

RESEARCH ARTICLE

Whole genome sequences of *Treponema pallidum* subsp. *endemicum* isolated from Cuban patients: The non-clonal character of isolates suggests a persistent human infection rather than a single outbreak

Eliška Vrbová¹, Angel A. Noda², Linda Grillová¹, Islay Rodríguez², Allyn Forsyth^{3,4}, Jan Oppelt⁵, David Šmajs^{1*}

1 Department of Biology, Faculty of Medicine, Masaryk University, Brno, Czech Republic, **2** Department of Mycology-Bacteriology, Institute of Tropical Medicine “Pedro Kourí”, Havana, Cuba, **3** GeneticPrime Dx, Inc., La Jolla, California, United States of America, **4** San Diego State University, San Diego, California, United States of America, **5** Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, United States of America

* dsmajs@med.muni.cz



OPEN ACCESS

Citation: Vrbová E, Noda AA, Grillová L, Rodríguez I, Forsyth A, Oppelt J, et al. (2022) Whole genome sequences of *Treponema pallidum* subsp. *endemicum* isolated from Cuban patients: The non-clonal character of isolates suggests a persistent human infection rather than a single outbreak. PLoS Negl Trop Dis 16(6): e0009900. <https://doi.org/10.1371/journal.pntd.0009900>

Editor: Jonny Alejandro Garcia Luna, CIDEIM: Centro Internacional de Entrenamiento e Investigaciones Medicas, COLOMBIA

Received: October 13, 2021

Accepted: April 21, 2022

Published: June 10, 2022

Copyright: © 2022 Vrbová et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Genome sequence data of C77 and C279 are available from the Genbank database (accession number(s) CP081507, CP078090).

Funding: This work was supported by a grant from the Grant Agency of the Czech Republic [17-25455S] to D.S. The funders had no role in study

Abstract

Bejel (endemic syphilis) is a neglected non-venereal disease caused by *Treponema pallidum* subsp. *endemicum* (TEN). Although it is mostly present in hot, dry climates, a few cases have been found outside of these areas. The aim of this work was the sequencing and analysis of TEN isolates obtained from “syphilis patients” in Cuba, which is not considered an endemic area for bejel. Genomes were obtained by pool segment genome sequencing or direct sequencing methods, and the bioinformatics analysis was performed according to an established pipeline. We obtained four genomes with 100%, 81.7%, 52.6%, and 21.1% breadth of coverage, respectively. The sequenced genomes revealed a non-clonal character, with nucleotide variability ranging between 0.2–10.3 nucleotide substitutions per 100 kbp among the TEN isolates. Nucleotide changes affected 27 genes, and the analysis of the completely sequenced genome also showed a recombination event between *tprC* and *tprI*, in TP0488 as well as in the intergenic region between TP0127–TP0129. Despite limitations in the quality of samples affecting breadth of sequencing coverage, the determined non-clonal character of the isolates suggests a persistent infection in the Cuban population rather than a single outbreak caused by imported case.

Author summary

The incidence of venereal syphilis has greatly increased in the last years, however endemic syphilis (bejel) which have been considered as a disease restricted to dry arid areas such as the Sahel and the Middle East, remain as a neglected disease. In Cuba, which is a tropical country, several bejel cases were unexpectedly detected few years ago in “syphilis” patients

design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

with no records of travel abroad or sex with foreign partners. In this study, we explored the whole genome sequences from four of the Cuban *Treponema pallidum* subsp. *endemicum* (TEN) isolates and the substantial genetic diversity detected among them suggests a persistent infection of TEN within the human population rather than a single outbreak of a TEN isolate introduced from an area where it is typically endemic. This finding has significant implications on this neglected and also possibly tropical disease in terms of geographical/temporal distribution, and highlights the importance of keeping in mind neglected diseases in apparently non-endemic areas.

Introduction

Treponema pallidum ssp. *endemicum* (TEN) is the causative agent of endemic syphilis (bejel), a neglected non-venereal disease that is mostly present in hot, dry areas of the world. TEN treponemes are highly related (99.7% identity at the genome level) to the *Treponema pallidum* ssp. *pallidum* (TPA), the causative agent of syphilis [1].

Acute bejel infections are mostly found among children between two and 15 years. Like syphilis, bejel can be divided into disease stages. In the primary stage, a small, painless ulcer is usually found in the oral cavity or nasopharynx [2] and usually remains undetected. In the secondary stage, numerous lesions appear in several body areas. In the last stage, gummatous lesions or bone alterations can appear. In several documented cases, bejel treponemes also infected the nipples of nursing women or genital areas [3]. Transmission of this disease occurs typically through direct mucosal and skin contact or is transferred by eating utensils or drinking vessels [2].

While bejel is a disease typically associated with dry arid areas such as the Sahel and the Middle East [2], it has also been found in Canada [4], France [5], Japan [6,7], and Cuba [8]. Cases in Canada and France were explained by bejel being imported from endemic areas, Senegal and Pakistan, respectively [4,5]. On the other hand, the bejel cases in Japan and Cuba were originally identified in patients suspected of having syphilis and with no evidence of disease import and having a sexual route of transmission despite being considered as non-venereal disease [6–8].

In bacteriology, the strict definition of a clone tends to be loosened slightly, and clones are defined pragmatically as isolates that are indistinguishable or highly similar, using a particular molecular typing procedure [9]. Among TPA, where molecular typing is widely used, certain predominant genotypes are observed to infect the human population, i.e., the allelic profile 1.3.1. according to the MLST system [10] and 14d/f using enhanced CDC molecular typing [11]. In many cases, individual allelic profiles differs in only one SNV. In this paper, based on the previously determined treponemal genome diversities and treponemal mutation rates, we defined the non-clonal character as a difference between two isolates in more than two nucleotides, which is consistent with at least 10 years period of separate evolution of TEN isolates [10,12].

