A public database for the new MLST scheme for *Treponema pallidum* subsp. *pallidum*: surveillance and epidemiology of the causative agent of syphilis

Linda Grillova [Corresp., 1] , **Keith Jolley** [2] , **David Šmajs** [3] , **Mathieu Picardeau** [1]

[1] Department of Microbiology, Institut Pasteur, Paris, France

[2] Department of Zoology, University of Oxford, Oxford, United Kingdom

[3] Department of Microbiology, Masaryk University, Brno, Czech Republic

Corresponding Author: Linda Grillova
Email address: linda.grillova@pasteur.fr

Treponema pallidum subsp. *pallidum* is the causative agent of syphilis, a sexually transmitted disease with worldwide prevalence. Several different molecular typing schemes are currently available for this pathogen. To enable population biology studies of the syphilis agent and for epidemiological surveillance at the global scale, a harmonized typing tool needs to be introduced. Recently, we published a new Multi-Locus Sequence Typing (MLST) with the potential to significantly enhance the epidemiological data in several aspects (e.g., distinguishing genetically different clades of syphilis, subtyping inside these clades, and, finally, distinguishing different subspecies of non-cultivable pathogenic treponemes). In this short report, we introduce the PubMLST database for treponemal DNA data storage and for assignments of allelic profiles and sequencing types. Moreover, we have summarized epidemiological data of all treponemal strains (n=358) with available DNA sequences in typing loci and found several association between genetic groups and characteristics of patients. This study proposes the establishment of a single MLST of *Treponema pallidum* subsp. *pallidum* and encourages researchers and public health communities to use this PubMLST database as a universal tool for molecular typing studies of the syphilis pathogen.

1   **A public database for the new MLST scheme for Treponema**
2   **pallidum subsp. pallidum: surveillance and epidemiology of**
3   **the causative agent of syphilis**

4

5   Linda Grillová[1], Keith A. Jolley[2], David Šmajs[3], Mathieu Picardeau[1]

6

7   [1] Department of Microbiology, Institut Pasteur, Paris, France

8   [2] Department of Zoology, University of Oxford, Oxford, UK

9   [3] Department of Biology, Masaryk University, Brno, Czech Republic

10

11   Corresponding Author:

12   Linda Grillová[1]

13   28 Rue du Docteur Roux, Paris, 75024, France

14   Email address: lindagrillova@gmail.com

15 **ABSTRACT**

16 *Treponema pallidum* subsp. *pallidum* is the causative agent of syphilis, a sexually transmitted
17 disease with worldwide prevalence. Several different molecular typing schemes are currently
18 available for this pathogen. To enable population biology studies of the syphilis agent and for
19 epidemiological surveillance at the global scale, a harmonized typing tool needs to be
20 introduced. Recently, we published a new Multi-Locus Sequence Typing (MLST) with the
21 potential to significantly enhance the epidemiological data in several aspects (e.g., distinguishing
22 genetically different clades of syphilis, subtyping inside these clades, and, finally, distinguishing
23 different subspecies of non-cultivable pathogenic treponemes). In this short report, we introduce
24 the PubMLST database for treponemal DNA data storage and for assignments of allelic profiles
25 and sequencing types. Moreover, we have summarized epidemiological data of all treponemal
26 strains (n=358) with available DNA sequences in typing loci and found several association
27 between genetic groups and characteristics of patients. This study proposes the establishment of
28 a single MLST of *Treponema pallidum* subsp. *pallidum* and encourages researchers and public
29 health communities to use this PubMLST database as a universal tool for molecular typing
30 studies of the syphilis pathogen.

INTRODUCTION

Syphilis is a bacterial infection caused by *Treponema pallidum* subsp. *pallidum* (TPA). Despite the availability of effective treatments, the annual worldwide incidence recently reached more than 5.6 million cases (WHO, 2016).

The proper discrimination of strains of infectious pathogens is crucial for epidemiological and surveillance analyses, for description of population structure and dynamics and for the improvement of public health control strategies. More specifically, the association of specific strains of TPA with different groups of patients can result in a better understanding of syphilis epidemiology. Association of macrolide resistance/sensitivity and allelic profiles are particularly important for a better understanding of the emergence of antibiotic-resistant strains. Finally, the molecular typing of syphilis-causing strains can help to determine syphilis diagnosis, especially in seronegative patients and cases with atypical symptoms (Grange et al., 2016; Noda et al., 2018).

