423. Smajs D, Norris SJ, Weinstock GM. Genetic diversity in *Treponema pallidum*: implications for pathogenesis, evolution and molecular diagnostics of syphilis and yaws. *Infect Genet Evol*. 2012;12: 191–202.
424. de la Haba RR, Arahall DR, Márquez MC, Ventosa A. Phylogenetic relationships within the family *Halomonadaceae* based on comparative 23S and 16S rRNA gene sequence analysis. *Int J Syst Evol Microbiol*. 2010;60: 737–748.

## References

---

425. Yang B, Wang Y, Qian P-Y. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics*. 2016;17: 135.
426. Martijn J, Lind AE, Schön ME, Spiertz I, Juzokaite L, Bunikis I, et al. Confident phylogenetic identification of uncultured prokaryotes through long read amplicon sequencing of the 16S-ITS-23S rRNA operon. *Environ Microbiol*. 2019;21: 2485–2498.
427. Rodríguez F, Oliver JL, Marín A, Medina JR. The general stochastic model of nucleotide substitution. *J Theor Biol*. 1990;142: 485–501.
428. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. *PLoS Biol*. 2006;4: e88.
429. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol*. 2005;22: 1185–1192.
430. Schuenemann VJ, Avanzi C, Krause-Kyora B, Seitz A, Herbig A, Inskip S, et al. Ancient genomes reveal a high diversity of *Mycobacterium leprae* in medieval Europe. *PLoS Pathog*. 2018;14: e1006997.
431. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer; 2016.
432. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. Ggtree : An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol*. 2017;8: 28–36.



— APPENDICES —



## Appendices

### Supplementary Notes

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

These archeological analysis was performed in collaboration with Dr. Paweł Dąbrowski, Dr. Joanna Grzelak, Dr. Maciej Oziembłowski.

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 1a)
- 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 sphenoid-occipital synchondrosis and the degree of tooth wear, as well as the condition of the metaphyses of the long bones, the condition of the pelvic ear surface. The age at the time of death of an 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 1a).



**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:

## Appendices

---

- 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. Photo: J.Grzelak

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 Figure 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 [400]). 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. Molecular clock dating from chapter 2**

An uncorrelated lognormal relaxed molecular clock was calibrated using the ages of the samples (see Supplementary Table 10) with a diffuse prior (uniform  $10^{-12}$  -  $10^{-4}$  for the mean rate and a gamma distribution ( $\alpha=0.5396$ ,  $\beta=0.3819$ ) for the rates' standard deviation). Ancient sample 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 [401]. A coalescent Bayesian Skyline tree prior with 5 groups was used as a simple model that is sufficiently flexible to fit many kinds of dynamics. We used **Modeltest** 1.2.1 [402] to average across all possible reversible substitution models. According to the results a model with  $\text{rac}=\text{rgt}$ , no gamma rate variation and no invariable sites received the most support. The MCMC chain was run for 2 billion steps with every 50'000<sup>th</sup> step sampled. The first 15% of samples were discarded as burn-in. Convergence and mixing were inspected using Tracer v1.7.1 [403]; the ESS of all parameters exceeded 100. The maximum clade credibility tree was generated using TreeAnnotator, a part of the **BEAST** 2.6.1 [149] software package and visualized using **FigTree** 1.4.4 (<http://tree.bio.ed.ac.uk/software/figtree/>).

### **Supplementary Note 4: Archaeological information**

#### **The sambaquis of the Laguna Region**

A sambaqui is the prevalent type of archaeological site on the Brazilian coast: a human-built shell midden or shell mound of varying dimensions, located in rich resource areas such as lagoons, mangroves or estuaries. Sambaquis consist of inorganic sediment, mollusc shells, food debris and organic matter mixed in intricate stratigraphies associated with domestic and/or funerary functions [404]. More than 1000 sambaquis are mapped along the 7500 km long Brazilian coast and dated to between 7500 and 1000 y BP [404,405]. Recent archaeological research suggests that these shell mound building populations were sedentary, with an abundant and stable marine-based subsistence, high population growth [404,406], elaborate funerary rituals [407] and landscape appropriation [408].

#### **Jaboticabeira II excavation site**

Jaboticabeira II (UTM 22J - 0699479E; 6835488S) is a medium size shell mound (400 x 250 x 10 m in height), settled on a paleodune and located in the Laguna



region, the highest density area of sambaquis from the Brazilian Southern coast, 3 km from Laguna do Camacho, one of several water sources associated with a barrier-lagoon geological system formed during the Holocene. Jabuticabeira II, built during a nearly 1'000-year period, is one of 65 sambaquis mapped around the lagoon system. This large number of settlements and their chronologically overlapping occupation history attest to a fairly dense occupation and intense interactions of the sambaqui builders between 7500 and 900 cal BP [408]. According to stratigraphic studies, Jabuticabeira II is the result of incremental funerary rituals accumulated over centuries. Although Jabuticabeira II was not completely excavated, 204 burials containing the remains of 282 individuals were exhumed from a 373 m<sup>2</sup> area [407-409]. Radiocarbon dates of Jabuticabeira II stratigraphy suggest a long occupation period between 1214–830 cal BCE and 118–413 cal CE or 3137–2794 to 1860–1524 cal BP - 2 $\sigma$  [407,408], roughly in line with the radiocarbon datings from bone material of the four individuals in this study, ranging from 350 cal BCE to 573 cal CE.

The Jabuticabeira II material included single, double, and multiple burials, dispersed in clusters. The skeletons recovered were mostly incomplete, avoiding categorical diagnosis of age and sex or other osteological findings. The burial pattern was tightly flexed and suggested intentional treatment of the body prior to the internment. The small size of the graves suggested that the bodies suffered previous desiccation or degeneration of soft tissues, but not enough to produce disarticulation (hand and feet bones were found articulated). Primary burials were the most common. However, because many burials come from profiles, the number of secondary burials is unknown. Reopening of graves, some of them containing intrusive bones, was possibly common, but this evidence is debatable. The bones of several individuals are stained with red ochre [410], a common practice in archaeological sites of the Santa Catarina state [411,412]. Offerings are common in burial contexts and include adornments made with faunal material and lithic tools

in a wide range of forms, from debris to polished tools and zooliths, with differences in frequency of occurrence among different loci and strata [413].

Altogether 99 Jabuticabeira II individuals, with and without bone alterations suggestive of infection, were screened for pathogen DNA content. 37 samples deemed positive for treponemal DNA in the initial screening and four samples yielded sufficient data for *T. pallidum* genome reconstruction (Supplementary Table 21).

### **Paleopathological analysis of treponematoses**

Bioarchaeological analyses showed results compatible with increasing population growth and high population density in Jabuticabeira II, including high frequencies of nonspecific stress markers [414] and infant stress [415], but no evidence of trauma associated with interpersonal conflicts over resources or territory [414].

There is, however, evidence of communicable systemic diseases in Jabuticabeira II and other local sites in the region [68]. Eleven <sup>14</sup>C AMS dates obtained directly from the presumably treponematoses-affected individuals suggest that these diseases are very old on the east coast of South America, with a time-range between 6300- and 500-years BP. Among the possible treponemal cases based on osteological analysis, three came from Jabuticabeira II. However, these did not overlap with the individuals yielding the detected genetic evidence in this study.

### **Information on individuals**

#### *Individual 41A-L2.05-E4, sample ZH1390*

The individual is an adult male of robust build, with an estimated stature of 150.49 ± 2.6 cm [415]. Although fragmented, the bones of this individual comprised almost a complete skeleton (80%), articulated and buried in an oval shell-rich matrix in a hyper-flexed position. The bones of the individual showed signs of systemic infectious disease in the lower limbs. Femurs, tibias, and fibulas all show discrete generalised periostitis and osteoarthritis. A wideness in the lateral portion

of clavicles was also observed. According to Filippini *et al.* (2019) [207], applying the SPIRAL method [416], this individual's disease could be classified non-conclusively as syphilis, yaws or bejel. The sampling was performed on an active lesion on the tibia fragment.

*Individual FS9-L3-T2, sample ZH1540*

The sample comes from an assemblage of commingled bones, of probably more than one individual. The bones assigned to this individual consist of several skeletal elements, some with pathological alterations, such as severe osteomyelitis in the distal third of the right humerus, severe periostitis in the left ulna, periostitis in a fibula diaphysis, and two vertebral bodies with osteophytosis. The sample was taken from the fibula fragment, in the area with periostitis.

*Individual FS3B-L3-T4, sample ZH1541*

The sample comes from one of three separate individuals, found commingled. The skeletal elements belonging to this robust adult of unknown age and sex include a left radius with arthritis, a fragment of the left ulna (very robust), a fragment of the left humerus, fragments of a femur, a tibia, and a fibula and a first metatarsal. The sample was taken from a femur fragment, under the immediate surface of the bone, to best avoid the possible introduction of external contaminants.

*Individual 2B-L6-E3, ZH1557*

The sample comes from a likely adult male individual. The individual was articulated and in a flexed position with another, adult female individual buried on top. Osteopathological findings on the bones of the sampled individual included signs of degenerative joint disease, severe lumbar intervertebral osteoarthritis, scoliosis, and possible injuries to the patellae. However, no typical lesions suggestive of treponemal infection were observed. The sample was taken from a small piece of long bone, under the immediate surface of the bone, to best avoid the possible introduction of external contaminants.

### **Marine reservoir effect correction for C14-dating**

The diet of the Jabuticabeira II inhabitants, substantially consisting of marine food sources, produces a reservoir effect in the radiocarbon dates calculated as mean age of 247.8 ( $\sigma = 103.7$ ) years [417]. Considering the high contribution of marine carbon to bone collagen of individuals in Jabuticabeira II, the radiocarbon dates were modelled with **Calib Rev** 8.1.0 using the Mixed Marine SoHCal 20 calibration curve [418,419] and applying the estimated average local marine radiocarbon reservoir correction value ( $\Delta R$ ) of  $-126 \pm 29$  for the South coast of Brazil (Marine Reservoir Correction database, Reimer & Reimer, 2001: see <http://calib.org/marine/>) [420]. We considered the average relative contribution of marine carbon to collagen derived from Bayesian Mixing Models for Jabuticabeira II individuals, calculated at a mean value of 42.5% [420]. For the individual estimates for the samples, see Table 18.

### **Supplementary Note 5. Exploratory characterization of the 16S-23S genes**

*T. pallidum* contains two rRNA (*rrn*) operons, each of which encodes the 16S-23S-5S rRNA genes and intergenic spacer regions (ISRs). There is evidence that the random distribution of *rrn* spacer patterns in *T. pallidum* may be generated by reciprocal translocation of *rrn* operons mediated by a recBCD-like system found in the intergenic spacer regions (ISRs) [421]. In concordance with previous studies, [85,421–423] we found that the 16S–23S ISRs of the TPA strains contain the tRNA-Ile (tRNA-Ile-1; *tp0012*) and tRNA-Ala (tRNA-Ala-3; *tp00t15*) genes within the *rrn1* and *rrn2* operons, respectively. In contrast, the TPE genomes show an Ala/Ile spacer pattern, where the *tp0012* and *tp00t15* orthologues are located within the *rrn2* and *rrn1* operons, respectively.

We identified 68 SNPs in genes *r0001*, *r0002*, *r0004* and *r0005*, encoding the 16S-23S rRNA genes of the new ancient genome ZH1540, placing them among the most variable genes in our alignment and raising the potential that including them in the alignment could result in a biased phylogenetic reconstruction. Although the SNPs

found appear to be well supported by the reads obtained from the sequence mapping (Supplementary Table 24), their origin from possible contamination cannot be completely ruled out and further analyses would be necessary to confirm them.

Excluding these genes from the alignment, in addition to the recombinant genes and *tp0316*, *tp0317* and *tp0897*, did not result in any changes to the topology (Supplementary Figure 40), although branch lengths were altered. As these genes are known to have conserved regions in addition to variable regions used to explore the evolutionary relationships among pathogenic bacteria [424–426], we decided to retain them in the alignment for all subsequent analyses.

Finally, we note that the ZH1540 genome did not possess either of the two *T. pallidum* 23S ribosomal RNA gene mutations known to confer macrolide resistance (A2058G and A2069G). In contrast, four modern TEN strains from Japan possess the A2048G mutation, suggesting recent selection pressure for antibiotic resistance mutations.

### **Supplementary Note 6. Molecular clock dating from Chapter 3**

The analysis was performed under a GTR+G+I substitution model [427], with an uncorrelated exponentially distributed relaxed clock (UCED) model [428] and a Bayesian skyline plot [429] demographic model (tree-prior) with 10 groups. We placed a lognormal prior with a mean (in real space) of  $1 \times 10^{-7}$  substitutions per site per year and standard deviation 0.25 on the mean clock rate. This strong prior was used to compensate for the poor temporal signal among *T. pallidum* genomes and was calibrated on previous estimates of the substitution rate [92,346]. For all genomes where the sampling dates are not known exactly, we used uniform priors across the date ranges reported in the original studies to account for the uncertainty [78,92,253,254,430]. For ZH1540 we set the date range to 247-363 CE, in accordance with the radiocarbon dating results above. Default priors were used for all other model parameters. The same analysis was repeated without ZH1540 in order to assess the effect of our new ancient genome on the divergence dates.

For each analysis we ran four MCMC chains of  $5 \times 10^8$  steps each, sampling parameters and trees every 10'000 steps. After assessing convergence in **Tracer** 1.7 [403] and confirming that all four chains converged to the same posterior distribution, we combined the chains after discarding the first 10% of samples as burn-in. In the resulting combined chains, all parameters have ESS values  $>200$ . **TreeAnnotator** 2.6.7 [149] was used to compute MCC trees and the results were visualized using **ggplot2** [431], **ggtree** [432] and custom scripts.

### Supplementary Files

**Supplementary File 1.** In-house script A used to ensure that the final whole genome alignment obtained was correct.

Available at:

<https://drive.google.com/file/d/1FZ79aeBh-kj4akA19c7jae8pqGAiQdyS/view?usp=sharing>

**Supplementary File 2.** In-house script B used to ensure that the final whole genome alignment obtained was correct.

Available at:

[https://drive.google.com/file/d/1d2ivVakBIIDZ3t8Tv\\_C-AfAKca\\_Q4MTX/view?usp=sharing](https://drive.google.com/file/d/1d2ivVakBIIDZ3t8Tv_C-AfAKca_Q4MTX/view?usp=sharing)

**Supplementary File 3.** Tree with mutations mapped to branches in a nexus format file obtained by Treetime.

Available at:

[https://drive.google.com/file/d/1Uc\\_3dAPiHs92iwzLQHKzvoxclvdgwCDo/view?usp=sharing](https://drive.google.com/file/d/1Uc_3dAPiHs92iwzLQHKzvoxclvdgwCDo/view?usp=sharing)

**Supplementary File 4.** In-house script to compute the number of SNPs per gene.

Available at:

[https://drive.google.com/file/d/1-jE\\_Xy1Tj12Lv1DIFLTx5Gyb4I77A8gt/view?usp=sharing](https://drive.google.com/file/d/1-jE_Xy1Tj12Lv1DIFLTx5Gyb4I77A8gt/view?usp=sharing)

**Supplementary File 5.** In-house script used to determine the subspecies and/or sublineages distinguishable by each of these SNPs.

