1 # Title

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

4

5 # Authors, Affiliations

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

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

9

10 [1] Parasites and Microbes, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton,
11 Cambridgeshire, UK

12 [2] Clinical Research Department, Faculty of Infectious and Tropical Diseases, London School
13 of Hygiene & Tropical Medicine, London, UK

14 [3] Hospital for Tropical Diseases, London, UK

15 [4] Department of Neurology, University of Washington, USA

16 [5] Guy's & St Thomas' NHS Foundation Trust, London, UK

17 [6] The Mortimer Market Centre CNWL, Camden Provider Services, London, UK

18 [7] Departments of Medicine and Global Health, University of Washington, USA

19 [8] Department of Pathogen Molecular Biology, Faculty of Infectious and Tropical Diseases,
20 London School of Hygiene & Tropical Medicine, London, UK

21

22 Abstract

23 Syphilis is a sexually transmitted infection caused by *Treponema pallidum* subspecies
24 *pallidum* and may lead to severe complications. Recent years have seen striking increases in
25 syphilis in many countries. Previous analyses have suggested one lineage of syphilis, SS14,
26 may have expanded recently, indicating emergence of a single pandemic azithromycin-
27 resistant cluster. We used direct sequencing of *T. pallidum* combined with phylogenomic
28 analyses to show that both SS14- and Nichols-lineages are simultaneously circulating in
29 clinically relevant populations in multiple countries. We correlate the appearance of
30 genotypic macrolide resistance with multiple independently evolved SS14 sub-lineages and
31 show that genotypically resistant and sensitive sub-lineages are spreading
32 contemporaneously, incompatible with the notion that SS14-lineage expansion is driven
33 purely by macrolide resistance. These findings inform our understanding of the current
34 syphilis epidemic by demonstrating how macrolide resistance evolves in *Treponema*
35 subspecies and provide a warning on broader issues of antimicrobial resistance.

36 Introduction

37 Syphilis is a centuries-old, predominantly sexually transmitted infection (STI) caused by the
38 bacterium *Treponema pallidum* subspecies *pallidum* (TPA). If untreated, syphilis causes a
39 multi-system disease that can progress to severe cardiovascular and neurological
40 involvement, which can be potentially fatal. Syphilis caused a pandemic wave that swept

41 across Renaissance Europe over 500 years ago, and remained a problem until the
42 introduction of antibiotics in the post-World War II era[1]. Despite effective treatment with
43 benzathine benzylpenicillin G (BPG), syphilis transmission levels fluctuated but persisted
44 throughout the 20th century, until the AIDS crisis of the 1980s and 1990s, where changes in
45 sexual behaviour (and possibly AIDS-related mortality), led to an overall decline in incidence
46 in many western countries and populations[2,3].

47

48 Recent years have seen a sharp increase in syphilis cases in many high-income countries,
49 predominantly within sexual networks of men who have sex with men (MSM)[4,5]. In the
50 United Kingdom there was a 20% increase in reported new diagnoses between 2016 and
51 2017, and a 148% increase since 2008[6]. Similar trends have been reported in other
52 countries[4,7]. The reasons for this increase are complex and multifactorial, incorporating
53 changing behavioural patterns mediated by cultural, societal and technological changes in
54 our modern world[8], resulting in a perfect epidemiological storm. It is also possible that there
55 are bacterial changes either driving the current rise in syphilis incidence, or occurring as a
56 consequence of this increase. However, current knowledge of TPA is limited, largely because
57 the bacterium was, until recently, intransigent to *in vitro* culture[9]. Most current
58 understanding of TPA biology therefore comes from related species or from TPA cultured in
59 the *in vivo* rabbit testicular model[10]. Genomic analysis has also been limited due to low
60 levels of TPA pathogen load in patients and difficulty in readily isolating new strains.
61 Sequencing must be performed directly on clinical specimens or after passage through
62 rabbits, leading to substantial bottlenecks in genomic data generation. Recent advances
63 have enabled target enrichment of pathogen reads directly from clinical or cultured

64 specimens[11,12], and this was recently employed separately by different groups, including our
65 own, to sequence TPA and other *T. pallidum* subspecies directly from patient samples[13–16].

66

67 The availability of increasing numbers of genomic sequences enabled the first description of
68 the global TPA population structure using 31 near genome-length TPA sequences, along
69 with a small number derived from closely related species[14]. The authors described two
70 lineages within TPA; a Nichols-lineage found almost exclusively in North American
71 sequences exhibiting substantial nucleotide diversity, and a geographically widespread but
72 genetically homogeneous SS14-lineage, confirming previous analyses using multi-locus
73 sequence typing[17]. Of these two lineages, they found that 68% of tested TPA genomes
74 belonged to the SS14-lineage, and further analysis using a larger dataset of 1354 single-
75 locus molecular types (comprising 623 samples from South East Asia, 241 from the USA, 392
76 from Europe and a small number of other locations) also supported this view (94%
77 SS14-lineage).

78

79 Although penicillin resistance has never been reported in syphilis, increasing levels of
80 genotypic resistance, and clinical treatment failure, to macrolides such as azithromycin have
81 been reported[18,19], conferred by either one of two single nucleotide polymorphisms (SNPs)
82 in the 23S ribosomal sequence (A2058G and A2059G). Previous molecular typing studies
83 indicated that differential patterns of macrolide resistance correlate with geographical
84 location[18–20] and molecular type[21,22,19,23] , and Arora *et al* reported that 90% of whole
85 genome sequenced SS14-lineage genomes and 25% of Nichols-lineage genomes contained
86 SNPs conferring macrolide resistance; furthermore, it has been suggested that SS14-lineage

87 may represent a single pandemic azithromycin-resistant cluster[14] perhaps related to
88 azithromycin selective pressure[14,17].
89
90 In this study, we performed direct whole genome sequencing on 73 TPA samples from the
91 US and Europe, and combined these data with 49 publicly available genomes. We tested the
92 hypothesis that, as with other pathogens[24], expansion of the SS14-lineage might have been
93 driven by selection of macrolide resistance. We used phylogenetic analysis to delineate sub-
94 lineages within the both the SS14- and Nichols- lineages, providing an evolutionary
95 framework for the emergence of macrolide resistance in treponemas, and showing striking
96 patterns of the emergence and fixation of macrolide resistance SNPs that indicate
97 independent evolution and proliferation of resistance alleles. These findings have
98 implications for our understanding of the increasing incidence of syphilis and on the
99 potential of the WHO Yaws eradication campaign to drive further development of macrolide
100 resistance in both TPA and in the closely related *Treponema pallidum* subspecies *pertenue*
101 (TPP)[25,26].
102

103 ## Results

104 We sequenced eight genomes directly from clinical swabs collected in 2016 from patients in
105 the United Kingdom and 60 isolate genomes from low rabbit passage samples (no more
106 than two passages from the original patient sample; henceforth referred to as 'recently
107 clinically derived') originally collected from patients between 2001-2011 in the USA. We also
108 resequenced three clonally derived laboratory strains from the USA that have been