Multi-Locus Sequence Typing (MLST) is typically based on amplification and sequencing of about seven housekeeping gene fragments (loci). For each locus, the different sequence variants are assigned as distinct alleles and, for each isolate, the combination of alleles defines the allelic profile and sequence type (ST). Allelic profiles and/or STs are used for definition of the strains and these can be clustered into clonal complexes based on shared alleles (Maiden et al., 2013). The traditional MLST approach is based on allele-based comparisons, where each allelic change is counted as a single event, regardless of the number of nucleotide changes involved. This helps to mitigate the effects of horizontal genetic transfer in organisms prone to recombination. The MLST data are commonly stored in publicly available and curated databases, such as PubMLST, which uses the Bacterial Isolate Genome Sequence Database (BIGSdb) platform (Jolley & Maiden, 2010). These databases can be used to automatically identify and assign new allele variants and sequence types, and provide tools to analyze the origin and phenotypic metadata of samples with the genotypic information of the typed pathogens. Several functions are available within this software to analyze and visualize epidemiological data (Jolley & Maiden, 2010; Maiden et al., 2013). BIGSdb is currently hosting over 100 different bacterial species and is one of the most widely used tools for epidemiological purposes (Jolley et al., 2018).

Recently, we published a new MLST scheme of TPA (Grillová et al, 2018, Pospíšilová et al. 2018). The TPA MLST scheme is somewhat unusual compared to typical MLST schemes reflecting the special characteristics of TPA. The TPA MLST uses three highly variable loci and performs sequence-based analyses, i.e., it takes into account sequence differences among alleles.

i.) Even though some recent attempts to cultivate TPA were successful using TPA strains passed through cultures with rabbit epithelial cells (Edmondson et al, 2018), we are still far away from the *in vitro* cultivation of TPA strains isolated directly from patients. It is, therefore, necessary to perform MLST directly from clinical samples, which usually contain a very low number of TPA DNA copies ($10-10^2$ TPA DNA copies/µl) (Pinto et al., 2017). To ensure a high proportion of fully typed samples, the newly proposed MLST is based on three loci only, in comparison to the conventional seven-locus scheme used in other bacteria.

73	ii.)	MLST schemes are usually based on several housekeeping genes. However, this
74		approach is not suitable for monomorphic bacteria as TPA (Šmajs et al. 2012;
75		Radolf et al. 2016; Šmajs et al. 2018). In this case, every single nucleotide variant
76		(SNV) is important and valuable and the typing loci with the highest SNV density
77		should be used to ensure sufficient variability to map the overall population
78		diversity. However, the typing loci should still be stable over time. Therefore, the
79		new MLST was designed on the principle of the sequence analyses using 3 highly
80		variable loci - TP0136, TP0548 and TP0705 (Grillová et al, 2018).
81	iii.)	Because of the monomorphic character of TPA genome and the low number of
82		typing loci, the "sequence-based analyses" approach is highly recommended. This
83		approach, compared to the conventional "allele-based comparisons", allows the
84		researcher to distinguish alleles based on a number of nucleotide differences and
85		permits interpretation of the data using Maximum-likelihood phylogeny. Even
86		though the TPA MLST uses only three typing loci, the scheme has a high
87		resolution (more than 30% of the resolution achieved with genome-wide data).
88		This approach also allows distinction of TPA from other treponemal subspecies
89		and species, discriminates between the two TPA clades (TP0136, TP0548)
90		(Nechvátal et al., 2014), and can be used for differentiation of strains within each
91		of these clades (TP0136 for the Nichols clade, and TP0705 for the SS14 clade).

73 The purpose of this short report is to establish a single universal MLST scheme
93 applicable to TPA (Grillová et al, 2018) and introduce the PubMLST database for treponemal
94 DNA data storage and ST assignments.

MATERIALS & METHODS

Treponema pallidum subps. *pallidum* database hosted at PubMLST

We have created a *Treponema pallidum* subsp. *pallidum* database hosted at PubMLST. PubMLST uses the BIGSdb platform for the storage and analysis of sequence data from bacterial isolates. The input format can be fasta files, which contain small contigs assembled from dideoxy chain termination sequencing of the typing loci, or complete or multiple contigs generated from next-generation sequencing platforms. When the new sequences are submitted to BIGSdb, the BLAST algorithm incorporated into BIGSdb identifies already known sequence variants or marks an unknown variant for curator verification. After verification by a curator, a novel allele number is assigned (Maiden et al., 2013). Metadata of the patients may be submitted to the database as well, and includes some important characteristics of isolates, e.g., isolate identification name/number; country where the strain was isolated; biological source of sample; year of isolation; and resistance to macrolide antibiotics. However, we encourage the researchers to upload as much information about patients and samples as available (e.g., age, sex and sexual orientation of the patients, stage of the disease, serological results etc.). The database can be accessed at https://pubmlst.org/tpallidum/. The database is primary designed to be used as a tool for automatic ST assignments; for determination of new alleles; for storage of sample metadata; for identification of new associations between genetic types and metadata using various tools; and for user-friendly visualization of molecular typing data, particularly using GrapeTree (Zhou et al., 2018) and Microreact (Argimón et al., 2016) plugins. However, this database can also serve as a repository for complete/draft genomes and gene-based comparative genomic data (using the Genome comparator tool). Sequences and metadata can be exported from the database in multiple formats. The database is overseen by the curators, who will check the submitted data and who will be available for any additional help (see https://pubmlst.org/tpallidum/).

