Title

Genomic epidemiology of syphilis reveals independent emergence of macrolide resistance across multiple circulating lineages

Authors, Affiliations

Mathew A. Beale[1], Michael Marks[2,3], Sharon K. Sahi[4], Lauren C. Tantalo[4], Achyuta V. Nori[5], Patrick French[6], Sheila A. Lukehart[7], Christina M. Marra[4], Nicholas R. Thomson[1,8]

Corresponding authors: MAB, mathew.beale@sanger.ac.uk; NRT, nrt@sanger.ac.uk

[1] Parasites and Microbes, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, UK

[2] Clinical Research Department, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, UK

[3] Hospital for Tropical Diseases, London, UK

[4] Department of Neurology, University of Washington, USA

[5] Guy's & St Thomas' NHS Foundation Trust, London, UK

[6] The Mortimer Market Centre CNWL, Camden Provider Services, London, UK

[7] Departments of Medicine and Global Health, University of Washington, USA

[8] Department of Pathogen Molecular Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, UK

Abstract

Syphilis is an ancient sexually transmitted infection caused by the bacterium *Treponema pallidum* subspecies *pallidum* and may lead to severe clinical complications. Recent years have seen striking increases in syphilis diagnoses in many high income countries, with the UK reporting a 148% increase in new diagnoses over 10 years. The reasons for this rise are complex and multifactorial, including changing cultural, behavioural, and technological factors that influence sexual networks and transmission dynamics. Previous genomic analyses have suggested that one lineage of syphilis, called SS14, may have expanded recently, with most syphilis caused by this lineage, and that this expansion indicates emergence of a single pandemic azithromycin-resistant cluster. In this study, we used high throughput sequencing of *Treponema pallidum* performed on DNA extracted directly from clinical swab samples and clinically derived samples with minimal passage in the rabbit to more than double the number of publicly available whole genome sequences. We used phylogenomic and population genomic analyses to show that both SS14-lineage and Nichols-lineage *T. pallidum* are present in contemporary patients and that SS14 is a polyphyletic lineage. We further correlate the appearance of genotypic macrolide resistance with multiple SS14 sub-lineages, showing that both genotypically macrolide resistant and macrolide sensitive sub-lineages are spreading contemporaneously. These findings demonstrate that macrolide resistance has independently evolved multiple times in *T. pallidum*, that once evolved it becomes fixed in the genome and is transmissible, and that

these findings are not consistent with the hypothesis of SS14-lineage expansion purely due to macrolide resistance. Beyond relevance to our understanding of the current syphilis epidemic, these findings show how macrolide resistance evolves in *Treponema* subspecies. Furthermore, the evolution of macrolide resistance, despite not being first-line treatment, provides a warning on broader issues of antimicrobial resistance, and highlights the importance of stewardship and strategic planning to prevent the emergence of antimicrobial resistance.

Introduction

Syphilis is an ancient, predominantly sexually transmitted infection (STI) caused by the bacterium *Treponema pallidum* subspecies *pallidum* (TPA). If untreated, syphilis causes a multi-system disease that can progress to severe cardiovascular and neurological involvement, which can be potentially fatal. Syphilis caused a pandemic wave that swept across Renaissance Europe over 500 years ago, and remained a problem until the introduction of antibiotics in the post-World War II era[1]. Despite effective treatment with benzathine benzylpenicillin G (BPG), syphilis transmission levels fluctuated but persisted throughout the 20th century, until the AIDS crisis of the 1980s and 1990s, where changes in sexual behaviour (and possibly AIDS-related mortality), led to an overall decline in incidence in many western countries and populations[2,3].

Recent years have seen a sharp increase in syphilis cases in many high-income countries, predominantly within sexual networks of men who have sex with men (MSM)[4,5]. In the United Kingdom there was a 20% increase in reported new diagnoses between 2016 and

2017, and a 148% increase since 2008[6]. Similar trends have been reported in other countries[4,7]. The reasons for this increase are complex and multifactorial, incorporating changing behavioural patterns mediated by cultural, societal and technological changes in our modern world[8], resulting in a perfect epidemiological storm. It is also possible that there are bacterial changes either driving the current rise in syphilis incidence, or occurring as a consequence of this increase. However, current knowledge of TPA is limited, largely because the bacterium was, until recently, intransigent to *in vitro* culture[9]. Most current understanding of TPA biology therefore comes from related species or from TPA cultured in the *in vivo* rabbit testicular model[10]. Genomic analysis has also been limited due to low levels of TPA pathogen load in patients and difficulty in readily isolating new strains. Sequencing must be performed directly on clinical specimens or after passage through rabbits, leading to substantial bottlenecks in genomic data generation. Recent advances have enabled target enrichment of pathogen reads directly from clinical or cultured specimens[11,12], and this was recently employed separately by different groups, including our own, to sequence TPA and other *T. pallidum* subspecies directly from patient samples[13–15].

The availability of increasing numbers of genomic sequences enabled the first description of the global TPA population structure using 31 near genome-length TPA sequences, along with a small number derived from closely related species[14]. The authors described two lineages within TPA; a Nichols-lineage found almost exclusively in North American sequences exhibiting substantial nucleotide diversity, and a geographically widespread but genetically homogeneous SS14-lineage, confirming previous analyses using multi-locus sequence typing[16]. Of these two lineages, they found that 68% of tested TPA genomes belonged to the SS14-lineage, and further analysis using a larger dataset of 1354 single-locus molecular types (comprising 623 samples from South East Asia, 241 from the USA, 392 from Europe and a small number of other locations) also supported this view (94% SS14-lineage).

Although penicillin resistance has never been reported in syphilis, increasing levels of genotypic resistance, and clinical treatment failure, to macrolides such as azithromycin have been reported[17,18], conferred by either one of two single nucleotide polymorphisms (SNPs) in the 23S ribosomal sequence (A2058G and A2059G). Arora *et al* reported that 90% of sequenced SS14-lineage genomes and 25% of Nichols-lineage genomes contained SNPs conferring macrolide resistance; furthermore, they suggested that SS14-lineage may represent a single pandemic azithromycin-resistant cluster[14].

In this study, we performed direct whole genome sequencing on 73 TPA samples from the US and Europe, and combined these data with 49 publicly available genomes. We used phylogenetic analysis to delineate sub-lineages within the both the SS14- and Nichols-lineages, showing striking patterns of the emergence and fixation of macrolide resistance SNPs that indicate independent evolution and proliferation of resistance alleles. These findings have implications for our understanding of the increasing incidence of syphilis and on the potential of the WHO Yaws eradication campaign to drive further development of macrolide resistance in both TPA and in the closely related *Treponema pallidum* subspecies *pertenue* (TPP)[19,20].

Results

We sequenced eight genomes directly from clinical swabs collected in 2016 from patients in the United Kingdom and 60 isolate genomes from low rabbit passage samples (no more than two passages from the original patient sample; henceforth referred to as 'recently clinically derived') originally collected from patients between 2001-2011 in the USA. We also resequenced three clonally derived laboratory strains from the USA that have been previously sequenced (Nichols Houston E, Nichols Houston J, Nichols Houston O) but remain unpublished, and two strains for which the sequencing reads were not publicly available (Chicago[21], Seattle 81-4[22]). We combined our data with 49 high-quality genomes published previously[13,14,23–28], 41 of which were recently derived from clinical patients, yielding a dataset of 122 genomes (109 with limited passage from clinical patients). Combined, our sample set included 72 genomes from the USA (predominantly Seattle), 8 from the UK (exclusively London), 9 from China (predominantly Shanghai), 23 from Portugal (exclusively Lisbon), and a small number from other countries, all collected between 1912-2016 (Supplementary Table 1).