In the case of TEN, there are very few genomic analyses, and the two available whole genome sequences, differing in 37 single nucleotide variants, come from the reference strains Bosnia A [13] (CP007548) and Iraq B [14] (CP032303). In addition, nine recombinant loci have been detected in TEN isolates during typing studies [15].

In this work, we aimed to determine the whole genome sequences of four TEN isolates from Cuba in order to better assess the clonal character of isolates reported previously using MLST [8,16]. Moreover, the third TEN complete genome is reported.

Methods

Ethics statement

The study protocol was approved by the Research Ethics Committee of Institute of Tropical Medicine “Pedro Kourí” (CEI IPK 44–18), and it was conducted in compliance with the Declaration of Helsinki. All participants provided written consent to participate in the study.

Study design and clinical samples

This observational descriptive study includes four clinical samples from bejel patients collected in years 2014, 2015 and 2017. Patients attended to the Instituto de Medicina Tropical “Pedro Kourí,” Havana, Cuba and preliminary characteristics were previously published [8,16]. Descriptions of the analyzed samples are shown in Table 1. Patients resided in Havana city and did not refer sexual contact with foreigners in the last six months. DNA was isolated using QIAmp DNA mini kits (Qiagen, Hilden, Germany) according to the manufacturer’s instructions. Following DNA isolation, whole genome amplification was performed using REPLI-g Single Cell kits (Qiagen, Hilden, Germany) [15]. The number of TEN copies in samples was determined by real-time PCR using primers targeting *polA* [17]. Samples for sequencing were selected according to the (1) total volume of available sample (at least 7.5 µl), and (2) percent positivity of PCR amplification of 14 different *T. pallidum* intervals (at least 5 out of 14 positive) [18].

Pooled segment genome sequencing (PSGS)

Genomes of samples C178 and C279 were chosen for the PSGS method [13,18,19], reflecting most successful ratios in the amplification of 14 TPIs (*Treponema pallidum* intervals) of pool 1 (11 of 14 for C279 and six for C178). All primers are listed in S1 Table [13]. PCR products were purified using QIAquick PCR Purification Kits (QIAGEN, Hilden, Germany) and were divided into four pools to separate paralogous genes; this was to avoid later misassembly of these genes. TPI 11A containing *tprC* from pool 1 and 25B containing *tprE* from pool 2 was added to pool 4. For sample C178, all PCR products of TPI were mixed in one pool since there were no paralogous regions amplified. PCR products of individual pools were mixed in equimolar amounts. During construction of the DNA library using a Nextera XT DNA Sample Preparation Guide kit, these four pools were labeled with multiplex identifier (MID) adapters. Prepared pools were sequenced using NGS sequencing on an Illumina platform (<150 nt paired-end reads). The numbers of repetitions in the *arp* and TP0470 genes in sample C279 and the sequence of TP0488 in sample C178 were determined using Sanger sequencing.

Table 1. Characteristics of bejel patients and the corresponding samples.

Sample	Year of isolation	Sex	Sexual behavior	Serology	HIV	Localization of lesion	Sufficient amount of DNA ¹	qPCR results (no. of copies/µl)	Amplification of TPIs (positive/all)	Sequencing method
C75 [8,16]	2014	Man	Homosexual	Positive	Positive	Glans penis or foreskin	No	94.1	6/14	Direct sequencing
C77 [8,16]	2014	Man	Homosexual	Negative	Negative	Shaft of the penis	No	2.6	5/14	Direct sequencing
C178 [16]	2015	Man	Homosexual	Positive	Positive	Hands, feet	Yes	Undetectable	6/14	PSGS
C279 [16]	2017	Man	Heterosexual	Positive	Negative	Inguinal area	Yes	1859.6	11/14	PSGS

¹ For completion of genome amplification and sequencing, at least 7.5 µl of the original sample was required

Direct sequencing

Samples C75 and C77 were directly sequenced as described previously with <150 nt paired-end reads [18]. Sample C77 was sequenced with and without DNA enrichment, and data from both approaches were combined in the analysis. For *DpnI* enrichment, 10–40 μ l of the clinical sample was mixed with *DpnI*-coated beads to a final volume of 50 μ l in 1.7 mL Eppendorf tubes as previously described [20].

Bioinformatic analyses

For the PSGS method, reads were aligned to the reference genome Bosnia A (CP007548) using Lasergene software (DNASTAR, Madison, WI, USA). Final sequences were assembled from the consensus of individual pools and the genome sequence was finished with Sanger sequenced PCR products covering sequencing gaps.

The bioinformatic analysis for direct sequencing was performed according to the pipeline described previously [21]. The quality of raw reads was checked using FastQC (v0.11.5) [22]; pre-processing used Cutadapt (v1.15) [23] and Fastx-toolkits (v0.0.14) [24]. The pre-processed reads were mapped to the human genome reference (hg38), and the human-matching reads were removed using BMap (v37.25) [25]. The remaining reads were mapped to the TEN reference genome of Bosnia A (CP007548) using BWA MEM (v0.7.15) [26]. The post-processing of the mapping was performed using Samtools (v1.4) [27], Picard (v2.8.1) [28], GATK (v3.7) [29], and NGSUtils/bamutils (v0.5.9, commit a7f08f5) [30].

Final sequences were determined by at least two good-quality aligned reads matching the reference genome. To identify SNVs and other changes, at least three good-quality aligned reads were required. The length of homopolymeric region that was found in more than 70% of reads was used in the consensus sequence. When the length of homopolymeric region have predominant reads less frequent than 70% of reads, it was Sanger sequenced.