Available at:

<https://drive.google.com/file/d/1CE41v9r4bZXAO411wa-kYIbTvQBvzqI/view?usp=sharing>

**Supplementary File 6.** In-house script used to discard haplotypes that could not be considered due to missing data in the sequences employed.

Available at:

<https://drive.google.com/file/d/1p0J0e8AsHKmcvVJqsAmOqK6kdygpZ0L/view?usp=sharing>

**Supplementary File 7.** The number of different alleles obtained for the gene *tp0136*

Available at:

<https://drive.google.com/file/d/1IIVAyE1LyWCnlmMWE5xruvcfqNJZxJqm/view?usp=sharing>

**Supplementary File 8.** The number of different alleles obtained for the gene *tp0326*

Available at:

<https://drive.google.com/file/d/1IIVAyE1LyWCnlmMWE5xruvcfqNJZxJqm/view?usp=sharing>

**Supplementary File 9.** The number of different alleles obtained for the gene *tp0548*

Available at:

[https://drive.google.com/file/d/1KbTwP7loar3kD0MCR\\_wizGs9osNrun1M/view?usp=drive\\_link](https://drive.google.com/file/d/1KbTwP7loar3kD0MCR_wizGs9osNrun1M/view?usp=drive_link)

**Supplementary File 10.** The number of different alleles obtained for the gene *tp0705*

Available at:

<https://drive.google.com/file/d/1w4IE14rcdwkh6Ud-x3yh2wSMxJZuJWSr/view?usp=sharing>

**Supplementary File 11.** The number of different alleles obtained for the gene *tp0858*

Available at:

<https://drive.google.com/file/d/1PMZyCVXAJM8I9YmlgXAAHF1HonFn2Zt3/view?usp=sharing>

**Supplementary File 12.** The number of different alleles obtained for the gene *tp0865*

Available at:

[https://drive.google.com/file/d/1nexOiHX\\_a66a3ulYArucYlsUr\\_KVC3nm/view?usp=sharing](https://drive.google.com/file/d/1nexOiHX_a66a3ulYArucYlsUr_KVC3nm/view?usp=sharing)

**Supplementary File 13.** The number of different alleles obtained for the gene *tp1031*

Available at:

[https://drive.google.com/file/d/1cBD0dlolhcDt5EvZ\\_uCtp2zLSstNka7L/view?usp=sharing](https://drive.google.com/file/d/1cBD0dlolhcDt5EvZ_uCtp2zLSstNka7L/view?usp=sharing)

## Supplementary Figures

**Supplementary Figures 1-12.** Each figure presents the maximum likelihood tree obtained from the non-recombinant regions of the multiple alignment of 75 *T. pallidum* strains (Figure 15) using the Nichols genome as reference for mapping. In each figure, the inferred source and recipient clade/strain of the recombination events detected for the corresponding locus is represented. In addition, a detail of the multiple alignment with the SNPs and strains involved in the recombination event(s) is shown. See also Table 8. Available at:

[https://docs.google.com/presentation/d/1Cb\\_YHbivt4tBKNeexe3N2Jt5JX4ZM8tX/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true](https://docs.google.com/presentation/d/1Cb_YHbivt4tBKNeexe3N2Jt5JX4ZM8tX/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true)

**Supplementary Figure 13.** Comparison of the topologies of two maximum likelihood trees: that obtained after excluding the recombinant genes plus *tp0897* from the alignment of the NIC-mapped dataset (A, 1,117,857 bp) (Figure 15) and that obtained after excluding the recombinant genes plus *tp0897* from the alignment of the SS14-mapped dataset (A, 1117793 bp)

Available at:

[https://docs.google.com/presentation/d/1HDFesBvcQjETyTrEfG\\_49aYXB0Cv6A7P/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true](https://docs.google.com/presentation/d/1HDFesBvcQjETyTrEfG_49aYXB0Cv6A7P/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true)



**Supplementary Figure 14.** Comparison of the topologies of two maximum likelihood trees: that obtained after excluding the recombinant genes plus *tp0897* from the alignment of the NIC-mapped dataset (A, 1,117,857 bp) (Figure 15) and that obtained after excluding the recombinant genes plus *tp0897* from the alignment of the CDC2-mapped dataset (A, 1117857 bp)

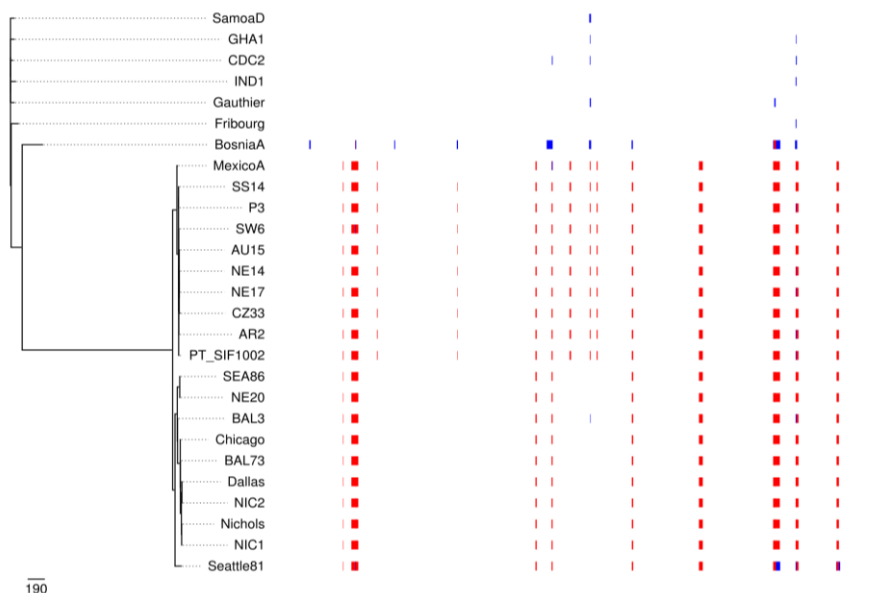
Available at:

[https://docs.google.com/presentation/d/1eUwo\\_IMBF13HuakqhYhO\\_iPNgix0aZ/W/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true](https://docs.google.com/presentation/d/1eUwo_IMBF13HuakqhYhO_iPNgix0aZ/W/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true)

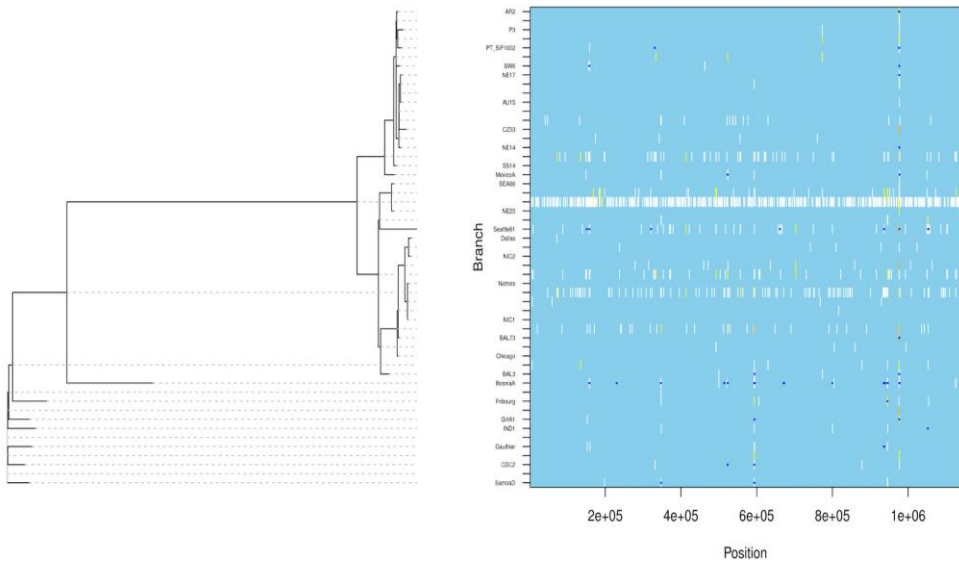
**Supplementary Figure 15.** Comparison of the topologies of two maximum likelihood trees: that obtained with all genes included in the alignment of the NIC-mapped dataset (A, 1,139,633 bp), and that obtained after excluding the recombinant genes plus *tp0897* from the alignment (B, 1,117,857 bp) (Figure 15).

Available at:

<https://docs.google.com/presentation/d/11t4DBfihMLDrNoeyuzS8QzgWLMoiOjPo/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true>

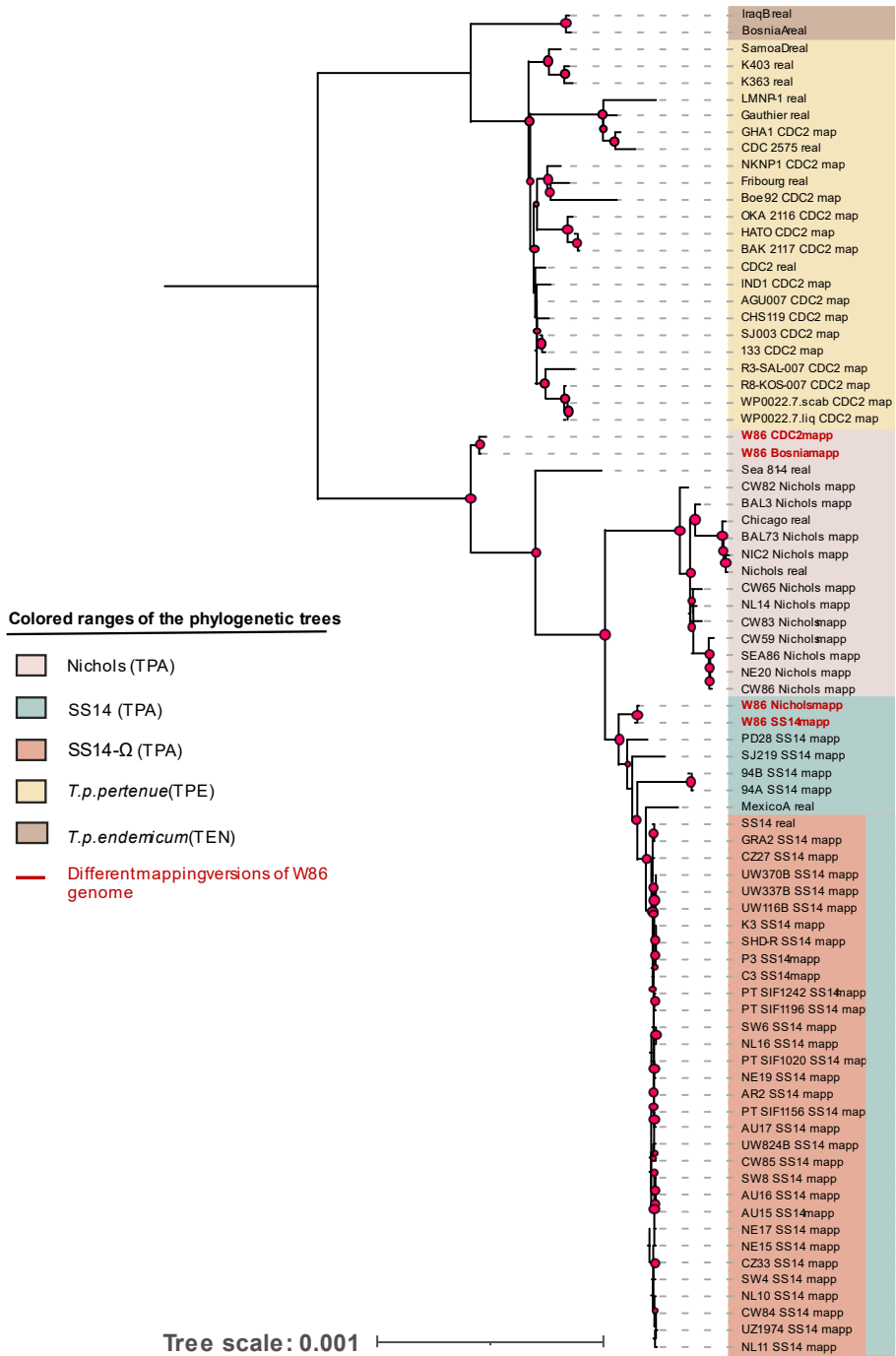


**Supplementary Figure 16.** Recombination events detected by Gubbins using a whole genome alignment of 27 representative *T. pallidum* strains. Regions identified as recombination events by Gubbins are represented by coloured blocks. Blue blocks are unique to a single isolate while red blocks are shared by multiple isolates. The horizontal position of the blocks represents their position in the alignment.



**Supplementary Figure 17.** Graphical representation of recombination events detected by ClonalFrameML. Each branch of the tree corresponds to a row of the heat map, horizontally aligned according to whole genome alignment of the 27 representative *T. pallidum* strains included in this analysis.