After removal of recombinant and repetitive sites (both by selective mapping and screening – see Methods and Supplementary Table 2), we performed phylogenomic analyses, using maximum likelihood and Bayesian methods to define lineages. In agreement with previous studies[14], we show the presence of two dominant lineages in our dataset (previously denoted SS14 and Nichols; Figure 1) that are separated by >70 non-recombining single nucleotide polymorphisms (SNPs). Of the 122 total samples included in this study, 105 (86%) belonged to the SS14-lineage, whilst of the 109 clinical samples included, 103 (94%) were from SS14-lineage. In contrast, only six Nichols-lineage samples were recently clinically derived, and most (11/17) Nichols-lineage genomes examined were historically passaged isolates, including those derived from the original Nichols strain isolated in 1912 and disseminated to different North American laboratories; some Nichols-lineage genomes represent clones of the parent strain derived *in vivo*. However, although we observed a strong bias towards clinically derived SS14-lineage samples in this dataset, not all recent clinical strains were of the SS14-lineage; six recent clinical samples belonged to the Nichols-lineage, three (of eight sequenced) clinical samples from the UK in 2016, one collected in the Netherlands in 2013, and two from the USA in 2004, indicating that transmission of this lineage is ongoing and potentially more widespread than previously thought.

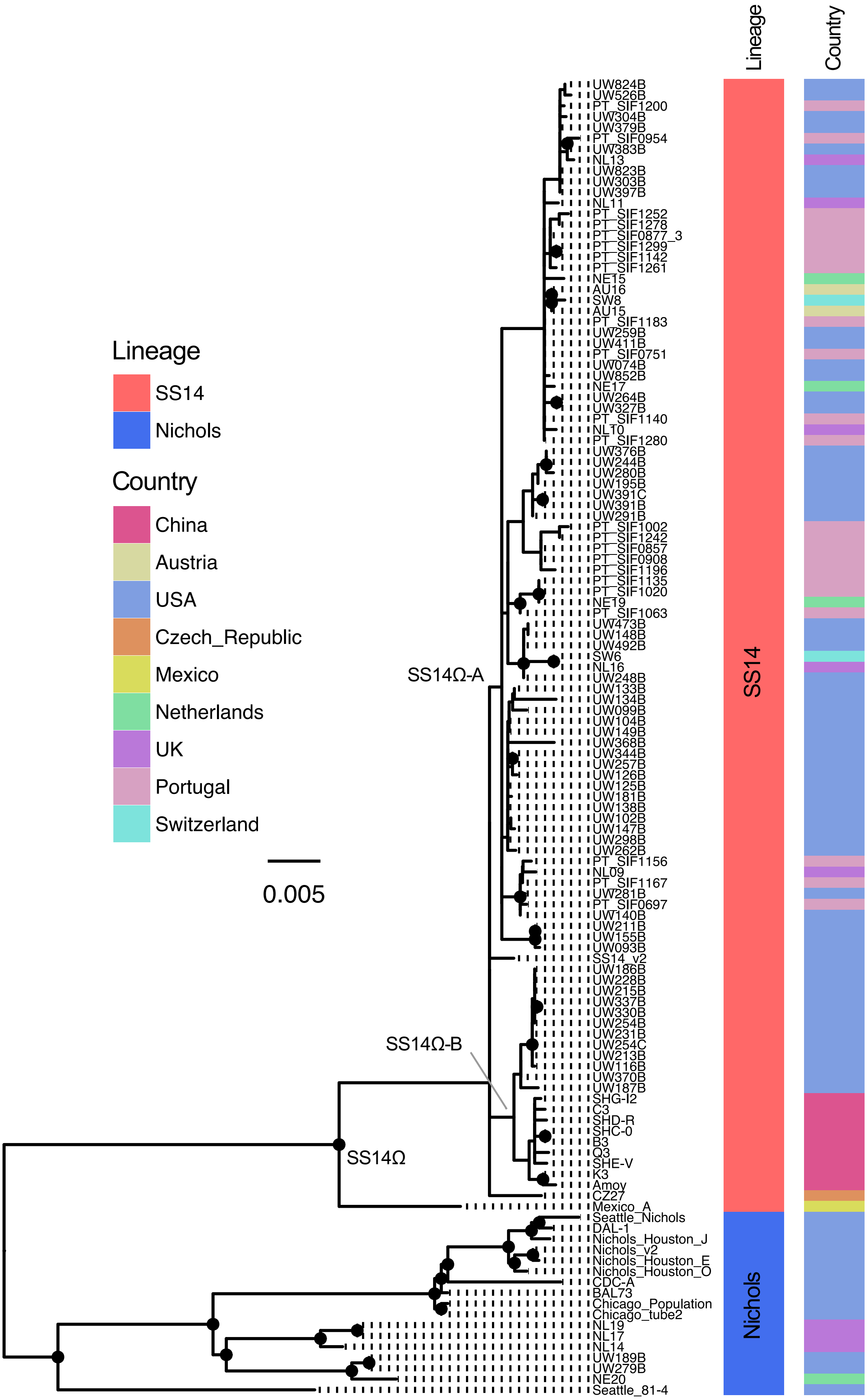

Lineage
Country
Lineage
SS14
Nichols
Country
China
Austria
USA
Czech_Republic
Mexico
Netherlands
UK
Portugal
Switzerland
0.005
SS14Ω-A
SS14Ω-B
SS14Ω
SS14
Nichols
UW824B
UW526B
PT_SIF1200
UW304B
UW379B
PT_SIF0954
UW383B
NL13
UW823B
UW303B
UW397B
NL11
PT_SIF1252
PT_SIF1278
PT_SIF0877_3
PT_SIF1299
PT_SIF1142
PT_SIF1261
NE15
AU16
SW8
AU15
PT_SIF1183
UW259B
UW411B
PT_SIF0751
UW074B
UW852B
NE17
UW264B
UW327B
PT_SIF1140
NL10
PT_SIF1280
UW376B
UW244B
UW280B
UW195B
UW391C
UW391B
UW291B
PT_SIF1002
PT_SIF1242
PT_SIF0857
PT_SIF0908
PT_SIF1196
PT_SIF1135
PT_SIF1020
NE19
PT_SIF1063
UW473B
UW148B
UW492B
SW6
NL16
UW248B
UW133B
UW134B
UW099B
UW104B
UW149B
UW368B
UW344B
UW257B
UW126B
UW125B
UW181B
UW138B
UW102B
UW147B
UW298B
UW262B
PT_SIF1156
NL09
PT_SIF1167
UW281B
PT_SIF0697
UW140B
UW211B
UW155B
UW093B
SS14_v2
UW186B
UW228B
UW215B
UW337B
UW330B
UW254B
UW231B
UW254C
UW213B
UW116B
UW370B
UW187B
SHG-I2
C3
SHD-R
SHC-0
B3
Q3
SHE-V
K3
Amoy
CZ27
Mexico_A
Seattle_Nichols
DAL-1
Nichols_Houston_J
Nichols_v2
Nichols_Houston_E
Nichols_Houston_O
CDC-A
BAL73
Chicago_Population
Chicago_tube2
NL19
NL17
NL14
UW189B
UW279B
NE20
Seattle_81-4