Phylogenetic analyses

For phylogenetic analyses, partial genomes and a whole-genome of isolate C279 were compared with TEN strains Bosnia A (CP007548), Iraq B (CP032303), isolate 11q/j (KY120774-KY120814), and with Japanese bejel isolates Kyoto-2017 (LC430601, LC430606, LC589183), Osaka-2017A (LC383801, LC430605, LC589184), Osaka-2017B (LC430602, LC430607, LC589185), Osaka-2018 (LC430603, LC430608, LC589186), Japan326e (CP073518), Japan 320e (CP073523), Japan 322e (CP073522), and Japan 346e (CP073506). Phylogenetic trees were constructed using the Maximum Likelihood Method [31] with bootstrapping [32] in MEGA7 software [33]. MEGA7 software [33] was also used for analysis of nucleotide diversity within genomes, which were calculated as number of SNVs per analyzed positions, and then modified to number of SNVs per 100 kbp.

Annotation of complete genomes

For gene annotation, Geneious software (2021.1.1.) was used. The *tprK* gene showed intra-strain variability in all samples, and the corresponding nucleotide positions were denoted as “N.” Genes from the TEN C279 strain were tagged with the TENDC279_ prefix. In C279, the original locus tag numbering corresponds to the tag numbering of orthologous genes annotated in the TEN Bosnia A genome. In the draft genome of TEN C77 genes were annotated based on TEN Bosnia A and tagged TENDC77. Genome sequences of isolates C279 and C77 were deposited in the NCBI database under the following accession numbers (raw data were deposited under BioProject numbers): CP078090 (PRJNA742865), and CP081507

Table 2. Sequencing results of four samples containing TEN isolated from Cuban patients.

Sample	Sequencing method	No. of mapped reads/no. of total reads	Final sequence length	Final average coverage	Breadth coverage*	No. of SNVs and indels compared to Bosnia A genome	Complete genome sequence
C279	PSGS [#]	29,337,833/29,894,321	1,137,501	3,843x	100%	157	Yes
C178	PSGS [#]	3,262,191/3,416,513	240,401	431x	21.1%	56	No
C75	Direct seq [†]	18,824/77,057,557	598,566	2x	52.6%	30	No
C77	Direct seq [†]	46,816/333,057,683	929,366	5x	81.6%	76	No

*when compared to Bosnia A genome

[#]Pooled segment genome sequencing—sequencing of genome divided into 4 pools consisting of multiple TPIs, which were individually amplified

[†]Whole genome sequencing without previous amplification, enriched by *DpnI*

<https://doi.org/10.1371/journal.pntd.0009900.t002>

(PRJNA743168), respectively. Genome sequences of isolates C75 (PRJNA743169) and C178 (PRJNA743172) are available in the Supplementary material of this article (S1 and S2 Text).

Results

The overall results of the whole genome sequencing are shown in Table 2. Comparison of individual genome breadth of coverage is in Fig 1.

Sequence diversity in the Cuban TEN genomes

Comparison of whole genome sequence of TEN C279 with the partially determined genome sequences of C77, C75, and C178 showed sequence diversity among the samples. Differences among genomes and pairwise nucleotide diversity values are summarized in Table 3 and Fig 2A. The phylogenetic relationship between C279 and the previously reported TEN isolates is presented in Fig 2B. For construction of the tree, regions TP0136 (158 208–159 279), TP0548 (591,226–592,285) and TP0856 (932,947–933,182) were used. C75, C77, and C178 were omitted due to unavailable sequences of used loci. For the same reason, Japanese isolate Osaka-2014 was not used. As a root, TPE strain Samoa D (CP002374) was used.

Compared to the complete genome of TEN Bosnia A, TEN Cuban genomes showed 3.9–39.4 nucleotide substitutions per 100 kb of the genome, varying by one order of magnitude.

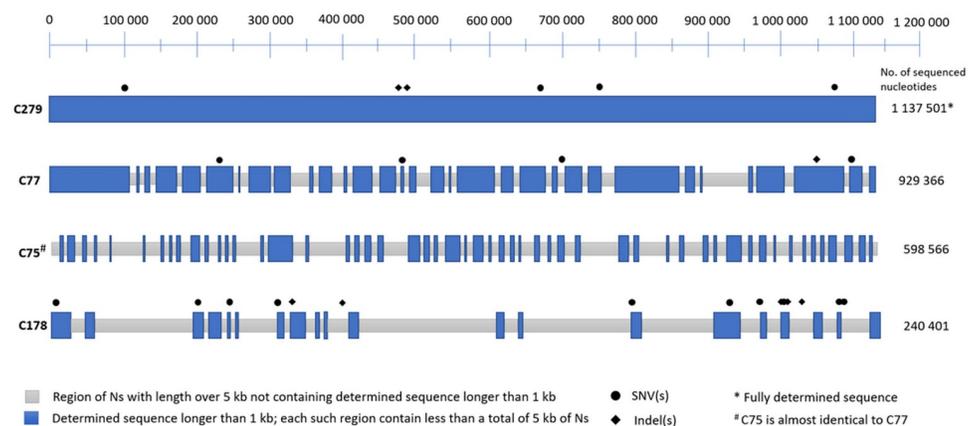


Fig 1. Comparison of breadth coverage of individual genomes. Sequenced genomic region of Cuban isolates with marked total sequenced genome length. Gray represents regions of Ns with length at least 5 kb, which were not interrupted with more than 1 kb of sequenced nucleotides. Regions containing SNVs and indels in Cuban bejel genomes are shown as symbols (•, ♦).

<https://doi.org/10.1371/journal.pntd.0009900.g001>

Table 3. Nucleotide diversity among Cuban bejel genomes. Nucleotide diversity was calculated as the number of nucleotide substitutions per 100,000 bp.

