Origin of modern syphilis and emergence of a pandemic *Treponema pallidum* cluster

Natasha Arora[1,2*], Verena J. Schuenemann[3], Günter Jäger[4], Alexander Peltzer[3,4†], Alexander Seitz[4], Alexander Herbig[3,4†], Michal Strouhal[5], Linda Grillová[5], Leonor Sánchez-Busó[6,7], Denise Kühnert[8], Kirsten I. Bos[3†], Leyla Rivero Davis[1‡], Lenka Mikalová[5], Sylvia Bruisten[9], Peter Komericki[10], Patrick French[11], Paul R. Grant[12], María A. Pando[13], Lucía Gallo Vaulet[14], Marcelo Rodríguez Fermepin[14], Antonio Martinez[15], Arturo Centurion Lara[16], Lorenzo Giacani[16], Steven J. Norris[17] David Šmajs[5], Philipp P. Bosshard[18], Fernando González-Candelas[6]*, Kay Nieselt[4]*, Johannes Krause[3†]* and Homayoun C. Bagheri[1§]*

Affiliations:

[1]Institute for Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland.

[2]Zurich Institute of Forensic Medicine, University of Zurich, Zurich, Switzerland.

[3]Institute for Archaeological Sciences, University of Tübingen, Tübingen, Germany.

[4]Center for Bioinformatics, University of Tübingen, Tübingen, Germany.

[5]Department of Biology, Faculty of Medicine, Masaryk University, Brno, Czech Republic.

[6]Unidad Mixta Infección y Salud Pública FISABIO/Universidad de Valencia. CIBER in Epidemiology and Public Health, Spain.

[7]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom

[8]Institute of Integrative Biology, Department of Environmental Systems Science, ETH Zürich, Switzerland.

[9]Public Health Laboratory, GGD Amsterdam, Department of Infectious Diseases, Amsterdam, the Netherlands.

[10]Department of Dermatology, Medical University of Graz, Graz, Austria.

[11]The Mortimer Market Centre CNWL, Camden Provider Services, UK.

[12]Department of Clinical Microbiology and Virology, University College London Hospitals NHS Foundation Trust, London, UK.

[13]Instituto de Investigaciones Biomédicas en Retrovirus y SIDA (INBIRS), Universidad de Buenos Aires-CONICET, Buenos Aires, Argentina.

[14]Universidad de Buenos Aires, Facultad de Farmacia y Bioquímica, Departamento de Bioquímica Clínica, Microbiología Clínica, Buenos Aires, Argentina.

[15]Servicio de Dermatología. Hospital General Universitario de Valencia, Spain.

[16]University of Washington, Department of Medicine, Division of Allergy and Infectious Diseases, and Department of Global Health, Seattle (WA), USA.

[17]Department of Pathology and Laboratory Medicine, UTHealth McGovern Medical School, Houston, TX USA.

[18]Department of Dermatology, University Hospital of Zurich, Zurich, Switzerland.

[†]Current address: Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany.

[‡]Current address: Department of Infectious Disease Epidemiology, Imperial College London, UK.

[§]Current address: Repsol Technology Center, Madrid, Spain

[*]Corresponding authors

Introductory paragraph: The abrupt onslaught of the syphilis pandemic starting in the late 15th century established this devastating infectious disease as one of the most feared in human history [1]. Surprisingly, despite the availability of effective antibiotic treatment since the mid-20th century, this bacterial infection caused by *Treponema pallidum* subsp. *pallidum* (TPA), has been re-emerging globally in the last few decades with an estimated 10.6 million cases in 2008 [2]. While resistance to penicillin has not yet been identified, an increasing number of strains fail to respond to the second-line antibiotic azithromycin [3]. Little is known about the genetic patterns in current infections or the evolutionary origins of the disease due to the low quantities of treponemal DNA in clinical samples, and difficulties to cultivate the pathogen [4]. Here we used DNA capture and whole genome sequencing to successfully interrogate genome-wide variation from syphilis patient specimens, combining it with laboratory samples of TPA and two other subspecies. Phylogenetic comparisons based on the sequenced genomes indicate that the TPA strains examined share a common ancestor after the 15th century, within the early modern era. Moreover, most contemporary strains are azithromycin resistant and members of a globally dominant cluster named here as SS14-Ω. This cluster diversified from a common ancestor in the mid-20th century subsequent to the discovery of antibiotics. Its recent phylogenetic expansion and global presence point to the emergence of a pandemic strain cluster.

Main Text: The first reported syphilis outbreaks in Europe occurred during the War of Naples in 1495 [5], prompting unresolved theories on a post-Columbian introduction [6,7]. Subsequently, the epidemic spread to other continents, remaining a severe health burden until treatment with penicillin five centuries later enabled incidence reduction. The striking present-day resurgence is poorly understood, particularly the underlying patterns of genetic diversity. Much of our molecular understanding of treponemes comes from propagating strains in laboratory animals to obtain sufficient DNA. The few published whole genomes were obtained after amplification through rabbit passage [4,8–10], and represent limited diversity for phylogenetic analyses. These sequences suggest that the TPA genome of 1.14 Mb is genetically monomorphic. Potential genetic diversity remains unexplored because clinical samples are mostly typed by PCR amplification of only 1-5 loci [11,12]. These epidemiological strain typing studies are motivated by the limitations of serologic or microscopic tests to distinguish among TPA strains or among the subspecies *Treponema pallidum* subsp. *pertenue* (TPE) and *Treponema pallidum* subsp. *endemicum* (TEN), which cause the diseases yaws and bejel, respectively. While all three diseases are transmitted through skin contact and show an overlap in their clinical manifestations, syphilis is geographically more widespread and generally transmitted sexually. The precise relationships among the bacteria are still debated, particularly regarding the evolutionary origin of syphilis.

The paucity of molecular studies and the focus on typing of a few genes means that we have limited information regarding the evolution and spread of epidemic TPA. In this study, we interrogated genome-wide variation across geographically widespread isolates. In total, we obtained 70 samples from 13 countries, including 52 syphilis swabs collected directly from patients between 2012 and 2013, and 18 syphilis, yaws, and bejel samples collected from 1912 onwards and propagated in laboratory rabbits (Supplementary Table 1). Through comparative genome analyses and phylogenetic reconstruction, we shed light on the evolutionary history of TPA and identify epidemiologically relevant haplotypes.