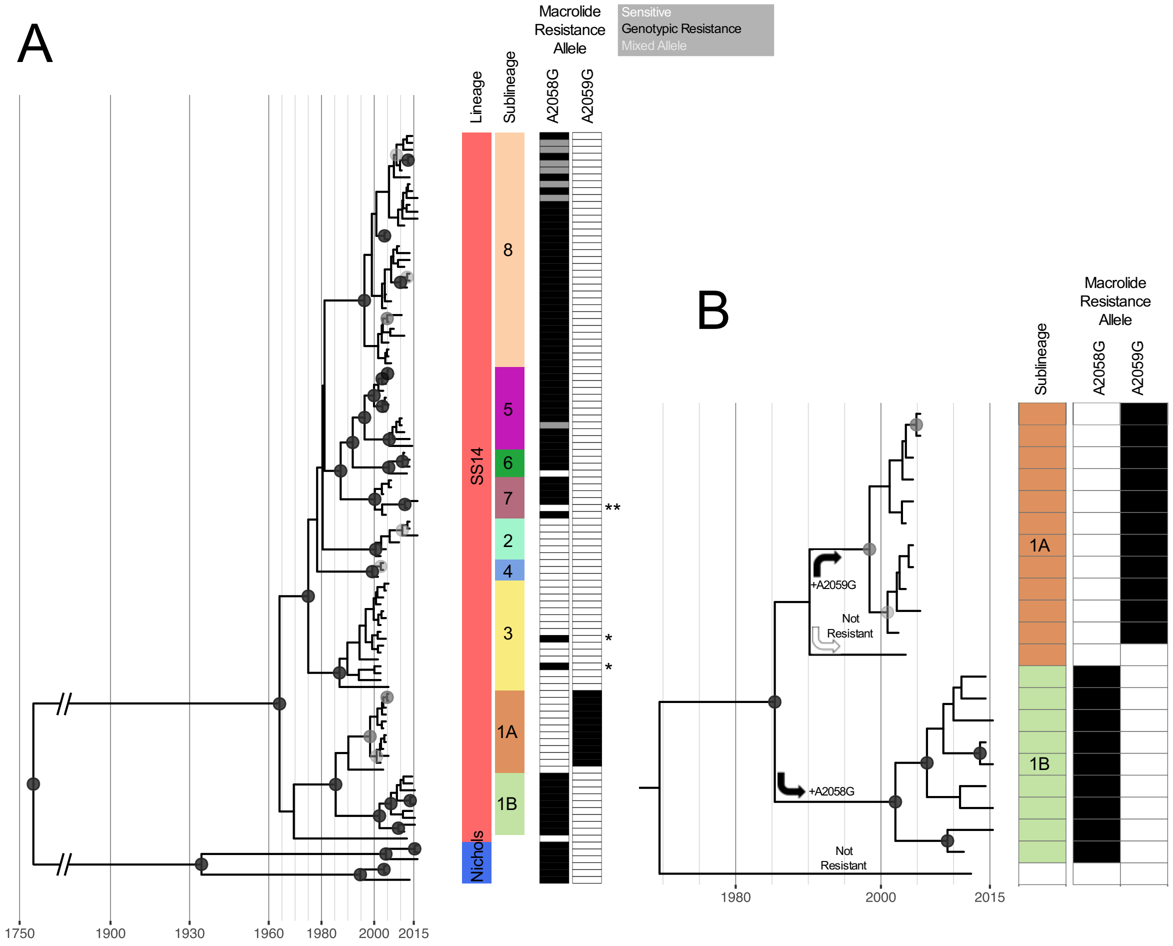
Figure 1. Maximum likelihood phylogeny of 122 high quality *T. pallidum* subspecies *pallidum* genomes (including clinical and non-clinically derived samples), showing lineage and country of origin. Ultra-Fast bootstrap values >=95% are labelled with black nodes points. Branches are scaled by mean nucleotide substitutions/site.

Bayesian phylogenetic reconstruction was used to date the time to most recent common ancestor (TMRCA) for the different TPA lineages. However, temporal analysis of heavily passaged laboratory strains (such as those derived from the original Nichols isolate) is problematic because the true mutational age may be unknown, meaning coalescent date is difficult to infer; we therefore removed extensively passaged strains or strains with no record of their passage history from this particular part of the analysis. This included the removal of the Nichols and SS14 reference strains, as well as the Mexico A strain that delineated the SS14Ω lineage described previously[14]. Root-to-tip regression analysis of the remaining 109 clinical genomes indicated that TPA possesses a clock-like signal (Supplementary Figure 1), and we performed Bayesian phylogenetic reconstruction and tip date analysis using BEAST[29] under a Strict Constant model, inferring a median molecular clock rate of 2.28 x10^{-7}, or 0.26 sites/genome/year (meaning we would expect TPA genomes on average to accumulate one SNP every four years by natural drift). We inferred a temporal timeline for the tree, and our analysis broadly supported previous estimates[14] dating the separation of Nichols- and SS14 at around the mid 18th Century (median date 1755, 95% HPD 1651–1835; Figure 2A).

SS14-lineage shows a polyphyletic structure

Within the SS14-lineage, the high number of full-length sequences enabled fine-scale description of phylogenetic sub-structure. In particular, we show partitioning of the SS14Ω centroid cluster previously defined[14] into two lineages; one composed of European and North American derived samples (SS14Ω-A), and another of Chinese and North American derived samples (SS14Ω-B) (Figure 1). While the American and European samples belonging to the former SS14Ω-A lineage are geospatially admixed, those of the latter SS14Ω-B lineage can be further separated between Chinese and North American samples. These partitions were well supported in our maximum likelihood (Figure 1) and Bayesian phylogenies (Figure 2A), as indicated by black node points. We used the rPinecone package[30] to formally classify these sub-lineages based on a defined root-to-tip SNP distance, identifying eight sub-lineages (one of which we further subdivided into sublineages 1A and 1B to aid analysis based temporal and geospatial divergence) within SS14-lineage that correlated well with the population structure described by the phylogeny (Figure 2A). Importantly, while some nodes close to the tips in our phylogeny are unsupported due to small numbers of differentiating SNPs, all sub-lineages defined by rPinecone are supported by >91% posterior support at the key nodes in our Bayesian phylogeny (Figure 2A).

186



187

188 Figure 2. Bayesian maximum credibility phylogeny of sequences recently derived from
189 clinical samples shows expansion of discrete sub-lineages within SS14-lineage, with
190 independent evolution of macrolide resistance. A – Time-scaled phylogeny of all clinical
191 genomes. Coloured tracks indicate lineage, sub-lineage, and presence of macrolide
192 resistance conferring 23S rRNA SNPs (black=present, white=absent, grey=mixed). Node
193 points are shaded according to posterior support (black ≥96%, dark grey >91%, light grey
194 >80%). *Sporadic (non-lineage associated) gain of resistance is highlighted in sub-lineage 3
195 (samples UW133B and UW262B). **Possible reversion from resistant to wildtype (sample

SW6). B – Expanded view of sub-lineages, showing independent acquisition and fixation of macrolide resistance alleles.

Macrolide resistance has evolved independently within SS14

The molecular basis for macrolide resistance has been well documented in *T. pallidum*, and is mediated by point mutations in the 23S ribosomal RNA gene at nucleotide positions 2058 and 2059[31–33]. The A2058G variant was first identified in *T. pallidum* Street Strain-14 (the prototype sample for the SS14-lineage), isolated as long ago as 1977[31], yet resistance has not previously been analysed in context with a detailed whole genome phylogeny. We used ARIBA[34] to perform localised assembly and variant calling of treponema-specific 23S ribosomal sequences from all genomes, and these data were used to infer the presence of both A2058G and A2059G 23S variants that confer macrolide resistance[31]. *T. pallidum* possesses two copies of the 23S ribosomal RNA gene, yet previous analyses have not identified heterozygosity between these two copies – where resistance alleles have been sequenced, they are homozygous between 23S copies, and it has been suggested a gene conversion unification mechanism may exist to facilitate this[18]. Of the 122 genomes, 83 showed evidence of genotypic resistance to macrolides, with 76 genomes showing >95% read support for either the A2058G or A2059G variant. Since it is not possible to discriminate between short reads originating from either copy of 23S because they are perfect repeats, this suggests that both copies carry the same resistance mutation. Seven clinically derived genomes (one UK sample from this study, six described by Pinto and colleagues[13]) showed a mixed 23S allelic profile. All of these samples had >179x read coverage for those sites, with only a fraction of reads (26% - 94%) possessing a resistance allele. In these cases it was not possible to clearly distinguish between a mixture of homozygous positive and negative bacteria in the same patient (either due to within-host evolution or coinfection with multiple strains) or heterozygous sequences from a single bacteria (different 23S rRNA alleles at each copy; heterozygosity in phase) (Figure 2A).