	C279 (100%)	C77 (81.69%)	C75 (52.61%)	C178 (21.13%)	Bosnia A (100%)
C279 (100%)*	-	1.1 [10 SNVs] 5 indels**	0.3 [2 SNVs] 3 indels	9.2 [23 SNVs] 7 indels	39.4 [145 SNVs] 12 indels
C77 (81.69%)	1.1 [10 SNVs] 5 indels	-	0.2 [1 SNV]	7.0 [13 SNVs] 5 indels	6.6 [71 SNVs] 5 indels
C75 (52.61%)	0.3 [2 SNVs] 3 indels	0.2 [1 SNV]	-	10.3 [13 SNVs] 3 indels	3.9 [28 SNVs] 2 indels
C178 (21.13%)	9.2 [23 SNVs] 7 indels	7.0 [13 SNVs] 5 indels	10.3 [13 SNVs] 3 indels	-	20.4 [47 SNVs] 9 indels

SNV, single nucleotide variation

*Breath of coverage of the corresponding whole genome sequence

**Indels were not used for nucleotide diversity calculations.

*Clonal genomes (according our definition at least 2 SNVs are needed for non-clonality).

<https://doi.org/10.1371/journal.pntd.0009900.t003>

The most divergent genomes were C279 and C178, with a nucleotide diversity of 39.4 and 20.4, respectively. On the other hand, diversity from Bosnia A was relatively small in samples C75 and C77. Genetic diversity within the Cuban samples was in the range 0.2–10.3, i.e., somewhat lower compared to the diversity between Cuban samples and Bosnia A.

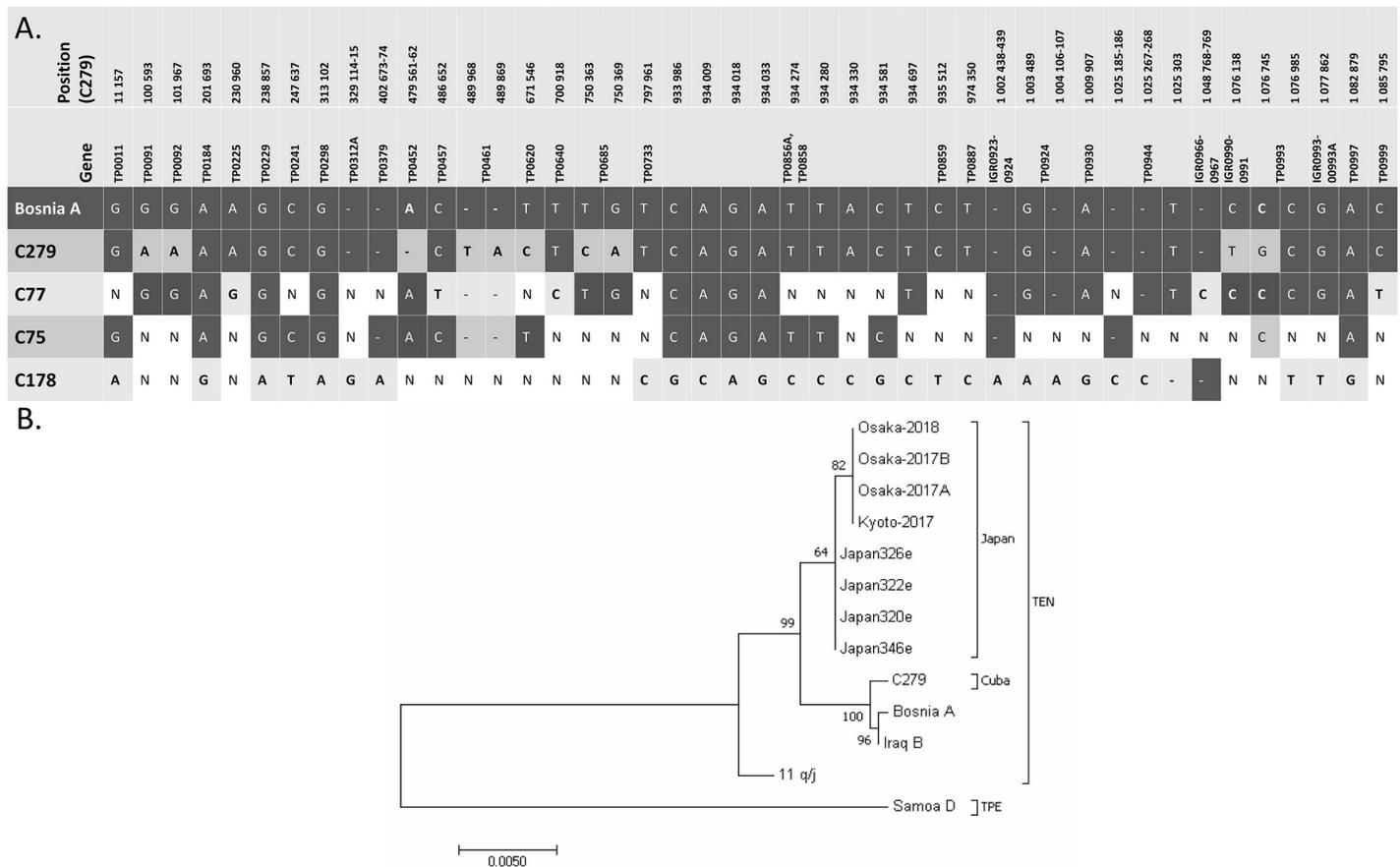


Fig 2. Genetic diversity among Cuban treponemal isolates. (A) Detailed visualization of genetic differences among Cuban bejel strains. N, the corresponding nucleotide was not determined. (B) Phylogenetic analysis of available TEN genomes or sequences. The evolutionary history was inferred using the Maximum Likelihood method based on the Tamura-Nei model [34]. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The percentage of trees in which the associated taxa clustered together is shown next to the branches. The analysis involved 13 nucleotide sequences. There were a total of 2299 positions in the final dataset with 119 variable positions.