## Appendices



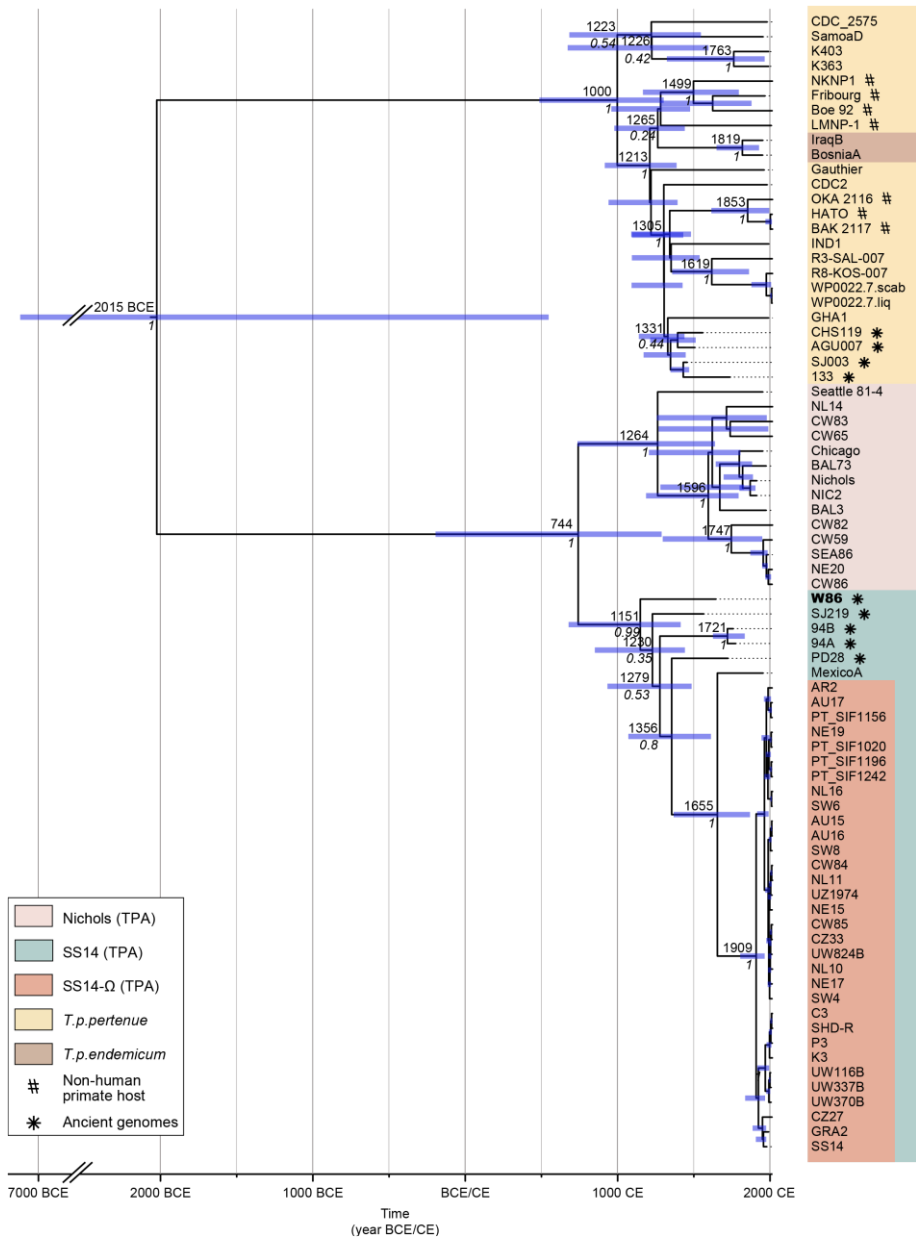
**Supplementary Figure 18.** Maximum likelihood tree (ML) obtained by IQ-TREE using GTR+G+I as the evolutionary model of the 76 *T. pallidum* strains included in this study plus the four different mapping versions of the W86 genome.

**Supplementary Figures 19-36:** A) Each figure presents the maximum likelihood tree obtained from the whole multiple alignment of 77 *T. pallidum* strains (Figure 22). In each figure, the inferred source and recipient clade/strain of the recombination events detected for the corresponding locus is represented. These were inferred by application of a maximum parsimony principle by which transfers are assumed to have occurred from an ancient to a recent branch in the evolution of *T. pallidum* after analyzing the distribution of SNPs in each event. B) Detail of the multiple sequence alignment with the SNPs and strains involved in the recombination event(s) is shown.

Available at:

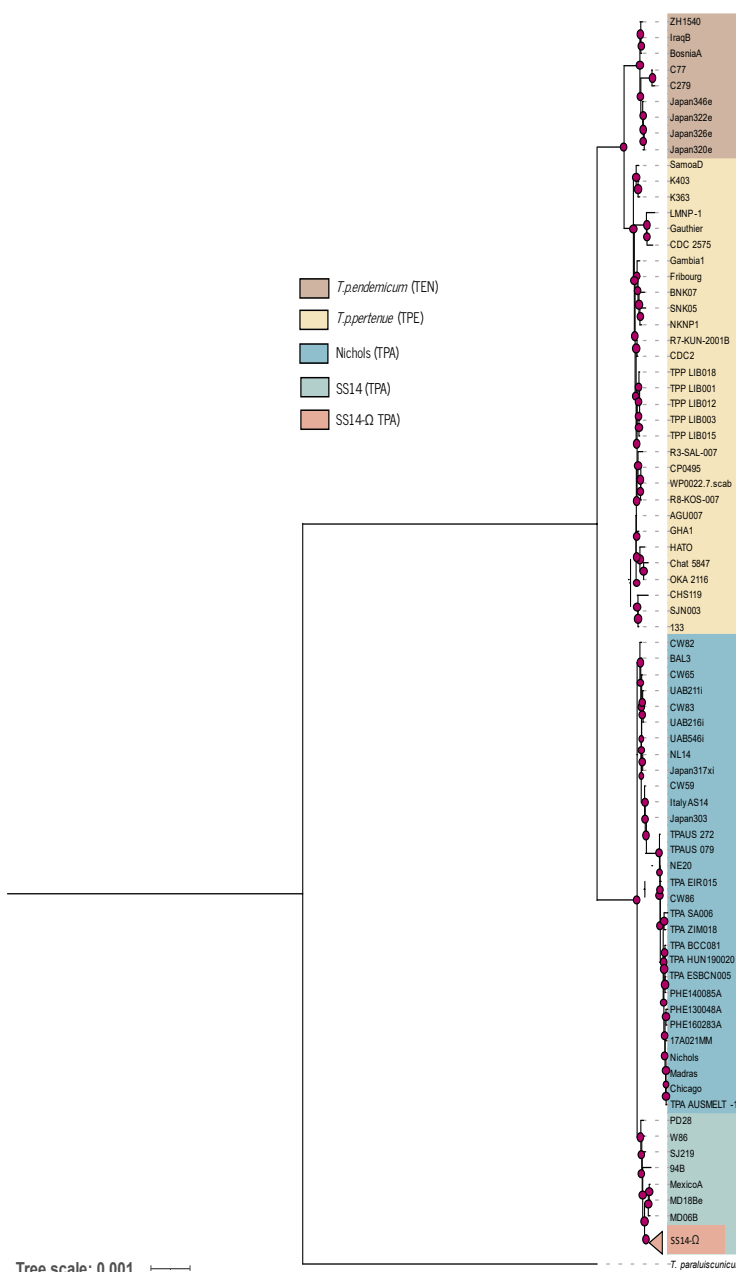
<https://drive.google.com/file/d/1nnnt5GQXLNq0Hf7uGTtbV5gFj18VDup3/view?usp=sharing>

## Appendices



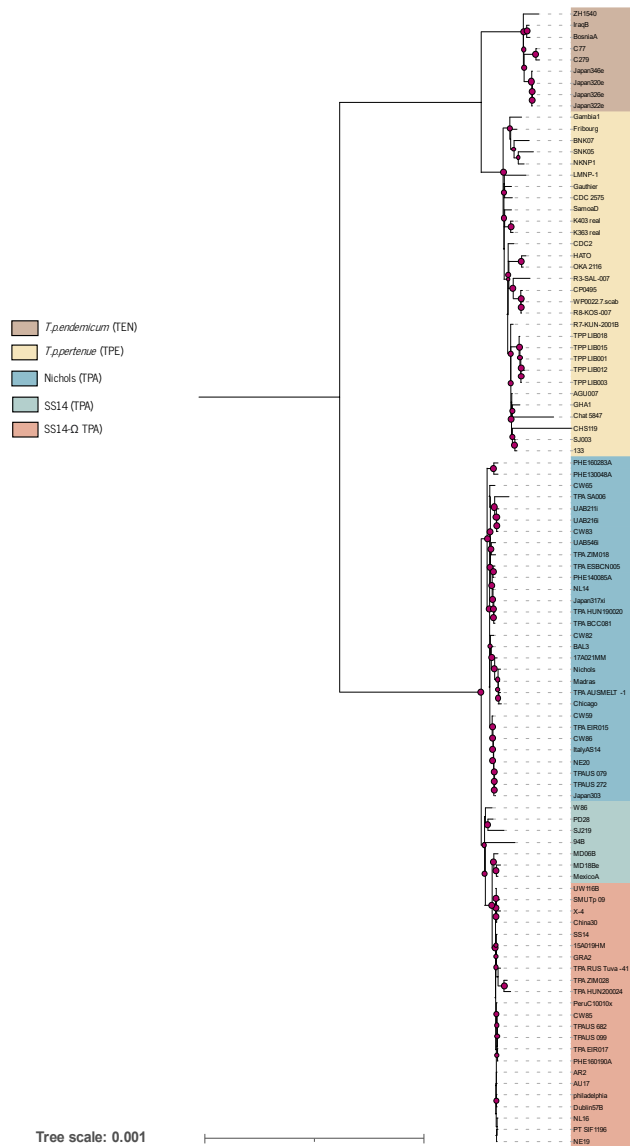
**Supplementary Figure 37.** Maximum clade credibility (MCC) tree of the dataset consisting of 68 modern and 9 ancient genomes (newly sequenced sample, W86, has been bolded) with recombinant and hypermutator genes removed, estimated in **BEAST2** 2.6.3 [149] under an uncorrelated lognormal relaxed clock model and a Bayesian skyline plot demographic model. Median age and Posterior Bayesian support estimates are provided for selected nodes. Blue bars show node age 95% HPD.

## Appendices



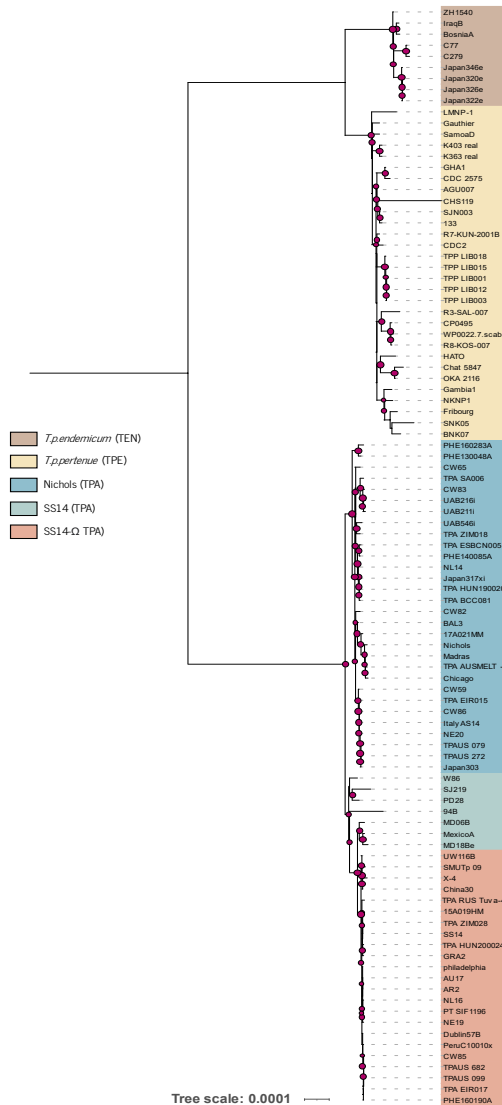
**Supplementary Figure 38.** Maximum likelihood tree with all genes included. ML tree from the multiple genome alignment, with *T. paraluiscuniculi* used as an outgroup and to root the phylogeny. The different clades corresponding to TPE and TEN, and the Nichols and SS14 lineages of TPA are colour-coded according to the legend. Bootstrap support values higher than 70% are indicated by red circles, with circle size proportional to bootstrap support percentage.

## Appendices



**Supplementary Figure 39.** Maximum likelihood tree with recombinant, and hypervariable genes excluded. ML tree obtained after excluding *tp0897*, *tp0316*, *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.

Available at: [https://drive.google.com/file/d/1a2cx7j5j0hN\\_IJfw-vHzuPL5UWDM0-Yu/view?usp=sharing](https://drive.google.com/file/d/1a2cx7j5j0hN_IJfw-vHzuPL5UWDM0-Yu/view?usp=sharing)



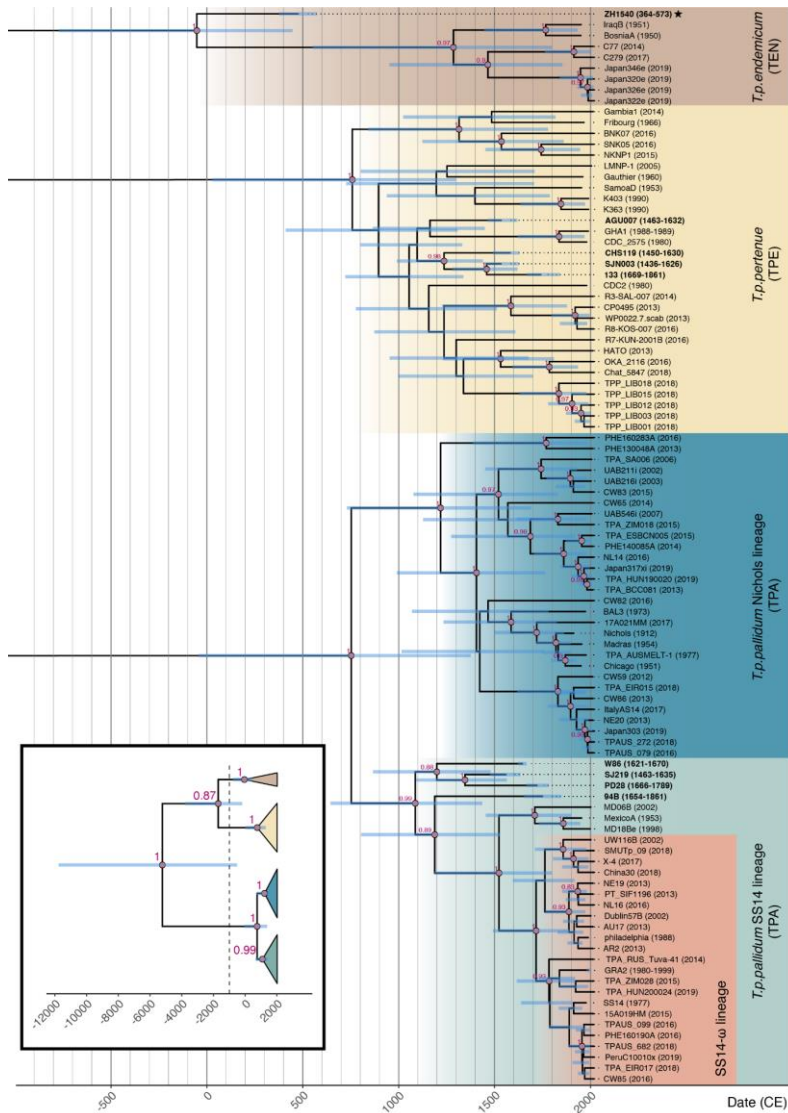
**Supplementary Figure 40.** Maximum likelihood tree with recombinant, hypervariable and 16S, 23S genes excluded. ML tree obtained after excluding *tp0897*, *tp0316*, *tp0317*, *16S*, *23S* and recombinant genes from the multiple genome alignment. The different clades corresponding to TPE and TEN, and 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.

