Lineage-specific *tprK* diversification and *Treponema pallidum* transmission dynamics in Buenos Aires, Argentina

Nicole A.P. Lieberman, PhD[1*], Luciana N. Garcia, PhD[2,3*], Shah A. K. Mohamed Bakhash, BS[1], Jeff Furlong, BA[1], B. Ethan Nunley, BS[1], Andrés Rabinovich, PhD[2,3], Patricia Fernandez Pardal, MD[4], Viviana Leiro, MD[4], Hong Xie, MD[1], Farhang Aghakhanian, PhD[5], Mauro Romero Leal Passos, PhD[6], Wilma Nancy Campos Arze, MD[6], Hugo Boechat Andrade, MD[6], Silver K. Vargas, PhD[7], Kelika A. Konda, PhD[7,8], Michael Reyes Diaz, MD[8], Carlos Caceres, MD[8], Jeffrey D. Klausner, MD[9], Jonathan B. Parr, MD[5,10], Arlene Seña, MD[5,10], Ariyaratne Manathunge, MD[11], Lorenzo Giacani, PhD[12,13], Jaime Altcheh, PhD[2,3**], Alexander L. Greninger, MD[1,14**]

*equal contributions as first author

**equal contributions as senior author

Correspondence:

Nicole Lieberman, naliebe@uw.edu; Alex Greninger, agrening@uw.edu

[1]Department of Laboratory Medicine and Pathology, University of Washington School of Medicine, Seattle, WA, USA

[2]Servicio Parasitología- Chagas, Hospital de Niños Ricardo Gutierrez, Capital Federal, Buenos Aires, Argentina

[3]Instituto Multidisciplinario de Investigaciones en Patologías Pediátricas (IMIPP), CONICET, Buenos Aires, Argentina

[4]Hospital de Infecciosas F J Muñiz, Servicio de Dermatología, Buenos Aires City, Argentina

[5]Institute for Global Health and Infectious Diseases, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[6]Department of Microbiology and Parasitology, Biomedical Institute, Fluminense Federal University, Niterói, RJ, Brazil

[7]Center for Interdisciplinary Studies in Sexuality, AIDS and Society, Universidad Peruana Cayetano Heredia, Lima, Peru

[8]Division of Infectious Disease, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA

[9]Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

35 [10]Department of Medicine, Division of Infectious Diseases, Institute for Global Health and
36 Infectious Diseases, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
37 [11]National STD/AIDS Control Programme, Ministry of Health, Colombo, Sri Lanka
38 [12]Department of Medicine, Division of Allergy and Infectious Diseases, Harborview Medical
39 Center, University of Washington, Seattle, WA, USA
40 [13]Department of Global Health, Harborview Medical Center, University of Washington, Seattle,
41 WA, USA
42 [14]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center, Seattle, WA, USA
43

Abstract

Background: Syphilis rates are rising globally, with increases in congenital syphilis in South America particularly concerning. The characterization of contemporary South American *Treponema pallidum (Tp)* strains is crucial to syphilis vaccine development, yet few genomic epidemiology studies have focused on this region. Here, we performed whole genome sequencing of *Tp* from Buenos Aires, Argentina, as well as deep sequencing of the hypervariable *tprK* locus, which is critical to *Tp* immune evasion.

Methods: People with primary, secondary, or congenital syphilis were enrolled at two clinics in Buenos Aires between October 2018 and January 2023, including individuals associated with intra-household transmission. *Tp* DNA from swabs was quantified by *tp0574* qPCR, and whole-genome sequencing was performed on samples with sufficient treponemal burden. *Tp* reads were assembled to the SS14 strain reference genome, recombinant regions masked, and a core genome phylogeny was generated. Full-length *tprK* was sequenced using PacBio reads.

Findings: *Tp* genomes were recovered from 96 samples from 70 individuals in Buenos Aires and primarily belonged to globally dominant SS14 sublineage-1 and Nichols sublineage-8, as did *Tp* recovered from contemporary samples from Brazil (n=8). Peruvian samples (n=3) all belonged to sublineage-1. Two individuals from Buenos Aires had co-infections with Nichols- and SS14-lineage strains. Macrolide resistance via A2058G mutation occurred in 27/70 (39%) samples. Across 56 samples, *tprK* allelic diversity was significantly increased in secondary syphilis, oral lesions, and SS14-lineage strains compared to primary syphilis, anogenital lesions, and Nichols-lineage strains, respectively. Increased *tprK* diversity in SS14-lineage strains is driven by an enhanced repertoire of V7-specific donor sequences. Using multiple approaches, *tprK* sequences from intra-household transmission events were more similar than unrelated samples with identical core genomes.