<https://doi.org/10.1371/journal.pntd.0009900.g002>

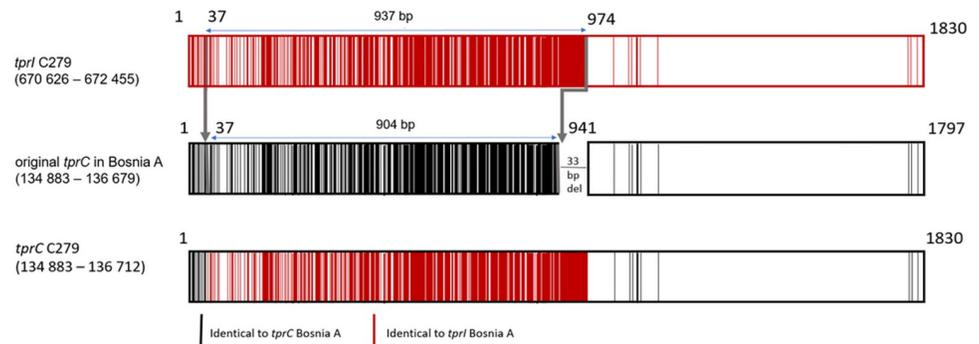


Fig 3. Overview of recombinant *tprC* in the C279 genome. Recombinant sites within the *tprC* gene and the resulting *tprCI* chimera of C279. Black lines represent nucleotides identical to *tprC*, while red lines nucleotides to *tprI*. White color represents nucleotides shared between both *tprC* and *tprI* genes.

<https://doi.org/10.1371/journal.pntd.0009900.g003>

New recombination events in the C279 genome

In addition to the inter-strain recombination events at the TP0488 and TP0548 loci described earlier [5]; the C279 genome contains (compared to Bosnia A) a total of 332 changes in the *tprC* gene (TP0117), which is, in part, sequentially identical to the *tprI* gene of Bosnia A (Fig 3). The *tprC* of C279 is thus a result of an intrastrain recombination event where the recombinant gene contains about half of *tprI* copied to *tprC*.

Between TP0127b and the pseudogene TP0128, the TEN genome C279 showed a 72 bp deletion followed by the duplication of a 53 bp-long sequence from TP0129. Similar to the *tprCI* chimera, the 53 bp-long sequence appears to be copied from TP0129 and inserted at the front of the TP0128 pseudogene. In other treponemal genomes, the region comprising TP0127-TP0129 was shown to be highly variable, including TPA Nichols, TEN Bosnia A, and TPeC strain Cuniculi A (Fig 4).

Discussion

While bejel typically occurs in dry arid areas such as the Sahel and the Middle East–Cuba, with its hot, humid weather, is not a typical region for bejel to occur [2]. In addition, Cuban bejel cases appear to be sexually transmitted, mostly among homosexual men, which is also not typical for bejel [2]. While in several countries, including Canada and France, bejel cases were found to have been imported, no such indication exists for bejel in Cuba [8,16]. This atypical occurrence and overlap of treponematoses are highlighting importance of molecular typing.

Whereas a previous paper on molecular typing of syphilis treponemes (TPA) revealed the presence of TEN, all of the nine independent isolates were found to be genetically indistinguishable [8,16]. A similar situation was recently published for Japanese bejel cases where it was found that all isolates were identical when two chromosomal loci were sequenced [6]. Unlike the Japanese isolates [6], Cuban genomes (C279, C178) did not contain mutations (A2058G) leading to macrolide resistance.

Since 2013, there have been numerous reports describing traces of genetic recombination events among pathogenic treponemes at more than a dozen genetic loci [34,35]. While there are examples of both intra- and interstrain recombination events [18], in the genomes of the Cuban strains, we identified two possible intrastrain recombination events. The first intrastrain recombination event between *tprG* and *tprJ* has also been described in TPA, where a sequence from *tprJ* was copied to the *tprG* gene [18]. Here, we have described a new gene conversion event where the sequence from *tprI* was copied to the *tprC* locus. Despite the donor

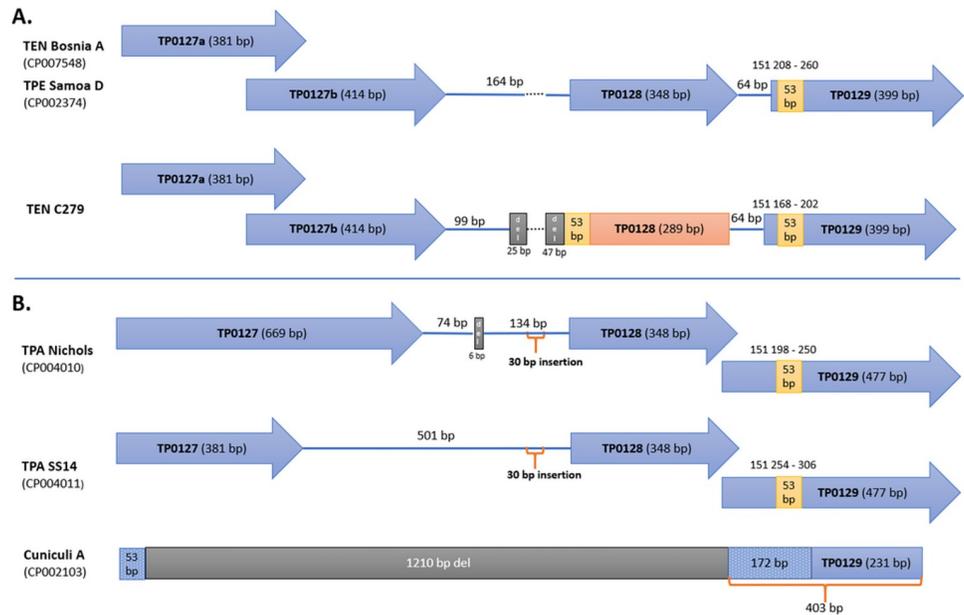


Fig 4. Comparison of region TP0127-TP0129. A. Comparison between TEN Bosnia A/TPE Samoa D and TEN C279. B. Region of TP0127-TP0129 in selected strains of other subspecies. TEN Bosnia A was used as the reference strain, and the TPE Samoa D strain was identical to TEN Bosnia A. The TEN genome C279 showed a 72 bp deletion and a 53 bp duplication compared to TP0129 of TEN Bosnia A. TPA Nichols strain in IGR showed a 6 bp deletion and a 30 bp insertion compared to Bosnia A. Most of the sequence of Cuniculi A showed deletions, except for 225 bp, which was sequentially unrelated to TP0127-TP0128 in the other strains.