Due to the large background of host DNA, samples were enriched for treponemal DNA prior to Illumina sequencing [13,14]. The resultant reads were mapped to the Nichols TPA reference genome (RefSeq NC_021490; Supplementary Table 3) [4,15]. Genomic coverage ranged from 0.13-fold to over 1000-fold. As expected, the highest mean coverage was found in strains propagated in rabbits, while high variation in mean coverage was observed in samples collected directly from patients (0.13-fold to 223-fold) (Supplementary Table 2). This heterogeneity could potentially affect our inferences. Therefore, we restricted the genome-wide analyses to the 28 samples where at least 80% of the genome was covered by a minimum of three reads (highlighted in Supplementary Table 2). Across the 28 samples, the average proportion of genome coverage with at least 3-fold or 10-fold depth was 97% and 82%, respectively (Supplementary Table 4).

De novo assemblies for the four highest covered syphilis swab samples (NE17, NE20, CZ27, AU15) and one Indonesian yaws isolate (IND1) show no significant structural changes in the five genomes (Fig. 1a; Supplementary Table 5), except for the deletion in IND1 of gene TP1030, which potentially encodes a virulence-factor [17]. The deletion was shared across all the yaws infection isolates (Supplementary Methods), consistent with other studies [18].

Prior to phylogenetic reconstruction we checked for signatures of recombination. While *T. pallidum* is considered to be a clonal species [19], previous studies suggest recombinant genes in a Mexican syphilis and a Bosnian bejel strain [10,16]. We screened for putative recombinants across the 978 annotated genes in our 28 sequenced genomes and the 11 publicly available genomes from laboratory strains (Supplementary Table 3). Genes were selected as candidates if they had unexpectedly high SNP densities, incongruent topologies with the genome-wide tree and more than 4 homoplasies in a pair of branches (Supplementary Methods). We identified 4 genes coding for outer membrane proteins (Supplementary Table 6), one of which (TP0136) is used in typing studies [8].

After excluding the 4 putative recombinant genes, the genome alignment for all 39 genomes contained 2,235 variable positions. We used the Bayesian framework implemented in BEAST [20] to reconstruct a phylogenetic tree (Fig. 1b). The tree topology revealed a marked separation between TPA and TPE/TEN (100% Bayesian posterior support), with TPA forming a monophyletic lineage. The distinction of the two lineages was robust even with the inclusion of putative recombinant genes (Supplementary Fig. 2). Analyses of divergence between the two lineages yielded an average mean distance of 1225 nucleotide differences. By contrast, within each of the lineages we found considerably less diversity (124.6 average pairwise mutations within the TPA lineage and 200.2 within TPE/TEN). A heat map (Supplementary Fig. 3) to show shared variation for pairs of samples with respect to the Nichols reference genome, confirms the divergence between the lineages. The underlying SNP matrix yielded 443 SNPs specific to TPA genomes and 1703 to TPE/TEN genomes. Previous studies have found cross-subspecies groupings when relying on a limited set of markers [21]. Our results, incorporating genome-wide data from clinical samples, not only establish a clear separation between the two lineages, in agreement with studies examining genomic data from rabbit propagated samples [10,18], but also illustrate the need for a careful choice of taxonomic markers when genome-wide data is not available.

Using the sample isolation dates as tip calibration and applying the Birth Death Serial Skyline model [23], we obtained a mean evolutionary rate of 3.6×10^{-4} (rate variance 3.8×10^{-8} ; 95% HPD 1.86×10^{-4} - 5.73×10^{-4}). This estimate is equivalent to scaled mean rate of 6.6×10^{-7} substitutions per site per year for the whole genome, in line with estimates for other clonal human pathogens such as *Shigella sonnei* (6.0×10^{-7}) and *Vibrio cholerae* (O1 lineage; 8.0×10^{-7}) [24,25]. Our divergence analyses for TPA samples provide a time to the most recent common ancestor (TMRCA) less than 500 years ago (mean calendar year 1744, 95% HPD 1611-1859; Fig. 1B).

Within the TPA lineage the samples group in two clades named after the SS14 and Nichols reference genomes (with 100% and 82% posterior support values respectively). The Nichols clade consists almost exclusively of samples collected from patients in North America from 1912 to 1986 and passaged in rabbits prior to sequencing, with the exception of one patient sample from 2013 (NE20). In contrast, the SS14 clade has a geographically widespread distribution, encompassing European, North American and South American samples collected from infections between 1951 and 2013. We investigated the TPA clades further by generating a median-joining (MJ) network to illustrate the mutational differences among the TPA samples (Fig. 2a). As underscored by distances in the network, greater nucleotide diversity is found within the Nichols clade (π=0.05) compared to the SS14 clade (π= 0.01). Three closely related sequences derive from the original Nichols sample isolated from the cerebrospinal fluid of a patient in 1912 and propagated in the lab: NIC_REF, the reference genome re-sequenced by Pětrošová et al. [15], and NIC-1 and NIC-2, which we sequenced following independent propagation of the strains in Houston and Seattle, respectively, during different time periods (Supplementary Table 1 and Supplementary Table 3). These three group together with another three sequences in a cluster labelled Nichols- α (Fig. 2a), with a TMRCA at the turn of the 19th century (Fig. 1a). The less diversified SS14 clade contains a dominant central haplotype (labelled as SS14-Ω) from which the other sequences radiate (Fig. 2A). Critically, the cluster associated with the SS14-Ω haplotype contains all but one of the recent patient samples from 2012-2013 (n=17) that were captured and sequenced directly, in addition to samples from 1977 (n=1) and 2004 (n=2). The genetic variation within the SS14-Ω cluster is found primarily as singleton mutations (95.5%), with no evidence for geographical structuring. Bayesian analyses estimate a median coalescence for the SS14-Ω cluster in 1963 (95% HPD 1948-1974; Fig. 1a), at a time when incidence was reduced due to the introduction of antibiotics. The star-like topology of this cluster observed in both the tree and the network is suggestive of a recent and rapid clonal expansion.