Sequences submitted to PubMLST

All published whole genome and draft genome sequences of TPA (Pětrošová et al, 2013; Giacani et al., 2010; Pětrošová et al., 2012; Zobaníková et al., 2012; Giacani et al., 2014; Tong et al., 2017; Arora et al., 2016; Sun et al., 2016; Pinto et al., 2016; Strouhal et al., 2018; Grillová et al. 2018), data available at GenBank, as well as sequences obtained by MLST (Grillová et al, 2018; Pospíšilová et al., 2018) were submitted to the database.

Phylogenetic analyses

The phylogenetic tree was generated using MEGA6 (Tamura et al., 2013) with the Maximum Likelihood bootstrap algorithm and Tamura Nei model based on concatenated sequences of typing loci of fully typed samples (n=286). The tree was visualized using iTOL (Letunic & Bork, 2016), available as an external plugin of the BIGSdb.

Statistical analyses

Statistical analyses were performed using STATISTICA software v.12 (StatSoft, Tulsa, OK, USA) using data extracted from the Two-field breakdown option. A Fisher's exact test was used to establish the correlations of patient's metadata and allelic variants and STs. Statistical significance was set at $p < 0.05$.

RESULTS

Treponema pallidum subsp. *pallidum* database

At the time of writing (September 2018), the databases contain 358 TPA strains, mostly isolated from Europe (70.87%), followed by North America (26.33%), Asia (2.52%) and South America (0.28%) (Table 1). More than 90% of samples were clinically acquired and 31 strains were passed through rabbits. 71 samples represented whole/draft genomes and 287 samples were typed using MLST and yielded sequences only in typing loci. The samples were collected from 1912 to 2017. There were 272 samples collected from males (183 samples were collected from Men who have sex with men; MSM) and only four samples from females. Most of the samples were isolated from patients diagnosed with primary syphilis, followed by secondary syphilis, patients in the border of primary and secondary stage, and there were also 3 cases of congenital syphilis. The patients were between 0 and 71 years old. Interestingly, 21 samples were found negative by serology, but positive by PCR and all of them were isolated from patients with primary syphilis by genital, anal, or throat swab (Table 1). Most of the samples (262, 73.39%) contained treponemes resistant to macrolide antibiotics caused by A2058G mutation (in 257 cases) and A2059G mutation (in 5 cases) in 23S rDNA gene.

Fully typed TPA strains (n=285) were divided into 40 allelic profiles or strain types (ST1-ST40), and further into two clonal complexes, where 31 STs belong to the clonal complex "SS14-like" and 9 STs to the clonal complex "Nichols-like" (Figure 1; Table 2). STs were divided based on the 137 variable sites present in the concatenated sequences of typing loci (2 584 bp), where 46 variable positions were found to be parsimony informative for distinction of the two clonal complexes. The majority of the samples were found to belong to the SS14-like clonal complex (92.3%), while only 7.7% of samples belonged to the Nichols-like clonal complex.

Association found using TPA BIGSdb

We have found several association between genetic data and metadata of the samples including patient characteristics. Nichols-like strains were found to be associated with susceptibility to macrolides, and, on the other hand, SS14-like strains were associated with the presence of mutations leading to macrolide resistance (p<0.0001) (Figure 2A). The data showed that Nichols-like strains appear to be spreading predominantly among MSM patients (Figure 2B). Moreover, Nichols-like strains were associated with an older population (35 years old and older, p=0.0104) (Figure 2C). Most of these association were already described previously (Grillova et al., 2014; Gallo Valuet et al., 2016; Read et al., 2016).

ST1 was found to be the predominant sequencing type (allelic profile: 1.3.1, clonal complex: SS14-like). However, with the increasing number of newly submitted TPA strains, the number of different STs also increased, suggesting that the diversity of the TPA strains is still not fully revealed. Interestingly, five STs were found exclusively among the strains passed through rabbits – ST20-ST24.

174 **DISCUSSION**

175   Since TPA is not cultivable in *vitro*, MLST needs to be performed directly from clinical
176 samples. To obtain fully typed samples enabling assignment of ST, the amplification efficiency
177 of all three loci needs to be as high as possible. The amplification efficiency depends on several
178 factors, including the type of material taken from patients, the time between sampling and DNA
179 isolation, the type of DNA isolation protocol as well as the type of amplification protocol.