<https://doi.org/10.1371/journal.pntd.0009900.g004>

tprI sequence is present also in other genomes, it is more likely that the donor sequence of *tprI* originated from the same genome. Previous studies also revealed changes at the TP0126–TP0127 loci in the TPA Nichols and TPE strains [35,36], showing that this region differs in gene prediction and region size. Moreover, the region comprising TP0129–TP0130 was previously described as one of the major donor sites for the variable regions of the *tprK* gene [37]. The emergence of the 53 bp-long sequences copied from the TP0129 by intrastrain gene conversion is therefore not surprising in this region.

The interstrain recombination events in the TP0488 and TP0548 were described earlier [5]. From a previous study, we know that in most of the samples the TP0488 gene sequence was the same as in the 11q/j isolate and therefore likely represented a recombinant allele [5]. Another similarity between C279 and isolate 11q/j was found in gene TP0865, which contains a 23 bp insertion in the same position as the 11q/j harbor 22 bp-long insertions. The size of insertions differs because of the different numbers of nucleotides in the homopolymeric tract. In both cases (i.e., in C249 and in 11q/j), the gene is considered to be non-functional, containing a frameshift mutation. In the C279 genome, the recombinant allele also seems to be present at the TP0548 locus, similar to Bosnia A and Iraq B versions but different from the TEN 11q/j strain [5].

Whole-genome sequencing revealed a surprising amount of genetic diversity within the Cuban isolates ranging between 0.2–10.3 differences per 100 kb, which corresponds to an estimated 22.8 to 117.4 nucleotide differences between individual strains on a genome-wide level. However, these changes are not visible in phylogenetic trees due to the differences in sequenced genome segments and omission of undetermined positions (N nucleotides) from the phylogenetic tree. This extent of genetic diversity is comparable to differences between TEN Bosnia and Iraq B with an average value of nucleotide diversity of 3.1, with 37 single-

nucleotide differences, four indels, two differences in the number of tandem repetitions, and 18 differences in the length of homopolymeric regions found in the Iraq B genome [14] compared to Bosnia A [13]. While TEN strain Iraq B was isolated in Iraq (the Middle East, south-western Asia) in 1951, the strain Bosnia A was isolated in 1950 in Bosnia, southern Europe. A similar extent of genetic diversity was also found between TPE strains at different times and different geographical regions from which whole genome sequences are available (Samoa D, Gauthier, Kampung Dalan 363, Sei Geringging K403, CDC-1, CDC-2, CDC 2575, and Ghana-051) [12,36,38] ranging between 0.0–20.5 differences per 100 kb. This fact further supports the non-clonal character of Cuban TEN strains and is consistent with the long-term evolution of each of the Cuban TEN isolates. For comparison, TPA SS14-like strains obtained by direct sequencing [18] have nucleotide diversity in a range of 0.2–4.6.

Uncultivable pathogenic treponemes including TPA, TPE, and TEN are monomorphic bacteria [1,39], which have extremely high sequence similarity; therefore, it is likely that these related treponemes show similar mutation rates [39]. While there are no studies on mutation rates in TEN, the mutation rate in TPE and TPA have been estimated as 1.21×10^{-7} and 0.82×10^{-7} per nucleotide site per year or lower, respectively [10,12]. Assuming that the mutation rate is similar in TEN as in TPA and TPE, the estimated 22.8 to 117.4 nucleotide differences between individual TEN strains detected in Cuba suggest at least several hundreds of years of separate evolution in the human population and, therefore, the long-term existence of the different TEN strains in the Cuban or related population. On the other hand, several studies presented higher treponemal mutation rates, e.g., 3.02×10^{-7} [7] or 6.6×10^{-7} [40] per nucleotide site per year, which would suggest several decades of separate evolution considering the lowest nucleotide difference of 22.8 nt.

Altogether, the findings presented in this study suggest that there are several different sequence types of TEN strains circulating in the population of Cuba. Evidence suggests that they are being transferred mainly through sexual transmission since all source patients were suspected of having syphilis, and most of them had ulcerations in the genital area. Unlike the established concept of treponemal subspecies and corresponding diseases, this study points to the fact that at least in the early stages of the disease, both bejel and syphilis treponemes produce symptoms that are indistinguishable [2,8,16].

The authors recognize that the main limitation of the study was the low number of patients with high-quality samples suitable for sequencing; better samples would have allowed a more robust sequence analysis.

Although, we cannot fully exclude parallel recent TEN introductions to Cuba from an endemic area(s), the non-clonal character of the Cuban TEN isolates suggests that this bacterium cause persistent infections within the Cuban human population.

Supporting information

S1 Table. Primers used in the Pooled Segment Genome Sequencing.
(XLSX)

S1 Text. Sequence of C75.
(DOCX)

S2 Text. Sequence C178.
(DOCX)

Acknowledgments

We thank Thomas Secrest (Secrest Editing, Ltd.) for the English editing of the manuscript.

Author Contributions

Conceptualization: David Šmajš.

Data curation: Eliška Vrbová.

Formal analysis: Eliška Vrbová, Linda Grillová, Islay Rodríguez, Allyn Forsyth, Jan Oppelt, David Šmajš.

Funding acquisition: David Šmajš.

Investigation: Eliška Vrbová.

Methodology: Eliška Vrbová, Linda Grillová, Allyn Forsyth, Jan Oppelt.

Resources: Angel A. Noda, Islay Rodríguez.

Software: Jan Oppelt.

Supervision: David Šmajš.