To determine whether the dominance of SS14 clade sequences applies across other countries for which genetic data is available, we examined sequences from the widely typed TP0548 gene in worldwide epidemiological studies [11]. Phylogenies for the TP0548 typing regions separate the SS14 from the Nichols clade for the TPA samples, while not distinguishing the TPA and TPE/TEN lineages (Supplementary Methods; Supplementary Fig. 3). Across 1353 worldwide TP0548 sequences from clinical samples, including the 78 from patients in this study, we found that 94% of them grouped in the SS14 clade (Supplementary Tables 8-9; Supplementary Fig. 5), consistent with a probable recent spread of the epidemic cluster. The wide geographical distribution of the SS14 clade establishes it as representative of the present worldwide epidemic. While studies to date have focused on the Nichols strain [26,27], our results indicate that further work on the SS14 clade is warranted.

Critically, typing of samples over multiple years in the Czech Republic, San Francisco, British Columbia and Seattle indicate that macrolide antibiotic resistance has increased over time [3,12,28–30]. We queried the presence of the two mutations (A2058G and A2059G) in the 23S rRNA genes associated with azithromycin resistance [3,31,32]. As observed in the MJ network, the resistance marker is a dominant characteristic of the SS14-Ω cluster (Fig. 2a), although it is also found in a recent patient sample (NE20) of the Nichols clade. Extending our analyses of the 23S rRNA gene to all sequenced samples from our study, including the 42 with lower coverage, revealed the mutations in 90% of the SS14 (n=51) and 25% of the Nichols (n=12) samples, indicating that neither resistance nor sensitivity is clade-specific (Supplementary Table 8). Hence resistance was probably not an ancestral characteristic of the SS14 clade. A likely scenario is that the extensive usage of azithromycin to treat syphilis and a wide range of bacterial infections, including co-infections with other sexually-transmitted diseases (STDs) such as chlamydia, has played an important role in the selection and subsequent spread of resistance [33,34].

Results here represent the first reported set of whole genome sequences successfully obtained directly from syphilis patients, enabling us to disentangle evolutionary relationships at high resolution, and paving the way for further clinical sequencing from current epidemics. Given our identification of putative recombinant genes in *Treponema*, and previous reports on genes involved in homologous recombination [4,35], further detailed analyses on the potential mechanisms of recombination will be necessary. Our phylogenetic reconstruction indicates that all TPA samples examined to date share a common ancestor that was infecting populations in the 1700s, within the early centuries of the modern era, and that was successful in leaving descendants until today. This date is posterior to the colonization of the Americas, and therefore potentially compatible with the post-Columbian model for the emergence of syphilis in Europe. Nonetheless, our work does not exclude the possibility that older TPA lineages had previously existed in Europe but went extinct. Obtaining more patient sample genomes with high coverage could potentially refine our detection of putative recombinants and our phylogenetic inferences. In addition, sequencing from ancient skeletal material would help to further ascertain the history of syphilis. Interestingly, we observed a time difference between the first reported syphilis outbreak in 1495 and the last common ancestor of modern strains dated to the 1700s. While this difference could stem from imprecision in the divergence estimates, an alternative scenario is the eventual establishment of a specific lineage due to selection. For instance, it has been hypothesized that the symptoms of syphilis became less severe after the first reported outbreaks in Europe because of the evolution of strains with lower virulence and higher transmission rates [36]. In this scenario, the 18th century provided the context for the origin and propagation of a lineage that successfully outcompeted other lineages.

Critical to our epidemiological understanding of contemporary syphilis is our observation of an epidemic cluster (SS14-Ω) that emerged after the discovery of antibiotics. The relatively recent phylogenetic expansion of the SS14-Ω cluster and its global presence point to the emergence of a pandemic azithromycin-resistant cluster. The genome-wide data in this study will be useful to determine a suitable set of typing loci, since typing remains a more accessible method for most laboratories. Further characterization of the genomic diversity of TPA across the globe can prove instrumental in understanding the genetic and epidemiological basis for the spread of SS14-Ω strains.

Methods

Sample collection, DNA extraction and library preparation

Samples from 64 syphilis infections, 5 yaws infections and 1 bejel infection were collected from numerous countries across the globe (Supplementary Table 1). Syphilis infection samples were classified as either clinical, if obtained from patients directly, or as laboratory strains, if passaged in rabbits after isolation from patients. Clinical samples were obtained after swabbing lesions from patients at sexual health clinics, dermatological clinics or hospitals. Flocked swabs (from Copan Diagnostics, Brescia, Italy) or Nylon swabs were used according to local laboratory instructions. Laboratory strains were obtained as DNA extracts from Masaryk University (Brno, Czech Republic) and the University of Washington (Seattle, USA).DNA extractions were carried out in the participating laboratories using in-house protocols. At the University of Zurich the QIAmp DNA mini kit and QIAmp DNA blood min kit (Qiagen) were used following the manufacturer’s protocols.

Library preparation was conducted following a modified Illumina protocol for ancient DNA [14,37], at the University of Tübingen (Supplementary Materials and Methods). Libraries were barcoded with double indices.

Genome-wide enrichment and sequencing

Target enrichment for *Treponema pallidum* subsp. *pallidum* was carried out through two rounds of capture hybridization on a 1 million Agilent SureSelect array following the protocol detailed by Hodges et al. [13]. The probes on the array were based on two reference genomes (Nichols, here abbreviated as NIC_REF, GenBank ID CP004010.2/RefSeq ID NC_021490.2, and SS14, GenBank ID CP000805.1/RefSeq ID NC_010741.1). High-throughput sequencing of the enriched libraries was performed on an Illumina Hiseq 2500 platform.

Sequencing analyses and genome reconstruction

We applied EAGER [38], our own developed pipeline for read preprocessing (adapter clipping, merging of corresponding paired-end reads in the overlapping regions and quality trimming), mapping, variant identification and genome reconstruction, to all sequenced samples (for full details see Supplementary Materials and Methods). All reads (merged and unmerged) were treated as single-end reads and mapping was performed using the BWA-MEM algorithm [39] with default parameters, using the Nichols genome as a reference. Subsequently, we selected the samples which had at least 80% coverage of the Nichols genome and a minimum of 3 reads (n= 28 samples, Supplementary Table 3). For each of these samples, we used the Genome Analysis Toolkit (GATK) [40] to generate a mapping assembly, applying the UnifiedGenotyper module of GATK to call reference bases and variants from the mapping. The reference base was called if the genotype quality of the call was at least 30 and the position was covered by at least 3 reads. A variant position (SNP) was called if the following criteria 3 were met: i) the position was covered by at least 3 reads; ii) the genotype quality of the call was at least 30 and iii) the minimum SNP allele frequency was 90%. If neither of the requirements for a reference base call nor the requirements for a variant call were met, the character ’N’ was inserted at the respective position. For the generation of draft genome sequences we used an in-house tool (VCF2Genome), which reads a VCF file such as produced by the GATK UnifiedGenotyper and incorporates for each row, and thus for each call, one nucleotide into the new draft sequence.