Within the 122 genomes included in this study, we observed that 67% (70/105) of SS14-lineage samples and 35% (6/17) of Nichols-lineage samples were homozygous for either A2058G or the A2059G 23S rRNA allele. In the SS14-lineage, samples possessed either the A2058G (n=59) or the A2059G (n=11) variant. In the Nichols-lineage, six possessed a resistance allele, of which all showed the A2058G variant, and all were from recent clinically derived samples.

To explore the emergence of macrolide resistance, we correlated the taxa in our time-scaled phylogeny with the presence of resistance alleles (Figure 2A). We observed a strong correlation between our well supported sub-lineages and genotypic macrolide resistance or sensitivity, such that resistance appears to have evolved on multiple occasions in a stepwise manner (Figure 2A). For example, Figure 2B shows how the wildtype ancestor of sub-lineage 1B sequences evolved the A2058G between the late 1980s and late 1990s, contrasting with sub-lineage 1A sequences which did not gain A2058G, but subsequently and independently evolved the A2059G variant.

More broadly in the phylogeny, we observe that similar independent 23S rRNA mutations have occurred on at least four occasions (Figure 2A), with separate sub-lineages characterised by either predominantly macrolide resistant or sensitive genotypes; six of the nine sub-lineages were predominantly resistant to macrolides, with three sub-lineages being predominantly sensitive. Several of the macrolide sensitive sub-lineages (in particular sub-lineage 3) contained similar numbers of samples as many of the resistant sub-lineages (Figure 3), suggesting ongoing selection for macrolide resistance has not influenced the expansion of these sub-lineages.

To examine sub-lineage expansion in greater detail, we extracted sampling dates according to sub-lineage, and correlated these data with the predicted time to most recent common ancestor (TMRCA) (Figure 3). Whilst all clinical sequences included in this analysis were sampled after 2000, our analysis indicates that the origins of most of the sub-lineages predated this time and likely arose during the 1990s (Figure 3). Regarding whole sub-lineage associated resistance, we also observed more recent sporadic appearance of resistance mutations, with two separate A2058G variants detected in sub-lineage 3 (an otherwise macrolide sensitive sub-lineage)(Fig. 2).

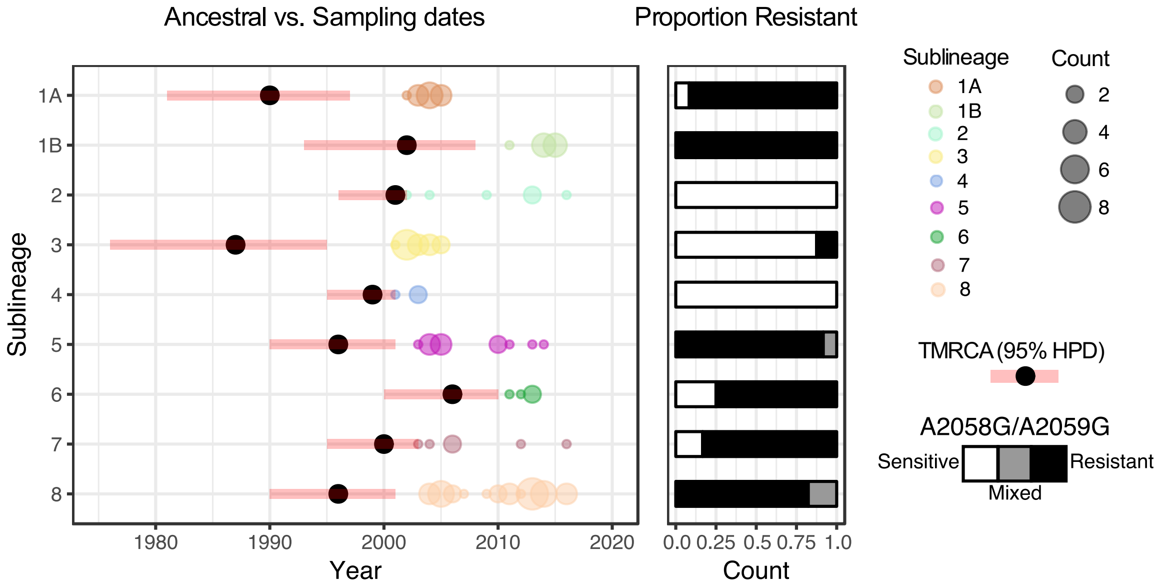


Figure 3. Macrolide resistant and sensitive SS14 sub-lineages evolved independently prior to 2006 and expanded equally regardless of resistance genotype. Figure 3 shows sample collection dates grouped by sub-lineage, with size of circle proportional to number of sequences, and showing predicted time to most recent common ancestor (TMRCA) with 95% highest posterior density (HPD), and proportion of genotypically macrolide resistant, sensitive and mixed samples.

Discussion

Compiling the largest *Treponema pallidum* subspecies *pallidum* sequence collection to date, we show that the majority of the contemporary samples sequenced here were from the SS14-lineage, consistent with other reports[14]. However, three of eight UK genomes (38%) belonged to the Nichols-lineage, showing that these two lineages are still circulating concurrently, and that the prevalence of Nichols-lineage strains may vary by sampling population.

We were able to reconstruct a time-scaled phylogeny using only recently clinically derived samples, and the increase in whole genome sequence numbers combined with the removal of heavily passaged samples contrasts with previous approaches[14,23]. Arora *et al*. reported a mean evolutionary rate of 6.6 x10^{-7} substitutions/site/year for *T. pallidum*[14], comparable with free-living bacterial pathogens with environmental life cycles such as *Vibrio cholerae* (6.1 x 10^{-7})[35] and *Shigella sonnei* (6.0 x10^{-7}). However, *T. pallidum* is a host-restricted pathogen with substantial periods of latency, and as such we would expect a molecular clock rate more similar to that of *Chlamydia trachomatis* (2.15 x10^{-7})[36]. Our inferred rate for TPA (2.28 x10^{-7}) is consistent with this expectation, as well as with other observations that suggest *T. pallidum* has a low evolutionary rate[37].

Within the SS14-lineage, we defined nine well supported sub-lineages that all diverged from their most recent common ancestors prior to 2006, with the earliest (sub-lineage 3) potentially emerging at the end of the 1980s. We observed clear associations between these nine sub-lineages and the presence of macrolide resistance conferring SNPs, with each sub-lineage dominated by either macrolide resistant (n=6) or macrolide sensitive (n=3) samples; there were no sub-lineages representing an even mix of resistance genotypes. Such observations are not consistent with the hypothesis of an ancestrally resistant SS14-lineage driven to high frequency in the population due to a fitness advantage conferred by macrolide resistance, where we would expect to see expansion of a single resistant lineage. Rather, we see evidence of multiple sub-lineages independently evolving macrolide resistance alleles, as a likely consequence of intermittent selective pressure from macrolide treatment, consistent with molecular typing data from Seattle[38]. Phylogenetic reconstruction shows *de novo* evolution of macrolide resistance in syphilis is not a rare event, and furthermore, when resistance evolves in a lineage, it persists in descendants, resulting in transmission from person to person. That the variants appear stable within lineages, with only a single instance in the phylogeny that might represent reversion to a wildtype state, suggests that there is no strong fitness cost associated with possessing these macrolide resistance mutations.