Available at:

<https://drive.google.com/file/d/14zWjcdOVsT2YyICofFVGgd0nA4LU1zPT/view?usp=sharing>



## Appendices

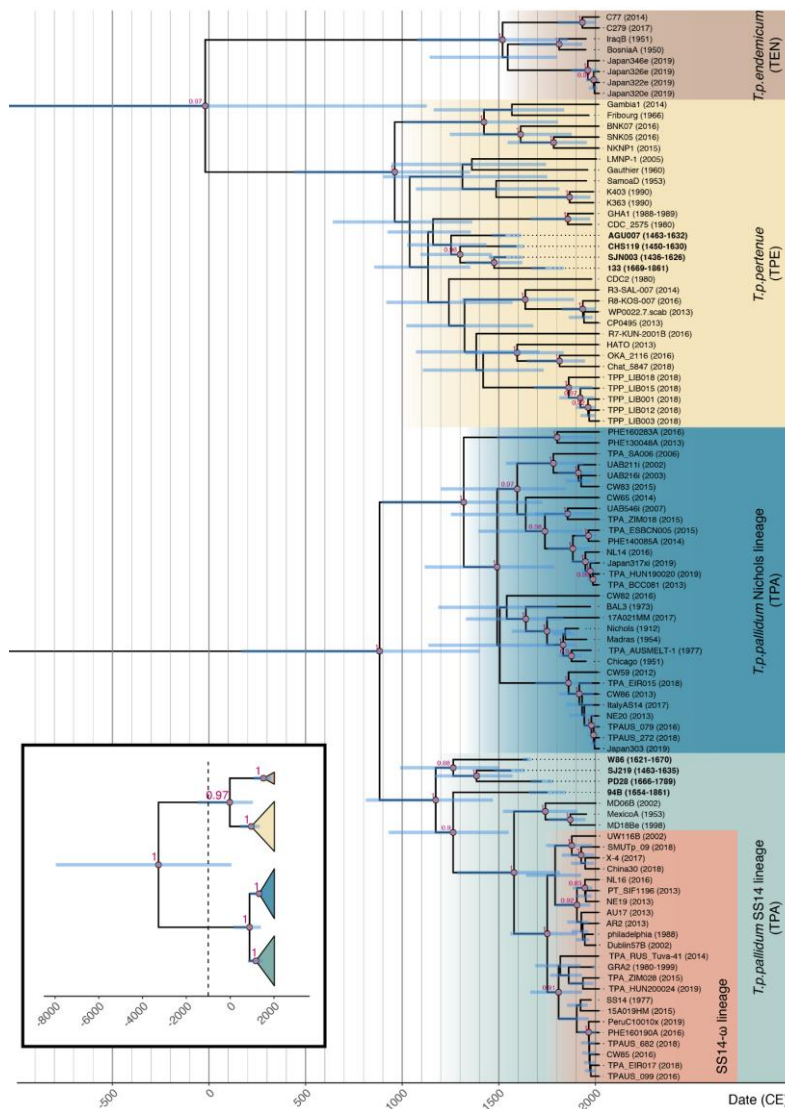


**Supplementary Figure 41.** Maximum-clade credibility (MCC) tree from the molecular clock dating analysis of the 98 genome context dataset with ZH1540 included ( $n=99$ ). The inset shows a simplified view of the entire tree, with the dashed line indicating the part of the tree shown in the main figure. Ancient genomes are labeled in bold text and ZH1540 is marked by a star. Blue bars indicate the 95% HPD intervals of node ages and red text the posterior probability that a clade is monophyletic (only shown for nodes with posterior probability  $> 0.8$ ).

Available at:

[https://drive.google.com/file/d/13k8MW3jA6YXyoRKxrfCpIr\\_rgh0eLnAa/view?usp=sharing](https://drive.google.com/file/d/13k8MW3jA6YXyoRKxrfCpIr_rgh0eLnAa/view?usp=sharing)

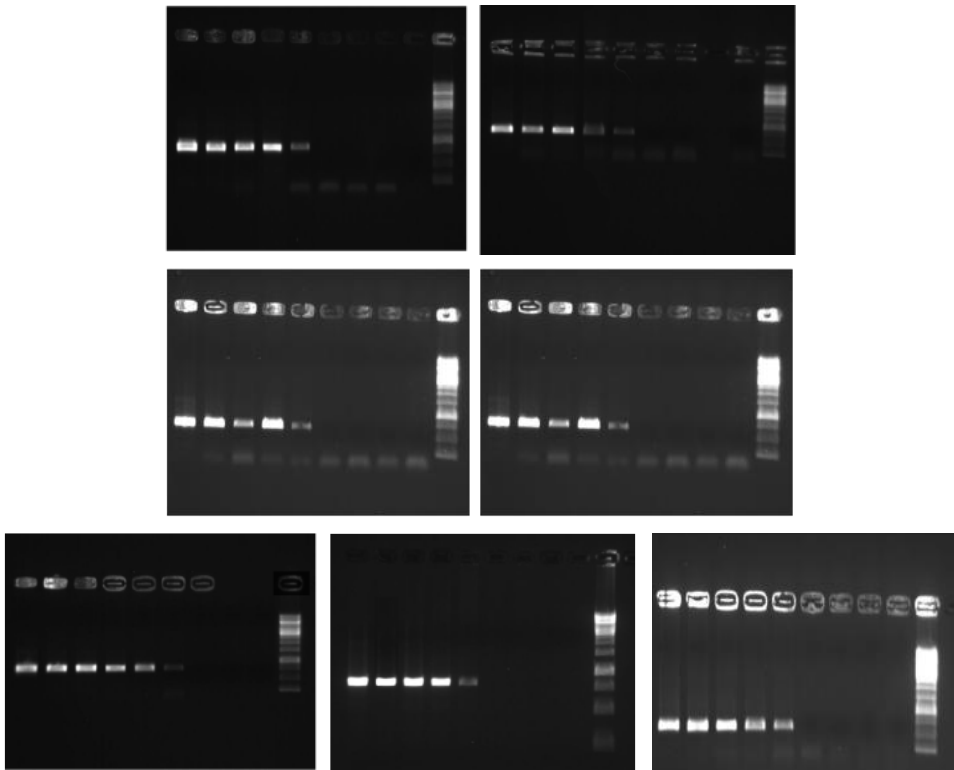
## Appendices



**Supplementary Figure 42.** Maximum-clade credibility (MCC) tree from the molecular clock dating analysis of the 98 genome context dataset with ZH1540 excluded (n=98). The inset shows a simplified view of the entire tree, with the dashed line indicating the part of the tree shown in the main figure. Ancient genomes are labelled in bold text. Blue bars indicate the 95% HPD intervals of node ages and red text the posterior probability that a clade is monophyletic (only shown for nodes with posterior probability > 0.8).

Available at:

[https://drive.google.com/file/d/12FLI-H6\\_GRz0vJLuYXwcIA-9babrj95\\_/view?usp=sharing](https://drive.google.com/file/d/12FLI-H6_GRz0vJLuYXwcIA-9babrj95_/view?usp=sharing)



**Supplementary Figure 43.** Results of PCR performed for the new set of primers designed for the new MLST scheme of *T. pallidum*. Each image corresponds to the following genes in the same order: *tp0136*, *tp0326*, *tp0548*, *tp0795*, *tp0858*, *tp0865*, *tp1031*. Likewise, for each gene, seven samples corresponding to a serial dilution of the Nichols sample were amplified, arranged in each image in the following order: Nichols,  $10^{-1}$ ,  $10^{-2}$ ,  $10^{-3}$ ,  $10^{-4}$ ,  $10^{-5}$ ,  $10^{-6}$ ,  $10^{-7}$ , and the negative control.

## Supplementary Tables

**Supplementary Table 1.** Detailed results of SNP calling results for each sample and reference with a coverage threshold of 3.0 and with a Minimum homozygous SNP allele frequency of 0.9 (90%).

Available at:

[https://docs.google.com/spreadsheets/d/1i\\_FeILi1vvGTV\\_6unGRVWkd5fepqEXw/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/1i_FeILi1vvGTV_6unGRVWkd5fepqEXw/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true)

**Supplementary Table 2.** Complete results of the likelihood mapping test for the MSA using Nichols as reference genome for mapping. Of the 978 protein-coding genes, 380 showed some phylogenetic signal and were retained for the ensuing analyses. It was not possible to perform this test for 23 genes due to the large number of undetermined positions in the corresponding multiple alignments. (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.).

Available at:

[https://docs.google.com/spreadsheets/d/1Dwrl9uYUOIODQlduf032xxSxEmLS\\_uTc/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/1Dwrl9uYUOIODQlduf032xxSxEmLS_uTc/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true)

**Supplementary Table 3.** Detailed results of the likelihood mapping test for the MSA using SS14 as reference genome for mapping. Of the 975 protein-coding genes, 498 genes showed some phylogenetic signal and were retained for the ensuing analyses. It was not possible to perform this test for 5 genes due to the large number of undetermined positions in the corresponding multiple alignments.

Available at:

[https://docs.google.com/spreadsheets/d/1T3rZrd0Ao\\_Aim7cjb\\_e5YL0gtwHKAgEt/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/1T3rZrd0Ao_Aim7cjb_e5YL0gtwHKAgEt/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true)

**Supplementary Table 4.** Detailed results of the likelihood mapping test for the MSA using CDC2 as reference genome for mapping. Of the 1066 protein-coding genes, 535 genes showed some phylogenetic signal and were retained for the ensuing analyses. It was not possible to perform this test for 11 genes due to the large number of undetermined positions in the corresponding multiple alignments.

Available at:

[https://docs.google.com/spreadsheets/d/1IMaVK0SsWbl\\_GtN39qPRs4iYhh\\_ia3-l/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/1IMaVK0SsWbl_GtN39qPRs4iYhh_ia3-l/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true)

**Supplementary Table 5.** Topology test results for the protein coding genes with the multiple alignments obtained with the three different genomes (Nichols, SS14, and CDC-2) as references for mapping. The table presents the results of the two tests (p-values for SH, and a posterior weight for ELW) for the final genes selected in the topology tests for the three references simultaneously. 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 whole genome alignment rejected the gene topology (reciprocal incongruence,  $p < 0.05$  and weight value close to 0). Dashes indicate that the corresponding gene was not detected in the topology test for that reference.

Available at:

[https://docs.google.com/spreadsheets/d/1CdEbxzVvnBIpE3WVHSJ\\_III03dG8OmlC/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/1CdEbxzVvnBIpE3WVHSJ_III03dG8OmlC/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true)

**Supplementary Table 6.** Recombination events detected using Gubbins, with the initial and last position referred to TPA Nichols strain coordinates and length of each event.

Available at:

<https://docs.google.com/spreadsheets/d/1Ng59rDcPNNEfXFemnBq-ckJ8nW-AafZu/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true>

**Supplementary Table 7.** Recombination events detected using ClonalFrameML, with the initial and last position referred to TPA Nichols strain coordinates and length of each event.

Available at:

[https://docs.google.com/spreadsheets/d/1rO3SqvcEKB\\_iVNHriQ0sjlZG7TJbMvvb/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/1rO3SqvcEKB_iVNHriQ0sjlZG7TJbMvvb/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true)

**Supplementary Table 8.** Protein-coding genes according to Nichols strain used in the Codeml and SNPeff analyses. The table shows the gene substitutions (obtained by SNPeff) and sites (obtained by Codeml), and the dN, dS, dN/dS, expected SNPs, and the difference between observed and expected SNPs divided by the expected ones.

Available at:

[https://docs.google.com/spreadsheets/d/1RGnOhWXs40\\_IVXsxb-dTWkE2Ev3gmhyf/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/1RGnOhWXs40_IVXsxb-dTWkE2Ev3gmhyf/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true)

**Supplementary Table 9.** Results of the BUSTED test, using Hyphy, to detect the action of natural selection on recombinant genes (Table 1), non-recombinant genes with evidence of positive selection (Supplementary Table 9), and genes in which the number of observed SNPs exceeded twice the number of expected ones (Supplementary Table 10). The genes with - were not tested due to an excess of missing data.

Available at:

[https://docs.google.com/spreadsheets/d/1vc4\\_MiOttH3OCdrZKNC19P3\\_C1pgJS8r/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/1vc4_MiOttH3OCdrZKNC19P3_C1pgJS8r/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true)

**Supplementary Table 10.** Genome reconstruction details for all published samples and genomes used in the analyses

Available at:

<https://docs.google.com/spreadsheets/d/1oYDSuTTEue88NnqNI4kg9gJWIKT9DdM53k1OcCJBMfg/edit?usp=sharing>

**Supplementary Table 11.** Orthology analysis results obtained by Proteinortho showing the detected orthologous genes in the four reference genomes employed.

Available at:

<https://docs.google.com/spreadsheets/d/1T42-pTIVcJX-ZRwNm2FwL5xhdcZNrfS7/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true>

**Supplementary Table 12.** 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 funded and their presence or absence in each reference genome.

Available at:

<https://docs.google.com/spreadsheets/d/1bxrEVtaoOFyyy1Zfl2Xf069dOe9xBkba/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true>

**Supplementary Table 13.** The number of SNPs detected per gene present in at least one of the four different reference genomes employed in the mapping analysis.

Available at:

[https://docs.google.com/spreadsheets/d/1OiGT\\_hSISwmk\\_QobPefX\\_6gMAoXqgf-n/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/1OiGT_hSISwmk_QobPefX_6gMAoXqgf-n/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true)

**Supplementary Table 14.** 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.)

Available at:

<https://docs.google.com/spreadsheets/d/1QclqpbvZgVMAiyvrAihrNkxxCLYHtH0D/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true>

**Supplementary Table 15.** 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 whole genome alignment rejected the gene topology (reciprocal incongruence,  $p < 0.05$  and weight value close to 0).