In order to apply our analysis pipeline also to those samples for which complete genomic sequences are available in GenBank (Supplementary Table 2), we produced artificial reads in these cases using an in-house tool (Genome2Reads), and then applied the same mapping, SNP calling and genome reconstruction procedure as for the sequenced samples in order to obtain consistent and comparable results.

To investigate conservation of structure and gene order in the genomes, in addition to the mapping assembly, we also performed a *de novo* assembly for the 5 samples with highest coverage (Supplementary Table 5). Our *de novo* assembly pipeline started with the merged reads and in a first step utilized the short read assembler software SOAPdenovo2 using ten different k-mer sizes ($k = 37 + i \cdot 10$, $i=0,\ldots,9$). Different k-mer sizes were used because merging of read pairs into one single read results in very different lengths (between 30 and 190 bases). Next, all input reads were mapped back against the resulting contigs using BWA-MEM [39]. Contigs that were not supported by any reads (no read mapped against these contigs) were removed. In order to assemble the contigs resulting from the different k-mers, the remaining contigs were subject to the overlap-based String Graph Assembler (SGA) [41]. Finally, contigs smaller than 1,000 bp were removed before these contigs were mapped against the Nichols reference genome for comparison of genome architectures.

Analyses to detect recombinants and reconstruct evolutionary relationships using genome-wide variation were conducted for the 28 sequenced samples meeting our genome-wide coverage criteria (highlighted in the Supplementary Table 3) as well as the 11 published genomes (Supplementary Table 2). Across the 39 whole genomes and draft genomes, 31 were TPA, 8 TPE and 1 TEN.

Recombination detection

Tests for the non-vertical transmission of genes were carried out on the TPA, TPE and TEN genomes (n= 39) by identifying those genes that i) had an unexpectedly high number of SNPs and ii) displayed patterns of transmission (i.e., phylogenies) incongruent with most other genes. First, an expected substitution rate was computed by dividing the total number of observed SNPs in the 978 annotated genes (n=2,098) by the total length of these genes (1,046,421 bp). This rate was then used to calculate the expected number of polymorphisms per gene according to its length. A total of 87 genes displayed at least twice the expected number of polymorphisms. Second, for each of these 87 genes the gene sequence alignment and the gene tree topology were tested against the maximum likelihood tree topology of the draft genome in TREE-PUZZLE v5.2 [42,43]. Genes for which both the Expected Likelihood Weight [44] and the Shimodaira-Hasegawa [45] test rejected the genome tree ($p < 0.05$) were examined more closely. Third, genes within which we identified a minimum of 4 homoplasies (identical mutations in separate lineages) in at least 2 branches of the tree were marked as putative recombinants (Supplementary Table 6).

Genome-wide variation and phylogenetic analyses

We investigated genome-wide patterns of polymorphism and divergence using MEGA 6 [46] and DnaSP v.5.10 to compute various measures of diversity including the average pairwise nucleotide differences, Nei's Pi (π), and the number of singletons in each group. We also estimated the number of SNPs private to particular groups. A comparison of the TPA and TPE/TEN genomes revealed between 1 (NIC1) and 339 (AR2) SNPs observed in the TPA samples and between 1091 (GHA1) and 1443 (Bosnia A) SNPs in the TPE/TEN strains (Supplementary Table 4). Furthermore, we produced a heat map to display the number of SNPs that any two genomes share (Supplementary Fig. 3).

The molecular clock hypothesis was tested with the maximum likelihood analysis in MEGA 6.0 [46]. Tests were conducted for all TPA, TPE and TEN genomes (39 samples) using i) multiple whole genome alignments and ii) alignments with only the variable positions, in both cases excluding the 4 putative recombinant genes. The molecular clock hypothesis was rejected at the 5% significance level.

Bayesian phylogenetic trees were produced in BEAST 2.3 [47] for the 28 sequenced samples and the 11 published samples. We compared the trees generated with the alignment of all variable positions in the TPA, TPE and TEN genomes (2,506) and the tree generated with the set of variable positions after excluding the 4 putative recombinant genes (2,235 positions). Additionally, rooted trees were generated with Maximum Parsimony by including *Treponema paraluiscuniculi* (NC_015714) as the outgroup.

As a calibration for the BEAST trees we used tip dates, that is, the isolation years of all samples. When not known with precision, we provided a range (for NIC_REF, NIC1, NIC2, and GAU). The two demographic models (coalescent tree prior under Constant Size and the Birth-Death Serial Skyline model (BDSS)) resulted in consistent parameter estimates. The relaxed clock model was chosen over the strict clock model based on marginal likelihood estimates obtained with PathSampler [47,48]. We provide results for the BDSS model run with the following specifications: uncorrelated lognormal relaxed clock-clock model, GTR plus gamma substitution model, 50 million generations with parameter sampling every 5,000 generations. The log file was viewed in Tracer 1.6 [49] to determine the appropriate burn-in period for adequate effective sample sizes. The annotated maximum clade credibility tree was visualized and edited using Figtree v1.4.2 [50]. Because TPA samples are the focus of this study and therefore more extensively sampled, we report mean branch rate and divergence estimates for the TPA lineage. The mean branch rate estimate obtained is in line with the number of mutations that differed between the samples NIC_REF and NIC 2 (n=15), which were isolated 15-20 years apart following continuous rabbit propagation. We also checked that a run with the same specifications but with only TPA samples (n=31) produced consistent results.

The phylogenetic relationships among the closely related TPA samples (n=31) were examined and visualized through a median joining (MJ) network analysis in Network 4.6 and Network Publisher [51,52] using all variable positions after excluding the putative recombinant loci and sites with missing data (resulting in a total of 628 variable positions).