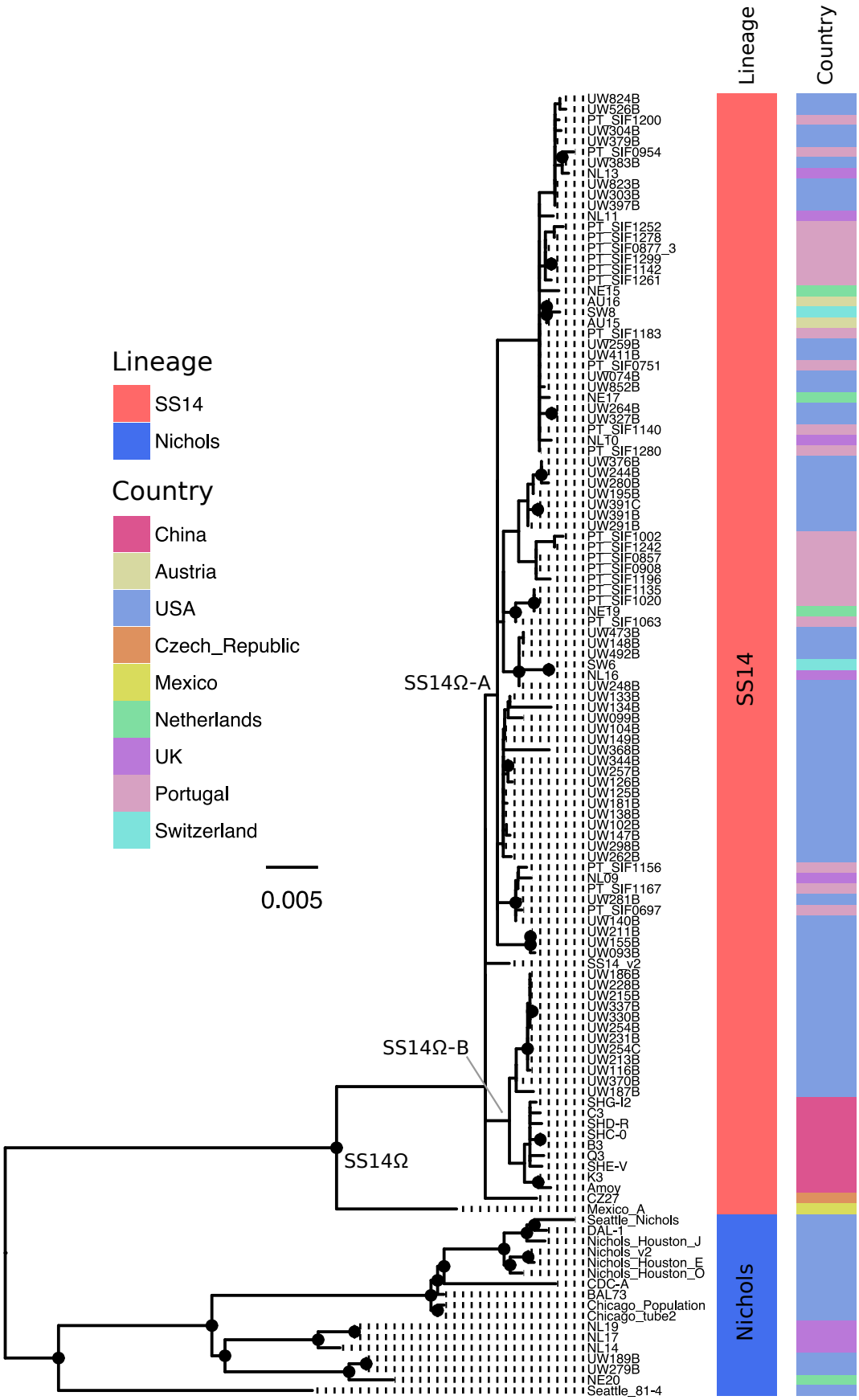
109 previously sequenced (Nichols Houston E, Nichols Houston J, Nichols Houston O) but remain
110 unpublished, and two strains for which the sequencing reads were not publicly available
111 (Chicago[27], Seattle 81-4[28]). We combined our data with 49 high-quality genomes published
112 previously[13,14,29–34], 41 of which were recently derived from clinical patients, yielding a
113 dataset of 122 genomes (109 with limited passage from clinical patients). Combined, our
114 sample set included 72 genomes from the USA (predominantly Seattle), 8 from the UK
115 (exclusively London), 9 from China (predominantly Shanghai), 23 from Portugal (exclusively
116 Lisbon), and a small number from other countries, all collected between 1912-2016
117 (Supplementary Table 1).

118

119 After removal of recombinant and repetitive sites (both by selective mapping and screening
120 – see Methods and Supplementary Table 2), we performed phylogenomic analyses, using
121 maximum likelihood and Bayesian methods to define lineages. In agreement with previous
122 studies[14], we show the presence of two dominant lineages in our dataset (previously
123 denoted SS14 and Nichols; Figure 1) that are separated by >70 non-recombining single
124 nucleotide polymorphisms (SNPs). Of the 122 total samples included in this study, 105 (86%)
125 belonged to the SS14-lineage, whilst of the 109 recently clinically derived samples included,
126 103 (94%) were from SS14-lineage. In contrast, only six Nichols-lineage samples were
127 recently clinically derived, and most (11/17) Nichols-lineage genomes examined were
128 historically passaged isolates, including those derived from the original Nichols strain
129 isolated in 1912 and disseminated to different North American laboratories; some Nichols-
130 lineage genomes represent clones of the parent strain derived *in vivo*. However, although
131 we observed a strong bias towards clinically derived SS14-lineage samples in this dataset,

132 not all recent clinical strains were of the SS14-lineage; six recent clinical samples belonged
133 to the Nichols-lineage, three (of eight sequenced) clinical samples from the UK in 2016, one
134 collected in the Netherlands in 2013, and two from the USA in 2004, indicating that
135 transmission of this lineage is ongoing and potentially more widespread than previously
136 thought.

137


Lineage
SS14
Nichols
Country
China
Austria
USA
Czech_Republic
Mexico
Netherlands
UK
Portugal
Switzerland
0.005
SS14Ω-A
SS14Ω-B
SS14Ω
Lineage
Country
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
C227
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

139 Figure 1. Maximum likelihood phylogeny of 122 high quality *T. pallidum* subspecies *pallidum*

140 genomes (including recently clinically derived samples and those with more extensive

141 laboratory passage), showing lineage and country of origin. Ultra-Fast bootstrap values

142 >=95% are labelled with black nodes points. Branches are scaled by mean nucleotide

143 substitutions/site.

144

145 Bayesian phylogenetic reconstruction was used to date the time to most recent common
146 ancestor (TMRCA) for the different TPA lineages. However, temporal analysis of heavily
147 passaged laboratory strains (such as those derived from the original Nichols isolate) is
148 problematic because the true mutational age may be unknown, meaning coalescent date is
149 difficult to infer; we therefore removed extensively passaged strains or strains with no
150 record of their passage history from this particular part of the analysis. This included the
151 removal of the Nichols and SS14 reference strains, as well as the Mexico A strain that
152 delineated the SS14Ω lineage described previously[14]. Root-to-tip regression analysis of the
153 remaining 109 recently clinically derived genomes indicated that TPA possesses a clock-like
154 signal (Supplementary Figure 1), and we performed Bayesian phylogenetic reconstruction
155 and tip date analysis using BEAST[35] under a Strict Constant model, inferring a median
156 molecular clock rate of 2.28 $x10^{-7}$ (95% Highest Posterior Density (HPD) 1.58 $x10^{-7}$ – 3.05
157 $x10^{-7}$), or 0.26 sites/genome/year (meaning we would expect TPA genomes on average to
158 accumulate one fixed SNP every four years). We inferred a temporal timeline for the tree,
159 and our analysis broadly supported previous estimates[14] dating the separation of Nichols-
160 and SS14 at around the mid 18th Century (median date 1755, 95% HPD 1651–1835; Figure
161 2A).

162 <u>SS14-lineage shows a polyphyletic structure</u>

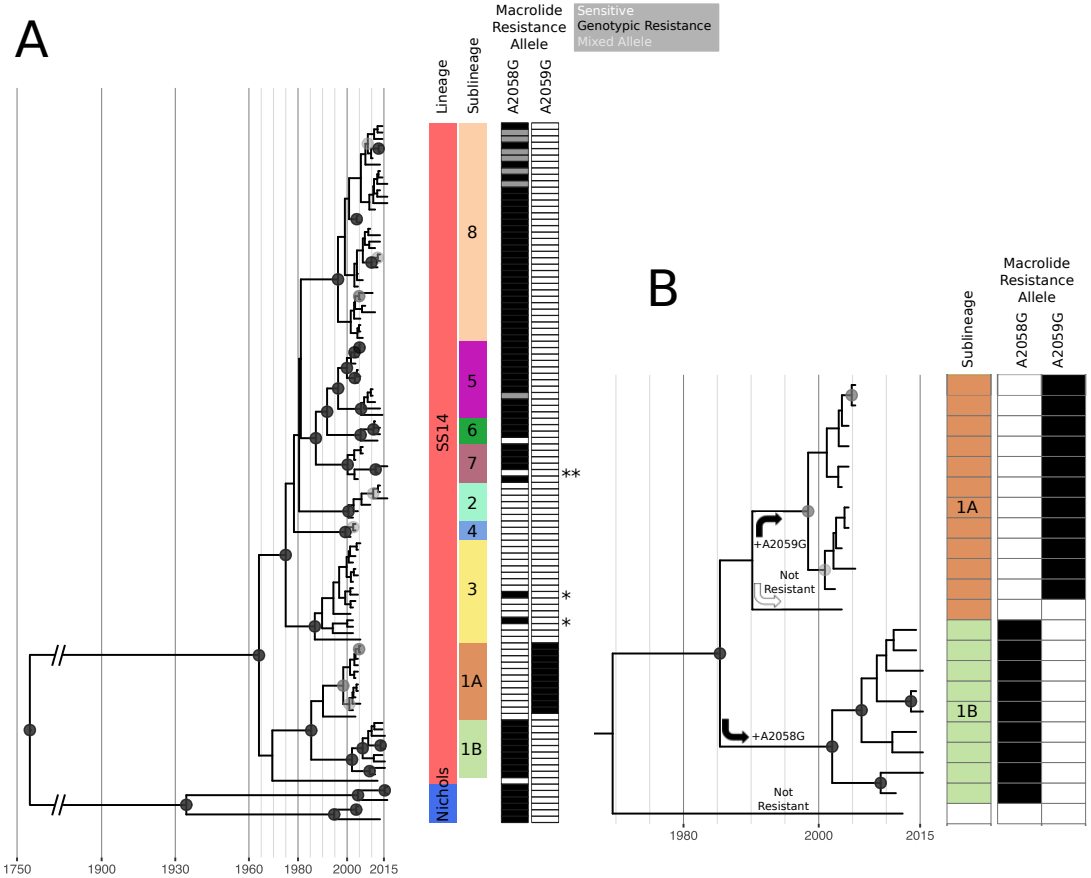
163 Within the SS14-lineage, the high number of full-length sequences enabled fine-scale
164 description of phylogenetic sub-structure. In particular, we show partitioning of the SS14Ω
165 centroid cluster previously defined[14] into two lineages; one composed of European and
166 North American derived samples (SS14Ω-A), and another of Chinese and North American
167 derived samples (SS14Ω-B) (Figure 1). While the American and European samples belonging