180   As reported in many studies, the highest concentration of the treponemal DNA is present in
181 swabs taken from the primary chancre, followed by blood-derived samples, cerebrospinal fluid
182 and others (Peng et al., 2011; Pinto et al., 2017). E.g., the swabs taken from primary chancre
183 represent the samples with the highest concentration of treponemal DNA and are the most
184 suitable candidates for MLST. However, MLST profiles revealed from parallel samples taken
185 from the same patient (e.g., primary chancer swab and whole blood) are also important with
186 respect to the stability of the used typing loci. For instance, the typing loci used in the enhanced
187 CDC-typing (based on restriction fragment length polymorphism analyses of the *tpr*EGJ genes
188 and a number of repeats in the *arp* gene (Marra et al., 2010) were identified to be genetically
189 unstable (Mikalová et al., 2013). This genetic instability was shown by revealing the different
190 subtypes in samples isolated from swab and whole blood, both taken from the same patient. Loci
191 selected for MLST (TP0136, TP0548 and TP0705) are considered to be genetically stable,
192 showing the same allelic profiles in parallel samples as well as in samples collected from
193 different rabbit passages (Grillová et al, 2018).

194   Another important factor to obtain high-quality DNA is the time between sampling of
195 patients and the DNA isolation. This time period should be as short as possible. If immediate
196 DNA isolation is not possible, the swab extracts (PBS) should be stored at -20 °C. The swab
197 extracts should be centrifuged at the lowest possible speed for 10 minutes and the supernatant
198 should be used for the subsequent isolation of DNA. Even though it is time consuming, nested
199 PCR should be used for target loci amplification to achieve the highest amplification efficiency.
200 Moreover, we recommend looking for the presence or absence of A2058G and A2059G
201 mutations in 23S rDNA, which encodes the macrolide resistance (Grillová et al, 2018).

202   Most of TPA strains analyzed to date from experimentally infected rabbits belong to the
203 Nichols-like group, while most clinical TPA isolates obtained directly from patients belong to
204 the SS14-like group (Šmajs et al., 2016). We encourage researchers and public health
205 communities to use MLST, which is able to distinguish these two clonal complexes. Moreover,
206 we recommend submitting the sequencing data to PubMLST along with the metadata (including
207 data from the serological tests). The expansion of the sample set in the PubMLST with syphilis
208 isolates from different geographical areas can lead to a better understanding of the epidemiology
209 of the two genetically distinct groups of syphilis treponemes.

CONCLUSIONS

Here, we present the PubMLST BIGSdb database of *Treponema pallidum* subsp. *pallidum*, which represents the first publicly available data storage of TPA sequences connected to metadata of patients. The database identifies specific regions of uploaded sequencing data and their genetic variation in order to reveal the sequencing types of TPA isolates based on the newly developed MLST (Grillová et al, 2018). We hope that this tool will open new opportunities in epidemiology of syphilis thereby allowing studies of sequencing types in different locations, tracking of syphilis infections, and finding the association of particular strains with specific groups of patients. All these data should improve our understanding of syphilis epidemiology.

ACKNOWLEDGEMENT

REFERENCES

Argimón, S., Abudahab, K., Goater, R.J.E., Fedosejev, A., Bhai, J., Glasner, C., Feil, E.J., Holden, M.T.G., Yeats, C.A., Grundmann, H., Spratt, B.G., Aanensen, D.M., 2016. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. Microb Genom 2, e000093. https://doi.org/10.1099/mgen.0.000093

Arora, N., Schuenemann, V.J., Jäger, G., Peltzer, A., Seitz, A., Herbig, A., Strouhal, M., Grillová, L., Sánchez-Busó, L., Kühnert, D., Bos, K.I., Davis, L.R., Mikalová, L., Bruisten, S., Komericki, P., French, P., Grant, P.R., Pando, M.A., Vaulet, L.G., Fermepin, M.R., Martinez, A., Centurion Lara, A., Giacani, L., Norris, S.J., Šmajs, D., Bosshard, P.P., González-Candelas, F., Nieselt, K., Krause, J., Bagheri, H.C., 2016. Origin of modern syphilis and emergence of a pandemic Treponema pallidum cluster. Nat Microbiol 2, 16245. https://doi.org/10.1038/nmicrobiol.2016.245

Edmondson, D.G., Hu, B., Norris, S.J., 2018. Long-Term In Vitro Culture of the Syphilis Spirochete Treponema pallidum subsp. pallidum. MBio 9. https://doi.org/10.1128/mBio.01153-18

Gallo Vaulet, L., Grillová, L., Mikalová, L., Casco, R., Rodríguez Fermepin, M., Pando, M.A., Šmajs, D., 2017. Molecular typing of Treponema pallidum isolates from Buenos Aires, Argentina: Frequent Nichols-like isolates and low levels of macrolide resistance. PLoS ONE 12, e0172905. https://doi.org/10.1371/journal.pone.0172905

Giacani, L., Iverson-Cabral, S.L., King, J.C.K., Molini, B.J., Lukehart, S.A., Centurion-Lara, A., 2014. Complete Genome Sequence of the Treponema pallidum subsp. pallidum Sea81-4 Strain. Genome Announc 2. https://doi.org/10.1128/genomeA.00333-14