Clade classification

Samples from this study: From the 70 TPA, TPE and TEN samples sequenced in this study, 28 fulfilled our criteria for genome-wide analyses (minimum 80% genome covered with at least 3 reads). For the remaining 42 samples, we implemented two classification strategies. First, we generated a new clade prediction strategy based on NGS reads to classify the genomes according to lineage (TPA or TPE/TEN), and within the TPA lineage, as part of the SS14 or the Nichols clade (details provided in the Supplementary Information). Second, we used a classification scheme based on the TP0548 gene. For the TP0548 classification scheme we carried out PCR and Sanger sequencing of the TP0548 gene region following the protocols and primers of Matějková et al.[31]. Single nucleotide polymorphisms (SNPs) in the TP0548 typing regions enable the distinction of an SS14 clade versus a Nichols clade. Indels enable the classification of TPE and TEN. Our NGS prediction strategy (detailed in the Supplementary Materials and Methods) was congruent with the TP0548 classification scheme wherever prediction strength was above 0.4, with the exception of 1 TEN sample.

Samples from typing studies: We put together all publicly available TP0548 sequences obtained in typing studies of syphilis infections around the world [12,53–60]. We additionally incorporated TP0548 sequences obtained for 34 Argentinian clinical samples by LGV at the

University of Buenos Aires, Argentina (Supplementary Table 8). All TP0548 sequences were classified as part of the SS14 clade or part of the Nichols clade based on an ML tree (Supplementary Fig. 5). Subtypes were distinguished through visual inspection (Supplementary Table 8).

Antibiotic resistance

The two mutations associated with resistance to the macrolide azithromycin, A2058G and A2059G on the 23S ribosomal RNA operon (with positions referring to coordinates in the 23S ribosomal RNA gene of *Escherichia coli*), were investigated in separate analyses. Since the operon contains two copies of the gene, mapping of reads with BWA was carried out independently for each of the genes, including a flanking region of 200 bases on both the 5' and 3' end of each genes. Following variant calling, the presence/absence of each of the two mutations was recorded for each sample. The two operons could not, however, be distinguished.

In addition, we used primers specific for each of the two operons to carry out PCR amplifications as well as Sanger sequencing on the samples, following the protocol in Matějková et al.[31]. Details on the samples sequenced, as well as resistance or sensitivity to the macrolide as determined by the presence or absence of the associated mutations are given in Supplementary Table 7.

Data availability

All samples sequenced in this study are available in an NCBI Bioproject under accession number PRJNA313497. Raw sequencing reads in FASTQ format were uploaded to the Short Read Archive (SRA). All accession codes are listed in Supplementary Table 2. Code for the in-house scripts developed for some of the analyses are available upon request from the authors.

Reference List

1. Gall, G. E. C., Lautenschlager, S. & Bagheri, H. C. Quarantine as a public health measure against an emerging infectious disease: syphilis in Zurich at the dawn of the modern era (1496–1585). *GMS Hyg Infect Control* **11,** (2016).
2. Rowley, J. *et al. Global incidence and prevalence of selected curable sexually transmitted infections, 2008*. (World Health Organization, 2012).
3. Stamm, L. V. Global Challenge of Antibiotic-Resistant *Treponema pallidum*. *Antimicrob. Agents Chemother.* **54,** 583–589 (2010).
4. Fraser, C. M. *et al.* Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281,** 375–88 (1998).
5. Quétel, C. *History of syphilis*. (Johns Hopkins University Press, 1990).
6. Fernandez de Oviedo y Valdes, G. *Sumario de la natural historia de las Indias.* (Fondo de Cultura Economico, 1526).
7. Harper, K. N., Zuckerman, M. K., Harper, M. L., Kingston, J. D. & Armelagos, G. J. The origin and antiquity of syphilis revisited: an appraisal of Old World pre-Columbian evidence for treponemal infection. *Am. J. Phys. Anthropol.* **146 Suppl 53,** 99–133 (2011).
8. Šmajs, D., Norris, S. J. & Weinstock, G. M. Genetic diversity in *Treponema pallidum*: Implications for pathogenesis, evolution and molecular diagnostics of syphilis and yaws. *Infect. Genet. Evol.* **12,** 191–202 (2012).
9. Giacani, L. *et al.* Complete Genome Sequence of the *Treponema pallidum* subsp. *pallidum* Sea81-4 Strain. *Genome Announc.* **2,** e00333-14-e00333-14 (2014).
10. Štaudová, B. *et al.* Whole Genome Sequence of the Treponema pallidum subsp. endemicum Strain Bosnia A: The Genome Is Related to Yaws Treponemes but Contains Few Loci Similar to Syphilis Treponemes. *PLoS Negl. Trop. Dis.* **8,** e3261 (2014).
11. Marra, C. M. *et al.* Enhanced Molecular Typing of *Treponema pallidum*: Geographical Distribution of Strain Types and Association with Neurosyphilis. *J. Infect. Dis.* **202,** 1380–1388 (2010).
12. Grillová, L. *et al.* Molecular Typing of *Treponema pallidum* in the Czech Republic during 2011 to 2013: Increased Prevalence of Identified Genotypes and of Isolates with Macrolide Resistance. *J. Clin. Microbiol.* **52,** 3693–3700 (2014).
13. Hodges, E. *et al.* Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat. Protoc.* **4,** 960–974 (2009).

14. Meyer, M. & Kircher, M. Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harb. Protoc.* **2010,** pdb.prot5448-prot5448 (2010).

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

16. Petrosova, H. *et al.* Whole genome sequence of *Treponema pallidum* ssp. *pallidum*, strain Mexico A, suggests recombination between yaws and syphilis strains. *PLoS Negl Trop Dis* **6,** e1832 (2012).

17. Centurion-Lara, A. *et al.* Fine Analysis of Genetic Diversity of the tpr Gene Family among Treponemal Species, Subspecies and Strains. *PLoS Negl. Trop. Dis.* **7,** e2222 (2013).

18. Mikalova, L. *et al.* Genome analysis of *Treponema pallidum* subsp. *pallidum* and subsp. pertenue strains: most of the genetic differences are localized in six regions. *PLoS One* **5,** e15713 (2010).

19. Achtman, M. Evolution, Population Structure, and Phylogeography of Genetically Monomorphic Bacterial Pathogens. *Annu. Rev. Microbiol.* **62,** 53–70 (2008).