168 to the former SS14Ω-A lineage are geospatially admixed, those of the latter SS14Ω-B lineage
169 can be further separated between Chinese and North American samples. These partitions
170 were well supported in our maximum likelihood (Figure 1) and Bayesian phylogenies (Figure
171 2A), as indicated by black node points. We used the rPinecone package[36] to formally classify
172 these sub-lineages based on a defined root-to-tip SNP distance, identifying eight sub-
173 lineages (one of which we further subdivided into sub-lineages 1A and 1B to aid analysis
174 based on temporal and geospatial divergence) within SS14-lineage that correlated well with
175 the population structure described by the phylogeny (Figure 2A). Importantly, while some
176 nodes close to the tips in our phylogeny are unsupported due to small numbers of
177 differentiating SNPs, all sub-lineages defined by rPinecone are supported by >91% posterior
178 support at the key nodes in our Bayesian phylogeny (Figure 2A). We considered the
179 possibility of geographical bias in our data, and whilst four sub-lineages are composed
180 entirely of samples from Seattle (sub-lineages 1A, 3, 4) or China (sub-lineage 1B), four
181 others (sub-lineages 2, 5, 7, 8) are geospatially admixed, sharing samples from both the USA
182 and Europe, with the last (sub-lineage 6) composed of samples from two European
183 countries, Portugal and the Netherlands (Supplementary Figure 2, Supplementary Table 1).

184

185

186


A
Macrolide Resistance Allele
Sensitive
Genotypic Resistance
Mixed Allele
Lineage
Sublineage
A2058G
A2059G
SS14
Nichols
8
5
6
7
2
4
3
1A
1B
**
*
*
1750
1900
1930
1960
1980
2000
2015
B
Macrolide Resistance Allele
Sublineage
A2058G
A2059G
1A
1B
+A2059G
Not Resistant
+A2058G
Not Resistant
1980
2000
2015

188

189 Figure 2. Bayesian maximum credibility phylogeny of sequences recently derived from
190 clinical samples shows expansion of discrete sub-lineages within SS14-lineage, with
191 independent evolution of macrolide resistance. A – Time-scaled phylogeny of all recently
192 clinically derived genomes. Coloured tracks indicate lineage, sub-lineage, and presence of
193 macrolide resistance conferring 23S rRNA SNPs (black=present, white=absent, grey=mixed).
194 Node points are shaded according to posterior support (black ≥96%, dark grey >91%, light
195 grey >80%). *Sporadic (non-lineage associated) gain of resistance is highlighted in sub-
196 lineage 3 (samples UW133B and UW262B). **Possible reversion from resistant to wildtype
197 (sample SW6). B – Expanded view of sub-lineages, showing independent acquisition and
198 fixation of macrolide resistance alleles.

199

200 Macrolide resistance has evolved independently within SS14

201 The molecular basis for macrolide resistance has been well documented in *T. pallidum*, and
202 is mediated by point mutations in the 23S ribosomal RNA gene at nucleotide positions 2058
203 and 2059[37–39]. The A2058G variant was first identified in *T. pallidum* Street Strain-14 (the
204 prototype sample for the SS14-lineage), isolated as long ago as 1977[37], yet resistance has
205 not previously been analysed in context with a detailed whole genome phylogeny. We used
206 ARIBA[40] to perform localised assembly and variant calling of treponema-specific 23S
207 ribosomal sequences from all genomes, and these data were used to infer the presence of
208 both A2058G and A2059G 23S variants that confer macrolide resistance[37]. *T. pallidum*
209 possesses two copies of the 23S ribosomal RNA gene, yet previous analyses of *T. pallidum*
210 have not identified heterozygosity between these two copies (although heterozygosity was
211 described for the distantly related *T. denticola*[41]) – where resistance alleles have been
212 sequenced, they are homozygous between 23S copies, and it has been suggested a gene
213 conversion unification mechanism may exist to facilitate this[19]. Of the 122 genomes, 83
214 showed evidence of genotypic resistance to macrolides, with 76 genomes showing >95%
215 read support for either the A2058G or A2059G variant. Since it is not possible to
216 discriminate between short reads originating from either copy of 23S because they are
217 perfect repeats, this suggests that both copies carry the same resistance mutation. Seven
218 clinically derived genomes (one UK sample from this study, six described by Pinto and
219 colleagues[13]) showed a mixed 23S allelic profile (replicating the result of a previous analysis
220 for three of six samples[13]). All of these samples had >179x read coverage for those sites,
221 with only a fraction of reads (26% - 94%) possessing a resistance allele. In these cases it was
222 not possible to clearly distinguish between a mixture of homozygous positive and negative
223 bacteria in the same patient (either due to within-host evolution or coinfection with

224 multiple strains) or heterozygous sequences from a single bacterium (different 23S rRNA
225 alleles at each copy; heterozygosity in phase) (Figure 2A).
226
227 Within the 122 genomes included in this study, we observed that 67% (70/105) of SS14-
228 lineage samples and 35% (6/17) of Nichols-lineage samples were homozygous for either
229 A2058G or the A2059G 23S rRNA allele. In the SS14-lineage, samples possessed either the
230 A2058G (n=59) or the A2059G (n=11) variant. In the Nichols-lineage, six possessed a
231 resistance allele, of which all showed the A2058G variant, and all were from recent clinically
232 derived samples.
233
234 To explore the emergence of macrolide resistance, we correlated the taxa in our time-scaled
235 phylogeny with the presence of resistance alleles (Figure 2A). We observed a strong
236 correlation between our well supported sub-lineages and genotypic macrolide resistance or
237 sensitivity, such that resistance appears to have evolved on multiple occasions in a stepwise
238 manner (Figure 2A). For example, Figure 2B shows how the wildtype ancestor of sub-lineage
239 1B sequences evolved the A2058G between the late 1980s and late 1990s, contrasting with
240 sub-lineage 1A sequences which did not gain A2058G, but subsequently and independently
241 evolved the A2059G variant.
242
243 More broadly in the phylogeny, we observe that similar independent 23S rRNA mutations
244 have occurred on at least six occasions (Figure 2A), with separate sub-lineages characterised
245 by either predominantly macrolide resistant or sensitive genotypes; six of the nine sub-

246 lineages were predominantly resistant to macrolides, with three sub-lineages being
247 predominantly sensitive. Several of the macrolide sensitive sub-lineages (in particular sub-
248 lineage 3) contained similar numbers of samples as many of the resistant sub-lineages
249 (Figure 3), suggesting ongoing selection for macrolide resistance has not influenced the
250 expansion of these sub-lineages.

251

252 To examine sub-lineage expansion in greater detail, we extracted sampling dates according
253 to sub-lineage, and correlated these data with the predicted time to most recent common
254 ancestor (TMRCA) (Figure 3). Whilst all clinical sequences included in this analysis were
255 sampled after 2000, our analysis indicates that the origins of most of the sub-lineages
256 predated this time and likely arose during the 1990s (Figure 3). Regarding whole sub-lineage
257 associated resistance, we also observed more recent sporadic appearance of resistance
258 mutations, with two separate A2058G variants detected in sub-lineage 3 (an otherwise
259 macrolide sensitive sub-lineage)(Figure 2). We also observe an instance of possible
260 reversion from resistant to wildtype allele in sub-lineage 7, with a wildtype sample (SW6)
261 descending from multiple putatively resistant MRCA nodes; this contrasts with sub-lineages
262 1A and 6, where isolated sensitive samples diverge from the rest of the sub-lineage prior to
263 the likely MRCA in which resistance evolved.