Giacani, L., Jeffrey, B.M., Molini, B.J., Le, H.T., Lukehart, S.A., Centurion-Lara, A., Rockey, D.D., 2010. Complete genome sequence and annotation of the Treponema pallidum subsp. pallidum Chicago strain. J. Bacteriol. 192, 2645–2646. https://doi.org/10.1128/JB.00159-10

Grange, P.A., Mikalová, L., Gaudin, C., Strouhal, M., Janier, M., Benhaddou, N., Šmajs, D., Dupin, N., 2016. Treponema pallidum 11qj Subtype May Correspond to a Treponema pallidum Subsp. Endemicum Strain. Sex Transm Dis 43, 517–518. https://doi.org/10.1097/OLQ.0000000000000474

Grillová, L., Bawa, T., Mikalová, L., Gayet-Ageron, A., Nieselt, K., Strouhal, M., Sednaoui, P., Ferry, T., Cavassini, M., Lautenschlager, S., Dutly, F., Pla-Díaz, M., Krützen, M., González-Candelas, F., Bagheri, H.C., Šmajs, D., Arora, N., Bosshard, P.P., 2018. Molecular characterization of Treponema pallidum subsp. pallidum in Switzerland and France with a new multilocus sequence typing scheme. PLoS ONE 13, e0200773. https://doi.org/10.1371/journal.pone.0200773

Grillová L, Pĕtrošová, H., Mikalová, L.,Strnadel et al. 2014. Molecular typing of Treponema pallidum subsp. pallidum in the Czech Republic during 2011 to 2013: Increased prevalence of identified genotypes and of isolates with macrolide resistance. J Clin Microbiol. 52(10):3693-37000.

Jolley, K.A., Maiden, M.C.J., 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 11, 595. https://doi.org/10.1186/1471-2105-11-595

Letunic, I., Bork, P., 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 44, W242-245. https://doi.org/10.1093/nar/gkw290

268   Maiden, M.C.J., Jansen van Rensburg, M.J., Bray, J.E., Earle, S.G., Ford, S.A., Jolley, K.A.,
269   McCarthy, N.D., 2013. MLST revisited: the gene-by-gene approach to bacterial
270   genomics. Nat. Rev. Microbiol. 11, 728–736. https://doi.org/10.1038/nrmicro3093
271   Marra, C.M., Sahi, S.K., Tantalo, L.C., Godornes, C., Reid, T., Behets, F., Rompalo, A.,
272   Klausner, J.D., Yin, Y.P., Mulcahy, F., Golden, M.R., Centurion-Lara, A., Lukehart,
273   S.A., 2010. Enhanced molecular typing of treponema pallidum: geographical distribution
274   of strain types and association with neurosyphilis. J. Infect. Dis. 202, 1380–1388.
275   https://doi.org/10.1086/656533
276   Mikalová, L., Pospíšilová, P., Woznicová, V., Kuklová, I., Zákoucká, H., Šmajs, D., 2013.
277   Comparison of CDC and sequence-based molecular typing of syphilis treponemes: tpr
278   and arp loci are variable in multiple samples from the same patient. BMC Microbiol. 13,
279   178. https://doi.org/10.1186/1471-2180-13-178
280   Nechvátal, L., Pětrošová, H., Grillová, L., Pospíšilová, P., Mikalová, L., Strnadel, R., Kuklová,
281   I., Kojanová, M., Kreidlová, M., Vaňousová, D., Procházka, P., Zákoucká, H.,
282   Krchňáková, A., Šmajs, D., 2014. Syphilis-causing strains belong to separate SS14-like
283   or Nichols-like groups as defined by multilocus analysis of 19 Treponema pallidum
284   strains. Int. J. Med. Microbiol. 304, 645–653. https://doi.org/10.1016/j.ijmm.2014.04.007
285   Noda, A.A., Grillová, L., Lienhard, R., Blanco, O., Rodríguez, I., Šmajs, D., 2018. Bejel in
286   Cuba: molecular identification of Treponema pallidum subsp. endemicum in patients
287   diagnosed with venereal syphilis. Clin. Microbiol. Infect.
288   https://doi.org/10.1016/j.cmi.2018.02.006
289   Peng, R.-R., Wang, A.L., Li, J., Tucker, J.D., Yin, Y.-P., Chen, X.-S., 2011. Molecular typing of
290   Treponema pallidum: a systematic review and meta-analysis. PLoS Negl Trop Dis 5,
291   e1273. https://doi.org/10.1371/journal.pntd.0001273
292   Pětrošová, H., Pospíšilová, P., Strouhal, M., Čejková, D., Zobaníková, M., Mikalová, L.,
293   Sodergren, E., Weinstock, G.M., Šmajs, D., 2013. Resequencing of Treponema pallidum
294   ssp. pallidum strains Nichols and SS14: correction of sequencing errors resulted in
295   increased separation of syphilis treponeme subclusters. PLoS ONE 8, e74319.
296   https://doi.org/10.1371/journal.pone.0074319
297   Pětrošová, H., Zobaníková, M., Čejková, D., Mikalová, L., Pospíšilová, P., Strouhal, M., Chen,
298   L., Qin, X., Muzny, D.M., Weinstock, G.M., Šmajs, D., 2012. Whole genome sequence
299   of Treponema pallidum ssp. pallidum, strain Mexico A, suggests recombination between
300   yaws and syphilis strains. PLoS Negl Trop Dis 6, e1832.
301   https://doi.org/10.1371/journal.pntd.0001832
302   Pinto, M., Antelo, M., Ferreira, R., Azevedo, J., Santo, I., Borrego, M.J., Gomes, J.P., 2017. A
303   retrospective cross-sectional quantitative molecular approach in biological samples from
304   patients with syphilis. Microb. Pathog. 104, 296–302.
305   https://doi.org/10.1016/j.micpath.2017.01.059
306   Pinto, M., Borges, V., Antelo, M., Pinheiro, M., Nunes, A., Azevedo, J., Borrego, M.J.,
307   Mendonça, J., Carpinteiro, D., Vieira, L., Gomes, J.P., 2016. Genome-scale analysis of
308   the non-cultivable Treponema pallidum reveals extensive within-patient genetic
309   variation. Nat Microbiol 2, 16190. https://doi.org/10.1038/nmicrobiol.2016.190
310   Pospíšilová, P., Grange, P.A., Grillová, L., Mikalová, L., Martinet, P., Janier, M., Vermersch, A.,
311   Benhaddou, N., Del Giudice, P., Alcaraz, I., Truchetet, F., Dupin, N., Šmajs, D., 2018.
312   Multi-locus sequence typing of Treponema pallidum subsp. pallidum present in clinical
313   samples from France: Infecting treponemes are genetically diverse and belong to 18