Available at:

[https://docs.google.com/spreadsheets/d/1B1\\_41HilnU-Gw6km6pXmkI3k1GKuerrW/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/1B1_41HilnU-Gw6km6pXmkI3k1GKuerrW/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true)

**Supplementary Table 16.** The results of the ancestral sequence reconstruction of the whole genome-based tree with the whole-genome alignment using TreeTime, detailed for the polymorphisms specific to either TPE or TEN and on the possible reversions occurring in the ancestral nodes of TPE and TEN strains, for the 18 putative recombinant genes and to loci *tp0897*, *tp0316* and *tp0317* removed from the multiple genome alignment to obtain the vertical-inheritance genome phylogeny. For each gene, it shows the possible reversions detected highlighted in red and the particular polymorphisms that affect the phylogenetic placement of the TEN genomes without those 21 corresponding genes.

Available at:

[https://docs.google.com/spreadsheets/d/1HMH4VFgbI49JjwT9Z6P\\_nVAJUKwAFG3g/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/1HMH4VFgbI49JjwT9Z6P_nVAJUKwAFG3g/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true)

**Supplementary Table 17.** The results of the ancestral sequence reconstruction of the whole genome-based tree with the whole-genome alignment using TreeTime, detailed for the polymorphisms specific to the ancestral node of TEN strains causing the long branch for this specific lineage.

Available at:

<https://docs.google.com/spreadsheets/d/1U9MolvMmFqdfEKdzV0xq880CdMS2LfTn8CSeWrbrjQ4/edit?usp=sharing>

**Supplementary Table 18.** The marginal probabilities for the polymorphism of the ancestral nodes of TPE and TEN subspecies as well as one containing guesses (obtaining the maximum probability) or the actual ancestral sequences calculated by RaxML, detailed for the recombinant genes and *tp0897*, *tp0316* and *tp0317*.

Available at:

<https://docs.google.com/spreadsheets/d/1K1oMIq2ZlAybvRG5497piVL6liU2WT6TfWiyExEDXiM/edit?usp=sharing>

**Supplementary Table 19.** 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 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.

Available at:

<https://docs.google.com/spreadsheets/d/1ZhiF8tTz5wt0oieJxjP-1nH7JNk5GY7e/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true>



**Supplementary Table 20.** Functional significance of the genes detected as recombinant and/or under positive selection according to Uniprot or in the literature.

Available at:

[https://docs.google.com/spreadsheets/d/1ydYaWjrfYzrH\\_Ej\\_cEvpLuPeg-9VrkAtB1Oh-4o7OVw/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1ydYaWjrfYzrH_Ej_cEvpLuPeg-9VrkAtB1Oh-4o7OVw/edit?usp=sharing)

**Supplementary Table 21.** Metadata for all published samples and genomes used in the analyses.

Available at:

[https://docs.google.com/spreadsheets/d/15KiVXoWck\\_TfX8ZtDU4ijNurRINXp1w\\_vyrvEvW3EzU/edit?usp=sharing](https://docs.google.com/spreadsheets/d/15KiVXoWck_TfX8ZtDU4ijNurRINXp1w_vyrvEvW3EzU/edit?usp=sharing)

**Supplementary Table 22.** Radiocarbon dating results for the four samples used for genome reconstruction Raw (unmodelled) values are denoted with light grey, and calibrated, marine reservoir effect corrected values in black.

Available at:

<https://docs.google.com/spreadsheets/d/18K4sY9mJ8DTfHxoK-fZ5iBHv65HuuSUKUqchOjhNsrY/edit?usp=sharing>

**Supplementary Table 23.** Radiocarbon dating results for the four samples used for genome reconstruction. Raw (unmodelled) values are denoted with light grey, and calibrated, marine reservoir effect corrected values in black.

Available at:

[https://docs.google.com/spreadsheets/d/1c-G9BpCxF\\_rcHDTqqRkw3Stp\\_X20so0fSrf9G2DiMKY/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1c-G9BpCxF_rcHDTqqRkw3Stp_X20so0fSrf9G2DiMKY/edit?usp=sharing)

**Supplementary Table 24.** Post target-enrichment results (samples with successful genome reconstruction marked with \*).

Available at:

[https://docs.google.com/spreadsheets/d/1G-V0srv7tXbtd4KkZCliwHmdh\\_HOP7tr7uN6FE9kx8Y/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1G-V0srv7tXbtd4KkZCliwHmdh_HOP7tr7uN6FE9kx8Y/edit?usp=sharing)

**Supplementary Table 25.** Genomic details for the 125 SNPs identified for the reconstructed ancient genome ZH1540. Recombinant genes are denoted in purple.

Available at:

<https://docs.google.com/spreadsheets/d/1KbIE83rIJ0kE72uQJ45Me3jxUXm5mgVL7jSI16OT8wI/edit?usp=sharing>

**Supplementary Table 26.** Genomic coordinates for each gene present in at least one of the four reference genomes, according to their location in the final merged alignment. Acronyms from Proteinortho (TPASS = SS14, TPANIC = Nichols, TENBA = BosniaA and TPECDC2 = CDC2) were used to create summary IDs for each gene present in the respective reference genome.

Available at:

[https://docs.google.com/spreadsheets/d/1iiDu4advjsivc6SpiekP\\_EigX4DxgApIsuPdf3--rCw/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1iiDu4advjsivc6SpiekP_EigX4DxgApIsuPdf3--rCw/edit?usp=sharing)

**Supplementary Table 27.** The number of SNPs detected per gene, present in the reference genomes mentioned (TPASS = SS14, TPANIC = Nichols, TENBA = BosniaA and TPECDC2 = CDC2).

Available at:

[https://docs.google.com/spreadsheets/d/1S62bAt\\_08gFY4nrw-b1sX7udlVcpA\\_dcPNwnKFeKb4k/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1S62bAt_08gFY4nrw-b1sX7udlVcpA_dcPNwnKFeKb4k/edit?usp=sharing)

**Supplementary Table 28.** Results obtained from the likelihood-mapping test: The 344 genes showing phylogenetic signal. (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.) 36 cases could not be included in the LM test due to a large amount of missing data.

Available at:

<https://docs.google.com/spreadsheets/d/16lmrtqwJY5B7u7q1CzGNB1ZVLaH3QpEbx2zfQ0VyFmI/edit?usp=sharing>

**Supplementary Table 29.** 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 the 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 whole genome alignment rejected the gene topology (reciprocal incongruence,  $p < 0.05$  and weight value close to 0).

Available at:

<https://docs.google.com/spreadsheets/d/1IbCdGRzxSd7lpt-jJgI-0CoUmEwjmlP3pO4s1cfpUSw/edit?usp=sharing>

**Supplementary Table 30.** Recombination events detected. 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. The arrows separate involved donor strains from the recipient strains. Bold font is used to highlight novel genes identified as recombinant in this study.

Available at:

<https://docs.google.com/spreadsheets/d/12FizcSaRzBoVR3cT7ygbfJYR4Lk2Ex0YLDpyOxT0asw/edit?usp=sharing>

**Supplementary Table 31.** The 121 *in silico* genomes employed for the design of the new MLST scheme for *T. pallidum*.

Available at:

<https://docs.google.com/spreadsheets/d/1gxLqP71I4zadSTzBy3tYcJT2idGadDuZ2PIUozgAGWg/edit?usp=sharing>

**Supplementary Table 32.** Complete results of the likelihood mapping test for the MSA using Nichols as reference genome for mapping. Of the 978 protein-coding genes, 332 showed some phylogenetic signal and were retained for the ensuing analyses. It was not possible to perform this test for 198 genes due to the large number of undetermined positions in the corresponding multiple alignments. (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).

Available at:

<https://docs.google.com/spreadsheets/d/1i1m0Q3G6RdlMfiEL6JUEx4yY0ojLpgAl/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true>

**Supplementary Table 33.** Number of SNPs obtained per each gene selected in the likelihood mapping test.

Available at:

[https://docs.google.com/spreadsheets/d/1X\\_Nx\\_s-KFgWwrFU10yMZFelbgEZtFIj7/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/1X_Nx_s-KFgWwrFU10yMZFelbgEZtFIj7/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true)

**Supplementary Table 34.** The results of the in-house script employed to know which subspecies and/or sublineages could be differentiated by each SNPs per gene. The excel file includes 21 different sheets, 20 corresponding the results of this script for the final candidate genes for primers design (Figure 25) plus the additional gene *tp0705*. The first column of each sheet represents the variable positions for that gene, and the following columns show which subspecies or sublineages it is possible to differentiate according to which SNPs and which strains have SNPs at that position. A summary graph of the results of that script by gene is also shown.

Available at:

<https://docs.google.com/spreadsheets/d/1UR0MKfxTER8cRIg6ZMBpArdjpAaCPOtsewoQdf1HxoU/edit?usp=sharing>

**Supplementary Table 35.** Minimum and theoretical number of haplotypes (ST) to differentiate the final 16 genes selected for the new MLST scheme. Note that some genes have more than one possible primer combination and all of them have been analyzed and included in this table. When the gene has more than one possible primer combination, it is indicated by different numbers. For each sample and locus, the different alleles determined are indicated in different colors, as well as the final haplotype (ST) corresponding to each allelic combination for the 16 loci. This table does not include the 23S gene. Alleles that could not be assigned due to the presence of Ns in the sequences of those genes for those samples are highlighted in black.

Available at:

<https://docs.google.com/spreadsheets/d/1KXlwGnMTUgOZIb2gw DTOxlii5-DFPn0/edit?usp=sharing&oid=105360711200422403961&rtpof=true&sd=true>

**Supplementary Table 36.** The minimum and theoretical number of haplotypes (STs) that allows differentiating the 7 final genes selected for the new MLST scheme. For each sample and locus, the different alleles determined are indicated in different colors as well as the final haplotype (ST) that corresponds to each allele combination for the 7 loci. This table does not include the 23S gene. Alleles that could not be assigned due to the presence of Ns in the sequences of those genes for those samples are highlighted in black. The haplotypes (STs) indicated with (?) are STs not properly determined due the indetermination in some of their alleles (NA) because of Ns in that gene's sequences.

Available at:

<https://docs.google.com/spreadsheets/d/1XDIAktVyZh68iGhJXmX6WvGt51uRQGn6/edit?usp=sharing&oid=105360711200422403961&rtopf=true&sd=true>

**Supplementary Table 37.** Final allelic profiles obtained per gene obtained for each sample. The table also includes information on the country where the sample was collected, the source of the sample, and whether the allelic profile was determined through in-silico analysis or experimentally. The symbol "-" indicates those alleles that could not be determined for a given sample.

Available at:

<https://docs.google.com/spreadsheets/d/1bzIAbWRHCRtxeGIKqXXcqUFPYBDGFDFit9d8YbaeTV4/edit?usp=sharing>

**Supplementary Table 38.** The STs obtained per each sample. The table also includes information on the country where the sample was collected, the source of the sample, and whether the allelic profile was determined through in-silico analysis or experimentally. It also specifies the subspecies or sublineage determined for each sample and the macrolide resistance profile determined.

Available at:

<https://docs.google.com/spreadsheets/d/1N8X5rXyjjD0puZstpF9q7UMoVHnPX3jFJcvAwyFoQ68/edit?usp=sharing>



— **RESUMEN EN CASTELLANO** —





## Resumen en castellano

### Introducción

*Treponema pallidum* es una bacteria gram negativa responsable de la sífilis, el pian y el bejel, también conocidas como enfermedades treponémicas. Concretamente, están causadas por tres subespecies estrechamente relacionadas, *T. pallidum* subsp. *pallidum* (TPA), *T. pallidum* subsp. *pertenue* (TPE) y *T. pallidum* subsp. *endemicum* (TEN), respectivamente. Otra treponematosi similar, la pinta, causada por *T. carateum*, es mucho menos conocida debido a la falta de aislados disponibles, lo que dificulta el análisis genético de este organismo. Actualmente, se han incrementado los casos de sífilis en todo el mundo, con más de 7 millones de infecciones notificadas anualmente. La incidencia del pian también ha aumentado en regiones donde su incidencia había disminuido sustancialmente debido a los esfuerzos de erradicación de la Organización Mundial de la Salud, y hay nuevos informes de infecciones por bejel en contextos clínicos inesperados. Este resurgimiento de las treponematosi se produce tras un descenso sustancial a mediados del siglo XX debido al descubrimiento de los antibióticos y a las campañas de tratamiento y prevención. Antes de eso, la sífilis había sido una enfermedad devastadora en Europa durante al menos 500 años. Su origen es desconocido, pero los primeros casos se reportaron como una serie de brotes repentinos en Nápoles y en toda Europa tras el regreso de Colón a finales del siglo XV. Los informes indican una rápida propagación posterior por todo el Viejo Mundo desde entonces. Sin embargo, no está claro si la sífilis, el pian o el bejel habían existido en Europa antes de esa fecha. Por todo ello, la historia de estas enfermedades ha intrigado a historiadores y científicos durante siglos.

Las tres treponematosi tienen síntomas clínicos muy similares: presentan tres estadios diferentes, se transmiten por contacto directo con la piel y forman úlceras en el lugar de transmisión antes de progresar a otros síntomas que se solapan con enfermedades no treponémicas. Sin embargo, existen diferencias demográficas y

geográficas: la sífilis se considera venérea y no está restringida geográficamente, mientras que el pian y el bejel se transmiten predominantemente a través del contacto cutáneo y afectan a comunidades rurales de los países tropicales, donde son endémicas.

Su diagnóstico suele lograrse mediante una combinación de examen clínico y patológico, con métodos serológicos y/o moleculares como la PCR o la PCR en tiempo real. Las pruebas serológicas tienen una sensibilidad y especificidad limitadas en comparación con los métodos moleculares en las etapas primarias de la infección, pero son más precisas para las etapas secundarias o latentes de la enfermedad. Sin embargo, estos métodos son incapaces de diferenciar entre las diferentes treponemosis. Además, la suposición de una distinción en el modo de transmisión puede ser problemática, ya que estudios recientes han demostrado que el bejel también puede transmitirse por contacto sexual, mediante el análisis genético de pacientes varones sexualmente activos diagnosticados de sífilis pero que en realidad estaban infectados por bejel.