264

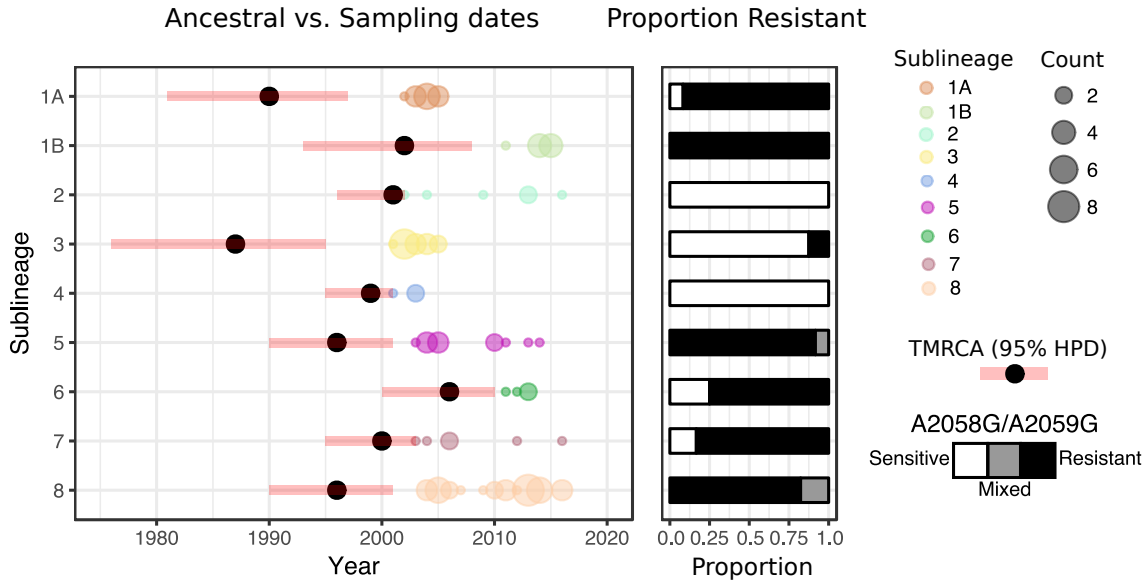


Figure 3. Macrolide resistant and sensitive SS14 sub-lineages evolved independently prior to 2006 and expanded equally regardless of resistance genotype. Figure 3 shows sample collection dates grouped by sub-lineage, with size of circle proportional to number of sequences, and showing predicted time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

272 Penicillin binding proteins are targeted by homoplasic changes

273 Previous analysis of Treponemal genomes have shown amino acid changes in penicillin
274 binding proteins[13,29], but large scale analysis of penicillin gene variants has not previously
275 been performed on a global dataset. We selected three penicillin binding protein genes[42,43]
276 (*pbp1*, *pbp2*, *mrcA*) and a putative β-lactamase[44] (*Tp47*), and used ARIBA to screen these
277 genes for novel variants across our full clinical dataset of 109 sequences, identifying 19 non-
278 synonymous SNPs (Supplementary Figure 3). Two variants (*pbp1* P564L; *Tp47* S394R) were
279 almost universally conserved in clinical SS14-lineage sequences, whilst three more (*mrcA*
280 I487L; *mrcA* M625V; *mrcA* G708S) showed strong association with SS14 sub-lineages. As no
281 phenotypic or clinical evidence exists for penicillin resistance or treatment failure in
282 Treponemes, we considered such widely prevalent SNPs unlikely to be of clinical relevance
283 or mediate a resistant phenotype.

284

285 Sun and colleagues[29] identified variants in *pbp2* (*pbp2* I415F/I415M) and *mrcA* (*mrcA*
286 A506V) in eight genomes from China, and whilst we confirmed the presence of the *pbp2*
287 variants, we did not find them in samples from other countries. However, in addition to the
288 *mrcA* A506V variant described by Sun et al, we also identified a SNP affecting the same
289 amino acid position (*mrcA* A506T), both in the recently clinically derived Nichols-lineage
290 genomes as well as in recently clinically derived SS14 sub-lineage 1A (comprising samples
291 from the USA). As with macrolide resistance, independent evolution of homoplasic variants
292 affecting the same amino acid are suggestive of selection, and we used SIFT[45] to infer the
293 likely functional impact of amino acid changes at *mrcA* A506 and *pbp2* I415. None of the

294 four variants detected showed a significant result (no SIFT scores were below the threshold
295 for functional effect of 0.05), leaving the clinical relevance of these SNPs unclear.

296

297 Discussion

298 Compiling the largest *Treponema pallidum* subspecies *pallidum* sequence collection to date,
299 we show that the majority of the contemporary samples sequenced here were from the
300 SS14-lineage, consistent with other reports[14]. However, three of eight recent UK genomes
301 (38%) belonged to the Nichols-lineage, showing that these two lineages are still circulating
302 concurrently, and that the prevalence of Nichols-lineage strains may vary by sampling
303 population. Moreover linking the phylogenetic and geospatial data of our samples shows
304 that a number of SS14 sub-lineages (2, 5, 7, 8) are comprised of both European and North
305 American samples (Supplementary Figure 2), suggesting that we have captured key lineages
306 circulating in these regions and that our data is not overly biased from localised sampling.
307 Notwithstanding the above, further studies will be needed to more accurately characterise
308 the genomic diversity and prevalence of lineages within different populations on a global
309 scale.

310 We were able to reconstruct a time-scaled phylogeny using only recently clinically derived
311 samples, and the increase in whole genome sequence numbers combined with the removal
312 of heavily passaged samples contrasts with previous approaches[14,29]. Arora *et al*. reported a
313 mean evolutionary rate of 6.6 $\times 10^{-7}$ substitutions/site/year for *T. pallidum*[14], comparable
314 with rapidly dividing, free-living bacterial pathogens with environmental life cycles such as
315 *Vibrio cholerae* (6.1 $\times 10^{-7}$)[46] and *Shigella sonnei* (6.0 $\times 10^{-7}$). However, *T. pallidum* is a slow

316 growing host-restricted pathogen with substantial periods of latency, and as such we would
317 expect a molecular clock rate more similar to other slow growing organisms such as
318 *Chlamydia trachomatis* (2.15 x10^{-7})[47] or *Mycobacterium tuberculosis* (1 ×10^{-7})[48]. Our
319 inferred rate for TPA (2.28 x10^{-7}) is consistent with this expectation, as well as with other
320 observations that suggest *T. pallidum* has a low evolutionary rate[49].

321

322 Within the SS14-lineage, we defined nine well supported sub-lineages that all diverged from
323 their most recent common ancestors prior to 2006, with the earliest (sub-lineage 3)
324 potentially emerging at the end of the 1980s. We observed clear associations between
325 these nine sub-lineages and the presence of macrolide resistance conferring SNPs, with
326 each sub-lineage dominated by either macrolide resistant (six sub-lineages) or macrolide
327 sensitive (three sub-lineages) samples. Such observations, previously documented to a
328 limited extent within a single country[13] but now extended to a broader sample, are not
329 consistent with a hypothesis of an ancestrally resistant SS14-lineage driven to high
330 frequency in the population due to a fitness advantage conferred by macrolide resistance,
331 where we would expect to see expansion of a single resistant lineage. Rather, we see
332 evidence of multiple sub-lineages independently evolving macrolide resistance alleles, as a
333 likely consequence of intermittent selective pressure from macrolide treatment, consistent
334 with molecular typing data from Seattle[21]. Phylogenetic reconstruction shows *de novo*
335 evolution of macrolide resistance in syphilis is not a rare event, and furthermore, when
336 resistance evolves in a lineage, it persists in descendants, resulting in transmission from
337 person to person. That the variants appear stable within lineages, with only a single instance
338 in the phylogeny that might represent reversion to a wildtype state, suggests that either

339 there is no strong fitness cost associated with possessing these macrolide resistance
340 mutations or that novel compensatory variants co-evolve to mitigate any fitness cost,
341 enabling variant stability; further sampling would be necessary in the future to address this.

342

343 Although our data strongly suggest that global expansion of the SS14-lineage is not
344 contingent on macrolide resistance, the global increases seen in macrolide resistant
345 syphilis[19], as well as the number of resistant lineages emerging in our data, are a cause for
346 concern. Macrolides such as azithromycin are not considered appropriate treatment for
347 syphilis, with WHO and US guidelines recommending treatment with BPG[50,51], with
348 doxycycline recommended as a secondary treatment option. In contrast, WHO now
349 recommends azithromycin rather than BPG as the treatment of choice for mass drug
350 administration in the eradication of yaws[25], caused by the closely related *T. pallidum*
351 subspecies *pertenue*. Macrolide resistance has recently been described in yaws[26] and our
352 data suggest that further independent evolution of azithromycin resistance is highly likely,
353 which would have significant implications for yaws eradication efforts. Worryingly, applying
354 azithromycin-based mass drug administration to populations infected with both TPP and
355 TPA could promote resistance in both species.