314   allelic profiles. PLoS ONE 13, e0201068. https://doi.org/10.1371/journal.pone.0201068

315 Radolf, J.D., Deka, R.K., Anand, A., Šmajs, D., Norgard, M.V., Yang, X.F., 2016. Treponema
316   pallidum, the syphilis spirochete: making a living as a stealth pathogen. Nat. Rev.
317   Microbiol. 14, 744–759. https://doi.org/10.1038/nrmicro.2016.141

318 Read, P., Tagg, K.A., Jeoffreys, N., Guy, R.J., Gilbert, G.L., Donovan, B., 2016. Treponema
319   pallidum Strain Types and Association with Macrolide Resistance in Sydney, Australia:
320   New TP0548 Gene Types Identified. J. Clin. Microbiol. 54, 2172–2174.
321   https://doi.org/10.1128/JCM.00959-16

322 Šmajs, D., Mikalova, L., Strouhal, M., Grillova, L., 2016. Why Are There Two Genetically
323   Distinct Syphilis-Causing Strains? FIDT 7.
324   https://doi.org/10.1615/ForumImmunDisTher.2017020184

325 Šmajs, D., Norris, S.J., Weinstock, G.M., 2012. Genetic diversity in Treponema pallidum:
326   implications for pathogenesis, evolution and molecular diagnostics of syphilis and yaws.
327   Infect. Genet. Evol. 12, 191–202. https://doi.org/10.1016/j.meegid.2011.12.001

328 Šmajs, D., Strouhal, M., Knauf, S., 2018. Genetics of human and animal uncultivable treponemal
329   pathogens. Infect. Genet. Evol. 61, 92–107. https://doi.org/10.1016/j.meegid.2018.03.015

330 Strouhal, M., Oppelt, J., Mikalová, L., Arora, N., Nieselt, K., González-Candelas, F., Šmajs, D.,
331   2018. Reanalysis of Chinese Treponema pallidum samples: all Chinese samples cluster
332   with SS14-like group of syphilis-causing treponemes. BMC Res Notes 11, 16.
333   https://doi.org/10.1186/s13104-017-3106-7

334 Sun, J., Meng, Z., Wu, K., Liu, B., Zhang, S., Liu, Y., Wang, Y., Zheng, H., Huang, J., Zhou, P.,
335   2016. Tracing the origin of Treponema pallidum in China using next-generation
336   sequencing. Oncotarget 7, 42904–42918. https://doi.org/10.18632/oncotarget.10154

337 Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: Molecular
338   Evolutionary Genetics Analysis version 6.0. Mol. Biol. Evol. 30, 2725–2729.
339   https://doi.org/10.1093/molbev/mst197

340 Tong, M.-L., Zhao, Q., Liu, L.-L., Zhu, X.-Z., Gao, K., Zhang, H.-L., Lin, L.-R., Niu, J.-J., Ji,
341   Z.-L., Yang, T.-C., 2017. Whole genome sequence of the Treponema pallidum subsp.
342   pallidum strain Amoy: An Asian isolate highly similar to SS14. PLoS ONE 12,
343   e0182768. https://doi.org/10.1371/journal.pone.0182768

344 Jolley KA, Bray JE and Maiden MCJ. Open-access bacterial population genomics: BIGSdb
345   software, the PubMLST.org website and their applications [version 1; referees: awaiting
346   peer review]. Wellcome Open Res 2018, 3:124. doi: 10.12688/wellcomeopenres.14826.1.