Sin embargo, gracias al desarrollo de tecnologías moleculares es posible identificar las diferentes subespecies de *T. pallidum* mediante el análisis genético de loci específicos, lo que permite un diagnóstico preciso y que además puede ayudar a la vigilancia y los análisis epidemiológicos de esta bacteria. En los últimos años incluso se han logrado obtener genomas completos de esta bacteria, gracias al proceso de enriquecimiento del ADN de las muestras, puesto que hasta la fecha, no existe un sistema de cultivo estandarizado que permita la obtención de genomas de una forma más económica y efectiva como con otras bacterias cultivables. Estos métodos de enriquecimiento son caros y laboriosos debido a las pequeñas cantidades de ADN de *T. pallidum* en las muestras clínicas. Esto subraya la necesidad de disponer de herramientas de genotipado que difieran de la secuenciación de genomas completos hasta que se facilite su obtención, para mejorar nuestra comprensión de la epidemiología de esta bacteria. Actualmente existen diferentes propuestas de esquemas MLST o de tipado para *T. pallidum*,

aunque tienen limitaciones en cuanto al número de loci utilizados y las dificultades técnicas en el análisis de muestras clínicas. Además, estos esquemas no están diseñados para captar toda la variabilidad genética dentro y entre las subespecies. Todo esto pone de manifiesto la necesidad de desarrollar un nuevo esquema MLST diseñado para aplicarse a cualquiera de las tres subespecies, que se diseñe teniendo en cuenta la información de los genomas completos disponibles actualmente y que salvaguarde las dificultades técnicas que presentan los anteriores métodos de tipado disponibles.

A pesar de las dificultades para obtener genomas completos de *T. pallidum*, el uso de tecnologías de secuenciación de alto rendimiento ha permitido generar numerosos genomas de *T. pallidum* en los últimos años. Esto ha aportado valiosos conocimientos sobre la historia evolutiva de esta bacteria, las relaciones genéticas entre subespecies, los procesos de recombinación y selección y los cambios en la dominancia de las cepas a lo largo del tiempo. Además, los avances tecnológicos y la cuidadosa selección de muestras con síntomas de patogénesis treponémica han permitido recuperar genomas antiguos de *T. pallidum* a partir de restos arqueológicos humanos, lo que ofrece oportunidades sin precedentes para estudiar su divergencia, origen y evolución. Todo esto abre la puerta a la investigación de cuestiones intrincadas como la evolución y el origen de las treponematoses, además de facilitar la obtención del limitado conocimiento epidemiológico que tenemos sobre estas tres enfermedades.

## Objetivos

Esta tesis doctoral pretende profundizar en el conocimiento de la evolución y genómica de *T. pallidum* mediante el estudio de los genomas antiguos y modernos de esta bacteria y su relevancia en la epidemiología de la treponematosi. Los objetivos específicos son los siguientes:

- **Investigar el papel de la selección natural y la recombinación en el genoma de *T. pallidum*.** Mediante el análisis de genomas completos, se examinará el impacto de diferentes procesos evolutivos en esta bacteria. La atención se centrará en la recombinación, porque tiene la capacidad de generar rápidamente nuevas combinaciones genéticas, y en la selección natural, que permite conocer los genes y proteínas que han sido objeto de selección adaptativa y purificadora en distintos linajes.
- **Inferir patrones de evolución y divergencia mediante el estudio de genomas antiguos de *T. pallidum*.** El análisis de genomas treponémicos antiguos ofrece la oportunidad de mejorar nuestra comprensión de las relaciones evolutivas y la diversificación en las subespecies treponémicas. Estas cepas históricas proporcionan una sólida inferencia filogenética que no está disponible de otro modo. Además, la búsqueda de pruebas genéticas de treponematosi anteriores al contacto en las Américas y el Viejo Mundo puede arrojar luz sobre el origen de esta bacteria.
- **Desarrollar un nuevo esquema MLST para *T. pallidum* basado en datos genómicos.** Debido a la falta de un sistema de cultivo estandarizado, es crucial disponer de un método de tipificación genética fiable para estudiar las muestras de *T. pallidum*. Por lo tanto, esta tesis pretende diseñar nuevos cebadores para la amplificación por PCR de loci adicionales para mejorar el sistema de tipado mediante secuenciación por Sanger. El objetivo es mejorar la discriminación entre linajes y cepas de TPA, particularmente en el linaje SS14, así como entre TPA, TPE y TEN.

La atención se centrará en maximizar la información genética (número de polimorfismos de nucleótido único y tipos de secuencia) con tamaños de fragmentos genómicos mínimos. La eficacia y las condiciones de uso de los cebadores diseñados se pondrán a prueba utilizando una colección de muestras de TPA y TPE, empezando por los grupos menos diversos y avanzando hacia los más diversos.

## Metodología

### 1. Obtención de genomas antiguos

En el Capítulo 2, en colaboración con un equipo de investigadores, se recogieron muestras antiguas de un yacimiento arqueológico además de realizar la exploración osteológica de los restos humanos obtenidos para detectar posibles lesiones causadas por *T. pallidum* en los dientes y los huesos. Este equipo estaba formado por la Dra. Hanna Panagiotopoulou, la Dra. Karolina Doa y el Dr. Wiesław Bogdanowicz, todos ellos de la Academia Polaca de Ciencias de Wrocław (Polonia), así como por el Dr. Paweł Dąbrowski, el Dr. Maciej Oziembłowski y la Dra. Joanna Grzelak, todos ellos de la Universidad Médica de Wrocław, también en Wrocław (Polonia). En el cribado genético inicial, que reveló la presencia de *Treponema* en una muestra, participaron la Dra. Martyna Molak, la Dra. Hanna Panagiotopoulou, la Dra. Karolina Doan y el Dr. Wiesław Bogdanowicz, de la Academia Polaca de Ciencias de Varsovia (Polonia). En el Capítulo 3, la recogida de muestras antiguas y la exploración osteológica fueron realizadas por el Dr. José Filippini y el Dr. Luis Pezo Lanfranco, ambos de la Universidad de São Paulo en São Paulo (Brasil). La datación por radiocarbono de las muestras antiguas fue realizada por el Dr. José Filippini y el Dr. Luis Pezo Lanfranco, y los detalles de este proceso de datación pueden encontrarse en la sección de Material Complementario del Capítulo 3.

El procesamiento de las muestras se llevó a cabo en colaboración con el grupo de la Dra. Verena J. W. Schünemann, de la Universidad de Basilea. Las muestras

incluían un diente premolar superior izquierdo y 99 muestras óseas. Para evitar la contaminación por ADN, se lijaron las superficies de las muestras antiguas y se lavaron con hipoclorito sódico, agua y etanol. A continuación, las muestras se irradiaron con UV y se pulverizaron. La extracción del ADN y la preparación de las bibliotecas se llevaron a cabo siguiendo protocolos establecidos en una sala blanca específica. El ADN extraído se utilizó para la secuenciación de alto rendimiento con una plataforma Illumina Nextseq para el cribado inicial de patógenos. El tratamiento con uracilo-ADN glicosilasa se utilizó para eliminar antiguos daños específicos del ADN antes de la preparación de bibliotecas adicionales. Las muestras candidatas se analizaron en busca de patógenos utilizando el programa MALT y Kraken2. Se obtuvo un perfil de daños para la autenticación del ADN antiguo y se realizó la captura del genoma completo de las muestras seleccionadas. El ADN capturado se secuenció utilizando una plataforma Illumina NextSeq. La resistencia a los antibióticos se investigó examinando mutaciones específicas en los operones del ARN ribosómico 23S.

## **2. Selección de conjuntos de datos, procesamiento de lecturas y generación de alineamientos genómicos múltiples**

En cada capítulo, generamos diferentes conjuntos de datos genómicos conformados en su mayoría por genomas de *T. pallidum* publicados en estudios previos, excepto los nuevos genomas antiguos obtenidos en los Capítulos 2 y 3. Concretamente, generamos un dataset de 75 genomas completos de *T. pallidum* para el Capítulo 1, un dataset de 77 genomas completos de *T. pallidum* en el Capítulo 2 y un dataset de genomas completos de 99 genomas en el Capítulo 3. En el Capítulo 4 se utilizó un dataset de 121 genomas completos para el diseño del nuevo esquema MLST, en el cual a su vez también se utilizó para el tipado *in silico* de dichos genomas. Es importante destacar que los dos dataset genómicos generados para los Capítulos 1 y 4 se obtuvieron en colaboración con la Dra. Kay Nieselt, de la University of Tübingen (Alemania).

Para generar los conjuntos de datos, descargamos los datos en bruto obtenidos en estudios anteriores del Archivo Europeo de Nucleótidos (ENA) y del Centro Nacional de Información Biotecnológica (NCBI). Sin embargo, algunos genomas sólo disponían de su secuencia de ensamblaje y sus datos brutos. Así, en los Capítulos 1 y 4, se simularon lecturas similares a las de la NGS basadas en los ensamblajes de los genomas utilizando la herramienta Genome2Reads (integrada en el pipeline de EAGER). En los Capítulos 2 y 3 utilizamos un enfoque diferente y, en lugar de simular lecturas, utilizamos el ensamblaje de los genomas completos.

Los datos brutos de cada nuevo genoma de *T. pallidum* obtenido y los descargados de estudios anteriores fueron procesados y mapeados contra una referencia genómica, utilizando diferentes estrategias de mapeo. En el Capítulo 1 se generaron 3 datasets de 75 genomas cada uno, en el que para su obtención se mapearon las lecturas usando tres referencias genómicas diferentes (CDC2, Nichols y SS14). En el Capítulo 2 y 3, se generó un único dataset conformado por 77 y 99 genomas, respectivamente, de los cuales cada uno se mapeó frente a su referencia genómica más próxima (Nichols, CDC2, SS14 y Bejel). En el Capítulo 4, se generó un dataset de 121 genomas completos de *T. pallidum* para el diseño del nuevo esquema MLST. A continuación, cada secuencia genómica obtenida se fusionó en un único archivo de genomas múltiples para cada conjunto de datos, que se alineó utilizando Mafft excepto en el caso de los conjuntos de datos de los Capítulos 2 y 3. En esos dos Capítulos, el archivo de genomas múltiples completos tuvo que ser corregido manualmente con Aliview después de emplear Mafft, debido a que algunas secuencias tienen una gran cantidad de datos faltantes (especialmente los genomas antiguos).

### **3. Detección de recombinación:**

Se desarrolló un nuevo método denominado PIM basado en la incongruencia filogenética, para detectar la recombinación en los genomas completos de *T. pallidum*. El método constaba de varios pasos: A) Se construyó un árbol de máxima

verosimilitud (ML) para el alineamiento múltiple de genomas. B) Se calculó el número de polimorfismos de nucleótido único (SNPs) para cada gen, excluyendo los genes con menos de tres SNPs. C) La señal filogenética en el alineamiento de cada gen se evaluó mediante el mapeo de verosimilitud. D) Se generaron árboles de máxima verosimilitud para los genes restantes y se comprobó la congruencia filogenética entre los árboles utilizando diferentes métodos (ELW y SH). E). Los genes que mostraban incongruencia recíproca se examinaron para identificar posibles eventos de recombinación. Asimismo, los eventos de recombinación se analizaron posteriormente para determinar los linajes/cepas donantes y receptores. Este método se aplicó a diferentes conjuntos de datos compuestos por genomas de *T. pallidum* en los Capítulos 1, 2 y 3 de la presente tesis.

Asimismo, en el Capítulo 1 también se analizó la recombinación en *T. pallidum* mediante el uso de dos herramientas ampliamente utilizadas en la literatura llamadas Gubbins y CLonalFrameML para comparar sus resultados obtenidos con los de PIM.

#### **4. Reconstrucción filogenética:**

En los Capítulos 1, 2 y 3, antes de los análisis filogenéticos finales, los genes identificados como recombinantes por PIM fueron excluidos de la alineación múltiple de genomas. Además, también se excluyó el conocido gen hipervariable *tp0897*. Sin embargo, en los Capítulos 2 y 3, para mantener una evaluación filogenética basada estrictamente en la herencia vertical, se eliminaron dos genes adicionales (*tp0316* y *tp0317*). Todos los árboles filogenéticos se construyeron utilizando el modelo evolutivo GTR+G+I.

#### **5. Mapeo de los SNPs que definen los nodos TPE y TEN en el árbol completo basado en el genoma**

Para analizar los cambios genéticos entre los nodos ancestrales de las subespecies TPE y TEN en la filogenia del Capítulo 2, se realizó una reconstrucción ancestral del alineamiento genómico múltiple y el árbol filogenético de referencia con



TreeTime. Utilizamos IcyTree para visualizar el árbol anotado con los cambios obtenidos en la reconstrucción de la secuencia ancestral. Extrajimos de este archivo la anotación de polimorfismos no compartidos entre los nodos ancestrales de las subespecies TPE y TEN, analizando en detalle los genes eliminados del alineamiento genómico múltiple antes de la reconstrucción filogenética, por ser recombinantes, hipervariables (*tp0897*), o considerados en conversión génica (*tp0316* y *tp0317*). Para inferir estados ancestrales, es necesario calcular las probabilidades condicionales (posteriores) dados los datos. La probabilidad de cada uno de estos polimorfismos se calculó utilizando RAxML.

## **6. Datación por reloj molecular**

Todos los análisis de datación molecular se realizaron con BEAST. En el Capítulo 2, los análisis de datación del reloj molecular fueron realizados por la Dra. Martyna Molak, afiliada al Centro de Nuevas Tecnologías de la Universidad de Varsovia y al Museo e Instituto de Zoología de la Academia Polaca de Ciencias, ambos situados en Polonia, mientras que fueron realizados por el Dr. Louis du Plessis, del Departamento de Ciencia e Ingeniería de Biosistemas de la ETH y del Instituto Suizo de Bioinformática (Suiza), en el Capítulo 3.

## **11. Diseño de un nuevo esquema MLST para *T. pallidum***

Con el fin de desarrollar un nuevo sistema de tipificación para *T. pallidum*, se identificaron loci candidatos utilizando información de 121 genomas completos de esta bacteria. La información filogenética se evaluó para cada gen codificador de proteínas en el genoma de referencia de Nichols mediante una prueba de mapeo de verosimilitud. Los genes que mostraban una señal filogenética se retuvieron para análisis posteriores. Basándose en la variación y el poder de discriminación, se seleccionaron los mejores genes para el diseño de cebadores. Los cebadores se diseñaron con criterios específicos, como estar en regiones conservadas, corresponder a la segunda posición de un codón, tener una longitud de unos 20 pb, generar un tamaño de amplicón de aproximadamente 650 pb, tener entre un 45% y

un 60% de contenido de GC, no ser complementarios entre sí y tener un valor mínimo de dG de -6 kcal/mol. Se comprobó el número de haplotipos que podían diferenciarse utilizando diferentes combinaciones de subconjuntos de cebadores, y se seleccionó el conjunto de cebadores con la mayor resolución para un esquema MLST.

El rendimiento de los siete pares de cebadores se comprobó mediante amplificación por PCR utilizando una dilución en serie de la muestra de Nichols y un control negativo. Se recogieron 183 muestras de *T. pallidum* para su tipificación con el nuevo esquema MLST. La mayoría de las muestras eran muestras clínicas obtenidas en colaboración con diversos grupos de investigación, mientras que cuatro muestras históricas se obtuvieron mediante cultivo en conejos. La extracción del ADN de las muestras fue realizada por los respectivos grupos de investigación, excepto las muestras proporcionadas por el Hospital General de València, que fueron extraídas por el grupo de investigación de los autores utilizando Nuclisens® easyMAG®.