Although our data strongly suggest that global expansion of the SS14-lineage is not contingent on macrolide resistance, the global increases seen in macrolide resistant syphilis[18], as well as the number of resistant lineages emerging in our data, are a cause for concern. Macrolides such as azithromycin are not considered frontline treatment for syphilis, with WHO and US guidelines recommending treatment with BPG[39,40], with doxycycline recommended as a secondary treatment option. In contrast, WHO now recommends azithromycin rather than BPG as the treatment of choice for mass drug administration and the eradication of yaws[19], caused by the closely related *T. pallidum* subspecies *pertenue*. Macrolide resistance has recently been described in yaws[20] and our data suggest that further independent evolution of azithromycin resistance is highly likely, which would have significant implications for yaws eradication efforts. Worryingly, applying azithromycin-based mass drug administration to populations infected with both TPP and TPA could promote resistance in both species.

Several factors are likely to have driven the repeated evolution and subsequent expansion of macrolide resistant lineages of TPA. Use of azithromycin, or other macrolides, for other indications is likely to have played a significant role[41]. Azithromycin entered global markets between 1988 and 1991 (marketed by Pfizer as Zithromax), and became one of the most widely used antibiotics in the United States for a wide variety of indications[42], including the treatment of respiratory tract infections and for the treatment of other sexually transmitted infections. In many cases, the dose used for treatment of these indications is lower than the recommended dose for the treatment of syphilis. Azithromycin and clarithromycin were also widely used prophylactically amongst individuals living with HIV prior to the widespread availability of combined anti-retroviral therapy. Off-target macrolide exposure is of particular concern because azithromycin has a long half-life[43] and may persist at subclinical concentrations in patients. Widespread use of macrolides for this broad range of indications might therefore have contributed to sub-therapeutic exposure of patients with incubating or early syphilis and ultimately selection of resistance.

The recent increase in incidence of syphilis in high income countries likely reflects changes in sexual behaviour[8]. The fact that we observe the expansion of both genotypically resistant and sensitive lineages highlights particular treatment issues. There have been significant global shortages of BPG[44] in low, middle, and high income nations, including the United States, with the global supply of BPG dependent on just three manufacturers of the active ingredient[44]. Pharmaceutical production of sterile, injectable β-lactam derived antimicrobials such as BPG is costly, yet as an older off-patent medication with declining demand in the face of growing antimicrobial resistance (AMR) in other organisms, the financial rewards for production are low[44]. This may be compounded by a misperception that BPG is an outdated drug that could be replaced by newer, more effective drugs[44]. In circumstances where azithromycin is used instead of doxycycline as second line treatment, a shortage of BPG leads inevitably to inadequate treatment of early infectious syphilis and contributes to ongoing, unchecked transmission. In China for example, studies have reported that despite high rates of macrolide resistance,[45] clinicians have inappropriately been resorting to macrolide treatment due to ongoing BPG shortages[46]. Thus although a well-established, highly effective treatment for syphilis (BPG) has been available since the mid-1950s, shortages in the present era contribute to suboptimal treatment strategies and continued use of drugs with a known resistance problem.

The epidemic of syphilis and the widespread problems of azithromycin resistance and BPG shortage require a multi-faceted response. This includes new strategies for treatment and reduction of transmission, finding ways to improve the security of the BPG supply chain, and strengthening molecular surveillance for antimicrobial resistance in *T. pallidum*[47]. Many authors have discussed the importance of rethinking the economics of antimicrobial development pipelines to ensure we are still able to treat infections[48,49]. In syphilis, we must rethink how we can protect the continued production of existing highly efficacious penicillins in the face of increasing antimicrobial resistance rendering them ineffective for other organisms, especially in the light of the recent increases in syphilis incidence in Europe, North America, and Asia.

Materials and Methods

Samples

UK samples consisted of residual DNA, extracted from clinical swabs using a QIAsymphony (Qiagen) from routine diagnostic samples obtained from patients presenting with clinical evidence of syphilis at the Mortimer Market Centre, London. Use of the UK samples was approved by the NHS Research Ethics Committee (IRAS Project ID 195816). US samples from Seattle were collected from individuals enrolled in a study of cerebrospinal fluid abnormalities in patients with syphilis, with ethical approval at the University of Washington (UW IRB # STUDY00003216). Specifically, 2.4-3.0 ml participant blood was inoculated into rabbit testes as previously described[50], and *T. pallidum* suspensions were collected after the second round of passage. Historical strains were propagated in rabbits and harvested from infected testes. *T. pallidum* suspensions were treated using a lysis buffer (10mM Tris pH 8.0, 0.1M EDTA pH 8.0, 0.5% SDS), freeze-thaw, and extraction using QIAamp Mini kit (Qiagen) according to the manufacturer’s instructions; in select cases the proteinase K incubation was extended overnight to improve DNA yield. Treponemal DNA was quantified using a qPCR targeting the Tp0574 gene that is conserved across all known members of the *T. pallidum* cluster, and compared to a standard curve derived from a plasmid containing the PCR amplicon. Samples with a concentration >2000 genome copies/μl were selected for sequencing; borderline samples with high volume and a pathogen load over 500 genome copies/μl were concentrated using a vacuum centrifuge. Samples were arranged in groups of 20 according to similar (within 2 C_T) treponemal load, with high concentration outlier samples diluted as necessary. We added 4μl pooled commercial human gDNA (Promega) to all samples to ensure total gDNA > 1μg/35μl, sufficient for library prep.

Sequencing

Genomic DNA was sheared to 100-400bp by ultrasonication, followed by adaptor ligation and index barcoding according to existing Illumina protocols. Samples were pooled in the preassigned groups of 20 to generate equimolar Total DNA pools. Each pool was then hybridised using SureSelect 120-mer RNA baits designed against published RefSeq examples of *T. pallidum* and *T. paraluiscuniculi* as described previously[15]. Libraries were subjected to 125bp paired end sequencing on Illumina HiSeq 2500 with version 4 chemistry according to established protocols. Raw sequencing reads were deposited at the European Nucleotide Archive (ENA) under project PRJEB20795; all accessions used in this project are listed in Supplementary Table 1.

Sequence analysis and phylogenetics

Treponemal sequencing reads were prefiltered using a Kraken[51] v0.10.6 database containing all bacterial and archaeal nucleotide sequences in RefSeq, plus mouse and human, to identify and extract those reads with homology to Treponema species, followed by adaptor trimming using Trimmomatic[52] v0.33. To reduce bias due to variable read depth, as well as make analysis computationally tractable, for samples with high read counts we used seqtk v1.0 (available at https://github.com/lh3/seqtk) to randomly down-sample the binned and trimmed reads to 2,500,000 unique treponemal read pairs. For publicly available genomes, raw sequencing reads were downloaded from SRA and subjected to the same binning and down-sampling pipeline. For a small minority of public genomes, raw sequencing reads were not available; for these we simulated 125bp PE perfect reads from the RefSeq closed genomes using Fastaq (available at https://github.com/sanger-pathogens/Fastaq).

For phylogenetic analysis, we used a reference mapping approach with a custom version of the SS14 v2 reference sequence (NC_021508.1) from which we first masked 14 highly repetitive or recombinogenic genes (12 repetitive Tpr genes A-L, arp and TPANIC_0470) using bedtools[53] v2.17.0 ‘maskfasta’. We mapped prefiltered sequencing reads to the reference using BWA mem[54] v0.7.17 (MapQ ≥ 20), followed by indel realignment using GATK[55] v3.4-46 IndelRealigner, deduplication with Picard MarkDuplicates v1.127 (available at http://broadinstitute.github.io/picard/), and variant calling and consensus pseudosequence generation using using samtools v1.2[56] and bcftools v1.2, requiring a minimum of three supporting reads per strand and eight in total to call a variant, and a variant frequency/mapping quality cut-off of 0.8; sites not meeting these criteria were masked to ‘N’ in the pseudosequence. Importantly, reads mapping to multiple genomic positions were marked and excluded from SNP calling, meaning repeated regions such as the duplicated 23S genes was not included in the multiple sequence alignment used to derive the phylogenies.