347 World Health Organization. WHO guidelines for the treatment of *Treponema pallidum*
348   (syphilis). 2016; Available at:
349   http://www.who.int/reproductivehealth/publications/rtis/syphilis-treatment-
350   guidelines/en/.

351 Zhou, Z., Alikhan, N.-F., Sergeant, M.J., Luhmann, N., Vaz, C., Francisco, A.P., Carriço, J.A.,
352   Achtman, M., 2018. GrapeTree: visualization of core genomic relationships among
353   100,000 bacterial pathogens. Genome Res. 28, 1395–1404.
354   https://doi.org/10.1101/gr.232397.117

355 Zobaníková, M., Mikolka, P., Cejková, D., Pospíšilová, P., Chen, L., Strouhal, M., Qin, X.,
356   Weinstock, G.M., Šmajs, D., 2012. Complete genome sequence of Treponema pallidum
357   strain DAL-1. Stand Genomic Sci 7, 12–21. https://doi.org/10.4056/sigs.2615838

Table 1(on next page)

Clinical characteristics of the TPA strains submitted to the BIGSdb

1

2 **Table 1: Clinical characteristics of the TPA strains submitted to the BIGSdb**

3

Continent (n, %)		Clinical source (n, %)		Stage (n, %)	
Europe	253 (70.87)	Genital swab	180 (50.42)	Primary	119 (33.33)
Norh America	94 (26.33)	Anal swab	56 (15.69)	Primary/Secondary	8 (2.24)
Asia	9 (2.52)	Throat swab	39 (10.92)	Secondary	40 (11.2)
South America	1 (0.28)	Skin lesion	17 (4.76)	Congenital	3 (0.84)
Country (n, %)		Blood	3 (0.84)	Unspecified	187 (52.38)
France	146 (40.90)	CSF	2 (0.56)	Macrolide resistance (n, %)	
Cuba	72 (20.17)	Amniotic fluid	1 (0.28)	Resistant	262 (73.39)
Switzerland	72 (20.17)	Tissue	1 (0.28)	Sensitive	62 (18.3)
Portugal	25 (7)	Other	1 (0.28)	Not done	33 (9.24)
USA	21 (5.88)	Unspecified	57 (15.97)	Mutations (n, %)	
China	9 (2.52)	Sex (n, %)		A2058G	257 (71.99)
Netherlands	5 (1.4)	Male	272 (76.19)	A2059G	5 (1.4)
Austria	3 (0.84)	Female	8 (2.24)	Unspecified	95 (26.61)
Czech Republic	2 (0.56)	Unspecified	77 (21.5)	Serology (n, %)	
Argentina	1 (0.280	Sexual orientation (n, %)		Positive	182 (50.98)
Mexico	1 (0.28)	MSM	183 (51.26)	Negative	21 (5.88)
Source (n, %)		MSW	27 (7.56)	Unspecified	154 (43.14)
Clinically acquired	326 (91.32)	WSM	3 (0.84)		
Experimental animals	31 (8.68)	Unspecified	144 (40.34)		

4 CSF - cerebrospinal fluid.

5 MSM - Men who have sex with men.

6 MSW - Men who have sex with woman.

7 WSM - Woman who have sex with men.

Table 2(on next page)

Sequencing types found among 286 TPA fully typed strains submitted to the BIGSdb

1 **Table 2: Sequencing types found among 286 TPA fully typed strains submitted to the**
2 **BIGSdb.**

ST	TP0136	TP0548	TP0705	Clonal complex	No. of samples
1	1	3	1	SS14-like	169
2	1	1	1	SS14-like	37
3	1	1	8	SS14-like	13
4	7	1	9	SS14-like	5
5	2	1	2	SS14-like	3
6	3	2	3	Nichols-like	3
7	1	4	1	SS14-like	3
8	1	5	1	SS14-like	2
9	4	3	1	SS14-like	1
10	1	11	8	SS14-like	3
11	1	1	3	SS14-like	4
12	5	3	8	SS14-like	1
13	6	3	1	SS14-like	1
14	1	8	1	SS14-like	1
15	1	3	5	SS14-like	1
16	1	3	7	SS14-like	1
17	13	1	1	SS14-like	1
18	1	9	1	SS14-like	1
19	1	1	10	SS14-like	2
20	10	14	10	Nichols-like	3
21	1	13	10	SS14-like	1
22	11	14	10	Nichols-like	2
23	12	15	10	Nichols-like	1
24	6	1	8	SS14-like	1
25	1	26	1	SS14-like	1
26	9	7	3	Nichols-like	6
27	1	1	9	SS14-like	2
28	1	17	9	SS14-like	2
29	1	18	1	SS14-like	1
30	1	19	1	SS14-like	1
31	9	20	3	Nichols-like	1
32	14	3	1	SS14-like	1
33	1	1	11	SS14-like	1
34	1	22	12	SS14-like	1
35	1	23	1	SS14-like	1
36	1	1	13	SS14-like	1
37	15	7	3	Nichols-like	4

3	38	9	24	8	Nichols-like	1
4	39	9	25	3	Nichols-like	1
	40	16	3	1	SS14-like	1
5						

Figure 1

Phylogeny of all fully typed samples (n=286) divided into 40 STs and two clonal complexes (Nichols-like and SS14-like), based on concatenated sequences of typing loci.