Los siete loci candidatos y el gen 23S rRNA se amplificaron mediante PCR primaria utilizando un protocolo touchdown. Los productos de la PCR se purificaron y secuenciaron por Sanger. Posteriormente, se realizaron análisis de las secuencias obtenidas para obtener información sobre alelos y tipos de secuencia (ST). En el caso del gen 23S rRNA, se comprobaron posiciones nucleotídicas específicas en busca de mutaciones asociadas a la resistencia a macrólidos.

### **13. Aplicación in silico del nuevo esquema MLST**

Se recopilaron todos los archivos de ensamblaje de *T. pallidum* de las bases de datos públicas y se obtuvo un total de 238 genomas de este patógeno. Estos genomas se fusionaron en un único archivo fasta y se alinearon para crear una alineación de genomas completos. De este alineamiento se extrajeron las regiones flanqueantes de los cebadores diseñados para cada locus en el nuevo esquema MLST. Se utilizó CD-HIT para identificar secuencias similares y generar clusters,

y los resultados se verificaron manualmente para garantizar su exactitud. Se compararon todas las secuencias génicas obtenidas, y mediante la identificación de SNPs o indels, se pudo asignar alelos diferentes en cada locus. La combinación de alelos en múltiples loci en un orden específico determinó el tipo de secuencia (ST) de cada muestra.

## **6. Diversidad genética**

En el Capítulo 4, comparamos la diversidad genética (D) entre las diferentes subespecies/sublineajes de *T. pallidum* utilizando la siguiente fórmula sobre diversidad de genotipos:  $F = 1 - \sum p_i^2$  (Nei 1977) donde p es la frecuencia de cada ST para cada subespecie/sublineaje de *T. pallidum*.

Adicionalmente, para este Capítulo, con el fin de observar si existía relación entre el origen geográfico de las muestras y los diferentes STs obtenidos, se examinaron las diferencias genéticas entre poblaciones utilizando DnaSP, considerando como población los países y continentes de origen de las muestras. Debido al reducido tamaño muestral disponible para las subespecies TPE y TEN, realizamos estos análisis únicamente para TPA. Para ello, agrupamos las muestras de TPA y las diferentes poblaciones conforme al origen geográfico de cada muestra. En concreto, dividimos las muestras de TPA, según su país de origen (nueve países diferentes), para analizar la diversidad genética entre países. Además, clasificamos las muestras de TPA según su continente de origen (cuatro continentes diferentes), para analizar la diversidad genética a nivel continental. Sin embargo, excluimos Oceanía como continente para analizar su divergencia genética, porque sólo tenía dos muestras. A continuación, tanto a nivel continental como nacional, estimamos el número medio de diferencias de nucleótidos (k) dentro de las poblaciones y entre ellas. Además, también estimamos la diversidad nucleotídica ( $\pi$ ) para examinar el grado de polimorfismo dentro de las poblaciones, y entre poblaciones utilizando el modelo Jukes-Cantor (Dxy(JC) de sustitución nucleotídica.

## Capítulo 1

Se analizaron 75 genomas de *T. pallidum* para investigar la recombinación y la selección en esta bacteria. Se utilizaron tres referencias genómicas diferentes para el mapeo y se emplearon tres métodos de detección de recombinación: PIM (un nuevo método desarrollado en la presente tesis), Gubbins y ClonalFrameML. Utilizando el método PIM, se identificaron 12 genes recombinantes que contenían 19 regiones recombinantes y estaban involucrados en al menos 21 eventos de recombinación diferentes. Los resultados obtenidos con PIM fueron más conservadores en la identificación de eventos y regiones de recombinación en comparación con los otros dos métodos utilizados. Sorprendentemente, se observó solo un evento de recombinación entre cepas de la misma subespecie de *T. pallidum* (TPA). La mayoría de los eventos de recombinación detectados correspondieron a transferencias entre subespecies (TPE/TEN a TPA). La razón detrás de este patrón no está clara y podría deberse a la dificultad para detectar transferencias intraespecíficas debido a los bajos niveles de variación genética en esta bacteria, o a mecanismos que favorecen la transferencia e incorporación de material genético de subespecies diferentes.

Se encontró una estrecha relación entre la recombinación y la selección en *T. pallidum*. Todos los genes recombinantes identificados mostraron señales fuertes de selección positiva o purificadora. La mayoría de estos genes codificaban proteínas ubicadas en la interfaz entre el patógeno y el huésped, lo que sugiere que las presiones selectivas de los huéspedes humanos impulsan la diversidad de las proteínas de la membrana externa de *T. pallidum*. Estos hallazgos son importantes para la selección de candidatos a vacunas y el diseño de una vacuna efectiva contra la sífilis.

Aunque este estudio proporciona nuevas perspectivas sobre la evolución de *T. pallidum*, todavía quedan preguntas por responder, como el mecanismo molecular exacto de la recombinación en esta bacteria. Sin embargo, la disponibilidad de

secuencias genómicas completas y los avances en el cultivo y la manipulación genética de *T. pallidum* pueden ayudar a abordar estas preguntas en el futuro.

## Capítulo 2

En este estudio, se secuenció y analizó un nuevo genoma antiguo de sífilis, denominado W86, con una alta cobertura (35X). Los restos arqueológicos de los que se obtuvo el genoma proceden de Polonia y datan del siglo XVII. Este nuevo genoma antiguo se contextualizó con un dataset genómico constituido por otros 76 genomas de *T. pallidum*. Se utilizó un enfoque de mapeo basado en cuatro genomas de referencia para asignar cada genoma reconstruido a su referencia más cercana. Se seleccionó un genoma de referencia de cada subespecie de *T. pallidum* (TPE y TEN) y de cada uno de los linajes principales de TPA (Nichols y SS14). Los resultados obtenidos demostraron la importancia de seleccionar la referencia más cercana, sobre todo en genomas antiguos, ya que la ubicación filogenética del genoma antiguo W86 variaba según el genoma de referencia utilizado.

Utilizando el método PIM, se identificaron 26 regiones recombinantes en 18 genes diferentes, incluyendo cinco nuevos genes recombinantes. Todos los eventos de recombinación detectados correspondieron a transferencias entre subespecies (TPE/TEN a TPA), excepto un evento de recombinación entre los linajes Nichols y SS14 de TPA. Según las cepas implicadas en cada evento de recombinación, se pudo determinar que posiblemente la recombinación entre las cepas TPE/TEN y TPA ocurrió en el Viejo Mundo, donde estas subespecies coexistieron y circulaban conjuntamente, ya que se encontró evidencia de que los genomas antiguos de los linajes TPE y TPA estuvieron involucrados en los eventos de recombinación detectados.

La datación molecular estableció que el último ancestro común de todas las cepas de *T. pallidum* existió alrededor del 2015 a.C., con un rango de incertidumbre (HPD) que va desde el 7100 a.C. hasta el 550 d.C. El tiempo estimado del ancestro

común para las cepas de TPA fue entre el 190 a.C. y el 1290 d.C., lo que sugiere que la aparición de esta subespecie fue anterior al brote inicial de sífilis en Europa en el siglo XV. Además, las estimaciones de tiempo de divergencia para TPE (entre 490 - 1305 d.C.) y TEN (entre 1650 - 1930 d.C.) indican una historia común de estas enfermedades en el Viejo Mundo, posiblemente relacionada con la propagación temprana de la sífilis venérea. Se encontraron diez eventos de recombinación antiguos entre los tres linajes en Europa antes del primer brote de sífilis conocido. Estos hallazgos sugieren una coexistencia compleja y prolongada de la sífilis y el pian con hábitats geográficos superpuestos en la Europa histórica.

También se descubrió que la mayoría de los genes detectados como recombinantes estaban bajo presión de selección positiva y desempeñan un papel importante en la virulencia, la evasión del sistema inmunitario del huésped o la defensa contra huésped. Sin embargo, las razones detrás de estas observaciones y las variantes génicas responsables de la transmisión y la virulencia de la bacteria aún no están claras.

Por otro lado, el nuevo enfoque de mapeo utilizado reveló un cambio drástico en la topología arbórea de las treponemosis endémicas una vez eliminados los loci recombinantes y altamente variables. Aunque las subespecies causantes de yaws y bejel son clados bien diferenciados, en la filogenia estrictamente vertical, las cepas TEN se agrupan dentro del clado TPE, lo que indica que ciertos genes recombinantes desempeñaron un papel importante en la diferenciación de estas subespecies. Además, los resultados obtenidos también sugieren que las treponemosis endémicas y la sífilis coexistieron en el Viejo Mundo antes de los proyectos modernos de erradicación. Las coinfecciones podrían haber proporcionado oportunidades para el intercambio genético entre las cepas. Además, existe la posibilidad de que el continente africano haya sido la fuente de las variantes treponémicas europeas, considerando las conexiones históricas y prehistóricas entre estas regiones.

### Capítulo 3

Este estudio presenta las primeras pruebas de ADN antiguo de una treponematosi precolombina en el Nuevo Mundo, concretamente en Brasil. Las hipótesis anteriores sobre el origen de la sífilis se basaban en pruebas paleopatológicas, pero este estudio aporta pruebas genéticas inequívocas al reconstruir un genoma de *T. pallidum* de alta cobertura (35X) a partir de restos humanos indígenas de 2.000 años de antigüedad denominado ZH1540. También se obtuvieron otros 3 genomas antiguos de *T. pallidum* pero de baja cobertura, potencialmente clasificables como TEN. Además, el nuevo genoma ZH1540 fue contextualizado con un dataset conformado por otros 98 genomas completos de *T. pallidum*. Es interesante resaltar, que el nuevo genoma ZH1540 es notablemente similar a las otras cepas actuales de TEN. Este hallazgo inesperado arroja luz sobre la evolución de los patógenos treponémicos y permite comprender mejor la divergencia de subespecies dentro de la familia de los treponémicos.

Los resultados obtenidos destacan la limitada comprensión de la historia de las treponematosi endémicas, ya que las hipótesis anteriores rara vez tenían en cuenta las formas endémicas de la enfermedad. Las pruebas genéticas presentadas en este estudio desafían la hipótesis unitaria, que trata todas las treponematosi como una única enfermedad antigua y global, al mostrar claras distinciones genéticas entre las subespecies causantes de la sífilis, el pian y el bejel. Sin embargo, la presencia de una cepa treponémica similar a la TEN en el antiguo Brasil apoya algunos aspectos de la visión unitaria, sugiriendo la capacidad de los treponemas para adaptarse a diversos climas y ubicaciones geográficas. Además, el nuevo genoma ZH1540 aporta información valiosa sobre la diversidad genética del bejel y la distribución mundial de las infecciones treponémicas. Los resultados cuestionan la idea actual de la asociación del bejel con las regiones áridas y sugieren una distribución más amplia en el pasado, posiblemente asociada a diferentes hábitats ambientales.

Los análisis de recombinación realizados revelan la importancia de la transferencia horizontal de genes y la recombinación en la diversificación de las subespecies. Se identificaron 27 genes recombinantes (algunos de ellos no detectados en estudios previos) en cuyos eventos de recombinación hay genomas antiguos implicados. Además, la mayoría de transferencias se producen entre las diferentes subespecies, lo que sugiere el intercambio de material genético entre las cepas causantes del bejel y las primeras formas de cepas causantes de la sífilis. Estos hallazgos apoyan la coexistencia y el intercambio genético entre los distintos linajes treponémicos.

Utilizando genomas antiguos con dataciones por radiocarbono robustas, se calibró un método que permitió estimar las fechas de divergencia a nivel de subespecie. Los resultados obtenidos parecen indicar que todas las tres subespecies de *T. pallidum* ya habían divergido antes de los viajes de Colón. Específicamente, el ancestro común más reciente de este patógeno se sitúa entre el 12,006 y 545 a.C., la aparición del clado TEN entre el 780 a.C. y 449 d.C., TPE entre el 28 y 1299 d.C., y TPA entre el 42 a.C. y 1376 d.C. Estas fechas de divergencia son mucho más antiguas que las estimaciones anteriores basadas, la mayoría, en genomas modernos. La inclusión del genoma precolombino en el análisis amplía los intervalos creíbles y destaca la importancia de incluir muestras más antiguas en los análisis del reloj molecular. El descubrimiento de una treponematosi similar al bejel de casi dos mil años de antigüedad en Sudamérica cuestiona las hipótesis previas sobre la distribución y propagación de *T. pallidum*. Sugiere la posibilidad de que un ancestro común de todas las cepas modernas de *T. pallidum* se originó en África o Asia y fue transportado a América por los primeros pobladores humanos. Sin embargo, se necesitan más pruebas para determinar si el *T. pallidum* americano surgió de forma endémica o acompañó a la colonización humana.

A pesar de los valiosos datos aportados por el genoma antiguo, aún quedan muchas preguntas por responder. Aún se desconoce el origen de la sífilis y los acontecimientos que condujeron a su aparición. Se necesitan más genomas de *T.*



*pallidum* anteriores al contacto procedentes de distintos continentes para fijar el origen de la sífilis en el Nuevo Mundo.

## Capítulo 4

Se ha diseñado un nuevo esquema de tipificación molecular para *T. pallidum* basado en la secuenciación de siete genes variables (*tp0136*, *tp0326*, *tp0548*, *tp0705*, *tp0858*, *tp0865* y *tp1031*) y el gen 23S rRNA. Estos genes fueron seleccionados debido a su alta variabilidad genética y su capacidad para proporcionar información sobre la diversidad de cepas de esta bacteria.

Se llevaron a cabo experimentos de amplificación y secuenciación utilizando 542 muestras clínicas, tanto de forma experimental como *in silico*. A pesar de algunas dificultades técnicas y limitaciones de calidad del ADN de las muestras, se logró una tasa global moderada de tipificación de muestras utilizando este nuevo esquema MLST. Se identificaron 98 perfiles alélicos distintos en un subconjunto de 386 muestras, lo que indica un alto nivel de discriminación y diversidad genética en *T. pallidum*.

Además, se observó una alta proporción de cepas resistentes a macrólidos entre las muestras analizadas, sobretudo en el sublinaje de SS14 de TPA, lo que plantea preocupaciones significativas sobre el tratamiento de la treponematosi y destaca la necesidad de investigaciones epidemiológicas adicionales para monitorear y caracterizar la diseminación de la resistencia.