Multiple sequence alignments were screened for evidence of recombination using Gubbins[57], generating recombination-masked full genome length and SNP-only alignments. Maximum likelihood phylogenies were calculated on SNP-only alignments using IQ-Tree v1.6.3[58], correcting for missing constant sites using the built in ascertainment bias correction[59], allowing the built-in model testing[60] to determine a K3P (three substitution types model and equal base frequencies) substitution model[61] with a FreeRate model of heterogeneity[62] assuming three categories, and performing 1000 UltraFast Bootstraps[63,64].

To determine SS14 sub-lineages, we recalculated a maximum likelihood tree as described above for SS14-clade sequences only (using the Mexico A strain, NC_018722.1 as outgroup), before performing a joint ancestral reconstruction[65] of SNPs on the tree branches using pyjar (available at https://github.com/simonrharris/pyjar). We then used the rPinecone package[64] (available at https://github.com/alexwailan/rpinecone), which applies a root-to-tip approach to defining clusters based on SNP distance relative to ancestral nodes. We used a cluster threshold of 10 SNPs, which proved optimal for describing the underlying phylogenetic structure of the tree, and yielded eight sub-lineages. Within sub-lineage 1, we found that pinecone clusters did not accurately represent the phylogeny, despite a clear phylogenetic separation, with one group of sequences from China associated with the A2058G allele, and the other group from the USA associated with the A2059G allele. We further manually clustered these sequences according to their shared ancestral nodes, naming them sublineages 1A and 1B.

We evaluated our maximum likelihood phylogeny for evidence of temporal signal using TempEst[66] v1.5, and this showed a correlation of 0.44 and R^2 of 0.20 for the whole tree (Supplementary Figure 1), whilst the SS14-lineage -only alignment showed a correlation of 0.66 and R^2 of 0.44; this indicated that there was sufficient evidence for temporal signal and we proceeded to BEAST analysis. BEAST[29] v1.8.2 was initially run on a recombination-masked SNP-only alignment containing 284 variable sites, applying a correction for invariant sites using the constantPatterns argument, in triplicate using an Uncorrelated Relaxed Clock model, assuming constant population size, lognormal population distribution, GTR substitution model, diffuse gamma distribution prior (shape 0.001, scale 1000), with a burnin of 10 million cycles followed by 100 million MCMC cycles. All MCMC chains converged, and on inspection of the marginal distribution of ucld.stdev we could not reject a Strict Clock. We therefore repeated the analysis using a Strict Clock model, using the same models and priors and assuming a starting molecular clock rate of 3.6×10^{-4} as described by others[14]. We used the Marginal Likelihood Estimates from the triplicate BEAST runs as input to Path Sampling and Stepping Stone Sampling analyses[67,68] and determined that the Strict Constant model was optimal for this dataset. To further confirm the temporal signal in our tree, we used the TIPDATINGBEAST package[69] in R[70] to resample tip-dates from our alignment, generating 20 new datasets with randomly assigned dates – BEAST analysis using the same Strict Clock prior conditions found no evidence of temporal signal in these replicates, indicating that the signal in our tree was not found by chance (Supplementary Figure 2).

Macrolide resistant SNPs were inferred using ARIBA[34], which performs localised assembly and mapping in comparison with a custom reference database containing 23S sequences from Nichols (NR_076156.1) and SS14 reference strains (NR_076531.1).

Processing of data, and all statistical analysis was performed in R[70] v3.4.1, primarily using the phytools and ape packages. Phylogenies were plotted using ggtree[71], and figures were produced using ggplot2[72]. All code used in the downstream analysis is available in Supplementary File 3.

Acknowledgements

The authors thank the sequencing team at Wellcome Sanger Institute; S Harris for access to local scripts; A Wailan for discussions around rPinecone; D Domman for helpful advice and discussion during analysis; M Fookes for initial pilot work and bait design for SureSelect; B Molini, and C Godornes at University of Washington for provision of strains; staff at University College London Hospital and clinical and sexual health staff at the Mortimer Market Clinic, London for UK sample provision. MM was supported by a Wellcome Trust Clinical Research Fellowship (102807). Strains from the University of Washington were supported by NIH grants R01 NS34235 to CMM, and R01 AI 34616 and R01 AI 42143 to SAL. MAB and NRT are supported by Wellcome funding to the Sanger Institute.

Author Contributions

Conceived and designed the study: NRT, SAL, MAB, AVN, MM. Collected and collated samples: SAL, CMM, MM, AVN, PF. Performed the laboratory work: MAB, SKS, LCT. Analysed the data: MAB. Wrote the initial draft of the manuscript: MAB. All authors viewed and contributed to the final manuscript.

Competing Interests

The authors have no competing interests to declare.

References

1. Tampa, M., Sarbu, I., Matei, C., Benea, V. & Georgescu, S. Brief History of Syphilis. *J. Med. Life* **7,** 4–10 (2014).

2. Chesson, H. W., Dee, T. S. & Aral, S. O. AIDS mortality may have contributed to the decline in syphilis rates in the United States in the 1990s. *Sex. Transm. Dis.* **30,** 419–424 (2003).

3. Fenton, K. A. *et al.* Infectious syphilis in high-income settings in the 21st century. *Lancet Infect. Dis.* **8,** 244–253 (2008).

4. Centers for Disease Control. Syphilis - 2016 STD Surveillance Report. (2017). Available at: https://www.cdc.gov/std/stats16/Syphilis.htm. (Accessed: 2nd July 2018)

5. Zhou, Y. *et al.* Prevalence of HIV and syphilis infection among men who have sex with men in China: a meta-analysis. *BioMed Res. Int.* **2014,** 620431 (2014).

6. Public Health England. Sexually transmitted infections and screening for chlamydia in England, 2017. (2018).

7. European Centre for Disease Prevention and Control. Sexually transmitted infections in Europe 2013. (2015).

8. Mohammed, H. *et al.* Increase in Sexually Transmitted Infections among Men Who Have Sex with Men, England, 2014. *Emerg. Infect. Dis.* **22,** 88–91 (2016).

9. Edmondson, D. G., Hu, B. & Norris, S. J. Long-Term In Vitro Culture of the Syphilis Spirochete Treponema pallidum subsp. pallidum. *mBio* **9,** e01153-18 (2018).

10. Lafond, R. E. & Lukehart, S. A. Biological basis for syphilis. *Clin. Microbiol. Rev.* **19,** 29–49 (2006).

11. Christiansen, M. T. *et al.* Whole-genome enrichment and sequencing of Chlamydia trachomatis directly from clinical samples. *BMC Infect. Dis.* **14,** 591 (2014).

12. Depledge, D. P. *et al.* Specific Capture and Whole-Genome Sequencing of Viruses from Clinical Samples. *PLoS ONE* **6,** e27805 (2011).

13. Pinto, M. *et al.* Genome-scale analysis of the non-cultivable Treponema pallidum reveals extensive within-patient genetic variation. *Nat. Microbiol.* **2,** 16190 (2016).

14. Arora, N. *et al.* Origin of modern syphilis and emergence of a pandemic Treponema pallidum cluster. *Nat. Microbiol.* **2,** 16245 (2016).

15. Marks, M. *et al.* Diagnostics for yaws eradication: insights from direct next generation sequencing of cutaneous strains of Treponema pallidum. *Clin. Infect. Dis.* (2017). doi:10.1093/cid/cix892

16. Nechvátal, L. *et al.* Syphilis-causing strains belong to separate SS14-like or Nichols-like groups as defined by multilocus analysis of 19 Treponema pallidum strains. *Int. J. Med. Microbiol.* **304,** 645–653 (2014).