20. Bouckaert, R. *et al.* BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput. Biol.* **10,** e1003537 (2014).

21. Lukehart, S. A. & Giacani, L. When Is Syphilis Not Syphilis? Or Is It?: *Sex. Transm. Dis.* **41,** 554–555 (2014).

22. Mikalova, L. *et al.* Genome analysis of Treponema pallidum subsp. pallidum and subsp. pertenue strains: most of the genetic differences are localized in six regions. *PLoS One* **5,** e15713 (2010).

23. Stadler, T., Kuhnert, D., Bonhoeffer, S. & Drummond, A. J. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc. Natl. Acad. Sci.* **110,** 228–233 (2013).

24. Holt, K. E. *et al.* Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nat Genet* **44,** 1056–1059 (2012).

25. Mutreja, A. *et al.* Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477,** 462–465 (2011).

26. Giacani, L. *et al.* Footprint of Positive Selection in T*reponema pallidum* subsp. *pallidum* Genome Sequences Suggests Adaptive Microevolution of the Syphilis Pathogen. *PLoS Negl. Trop. Dis.* **6,** e1698 (2012).

27. Strouhal, M. *et al.* Genome Differences between *Treponema pallidum* subsp. *pallidum* Strain Nichols and T. paraluiscuniculi Strain Cuniculi A. *Infect. Immun.* **75,** 5859–5866 (2007).

28. Marra, C. M. *et al.* Antibiotic selection may contribute to increases in macrolide-resistant Treponema pallidum. *J. Infect. Dis.* **194,** 1771–1773 (2006).

29. Mitchell, S. J. *et al.* Azithromycin-resistant syphilis infection: San Francisco, California, 2000–2004. *Clin. Infect. Dis.* **42,** 337–345 (2006).

30. Morshed, M. & Jones, H. Treponema pallidum macrolide resistance in BC. *CMAJ* 349 (2006).

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

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

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

34. Geisler, W. M. *et al.* Azithromycin versus Doxycycline for Urogenital *Chlamydia trachomatis* Infection. *N. Engl. J. Med.* **373,** 2512–2521 (2015).

35. Centurion-Lara, A. in *Pathogenic Treponema: Molecular and Cellular Biology* (eds. Radolf, J. D. & Lukehart, S. A.) 267–283 (Caister Academic Press, 2006).

36. Knell, R. J. Syphilis in renaissance Europe: rapid evolution of an introduced sexually transmitted disease? *Proc Biol Sci* **271 Suppl 4,** S174-6 (2004).

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

38. Peltzer, A. *et al.* EAGER: Efficient Ancient Genome Reconstruction. *Genome Biol.* **17,** 1 (2016).

39. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv Prepr. ArXiv13033997* (2013).

40. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

41. Simpson, J. T. & Durbin, R. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res.* **22,** 549–556 (2012).

42. Schmidt, H. A., Strimmer, K., Vingron, M. & von Haeseler, A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18,** 502–504 (2002).

43. Strimmer, K. & von Haeseler, A. Quartet Puzzling: A Quartet Maximum-Likelihood Method for Reconstructing Tree Topologies. *Mol. Biol. Evol.* **13,** 964 (1996).

44. Strimmer, K. & Rambaut, A. Inferring confidence sets of possibly misspecified gene trees. *Proc. R. Soc. B Biol. Sci.* **269,** 137–142 (2002).

45. Shimodaira, H. & Hasegawa, M. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Mol. Biol. Evol.* **16,** 1114 (1999).

46. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Mol. Biol. Evol.* **30,** 2725–2729 (2013).

47. Bouckaert, R. *et al.* BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput. Biol.* **10,** e1003537 (2014).

48. Baele, G. *et al.* Improving the Accuracy of Demographic and Molecular Clock Model Comparison While Accommodating Phylogenetic Uncertainty. *Mol. Biol. Evol.* **29,** 2157–2167 (2012).

49. Rambaut, A., Suchard, M., Xie, D. & Drummond, A. *Tracer v1.6*. (2014).

50. Rambaut, A. *FigTree v.1.4.2*. (2014).

51. Bandelt, H.-J., Forster, P. & Röhl, A. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16,** 37–48 (1999).

52. *www.fluxux-engineering.com*.

53. Dai, T. *et al.* Molecular Typing of Treponema pallidum: a 5-Year Surveillance in Shanghai, China. *J. Clin. Microbiol.* **50,** 3674–3677 (2012).

54. Flasarová, M. *et al.* Sequencing-based Molecular Typing of *Treponema pallidum* Strains in the Czech Republic: All Identified Genotypes are Related to the Sequence of the SS14 Strain. *Acta Derm. Venereol.* **92,** 669–674 (2012).

55. Grange, P. A. *et al.* Molecular Subtyping of Treponema pallidum in Paris, France: *Sex. Transm. Dis.* **40,** 641–644 (2013).

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

57. Peng, R.-R. *et al.* Molecular Typing of *Treponema pallidum* Causing Early Syphilis in China: A Cross-Sectional Study: *Sex. Transm. Dis.* **39,** 42–45 (2012).

58. Tian, H. *et al.* Molecular typing of *Treponema pallidum*: identification of a new sequence of tp0548 gene in Shandong, China. *Sex. Transm. Dis.* **41,** 551 (2014).

59. Tipple, C., McClure, M. O. & Taylor, G. P. High prevalence of macrolide resistant Treponema pallidum strains in a London centre. *Sex. Transm. Infect.* **87,** 486–488 (2011).

60. Wu, B.-R. *et al.* Multicentre surveillance of prevalence of the 23S rRNA A2058G and A2059G point mutations and molecular subtypes of *Treponema pallidum* in Taiwan, 2009–2013. *Clin. Microbiol. Infect.* **20,** 802–807 (2014).

Acknowledgments

Research in Zurich by N.A. and H.C.B. was funded by the *Forschungskredit* and the University of Zurich. A.H. was funded by an ERC Starting Grant. F.G.C. and L.S.B. were funded by MINECO (Spanish Government) and PROMETEO (Generalitat Valenciana). K.I.B. was funded by Social Sciences and Humanities Research Council of Canada. L.M. was funded by the Faculty of Medicine of Masaryk University. We are grateful to Stephan Lautenschlager for his guidance. We thank Alexei Drummond for input on BEAST. We thank Sheila Lukehart for providing HaitiB, Sea86-1, Bal3, Bal9, Bal73-1, and Grady1 strain DNA, and Christina Marra for providing UW249B and UW231B strain DNA. We are especially grateful to Antonio Messina and the S3IT at the University of Zurich for providing computational resources and services. Special thanks to Lukas Keller and his lab for their support.