356

357 Several factors are likely to have driven the repeated evolution and subsequent expansion
358 of macrolide resistant lineages of TPA. Use of azithromycin, or other macrolides, for other
359 indications is likely to have played a significant role[52]. Azithromycin entered global markets
360 between 1988 and 1991 (marketed by Pfizer as Zithromax), and became one of the most
361 widely used antibiotics in the United States for a wide variety of indications[53], including the

362 treatment of respiratory tract infections and for the treatment of other sexually transmitted
363 infections. In many cases, the dose used for treatment of these indications is lower than the
364 recommended dose for the treatment of syphilis. Azithromycin and clarithromycin were
365 also widely used prophylactically amongst individuals living with HIV prior to the widespread
366 availability of combined anti-retroviral therapy. Off-target macrolide exposure is of
367 particular concern because azithromycin has a long half-life[54] and may persist at subclinical
368 concentrations in patients. Widespread use of macrolides for this broad range of indications
369 might therefore have contributed to sub-therapeutic exposure of patients with incubating
370 or early syphilis and ultimately selection of resistance[52].

371

372 The recent increase in incidence of syphilis in high income countries likely reflects changes
373 in sexual behaviour[8]. The fact that we observe the expansion of both genotypically resistant
374 and sensitive lineages highlights particular treatment issues. There have been significant
375 global shortages of BPG[55] in low, middle, and high income nations, including the United
376 States, with the global supply of BPG dependent on just three manufacturers of the active
377 ingredient[55]. Pharmaceutical production of sterile, injectable β-lactam-derived
378 antimicrobials such as BPG is costly, yet as an older off-patent medication with declining
379 demand in the face of growing antimicrobial resistance (AMR) in other organisms, the
380 financial rewards for production are low[55]. This may be compounded by a misperception
381 that BPG is an outdated drug that could be replaced by newer, more effective drugs[55]. In
382 circumstances where azithromycin is used instead of doxycycline, a shortage of BPG leads
383 inevitably to inadequate treatment of early infectious syphilis and contributes to ongoing,
384 unchecked transmission. In China for example, studies have reported that despite high rates

385 of macrolide resistance,[20] clinicians have inappropriately been resorting to macrolide
386 treatment due to ongoing BPG shortages[56]. Thus although a well-established, highly
387 effective treatment for syphilis (BPG) has been available since the mid-1950s, shortages in
388 the present era contribute to suboptimal treatment strategies and continued use of drugs
389 with a known resistance problem.

390

391 Phenotypic or clinical penicillin resistance has never been observed in syphilis in more than
392 60 years of clinical use. In our analysis, we detected a number of variants affecting penicillin
393 binding proteins, although given that penicillin remains an effective treatment, many were
394 so widely distributed throughout the phylogeny as to be implausible for causing resistance.

395

396 The epidemic of syphilis and the widespread problems of azithromycin resistance and BPG
397 shortage require a multi-faceted response. This includes new strategies for treatment and
398 reduction of transmission, finding ways to improve the security of the BPG supply chain, and
399 strengthening molecular surveillance for antimicrobial resistance in *T. pallidum*[57]. Many
400 authors have discussed the importance of rethinking the economics of antimicrobial
401 development pipelines to ensure we are still able to treat infections[24,58]. In syphilis, we must
402 rethink how we can protect the continued production of existing highly efficacious
403 penicillins in the face of increasing antimicrobial resistance rendering them ineffective for
404 other organisms, especially in the light of the recent increases in syphilis incidence in
405 Europe, North America, and Asia.

406

407 ## Materials and Methods

408 Samples

409 UK samples consisted of residual DNA, extracted from clinical swabs using a QIAsymphony
410 (Qiagen) from routine diagnostic samples obtained from patients presenting with clinical
411 evidence of syphilis at the Mortimer Market Centre, London. Use of the UK samples was
412 approved by the NHS Research Ethics Committee (IRAS Project ID 195816). US samples
413 from Seattle were collected from individuals enrolled in a study of cerebrospinal fluid
414 abnormalities in patients with syphilis (a mixture of 21 patients with evidence of CSF
415 infection and 38 patients without), with ethical approval at the University of Washington
416 (UW IRB # STUDY00003216). Specifically, 2.4-3.0 ml participant blood was inoculated into
417 rabbit testes as previously described[59], and *T. pallidum* suspensions were collected after the
418 second round of passage. Historical strains were propagated in rabbits and harvested from
419 infected testes. *T. pallidum* suspensions were treated using a lysis buffer (10mM Tris pH 8.0,
420 0.1M EDTA pH 8.0, 0.5% SDS), freeze-thaw, and extraction using QIAamp Mini kit (Qiagen)
421 according to the manufacturer's instructions; in select cases the proteinase K incubation
422 was extended overnight to improve DNA yield. In some cases, samples had been used for
423 previous molecular tests, and in these cases residual DNA extracts were used for
424 sequencing. Treponemal DNA was quantified using a qPCR targeting the *Tp47*
425 (TPANIC_0574) gene that is conserved across all known members of the *T. pallidum* cluster,
426 and compared to a standard curve derived from a plasmid containing the PCR amplicon.
427 Samples with a concentration >2000 genome copies/μl were selected for sequencing;
428 borderline samples with high volume and a pathogen load over 500 genome copies/μl were
429 concentrated using a vacuum centrifuge. Samples were arranged in groups of 20 according

430 to similar (within 2 C_T) treponemal load, with high concentration outlier samples diluted as
431 necessary. We added 4μl pooled commercial human gDNA (Promega) to all samples to
432 ensure total gDNA > 1μg/35μl, sufficient for library prep.

433 Sequencing

434 Genomic DNA was sheared to 100-400bp (average size distribution ~ 150bp) by
435 ultrasonication (LE220, Covaris Inc), followed by library preparation (NEBNext Ultra II DNA
436 Library prep Kit, New England Biolabs), adaptor ligation and index barcoding (Sanger 168 tag
437 set). Index tagged samples were amplified (6 cycles of PCR, KAPA HiFi kit, KAPA Biosystems),
438 quantified (Accuclear dsDNA Quantitation Solution, Biotium) , then pooled in the
439 preassigned groups of 20 to generate equimolar Total DNA pools. 500 ng pooled material
440 was then hybridised using 120-mer RNA baits (SureSelect Target enrichment system ,
441 Agilent technologies) designed against published RefSeq examples of *T. pallidum* and *T.*
442 *paraluiscuniculi* (Bait design ELID ID 0616571) as described previously[15]. Libraries were
443 subjected to 125bp paired end sequencing on Illumina HiSeq 2500 with version 4 chemistry
444 according to established protocols. This sequencing method has been previously
445 described[15], and differs from that of others[13,14,16], in that samples are barcoded and pooled
446 before enrichment thus enabling multiplexing of hybridisation reactions. Raw sequencing
447 reads were deposited at the European Nucleotide Archive (ENA) under project PRJEB20795;
448 all accessions used in this project are listed in Supplementary Table 1.

449 Sequence analysis and phylogenetics

450 Treponemal sequencing reads were prefiltered using a Kraken[60] v0.10.6 database containing
451 all bacterial and archaeal nucleotide sequences in RefSeq, plus mouse and human, to
452 identify and extract those reads with homology to Treponema species, followed by adaptor

453 trimming using Trimmomatic[61] v0.33. To reduce bias due to variable read depth, as well as
454 make analysis computationally tractable, for samples with high read counts we used seqtk
455 v1.0 (available at https://github.com/lh3/seqtk) to randomly down-sample the binned and
456 trimmed reads to 2,500,000 unique treponemal read pairs. For publicly available genomes,
457 raw sequencing reads were downloaded from SRA and subjected to the same binning and
458 down-sampling pipeline. For eight public genomes (see Supplementary Table 1), raw
459 sequencing reads were not available; for these we simulated 125bp PE perfect reads from
460 the RefSeq closed genomes using Fastaq (available at https://github.com/sanger-
461 pathogens/Fastaq).