17. Lukehart, S. A. *et al.* Macrolide resistance in Treponema pallidum in the United States and Ireland. *N. Engl. J. Med.* **351,** 154–158 (2004).

18. Šmajs, D., Paštěková, L. & Grillová, L. Macrolide Resistance in the Syphilis Spirochete, Treponema pallidum ssp. pallidum: Can We Also Expect Macrolide-Resistant Yaws Strains? *Am. J. Trop. Med. Hyg.* **93,** 678–683 (2015).

19. WHO. Eradication of yaws – the Morges Strategy. *Wkly Epidemiol Rec* **87,** 189–194 (2012).

20. Mitjà, O. *et al.* Re-emergence of yaws after single mass azithromycin treatment followed by targeted treatment: a longitudinal study. *The Lancet* **391,** 1599–1607 (2018).

21. Giacani, L. *et al.* Complete Genome Sequence and Annotation of the Treponema pallidum subsp. pallidum Chicago Strain. *J. Bacteriol.* **192,** 2645–2646 (2010).

22. Giacani, L. *et al.* Complete Genome Sequence of the Treponema pallidum subsp. pallidum Sea81-4 Strain. *Genome Announc.* **2,** (2014).

23. Sun, J. *et al.* Tracing the origin of Treponema pallidum in China using next-generation sequencing. *Oncotarget* **7,** 42904–42918 (2016).

24. Matějková, P. *et al.* Complete genome sequence of Treponema pallidum ssp. pallidum strain SS14 determined with oligonucleotide arrays. *BMC Microbiol.* **8,** 76 (2008).

25. Pětrošová, H. *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* **8,** e74319 (2013).

26. Tong, M.-L. *et al.* Whole genome sequence of the Treponema pallidum subsp. pallidum strain Amoy: An Asian isolate highly similar to SS14. *PLoS ONE* **12,** (2017).

27. Čejková, D. *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. *PLoS Negl. Trop. Dis.* **6,** e1471 (2012).

28. Pětrošová, H. *et al.* Whole Genome Sequence of Treponema pallidum ssp. pallidum, Strain Mexico A, Suggests Recombination between Yaws and Syphilis Strains. *PLoS Negl. Trop. Dis.* **6,** e1832 (2012).

29. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4,** (2018).

30. Wailan, A. M. *et al.* rPinecone: Define sub-lineages of a clonal expansion via a phylogenetic tree. *bioRxiv* 404624 (2018). doi:10.1101/404624

31. Stamm, L. V. & Bergen, H. L. A Point Mutation Associated with Bacterial Macrolide Resistance Is Present in Both 23S rRNA Genes of an Erythromycin-Resistant Treponema pallidum Clinical Isolate. *Antimicrob. Agents Chemother.* **44,** 806–807 (2000).

32. Matějková, 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.* **58,** 832–836 (2009).

33. Molini, B. J. *et al.* Macrolide Resistance in Treponema pallidum Correlates With 23S rDNA Mutations in Recently Isolated Clinical Strains. *Sex. Transm. Dis.* **43,** 579–583 (2016).

34. Hunt, M. *et al.* ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb. Genomics* **3,** (2017).

35. Weill, F.-X. *et al.* Genomic history of the seventh pandemic of cholera in Africa. *Science* **358,** 785–789 (2017).

36. Hadfield, J. *et al.* Comprehensive global genome dynamics of Chlamydia trachomatis show ancient diversification followed by contemporary mixing and recent lineage expansion. *Genome Res.* **27,** 1220–1229 (2017).

37. Strouhal, M. *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.* **11,** e0005894 (2017).

38. Grimes, M. *et al.* Two Mutations associated with Macrolide Resistance in Treponema pallidum: Increasing Prevalence and Correlation with Molecular Strain Type in Seattle, Washington. *Sex. Transm. Dis.* **39,** 954–958 (2012).

39. WHO. WHO guidelines for the treatment of *Treponema pallidum* (syphilis). (2016).

40. Centers for Disease Control. Sexually Transmitted Diseases Treatment Guidelines, 2015. *Morb. Mortal. Wkly. Rep.* **64,** (2015).

41. Marra, C. M. *et al.* Antibiotic Selection May Contribute to Increases in Macrolide-Resistant Treponema pallidum. *J. Infect. Dis.* **194,** 1771–1773 (2006).

42. Hicks, L. A., Taylor, T. H. & Hunkler, R. J. U.S. Outpatient Antibiotic Prescribing, 2010. *N. Engl. J. Med.* **368,** 1461–1462 (2013).

43. Kong, F. Y. S. *et al.* Pharmacokinetics of a single 1g dose of azithromycin in rectal tissue in men. *PLOS ONE* **12,** e0174372 (2017).

44. Nurse-Findlay, S. *et al.* Shortages of benzathine penicillin for prevention of mother-to-child transmission of syphilis: An evaluation from multi-country surveys and stakeholder interviews. *PLOS Med.* **14,** e1002473 (2017).

45. Chen, X.-S. *et al.* High prevalence of azithromycin resistance to Treponema pallidum in geographically different areas in China. *Clin. Microbiol. Infect.* **19,** 975–979 (2013).

46. Lu, H. *et al.* High frequency of the 23S rRNA A2058G mutation of *Treponema pallidum* in Shanghai is associated with a current strategy for the treatment of syphilis. *Emerg. Microbes Infect.* **4,** e10 (2015).

47. Taylor, M. *et al.* Revisiting strategies to eliminate mother-to-child transmission of syphilis. *Lancet Glob. Health* **6,** e26–e28 (2018).

48. Baker, S., Thomson, N., Weill, F.-X. & Holt, K. E. Genomic insights into the emergence and spread of antimicrobial-resistant bacterial pathogens. *Science* **360,** 733–738 (2018).

49. WHO. Global action plan on antimicrobial resistance. (2015).

50. Lukehart, S. A. & Marra, C. M. Isolation and laboratory maintenance of Treponema pallidum. *Curr. Protoc. Microbiol.* **Chapter 12,** Unit 12A.1 (2007).

51. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15,** R46 (2014).

52. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).

53. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

54. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* (2013). doi:1303.3997v1 [q-bio.GN]

55. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al* **11,** 11.10.1-11.10.33 (2013).

56. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

57. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* gku1196 (2014). doi:10.1093/nar/gku1196

58. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **32,** 268–274 (2015).

59. Lewis, P. O. A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data. *Syst. Biol.* **50,** 913–925 (2001).

60. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Haeseler, A. von & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14,** 587–589 (2017).

61. Kimura, M. Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci.* **78,** 454–458 (1981).

62. Soubrier, J. *et al.* The Influence of Rate Heterogeneity among Sites on the Time Dependence of Molecular Rates. *Mol. Biol. Evol.* **29,** 3345–3358 (2012).

63. Hoang, D. T., Chernomor, O., Haeseler, A. von, Minh, B. Q. & Le, V. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *bioRxiv* 153916 (2017). doi:10.1101/153916

64. Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. Ultrafast Approximation for Phylogenetic Bootstrap. *Mol. Biol. Evol.* **30,** 1188–1195 (2013).

65. Pupko, T., Pe, I., Shamir, R. & Graur, D. A Fast Algorithm for Joint Reconstruction of Ancestral Amino Acid Sequences. *Mol. Biol. Evol.* **17,** 890–896 (2000).

66. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2,** (2016).

67. Baele, G. & Lemey, P. Bayesian evolutionary model testing in the phylogenomics era: matching model complexity with computational efficiency. *Bioinformatics* **29,** 1970–1979 (2013).