Author contributions

N.A. and H.C.B. conceived the investigation. N.A., L.G., S.J.N., D.S., P.B., F.G.C., K.N., J.K. and H.C.B. devised research and analyses. N.A., G.J., A.P., A.S., A.H., M.S., L.G., L.S.B, D.K., L.R.D., L.M., F.G.C., and K.N. analyzed data. N.A., V.J.S., M.S., L.G., K.I.B., L.R.D., L.G.V., and P.B. contributed to or performed experiments. M.S., L.G., S.B., P.K., P.F., P.R.G., M.A.P., L.G.V., M.R.F., A.M., D.S., P.B., and F.G.C. provided clinical samples and A.C.L, L.G., S.J.N. and D.S. provided lab samples. N.A and H.C.B. wrote the manuscript with significant contributions from M.S., L.G., L.S.B., D.K., K.I.B., L.R.D., L.M., S.B., L.G., S.J.N., D.S., P.B., F.G.C., K.N., J.K., and comments from all co-authors.

Additional information

Supplementary information is available for this paper. Correspondence and requests for materials should be addressed to N.A. (natasha.arora@uzh.ch), F.G.C. (fernando.gonzalez@uv.es),K.N. (kay.nieselt@uni-tuebingen.de), J.K. (johannes.krause@uni-tuebingen.de) or H.C.B (homayoun.bagheri@repsol.com).

Accession codes

All raw read files have been deposited in the trace archive of the NCBI Sequence Read Archive under accession number SRP072086.

Competing interests

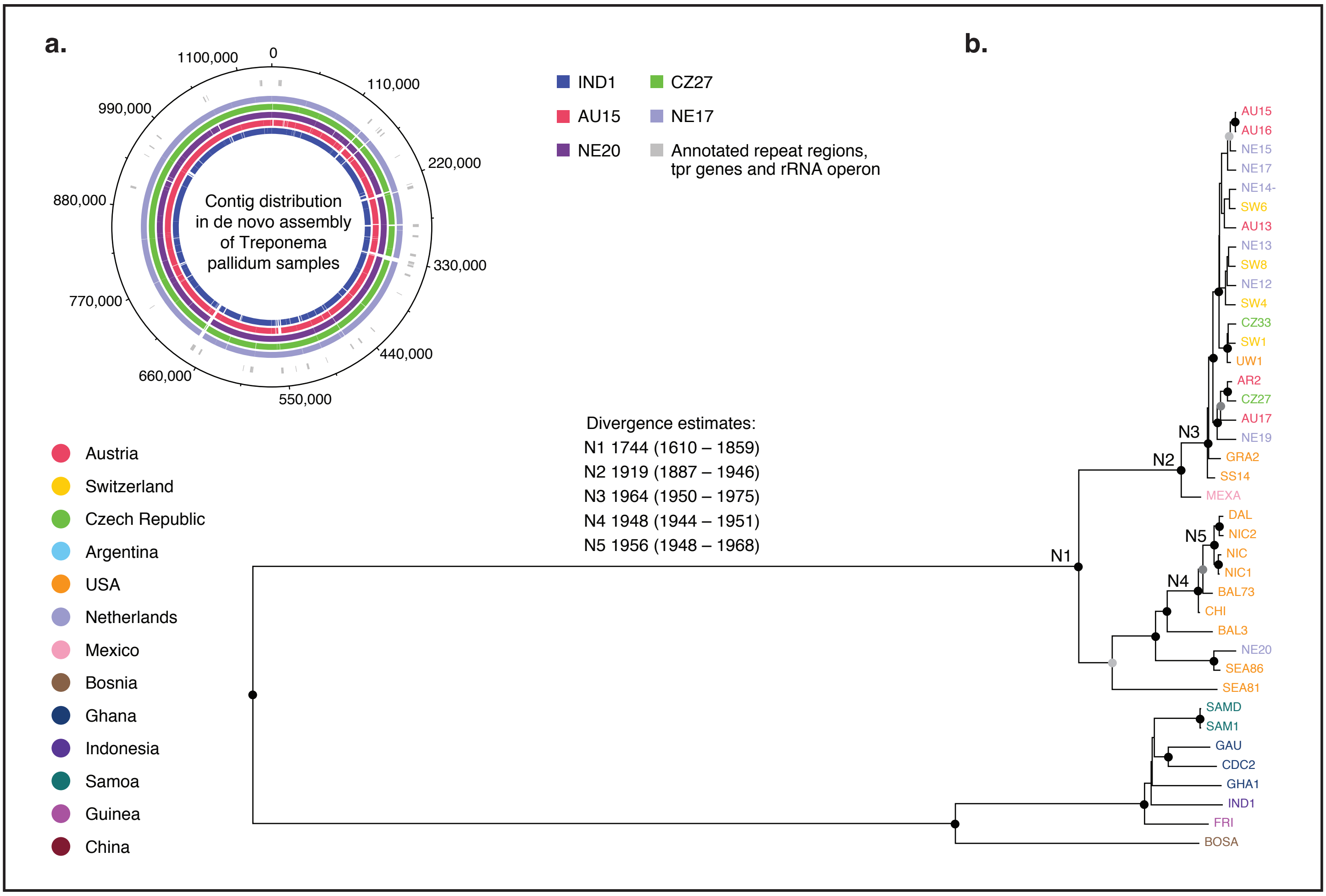
The authors declare no competing financial interests.

Figure Legends

Figure 1 | De novo genome assemblies and phylogenetic reconstruction. a, De novo genome assembly for four syphilis patient samples and one yaws strain, with color coded geographic origin (inset legend). Blank spaces correspond to gaps, overlapping with gene regions that are difficult to assemble from short reads such as the tpr subfamilies and rRNA operons (regions shown in the outermost ring in gray). **b**, BEAST tree for the 39 genomes (excluding putative recombinant genes), with black circles for nodes with ≥96% posterior probabilities (PP); dark gray circles for nodes with 91-95% PP; and white circles for nodes with 81-85% PP. Divergence date estimates (mean and 95% highest posterior density) for major well-supported TPA nodes are given in the legend.