462

463 For phylogenetic analysis, we used a reference mapping approach with a custom version of
464 the SS14 v2 reference sequence (NC_021508.1) from which we first masked 14 highly
465 repetitive or recombinogenic genes (12 repetitive Tpr genes A-L, arp and TPANIC_0470)
466 using bedtools[62] v2.17.0 'maskfasta'. We mapped prefiltered sequencing reads to the
467 reference using BWA mem[63] v0.7.17 (MapQ ≥ 20), followed by indel realignment using
468 GATK[64] v3.4-46 IndelRealigner, deduplication with Picard MarkDuplicates v1.127 (available
469 at http://broadinstitute.github.io/picard/), and variant calling and consensus
470 pseudosequence generation using using samtools v1.2[65] and bcftools v1.2, requiring a
471 minimum of three supporting reads per strand and eight in total to call a variant, and a
472 variant frequency/mapping quality cut-off of 0.8; sites not meeting these criteria were
473 masked to 'N' in the pseudosequence. Importantly, reads mapping to multiple positions
474 were marked and excluded from SNP calling, meaning repeated regions such as the

475 duplicated 23S genes was not included in the multiple sequence alignment used to derive
476 the phylogenies.

477

478 Multiple sequence alignments were screened for evidence of recombination using
479 Gubbins[66], generating recombination-masked full genome length and SNP-only alignments.
480 Maximum likelihood phylogenies were calculated on SNP-only alignments using IQ-Tree
481 v1.6.3[67], correcting for missing constant sites using the built in ascertainment bias
482 correction[68], allowing the built-in model testing[69] to determine a K3P (three substitution
483 types model and equal base frequencies) substitution model[70] with a FreeRate model of
484 heterogeneity[71] assuming three categories, and performing 1000 UltraFast Bootstraps[72,73].

485

486 To determine SS14 sub-lineages, we recalculated a maximum likelihood tree as described
487 above for SS14-clade sequences only (using the Mexico A strain, NC_018722.1 as outgroup),
488 before performing a joint ancestral reconstruction[74] of SNPs on the tree branches using
489 pyjar (available at https://github.com/simonrharris/pyjar). We then used the rPinecone
490 package[36] (available at https://github.com/alexwailan/rpinecone), which applies a root-to-
491 tip approach to defining clusters based on SNP distance relative to ancestral nodes. We used
492 a cluster threshold of 10 SNPs, which proved optimal for describing the underlying
493 phylogenetic structure of the tree, and yielded eight sub-lineages. Within sub-lineage 1, we
494 found that pinecone clusters did not accurately represent the phylogeny, despite a clear
495 phylogenetic separation, with one group of sequences from China associated with the
496 A2058G allele, and the other group from the USA associated with the A2059G allele. We

497 further manually clustered these sequences according to their shared ancestral nodes,
498 naming them sub-lineages 1A and 1B.
499
500 We evaluated our maximum likelihood phylogeny for evidence of temporal signal using
501 TempEst[75] v1.5, and this showed a correlation of 0.44 and R^2 of 0.20 for the whole tree
502 (Supplementary Figure 1), whilst the SS14-lineage -only alignment showed a correlation of
503 0.66 and R^2 of 0.44; this indicated that there was sufficient evidence for temporal signal and
504 we proceeded to BEAST analysis. BEAST[35] v1.8.2 was initially run on a recombination-
505 masked SNP-only alignment containing 284 variable sites, applying a correction for invariant
506 sites using the constantPatterns argument, in triplicate using an Uncorrelated Relaxed Clock
507 model, assuming constant population size, lognormal population distribution, GTR
508 substitution model, diffuse gamma distribution prior (shape 0.001, scale 1000), with a
509 burnin of 10 million cycles followed by 100 million MCMC cycles. All MCMC chains
510 converged, and on inspection of the marginal distribution of ucld.stdev we could not reject
511 a Strict Clock. We therefore repeated the analysis using a Strict Clock model, using the same
512 models and priors and assuming a starting molecular clock rate of 3.6×10^{-4} as described by
513 others[14]. We used the Marginal Likelihood Estimates from the triplicate BEAST runs as input
514 to Path Sampling and Stepping Stone Sampling analyses[76,77] and determined that the Strict
515 Constant model was optimal for this dataset. To further confirm the temporal signal in our
516 tree, we used the TIPDATINGBEAST package[78] in R[79] to resample tip-dates from our
517 alignment, generating 20 new datasets with randomly assigned dates – BEAST analysis using
518 the same Strict Clock prior conditions found no evidence of temporal signal in these

519 replicates, indicating that the signal in our tree was not found by chance (Supplementary
520 Figure 4).

521

522 Macrolide resistant SNPs were inferred using ARIBA[40], which performs localised assembly
523 and mapping in comparison with a custom reference database containing 23S sequences
524 from Nichols (NR_076156.1) and SS14 reference strains (NR_076531.1). Non-synonymous
525 penicillin binding protein gene variants in *pbp1* (TPANIC_0500), *pbp2* (TPANIC_0760), *mrcA*
526 (TPANIC_0705) and *Tp47* (TPANIC_0574) were inferred using a second ARIBA database
527 containing the corresponding amino acid sequences from the Nichols genome
528 (NC_021490.2). Functional inference of amino acid changes was performed by inputting the
529 gene amino acid sequences into the SIFT[45] server (available at http://sift.bii.a-star.edu.sg).
530 The treponema-specific *Tp47* protein (TPANIC_0574) did not return a result from SIFT due to
531 insufficient sequences being found during the PSIBLAST step.

532

533 Processing of data, and all statistical analysis was performed in R[79] v3.4.1, primarily using
534 the phytools and ape packages. Phylogenies were plotted using ggtree[80], and figures were
535 produced using ggplot2[81]. All code used in the downstream analysis is available in
536 Supplementary File 4, as well as being available along with source data at
537 doi:10.6084/m9.figshare.7688093.

538

539 ## Acknowledgements

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

549

550 ## Author Contributions

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

555

556 ## Competing Interests

557 The authors have no competing interests to declare.

558

559 ## References

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

562 2. Chesson, H. W., Dee, T. S. & Aral, S. O. AIDS mortality may have contributed to the
563 decline in syphilis rates in the United States in the 1990s. *Sex. Transm. Dis.* **30**, 419–424
564 (2003).

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

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

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

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

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

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

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

579 10. Lafond, R. E. & Lukehart, S. A. Biological basis for syphilis. *Clin. Microbiol. Rev.* **19**, 29–49
580 (2006).

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

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

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

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

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

592 16. Knauf, S. *et al.* Nonhuman primates across sub-Saharan Africa are infected with the
593 yaws bacterium Treponema pallidum subsp. pertenue. *Emerg. Microbes Infect.* **7**, 157
594 (2018).

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

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

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

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

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

608 22. Grillová, L. *et al.* Molecular Typing of Treponema pallidum in the Czech Republic during
609 2011 to 2013: Increased Prevalence of Identified Genotypes and of Isolates with
610 Macrolide Resistance. *J. Clin. Microbiol.* **52**, 3693–3700 (2014).

611 23. Vaulet, L. G. *et al.* Molecular typing of Treponema pallidum isolates from Buenos Aires,
612 Argentina: Frequent Nichols-like isolates and low levels of macrolide resistance. *PLOS*
613 *ONE* **12**, e0172905 (2017).

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

616 25. WHO. Eradication of yaws – the Morges Strategy. *Wkly Epidemiol Rec* **87**, 189–194
617 (2012).

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

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

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

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

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

628 31. Pětrošová, H. *et al.* Resequencing of Treponema pallidum ssp. pallidum Strains Nichols
629 and SS14: Correction of Sequencing Errors Resulted in Increased Separation of Syphilis
630 Treponeme Subclusters. *PLOS ONE* **8**, e74319 (2013).

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

633 33. Čejková, D. *et al.* Whole Genome Sequences of Three Treponema pallidum ssp.
634 pertenue Strains: Yaws and Syphilis Treponemes Differ in Less than 0.2% of the Genome
635 Sequence. *PLoS Negl. Trop. Dis.* **6**, e1471 (2012).

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

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

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

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

646 38. Matějková, P. *et al.* Macrolide treatment failure in a case of secondary syphilis: a novel
647 A2059G mutation in the 23S rRNA gene of Treponema pallidum subsp. pallidum. *J. Med.*
648 *Microbiol.* **58**, 832–836 (2009).

649 39. Molini, B. J. *et al.* Macrolide Resistance in Treponema pallidum Correlates With 23S
650 rDNA Mutations in Recently Isolated Clinical Strains. *Sex. Transm. Dis.* **43**, 579–583
651 (2016).

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

654 41. Lee, S. Y., Ning, Y. & Fenno, J. C. 23S rRNA point mutation associated with erythromycin
655 resistance in Treponema denticola. *FEMS Microbiol. Lett.* **207**, 39–42 (2002).

656 42. Fraser, C. M. *et al.* Complete Genome Sequence of Treponema pallidum, the Syphilis
657 Spirochete. *Science* **281**, 375–388 (1998).