68. Baele, G. *et al.* Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol. Biol. Evol.* **29,** 2157–2167 (2012).

69. Rieux, A. & Khatchikian, C. E. tipdatingbeast: an r package to assist the implementation of phylogenetic tip-dating tests using beast. *Mol. Ecol. Resour.* **17,** 608–613

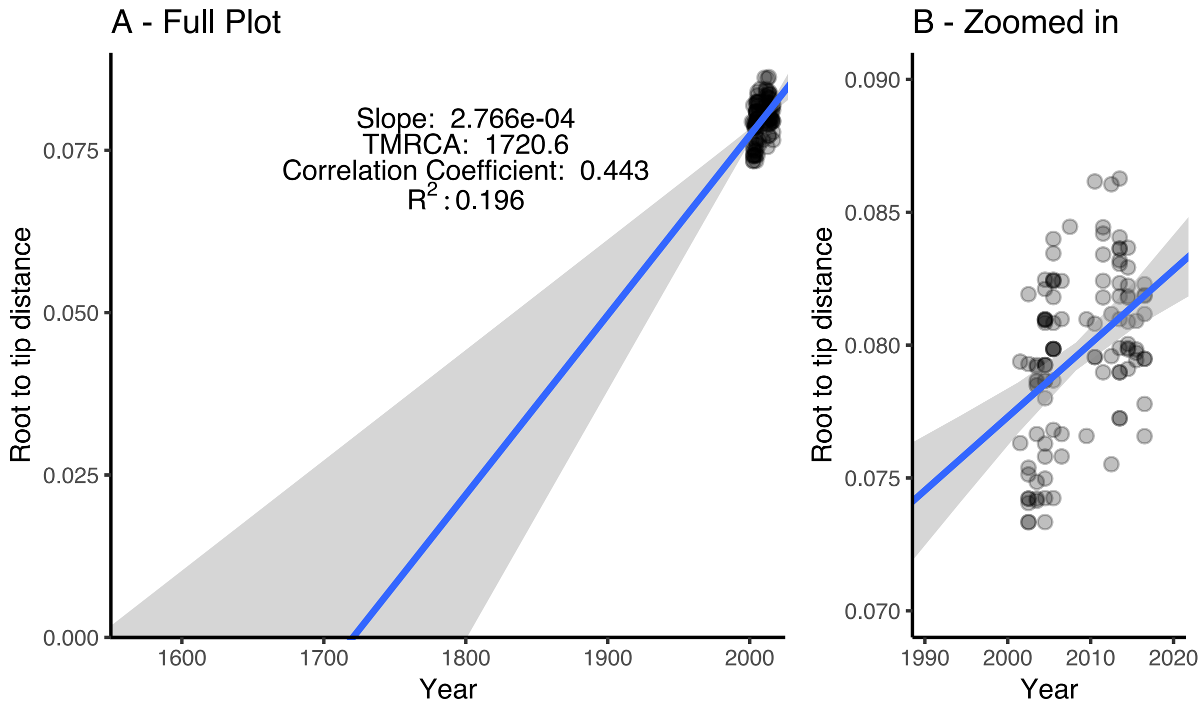
70. R Core Team. *R: A Language and Environment for Statistical Computing.* (R Foundation for Statistical Computing, 2014).

71. Yu, G., Smith, D., Zhu, H., Guan, Y. & Lam, T. T.-Y. *ggtree: an R package for visualization and annotation of phylogenetic tree with different types of meta-data*.

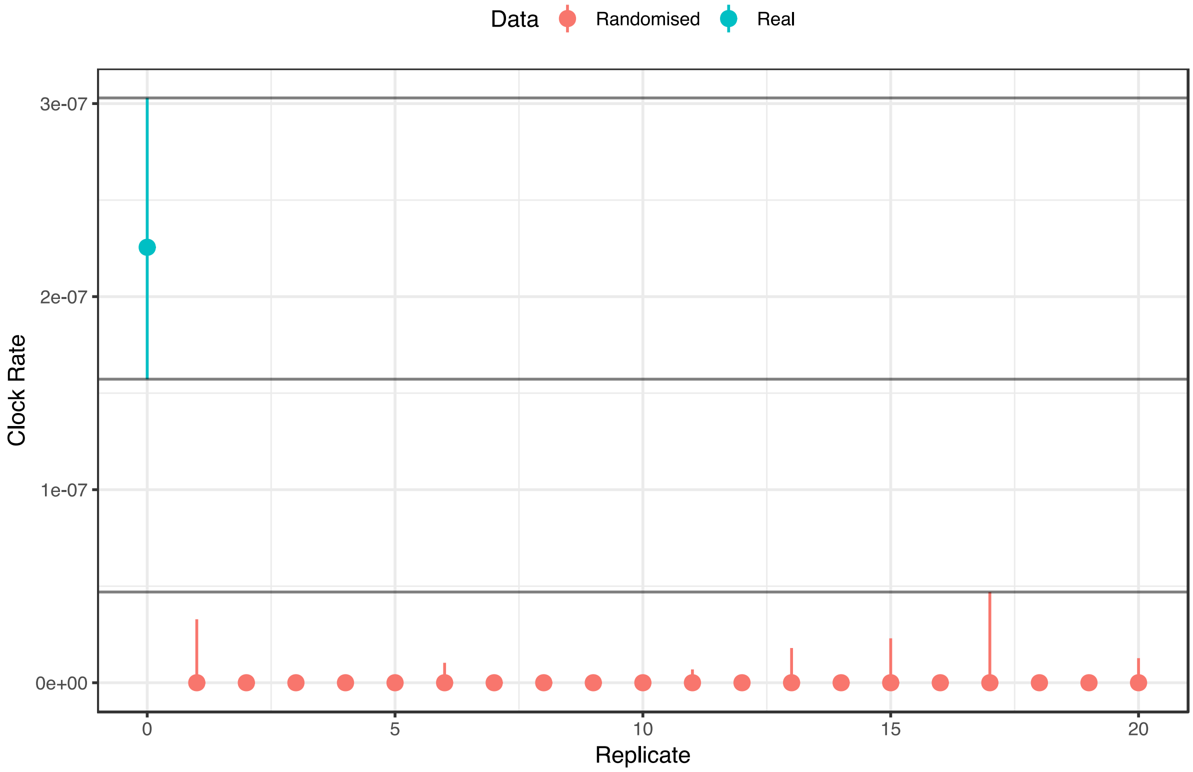
72. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag, 2009).

Supplementary Figures

Supplementary Figure 1. Root-to-tip regression analysis of tip dates against branch lengths showing a correlation of 0.443 and R2 of 0.196, providing evidence for temporal signal in the Maximum Likelihood tree, performed in TempEst using clinically derived genomes from both Nichols and SS14 lineages. Plots show tip points and linear regression (with standard error) for full timeline (A) and zoomed in to only include sampled tipdates (B). Each data point is coloured grey, with darker shading indicating multiple overlapping points.



Supplementary Figure 2. Tip date resampling analysis performed using twenty datasets with randomised tip dates generated from the original Strict Clock analysis and run in BEAST under the same conditions. Median clock rate for the real tree was 2.26 x 10^{-7}, whilst all randomly assigned datasets gave substantially lower clock rates, with the highest median clock rate obtained at 1.90 x 10^{-12}. This indicates that the temporal signal observed in our tree was not obtained by chance, and provides further evidence for a temporal signal in the multiple sequence alignment.



Supplementary Table 1. Full sample metadata (Excel Sheet) for this study, including sequence naming used in this paper, in other publications, and on GenBank, ENA Accessions for all new genomes, as well as results of lineage and sub-lineage typing and inference of genotypic macrolide resistance.

Supplementary Table 2. List of genomic regions behaving in a non-clocklike manner and masked due to hypervariable, recombining or repetitive elements.

Supplementary File 3. Rnotebook (HTML format) containing all downstream code used to generate primary figures and statistics.