La filogenia generada a partir de los datos de tipificación molecular mostró una clara diferenciación entre las subespecies y los sublinajes de *T. pallidum*, y fue congruente con las relaciones filogenéticas obtenidas a partir de datos de genomas completos. Esto indica que el nuevo esquema MLST propuesto proporciona una resolución casi equivalente a la de los genomas completos, lo que lo convierte en una herramienta eficaz y práctica para la tipificación de *T. pallidum* en laboratorios con acceso a recursos de secuenciación Sanger.

Asimismo, en este estudio, se analizaron las variaciones genéticas dentro y entre poblaciones de *T. pallidum*, teniendo en cuenta el origen geográfico de las muestras. Se encontraron variaciones genéticas dentro y entre poblaciones, revelando distintas estructuras poblacionales. Esto sugiere divergencias y patrones de transmisión localizados en diferentes regiones geográficas. Además, se observó una agrupación geográfica de cepas con variantes genéticas más prevalentes en áreas específicas. África y Asia mostraron mayor diversidad genética, posiblemente debido a tasas de transmisión más altas o endemicidad prolongada. Madagascar presentó un perfil genético único y conexiones epidemiológicas con múltiples regiones. Estos hallazgos destacan la importancia de considerar factores regionales y la dinámica de transmisión al abordar el control de *T. pallidum*, aunque se necesitan investigaciones adicionales para comprender mejor su propagación y fundamentar medidas de control efectivas.

## **Discusión**

La disponibilidad de más de 1.400 genomas de *T. pallidum* ha proporcionado a los investigadores oportunidades sin precedentes para explorar la diversidad genética, la evolución y la dinámica de poblaciones. En esta tesis, se examinan genomas antiguos y contemporáneos de *T. pallidum* a través de cuatro estudios diferentes.

El Capítulo 1 se centra en analizar 75 genomas contemporáneos para evaluar la recombinación y la selección mediante un nuevo método denominado PIM. Esta nueva metodología desarrollada, superó a Gubbins y ClonalFrameML, otras dos herramientas ampliamente utilizadas para la detección de recombinación. Aunque en los estudios epidemiológicos se sigue utilizando sobre todo Gubbins (debido a su capacidad para manejar grandes conjuntos de datos), de acuerdo con los resultados obtenidos, parece que sobreestima la recombinación en *T. pallidum*. Por ello, como la precisión en la detección de la recombinación es crucial para las reconstrucciones filogenéticas, en la presente tesis doctoral se aplicó PIM en todos los análisis realizados para detectar recombinación. El número de genes

recombinantes detectados varió entre los Capítulos 1, 2 y 3, debido a las diferentes composiciones genómicas y a las limitaciones de los genomas disponibles. Además, la filogenia obtenida en el Capítulo 2 mostró una topología de árbol distinta a la de otros capítulos cuando se excluyeron los genes recombinantes del alineamiento de genomas completos. La detección de un número de genes recombinantes diferente y su eliminación junto a la composición de genomas utilizados, causaron una variación genómica diferente que pudo originar la incongruencia observada. No obstante, todos los clados de las filogenias obtenidas presentaban un sólido apoyo en el contexto del dataset genómico utilizado.

Asimismo, se descubrió que la recombinación y la selección positiva están estrechamente relacionadas en *T. pallidum*, y que los genes recombinantes y los genes seleccionados positivamente desempeñan papeles cruciales en la defensa de este patógeno y, potencialmente, en su virulencia. Además, tal y cómo sucede para otras bacterias, se ha demostrado que la recombinación es importante en la evolución de *T. pallidum*, incluidas las fases iniciales de especiación y adaptación a nuevos nichos ecológicos.

Un resultado muy interesante fue que la recombinación se produce principalmente entre distintas subespecies de *T. pallium* y no dentro de la misma subespecie. Las razones de este patrón siguen siendo inciertas, y se necesita más investigación para comprender los mecanismos subyacentes. Además, la ausencia de eventos de recombinación recientes en *T. pallidum* tiene importantes implicaciones para la inclusión de genes recombinantes en los esquemas MLST (Multi-Locus Sequence Typing). La ausencia de intercambios genéticos entre cepas significa que es más probable que las variaciones alélicas observadas en estos genes surjan de nuevas mutaciones. Esto hace que las variaciones alélicas identificadas mediante el análisis MLST sean una representación fiable de la diversidad genética y las relaciones evolutivas entre las cepas de *T. pallidum*. Esta característica hace que el MLST sea

muy adecuado para construir árboles filogenéticos y realizar estudios de genética de poblaciones.

El Capítulo 4 de la tesis presenta un nuevo esquema MLST para *T. pallidum*, en el que se seleccionaron siete genes (detectados previamente como genes recombinantes) en función de la alta resolución y rendimiento que aportan en conjunto. El árbol filogenético generado utilizando el concatenado de las secuencias de los siete genes del esquema MLST resultó congruente con otros árboles obtenidos en la tesis y con los de otros estudios anteriores, lo que confirma aún más la idoneidad de los genes elegidos.

El nuevo esquema de MLST proporciona información valiosa sobre la epidemiología de *T. pallidum*, al permitir a los investigadores seguir la dinámica de transmisión y propagación entre sus diferentes subespecies. Además, el esquema MLST tiene una aplicación adicional que no se mencionó explícitamente en el Capítulo 4: puede ayudar en la selección de genomas apropiados para la secuenciación, facilitando los estudios evolutivos. Utilizando el esquema MLST, los investigadores pueden identificar y clasificar con precisión las cepas de *T. pallidum* en sus respectivas subespecies, asegurándose de que los genomas seleccionados representan una gama diversa de cepas de cada subespecie.

Asimismo, se han obtenido con éxito dos genomas completos antiguos de alta cobertura de *T. pallidum*. El primer genoma, conocido como W86, procede de Polonia y data del siglo XVII, perteneciendo a la subespecie TPA. El segundo genoma, identificado como ZH1540, procedía de restos humanos de Brasil y se remonta aproximadamente 2.000 años atrás, representando el primer genoma disponible de *T. pallidum* precolombino.

Al integrar estos genomas antiguos en diversos conjuntos de datos de genomas de *T. pallidum*, se pudieron descubrir diferentes genes recombinantes que no habían sido previamente identificados. Este descubrimiento mostró la posible ocurrencia

de eventos de recombinación entre cepas TPE/TEN y TPA en el Viejo Mundo, lo que sugiere la coexistencia y circulación de estas subespecies en la misma región. Además, la inclusión de estos genomas antiguos en la datación bayesiana de reloj molecular amplió los rangos de credibilidad y mejoró la exactitud y precisión de las estimaciones de la cronología evolutiva; obteniendo unas fechas de divergencia mucho más antiguas que las obtenidas hasta ahora. Esto subraya la importancia de incorporar muestras antiguas a los análisis de relojes moleculares.

No obstante, los métodos de datación utilizados para las dos muestras óseas de las que se obtuvieron los genomas antiguos difieren. La muestra ZH1540 fue datada mediante radiocarbono, mientras que la muestra W86 se basó en el contexto arqueológico. Esto se debe a que fiabilidad del radiocarbono disminuye para los periodos más recientes. El radiocarbono se desarrolló para periodos geológicos más largos y se basa en valores isotópicos medios. Además, el Mínimo de Maunder, un periodo de baja actividad solar entre 1645 y 1715, afectó la datación por radiocarbono al disminuir la intensidad de los rayos cósmicos en la atmósfera terrestre. Esto puede hacer que los materiales orgánicos del periodo del Mínimo de Maunder parezcan más antiguos de lo que realmente son cuando se datan con radiocarbono. Los efectos específicos del Mínimo de Maunder en la datación por radiocarbono pueden variar según la ubicación geográfica, el tipo de muestra y los métodos de calibración utilizados. Por lo tanto, cuando existen pruebas sólidas de que una muestra proviene de un periodo posterior a 1500, rara vez se utiliza la datación por radiocarbono. En estos casos, es posible obtener una datación más precisa mediante la presencia de elementos arqueológicos contextuales de donde se obtiene la muestra.

Además de *T. pallidum*, otras especies bacterianas han proporcionado genomas antiguos, siendo *Y. pestis* la que cuenta con el mayor número de genomas antiguos disponibles, seguida de *M. leprae* y *M. tuberculosis*. Sin embargo, *T. pallidum* presenta un desafío único, ya que es la única bacteria entre estas especies que no

puede cultivarse utilizando un sistema estandarizado. Esto plantea dificultades no sólo para estudiar los genomas de las cepas modernas, sino también para las cepas antiguas. La mayoría de las funciones proteicas conocidas en *T. pallidum* son hipotéticas o se deducen mediante comparaciones de secuencias con proteínas de función conocida, lo que dificulta establecer cambios en la patogenicidad o virulencia utilizando genomas antiguos. Sin embargo, el análisis de genomas antiguos de *T. pallidum* no ha indicado cambios importantes en su genoma, y los genes de virulencia han persistido durante siglos a pesar de las variaciones en la prevalencia en diferentes poblaciones y periodos de tiempo, de forma similar a lo observado en el caso de *M. leprae*.

En Valiente-Mullor *et al.* (2021), se exploró el impacto de utilizar un único genoma de referencia en el análisis genómico microbiano. El estudio concluye que el uso de un único genoma de referencia para mapear las lecturas de secuenciación puede introducir errores en los análisis posteriores, como en la detección de SNPs y la inferencia filogenética, especialmente cuando se trata de aislados genéticamente diversos.

En el Capítulo 1, se comparan tres referencias genómicas diferentes de *T. pallidum* y su efecto en el mapeo y el análisis de recombinación. Los resultados muestran que la elección del genoma de referencia no afectó significativamente a los resultados o conclusiones obtenidas, lo que contrasta con los resultados obtenidos por Valiente-Mullor *et al.* (2021). Esta diferencia se atribuye a la clonalidad de *T. pallidum*, donde las subespecies tienen una baja tasa de divergencia. Sin embargo, el número de genomas de *T. pallidum* disponibles para cada subespecie fue desigual o limitado, lo que puede sesgar las conclusiones obtenidas.

En el Capítulo 2 se presenta un enfoque metodológico novedoso para abordar el sesgo de referencia y garantizar la robustez de los resultados sin necesidad de realizar análisis repetitivos. Este enfoque consiste en seleccionar un genoma de referencia de cada subespecie y linaje principal de *T. pallidum*, y luego mapear las

lecturas de cada cepa contra el genoma de referencia más cercano. Además, se realizó un análisis adicional para evaluar la consistencia de los resultados, mapeando las lecturas del genoma antiguo (W86) utilizando los cuatro posibles genomas de referencia de *T. pallidum*. Los resultados demostraron que el nuevo genoma fue clasificado como perteneciente a la misma subespecie de *T. pallidum*, independientemente de la referencia genómica utilizada. Sin embargo, la posición de la cepa en el árbol filogenético obtenido varió según la referencia genómica utilizada. Los resultados de estos análisis demostraron claramente la importancia de seleccionar la referencia más cercana al analizar genomas antiguos.

El Capítulo 3 sigue un enfoque metodológico similar al del Capítulo 2, mientras que el Capítulo 4 utiliza exclusivamente la cepa de referencia Nichols para el diseño de un nuevo esquema MLST (Multi-Locus Sequence Typing). A pesar de los resultados obtenidos, es necesario seguir investigando para ampliar los resultados y cuantificar el verdadero impacto de la elección de la referencia genómica en la reconstrucción del genoma de *T. pallidum* y en los análisis evolutivos.

En conclusión, esta tesis doctoral contribuye significativamente a nuestra comprensión de la evolución, genómica y epidemiología de *T. pallidum*. La inclusión de genomas antiguos, el desarrollo de un método de mapeo innovador y la creación de un nuevo esquema MLST hacen avanzar la investigación sobre esta bacteria. Además, los hallazgos y recursos aportados sientan las bases para futuras investigaciones.

## Conclusiones

1. En este estudio, se desarrolló un nuevo método llamado PIM para detectar con precisión la recombinación en los genomas de *T. pallidum*, superando a las herramientas comúnmente utilizadas en términos de precisión. Además, se investigó el papel de la selección natural en este patógeno clonal, revelando que la recombinación y la selección son factores fundamentales en la evolución de *T. pallidum*, contribuyendo a la diversidad genética en las interacciones entre el huésped y el patógeno, y facilitando su evolución adaptativa.
2. Es interesante destacar que los genes recombinantes identificados mostraron señales fuertes de selección positiva o purificadora, lo que enfatiza su importancia funcional en las interacciones con el huésped, la virulencia y la evasión inmune. Sin embargo, aún quedan preguntas sin responder sobre los procesos moleculares que subyacen a la recombinación, así como sobre la frecuencia y los sitios específicos de las coinfecciones, que son elementos esenciales para que ocurra la recombinación.
3. La ausencia de eventos recientes de recombinación en *T. pallidum* tiene implicaciones en el diseño y el uso de los esquemas MLST. Los genes involucrados en eventos de recombinación pueden ser potencialmente utilizados en estos esquemas, ya que la mayoría de los alelos se generan a través de nuevas mutaciones.
4. En Capítulo 1, se compararon tres referencias genómicas diferentes y se encontró que la elección de la referencia no tuvo un impacto significativo en los resultados. Sin embargo, la disponibilidad limitada de genomas, especialmente para ciertas subespecies, dificultó un análisis más completo.



5. En el Capítulo 2, se desarrolló un enfoque de mapeo novedoso para mejorar la cobertura del genoma, reducir el sesgo de referencia y aumentar la precisión de la inferencia filogenética. Este enfoque eliminó la necesidad de comparaciones con múltiples genomas de referencia, simplificando los análisis posteriores. La aplicación de este nuevo enfoque reveló que la elección de la referencia afectó la ubicación filogenética de los genomas antiguos, pero no afectó a la clasificación de las cepas dentro de las subespecies.
6. Se utilizaron 121 genomas de *T. pallidum* para desarrollar un nuevo esquema MLST que incorpora siete genes variables y los genes 23S rRNA. Este esquema mejora significativamente la capacidad de distinguir entre diferentes cepas y puede aplicarse de manera efectiva a todas las subespecies de *T. pallidum*. Aunque la eficiencia de amplificación del esquema MLST fue moderada, se realizaron mejoras para reducir el tiempo y el costo. No obstante, se recomienda realizar pruebas adicionales con muestras de mayor calidad.
7. Destacablemente, nuestro esquema de tipificación identificó exitosamente la diversidad genética de *T. pallidum*. También demostró la presencia de resistencia a macrólidos en todas las subespecies y sublinajes de esta bacteria, con una prevalencia particularmente alta en el sublinaje SS14.