658 43. Stamm, L. V. Global Challenge of Antibiotic-Resistant Treponema pallidum. *Antimicrob.*
659 *Agents Chemother.* **54**, 583–589 (2010).

660 44. Cha, J. Y., Ishiwata, A. & Mobashery, S. A Novel β-Lactamase Activity from a Penicillin-
661 binding Protein of Treponema pallidum and Why Syphilis Is Still Treatable with Penicillin.
662 *J. Biol. Chem.* **279**, 14917–14921 (2004).

663 45. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function.
664 *Nucleic Acids Res.* **31**, 3812–3814 (2003).

665 46. Weill, F.-X. *et al.* Genomic history of the seventh pandemic of cholera in Africa. *Science*
666 **358**, 785–789 (2017).

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

670 48. Roetzer, A. *et al.* Whole Genome Sequencing versus Traditional Genotyping for
671 Investigation of a Mycobacterium tuberculosis Outbreak: A Longitudinal Molecular
672 Epidemiological Study. *PLOS Med.* **10**, e1001387 (2013).

673 49. Strouhal, M. *et al.* Complete genome sequences of two strains of Treponema pallidum
674 subsp. pertenue from Ghana, Africa: Identical genome sequences in samples isolated
675 more than 7 years apart. *PLoS Negl. Trop. Dis.* **11**, e0005894 (2017).

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

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

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

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

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

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

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

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

693 58. WHO. Global action plan on antimicrobial resistance. (2015).

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

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

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

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

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

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

707 65. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
708 2078–2079 (2009).

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

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

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

717 69. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Haeseler, A. von & Jermiin, L. S.
718 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods*
719 **14**, 587–589 (2017).

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

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

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

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

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

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

734 76. Baele, G. & Lemey, P. Bayesian evolutionary model testing in the phylogenomics era:
735 matching model complexity with computational efficiency. *Bioinformatics* **29**, 1970–
736 1979 (2013).

737 77. Baele, G. *et al.* Improving the accuracy of demographic and molecular clock model
738 comparison while accommodating phylogenetic uncertainty. *Mol. Biol. Evol.* **29**, 2157–
739 2167 (2012).

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

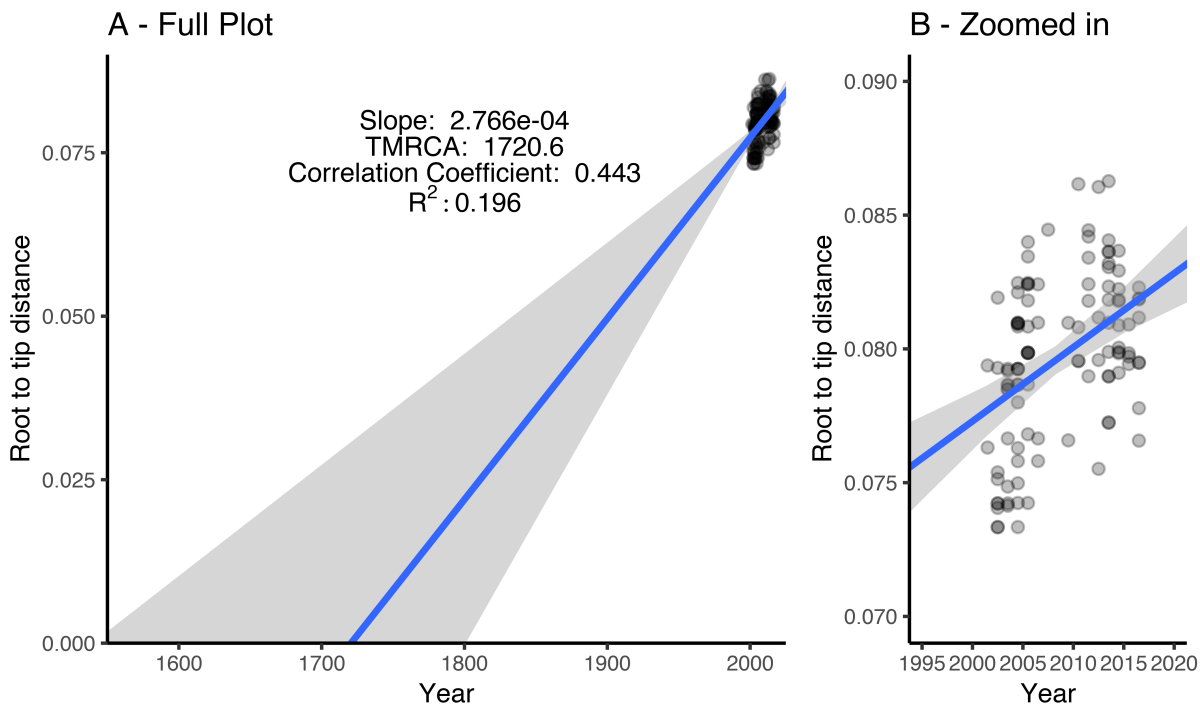
742 79. R Core Team. *R: A Language and Environment for Statistical Computing.* (R Foundation
743 for Statistical Computing, 2014).

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

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

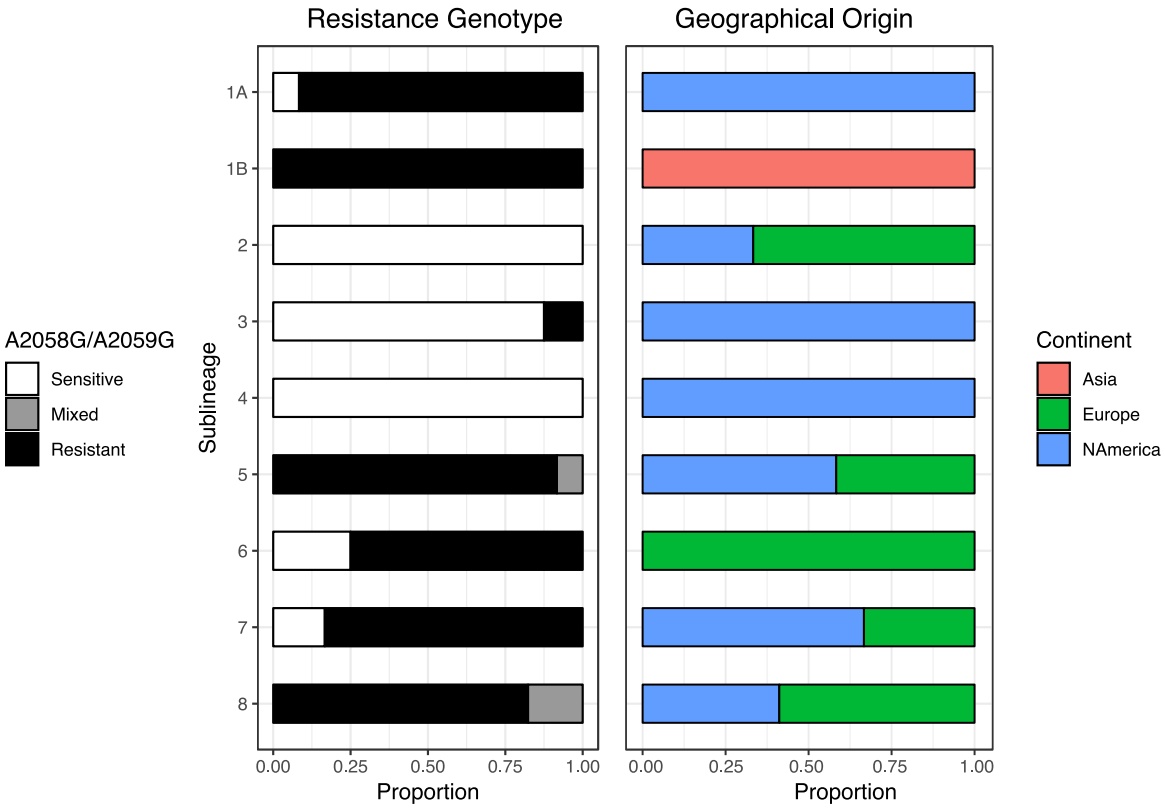
Supplementary Figures

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 colours indicating multiple overlapping points.