Validation: Linda Grillová, David Šmajš.

Visualization: Eliška Vrbová.

Writing – original draft: Eliška Vrbová, Angel A. Noda, David Šmajš.

Writing – review & editing: Eliška Vrbová, Angel A. Noda, David Šmajš.

References

1. Šmajš D, Norris SJ, Weinstock GM. Genetic diversity in *Treponema pallidum*: implications for pathogenesis, evolution and molecular diagnostics of syphilis and yaws. *Infect Genet Evol.* 2012; 12: 191–202. <https://doi.org/10.1016/j.meegid.2011.12.001> PMID: 22198325
2. Giacani L, Lukehart SA. The Endemic Treponematoses. *Clin Microbiol Rev.* 2014; 27: 89–115. <https://doi.org/10.1128/CMR.00070-13>
3. Grin EI. Endemic syphilis in Bosnia; clinical and epidemiological observations on a successful mass-treatment campaign. *Bull World Health Organ.* 1952; 7: 1–74. PMID: 13019544
4. Fanella S, Kadkhoda K, Shuel M, Tsang R. Local transmission of imported endemic syphilis, Canada, 2011. *Emerg Infect Dis.* 2012; 18: 1002–1004. <https://doi.org/10.3201/eid1806.111421> PMID: 22607961
5. Mikalová L, Strouhal M, Oppelt J, Grange PA, Janier M, Benhaddou N, et al. Human *Treponema pallidum* 11q/j isolate belongs to subsp. endemicum but contains two loci with a sequence in TP0548 and TP0488 similar to subsp. pertenuis and subsp. pallidum, respectively. *PLoS Negl Trop Dis.* 2017; 11: e0005434–e0005434. <https://doi.org/10.1371/journal.pntd.0005434> PMID: 28263990
6. Kawahata T, Kojima Y, Furubayashi K, Shinohara K, Shimizu T, Komano J, et al. Bejel, a Nonvenereal Treponematoses, among Men Who Have Sex with Men, Japan. *Emerg Infect Dis.* 2019; 25: 1581–1583. <https://doi.org/10.3201/eid2508.181690> PMID: 31310214
7. Lieberman NAP, Lin MJ, Xie H, Shrestha L, Nguyen T, Huang M-L, et al. *Treponema pallidum* genome sequencing from six continents reveals variability in vaccine candidate genes and dominance of Nichols clade strains in Madagascar. *PLoS Negl Trop Dis.* 2021; 15: e0010063. <https://doi.org/10.1371/journal.pntd.0010063> PMID: 34936652
8. Noda AA, Grillová L, Lienhard R, Blanco O, Rodríguez I, Šmajš D. Bejel in Cuba: molecular identification of *Treponema pallidum* subsp. endemicum in patients diagnosed with venereal syphilis. *Clinical Microbiology and Infection.* 2018; 24: 1210.e1–1210.e5. <https://doi.org/10.1016/j.cmi.2018.02.006> PMID: 29454847
9. Spratt BG. Exploring the Concept of Clonality in Bacteria. In: Woodford N, Johnson AP, editors. *Genomics, Proteomics, and Clinical Bacteriology: Methods and Reviews.* Totowa, NJ: Humana Press; 2004. pp. 323–352. <https://doi.org/10.1385/1-59259-763-7:323>
10. Grillová L, Bawa T, Mikalová L, Gayet-Ageron A, Nieselt K, Strouhal M, et al. Molecular characterization of *Treponema pallidum* subsp. pallidum in Switzerland and France with a new multilocus sequence typing scheme. *PLoS One.* 2018; 13: e0200773–e0200773. <https://doi.org/10.1371/journal.pone.0200773> PMID: 30059541