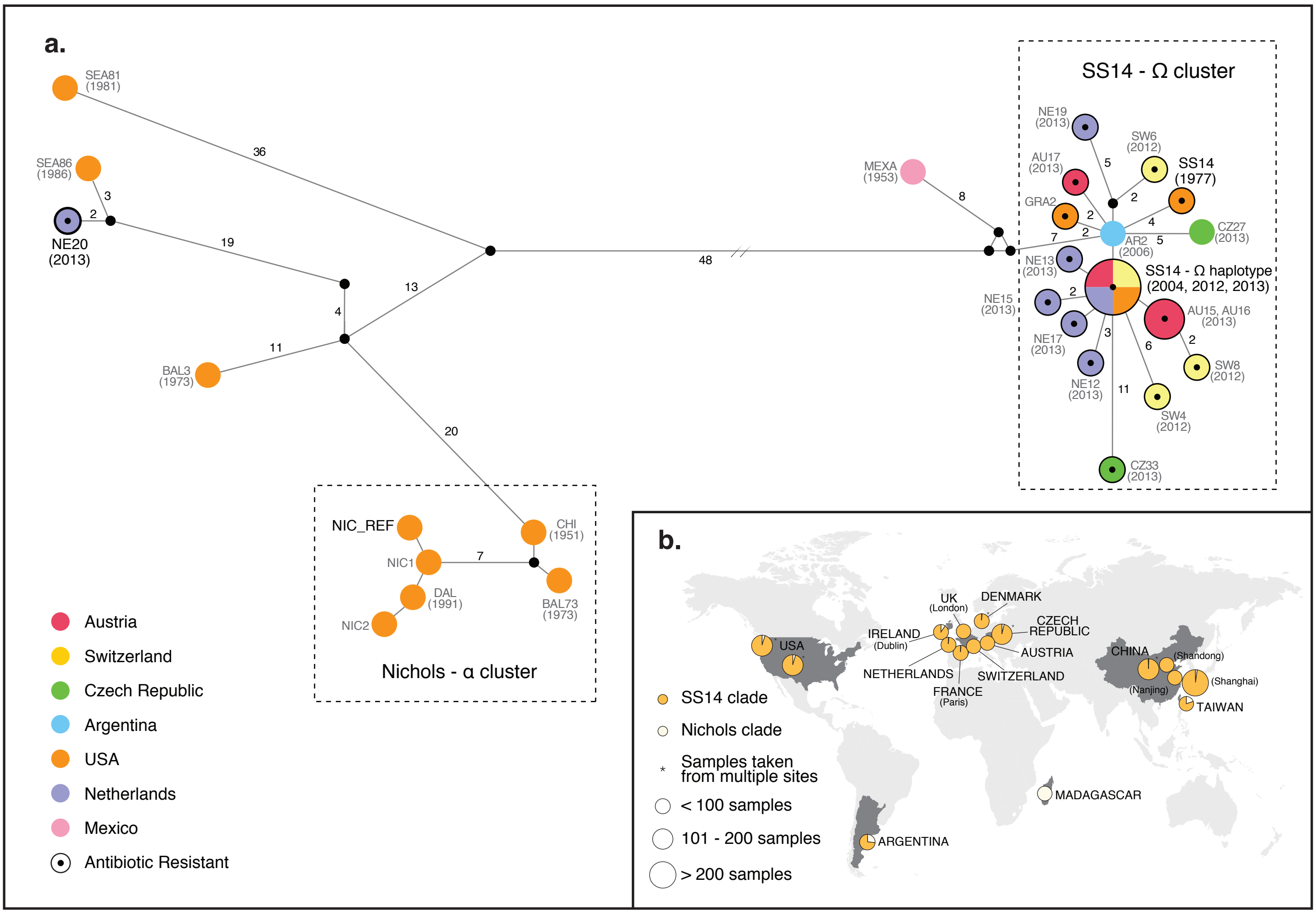
Figure 2│Median-joining (MJ) network analysis and geographic distribution of the SS14 and Nichols clades. **a**, Median-joining network for genome-wide variable positions after excluding sites with missing data (n=682). Circles represent haplotypes, with geographical origin color-coded. Number of mutations, when above one, is shown next to the lines. Inferred haplotypes (median vectors) are shown as black connecting circles. Central black circles within haplotypes indicate mutations associated with azithromycin resistance. **b**, Relative frequencies of SS14 versus Nichols clade isolates across the globe shown in the pie charts, with sizes proportional to sampling efforts. SS14 clade and Nichols classification are based on the TP0548 gene.

FIGURE 1



Divergence estimates:
N1 1744 (1610 – 1859)
N2 1919 (1887 – 1946)
N3 1964 (1950 – 1975)
N4 1948 (1944 – 1951)
N5 1956 (1948 – 1968)

FIGURE 2


a.
SEA81 (1981)
36
SEA86 (1986)
3
2
NE20 (2013)
19
13
4
11
BAL3 (1973)
20
48
MEXA (1953)
8
SS14 - Ω cluster
NE19 (2013)
SW6 (2012)
AU17 (2013)
5
SS14 (1977)
2
GRA2
2
4
CZ27 (2013)
7
2
AR2 (2006)
5
NE13 (2013)
SS14 - Ω haplotype (2004, 2012, 2013)
NE15 (2013)
2
AU15, AU16 (2013)
NE17 (2013)
3
6
2
SW8 (2012)
NE12 (2013)
11
SW4 (2012)
CZ33 (2013)
NIC_REF
CHI (1951)
NIC1
7
DAL (1991)
BAL73 (1973)
NIC2
Nichols - α cluster
Austria
Switzerland
Czech Republic
Argentina
USA
Netherlands
Mexico
Antibiotic Resistant
b.
UK (London)
DENMARK
IRELAND (Dublin)
CZECH REPUBLIC
USA
AUSTRIA
NETHERLANDS
SWITZERLAND
FRANCE (Paris)
CHINA
(Shandong)
(Shanghai)
(Nanjing)
TAIWAN
MADAGASCAR
ARGENTINA
SS14 clade
Nichols clade
* Samples taken from multiple sites
< 100 samples
101 - 200 samples
> 200 samples