Blue color represents the SS14-like clade and red color represents the Nichols-like clade. The green squares mark the macrolide sensitive strains and the orange squares mark the resistant strains.
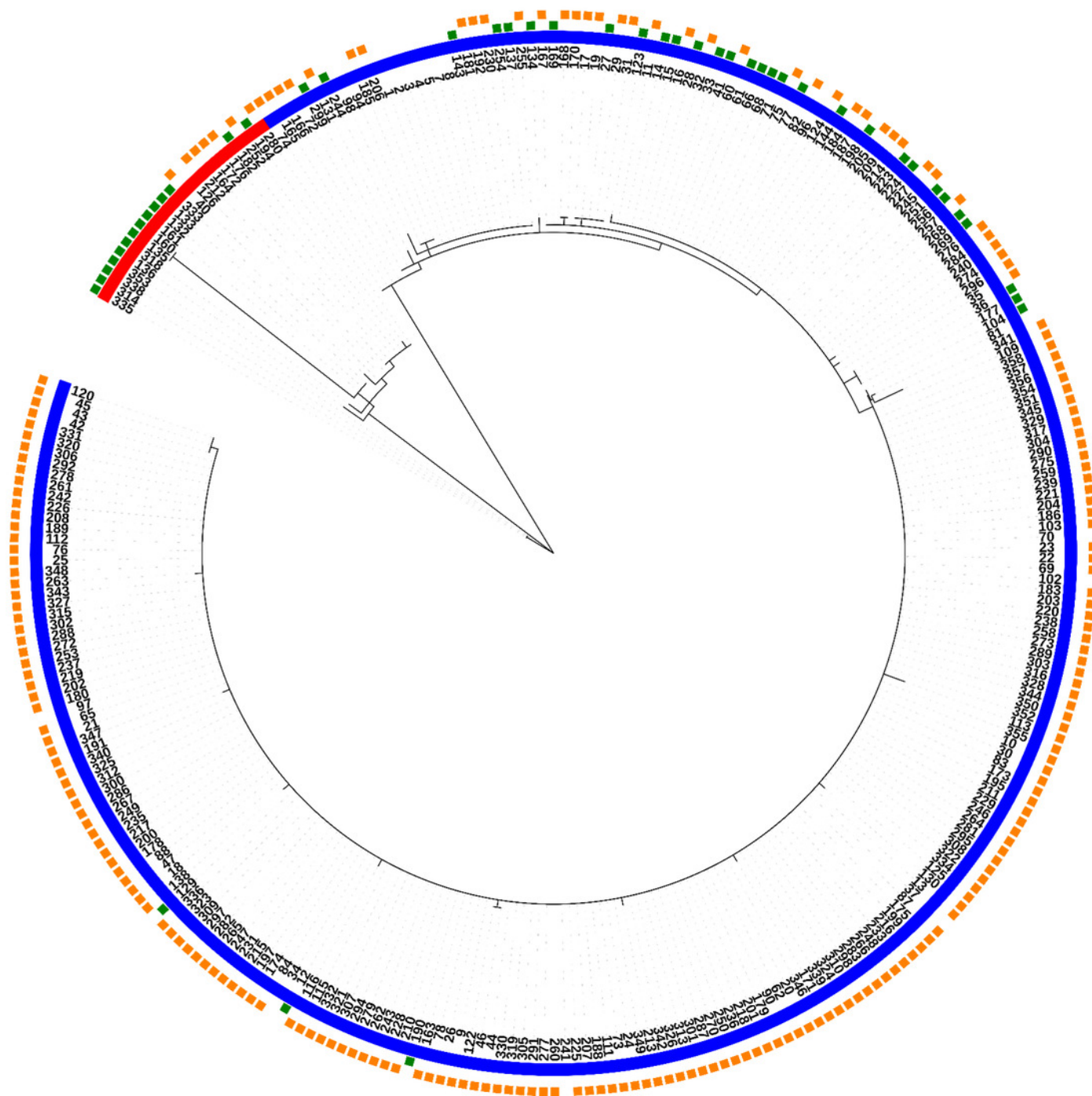
Figure 2

Clonal complexes associated with the macrolide sensitivity/resistance (A.); sexual orientation (B); and age of the patients (C). The figure was created based on the BIGSdb output.