11. Marra CM, Sahi SK, Tantaló LC, Godornes C, Reid T, Behets F, et al. Enhanced molecular typing of *Treponema pallidum*: geographical distribution of strain types and association with neurosyphilis. *J Infect Dis*. 2010; 202: 1380–1388. <https://doi.org/10.1086/656533> PMID: 20868271
12. Strouhal M, Mikalová L, Havlíčková P, Tenti P, Čejková D, Rychlík I, et al. Complete genome sequences of two strains of *Treponema pallidum* subsp. *pertenue* from Ghana, Africa: Identical genome sequences in samples isolated more than 7 years apart. *PLoS Negl Trop Dis*. 2017; 11: e0005894. <https://doi.org/10.1371/journal.pntd.0005894> PMID: 28886021
13. Staudová B, Strouhal M, Zobaníková M, Čejková D, Fulton LL, Chen L, et al. Whole genome sequence of the *Treponema pallidum* subsp. *endemicum* strain Bosnia A: the genome is related to yaws treponemes but contains few loci similar to syphilis treponemes. *PLoS Negl Trop Dis*. 2014; 8: e3261. <https://doi.org/10.1371/journal.pntd.0003261> PMID: 25375929
14. Mikalová L, Janečková K, Nováková M, Strouhal M, Čejková D, Harper KN, et al. Whole genome sequence of the *Treponema pallidum* subsp. *endemicum* strain Iraq B: A subpopulation of bejel treponemes contains full-length *tprF* and *tprG* genes similar to those present in *T. p.* subsp. *pertenue* strains. *PLoS ONE*. 2020; 15: e0230926. <https://doi.org/10.1371/journal.pone.0230926> PMID: 32236138
15. Noda AA, Méndez M, Rodríguez I, Šmajš D. Genetic recombination in *Treponema pallidum*: Implications for diagnosis, epidemiology, and vaccine development. *Sexually Transmitted Diseases*. 2021 [cited 20 Jul 2021]. <https://doi.org/10.1097/OLQ.0000000000001497> PMID: 34618784
16. Grillová L, Noda AA, Lienhard R, Blanco O, Rodríguez I, Šmajš D. Multilocus Sequence Typing of *Treponema pallidum* subsp. *pallidum* in Cuba From 2012 to 2017. *J Infect Dis*. 2019; 219: 1138–1145. <https://doi.org/10.1093/infdis/jiy604> PMID: 30325448
17. Dubourg G, Edouard S, Prudent E, Fournier P-E, Raoult D. Incidental Syphilis Diagnosed by Real-Time PCR Screening of Urine Samples. *J Clin Microbiol*. 2015; 53: 3707–3708. <https://doi.org/10.1128/JCM.01026-15> PMID: 26292307
18. Grillová L, Oppelt J, Mikalová L, Nováková M, Giacani L, Niesnerová A, et al. Directly Sequenced Genomes of Contemporary Strains of Syphilis Reveal Recombination-Driven Diversity in Genes Encoding Predicted Surface-Exposed Antigens. *Front Microbiol*. 2019; 10: 1691. <https://doi.org/10.3389/fmicb.2019.01691> PMID: 31417509
19. Weinstock GM, Smajs D, Hardham J, Norris SJ. From microbial genome sequence to applications. *Res Microbiol*. 2000; 151: 151–158. [https://doi.org/10.1016/s0923-2508\(00\)00115-7](https://doi.org/10.1016/s0923-2508(00)00115-7) PMID: 10865961
20. Barnes HE, Liu G, Weston CQ, King P, Pham LK, Waltz S, et al. Selective microbial genomic DNA isolation using restriction endonucleases. *PLoS One*. 2014; 9: e109061. <https://doi.org/10.1371/journal.pone.0109061> PMID: 25279840
21. Grillová L, Giacani L, Mikalová L, Strouhal M, Strnadl R, Marra C, et al. Sequencing of *Treponema pallidum* subsp. *pallidum* from isolate UZ1974 using Anti-Treponemal Antibodies Enrichment: First complete whole genome sequence obtained directly from human clinical material. *PLoS One*. 2018; 13: e0202619. <https://doi.org/10.1371/journal.pone.0202619> PMID: 30130365
22. Andrew S. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. 2010.
23. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal*. 2011; 17: 10–12. <https://doi.org/10.14806/ej.17.1.200>
24. Gordon A. (2014). FASTX-Toolkit: FASTQ/A Short-Reads Pre-Processing Tools. Available at: http://hannonlab.cshl.edu/fastx_toolkit/ (accessed November 29, 2016).
25. Bushnell B. (2017). BBMap. Available at: sourceforge.net/projects/bbmap/ (accessed May 25, 2017).
26. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*. 2014; 30: 2843–2851. <https://doi.org/10.1093/bioinformatics/btu356> PMID: 24974202
27. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009; 25: 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352> PMID: 19505943
28. Broad Institute (2015). Picard Toolkit. Available at: <http://broadinstitute.github.io/picard/> (accessed October 24, 2018).
29. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010; 20: 1297–1303. <https://doi.org/10.1101/gr.107524.110> PMID: 20644199
30. Breese MR, Liu Y. NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. *Bioinformatics*. 2013; 29: 494–496. <https://doi.org/10.1093/bioinformatics/bts731> PMID: 23314324

31. Tamura K, Nei M, Kumar S. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc Natl Acad Sci U S A*. 2004; 101: 11030–11035. <https://doi.org/10.1073/pnas.0404206101> PMID: 15258291
32. Felsenstein J. CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP. *Evolution*. 1985; 39: 783–791. <https://doi.org/10.1111/j.1558-5646.1985.tb00420.x> PMID: 28561359
33. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol*. 2016; 33: 1870–1874. <https://doi.org/10.1093/molbev/msw054> PMID: 27004904
34. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*. 1993; 10: 512–526. <https://doi.org/10.1093/oxfordjournals.molbev.a040023> PMID: 8336541
35. Šmajš D, McKeivitt M, Wang L, Howell JK, Norris SJ, Palzkill T, et al. BAC Library of *T. pallidum* DNA in *E. coli*. *Genome Res*. 2002; 12: 515–522. <https://doi.org/10.1101/gr.207302> PMID: 11875041
36. Čejková D, Zbaníková M, Chen L, Pospíšilová P, Strouhal M, Qin X, et al. Whole Genome Sequences of Three *Treponema pallidum* ssp. *pertenue* Strains: Yaws and Syphilis *Treponemes* Differ in Less than 0.2% of the Genome Sequence. *PLoS Negl Trop Dis*. 2012; 6: e1471. <https://doi.org/10.1371/journal.pntd.0001471> PMID: 22292095
37. Centurion-Lara A, LaFond RE, Hevner K, Godornes C, Molini BJ, Voorhis WCV, et al. Gene conversion: a mechanism for generation of heterogeneity in the *tprK* gene of *Treponema pallidum* during infection. *Molecular Microbiology*. 2004; 52: 1579–1596. <https://doi.org/10.1111/j.1365-2958.2004.04086.x> PMID: 15186410
38. Strouhal M, Mikalová L, Haviernik J, Knauf S, Bruisten S, Noordhoek GT, et al. Complete genome sequences of two strains of *Treponema pallidum* subsp. *pertenue* from Indonesia: Modular structure of several treponemal genes. *PLoS Negl Trop Dis*. 2018; 12: e0006867. <https://doi.org/10.1371/journal.pntd.0006867> PMID: 30303967
39. Šmajš D, Strouhal M, Knauf S. Genetics of human and animal uncultivable treponemal pathogens. *Infect Genet Evol*. 2018; 61: 92–107. <https://doi.org/10.1016/j.meegid.2018.03.015> PMID: 29578082
40. Arora N, Schuenemann VJ, Jäger G, Peltzer A, Seitz A, Herbig A, et al. Origin of modern syphilis and emergence of a pandemic *Treponema pallidum* cluster. *Nature Microbiology*. 2016; 2: 1–6. <https://doi.org/10.1038/nmicrobiol.2016.245> PMID: 27918528