Supplementary Figure 2. Distribution of macrolide resistance genotype and geographical origin according to SS14 sub-lineage of recently clinically derived genomes. Four sub-lineages (one comprising sensitive samples, three comprising predominantly resistant samples) are geospatially admixed, containing sequences from both North America and Europe. Three other sub-lineages are composed entirely of samples from Seattle (USA,

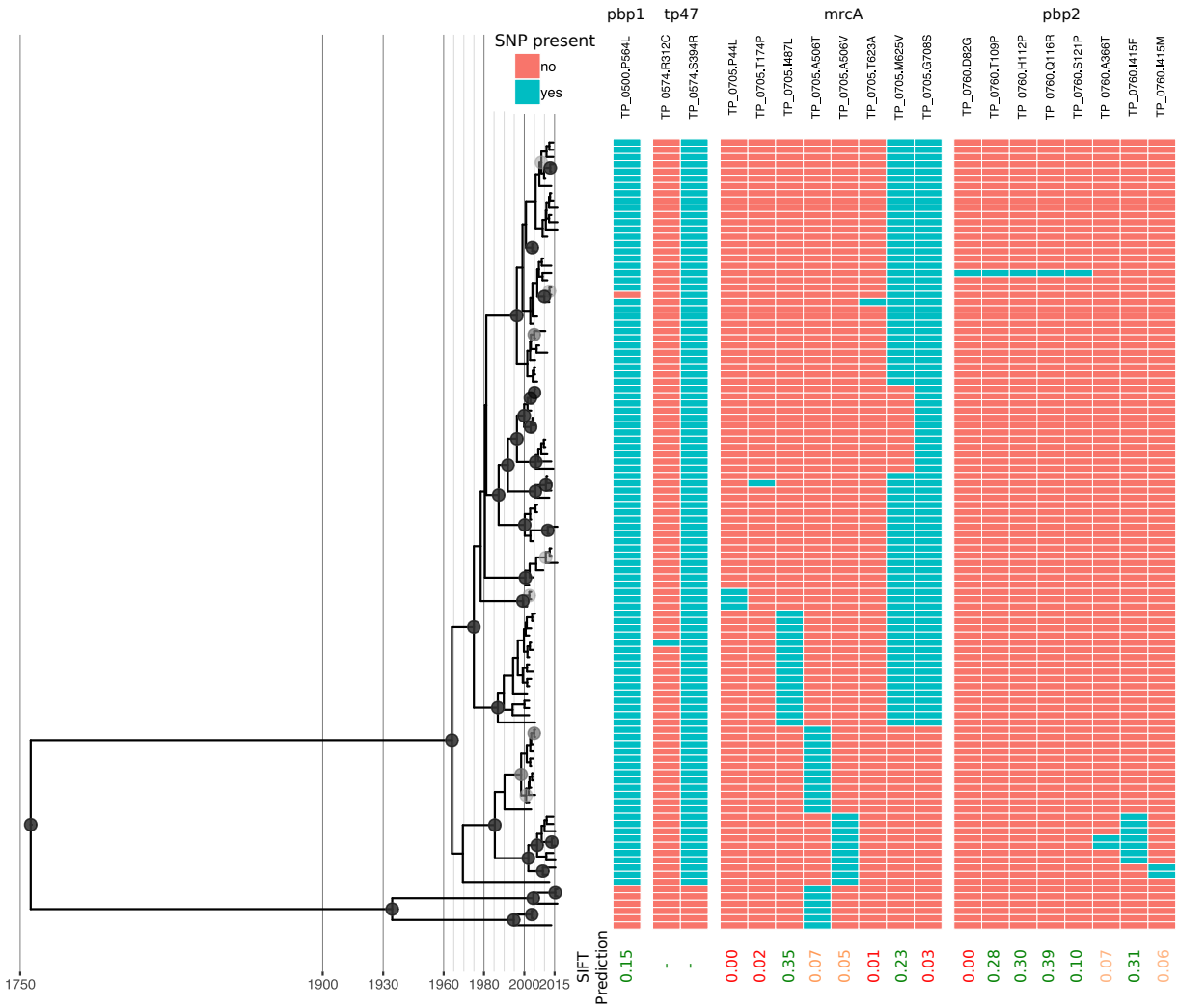
North America) including two predominantly macrolide sensitive lineages and one genotypically resistant lineage, as is one resistant lineage from China (Asia).



Supplementary Figure 3. Bayesian maximum credibility phylogeny of sequences recently derived from clinical samples with non-synonymous SNPs predicted in penicillin binding protein genes *pbp1* (TPANIC_0500), *pbp2* (TPANIC_0760), *mrcA* (TPANIC_0705) and *Tp47* (TPANIC_0574). SNPs are relative to the Nichols reference sequence (NC_021490.2), and blue indicates a SNP is present. SIFT predictions of functional impact are shown for all genes except *Tp47*, where scores below 0.05 (coloured red) indicate a likely functional impact of the amino acid change; variants above this threshold but below 0.1 are indicated in orange.

774 Many SNPs are conserved by lineage, with some isolated *de novo* mutation. Within *mrcA*,

775 amino acid position A506 is targeted by three independent non-synonymous SNPs.



776

777

778

779

780 Supplementary Figure 4. Tip date resampling analysis performed using twenty datasets with

781 randomised tip dates generated from the original Strict Clock analysis and run in BEAST

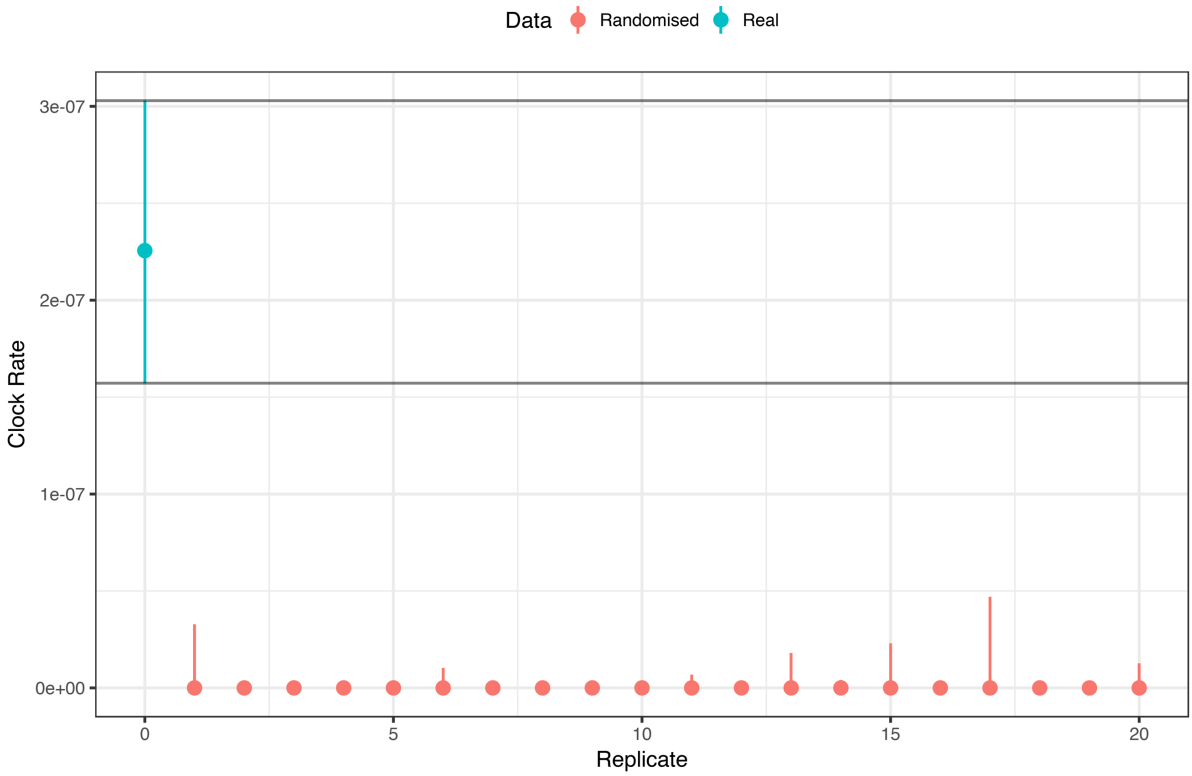
782 under the same conditions. Median clock rate for the real tree was 2.26 x 10^{-7}, whilst all

783 randomly assigned datasets gave substantially lower clock rates, with the highest median

784 clock rate obtained at 1.90 x 10^{-12}. This indicates that the temporal signal observed in our

785 tree was not obtained by chance, and provides further evidence for a temporal signal in the

786 multiple sequence alignment.



787

788

789 Supplementary Table 1. Full sample metadata (Excel Sheet) for this study, including

790 sequence naming used in this paper, in other publications, and on GenBank, ENA Accessions

791 for all new genomes, as well as results of lineage and sub-lineage typing and inference of

792 genotypic macrolide resistance.

793

794 Supplementary Table 2. List of genomic regions behaving in a non-clocklike manner and

795 masked due to hypervariable, recombining or repetitive elements.

Supplementary File 1. Rnotebook (HTML format) containing all downstream code used to generate primary figures and statistics.