Interpretation: Tp circulating in South America is closely related to dominant global sublineages. Increased *tprK* diversity in the SS14 lineage may influence *Tp's* ability to escape host adaptive immunity. We confirmed that *tprK* profiling is a promising tool to elucidate syphilis transmission networks. This study underscores the utility of genomics to yield insights into *Tp* pathogenesis.

Funding: Gates Foundation INV-036560 (to A.S.) and NIH/NIAID R01AI139265 (to J.D.K).

Research in context

Evidence before this study: Although whole genome sequencing (WGS) and genomic epidemiology have contributed to an understanding of the global diversity of *Treponema pallidum* (*Tp*), very few strains from South America have been sequenced to date. On January

6th, 2026, we performed a PubMed search including terms “syphilis genomic epidemiology”, “South America”, “*Treponema pallidum*”, and “Argentina”. We excluded studies of ancient *Tp* samples found in South American archeological sites. A single *Tp* sample from Argentina was originally sequenced in 2016 and included in subsequent analyses of global diversity. Nine Peruvian samples were previously sequenced in a study of global diversity on six continents. Thirty-three samples from Cali, Colombia were predominantly SS14-lineage and genotypic macrolide resistance was found in half of strains. Additionally, studies of *Tp* in Buenos Aires using a multi-locus sequence typing approach have shown circulation of strains belonging to both Nichols and SS14 lineages, with an increasing rate of macrolide resistance over time. Aside from a report of inter-strain recombination of *Tp* in an individual infected with both Nichols- and SS14-lineage strains, no studies have found intra-host evolution of *Tp* outside of the *T. pallidum* repeat (tpr) family or acidic repeat proteins. Five previous studies have examined *tprK* in clinical specimens and consistently shown increased *tprK* diversity in specimens associated with secondary syphilis compared to primary syphilis lesions. One prior study has shown that loss of *tprK* donor cassettes is associated with reduced *tprK* diversity, consistent with results from in vitro experiments. A single study has shown that *tprK* sequence content is more similar in samples with a suspected epidemiologic link, though no prior studies have looked at *tprK* diversity in the context of known syphilis transmission events.

Added value of this study: This study is the first to use WGS to comprehensively examine *Tp* transmission in a large South American city, yielding 96 samples from 70 individuals. We also add the first contemporary *Tp* genomes from Brazil. In contrast to findings from England, Australia, and other high-income countries, no *Tp* sublineages are associated with particular demographic groups or sexual networks in Buenos Aires. Two of seventy patients were co-infected with both Nichols- and SS14-lineage strains, showing the previously unappreciated frequency of conditions that permit inter-strain recombination-driven diversification of *Tp*. This study also reveals novel aspects of *Tp* pathogenesis, including recovery of specific intra-host *tp0006* point mutations from different specimen types and lineage-specific differences in *tprK* diversity. We also develop methods for the analysis of *tprK* relatedness between samples and demonstrate that *tprK* sequences are more similar in samples from individuals within intra-household syphilis transmission chains compared to those from epidemiologically unrelated individuals.

Implications of all the available evidence: *Tp* strains circulating in Buenos Aires,are genetically similar to those circulating worldwide and in Brazil and Peru but are noteworthy for the low (but rising) rate of macrolide resistance. Lineage-specific patterns of *tprK* antigenic variation could

result in differences between Nichols- and SS14-lineage strains' interactions with the host immune system. Finally, we show that *tprK* profiling holds promise to identify samples from within a syphilis transmission chain and could play an important role in public health.

Introduction

In 2022, the Region of the Americas accounted for 42% (3.37 million) of all new syphilis diagnoses worldwide, representing the highest incidence of any World Health Organization region. Low- and middle-income countries in Central and South America account for most cases. The rate of syphilis in pregnant women in the Region increased 28% between 2020 and 2022, with 68,000 cases of congenital syphilis annually[1] and in Argentina in 2023, 6% of women attending antenatal care tested positive for syphilis[2]. Until a vaccine that can prevent vertical transmission of *Treponema pallidum* subspecies *pallidum (Tp)* is available, reducing rates of congenital syphilis depends on identifying and breaking transmission chains, a process complicated by overlapping sexual networks and high rates of syphilis in vulnerable and hard-to-reach populations[3,4]. Despite the key role genomic epidemiology can play in understanding the spread and evolution of *Tp*, South American strains have been the subject of limited study.

Genomics-based studies of clinical samples can also shed light on the basic biology of *Tp*, which is difficult to study in the laboratory due to technically challenging cultivation and limited genetic tools[5,6]. *Tp* is notable for its minimalist 1.14 Mb genome and slow accumulation of genetic diversity, with fewer than 150 single nucleotide variants (SNVs) in the core genome separating strains of the Nichols and SS14 lineages of *Tp*[7,8], which are clinically indistinguishable[9]. In contrast, ongoing gene conversion in *tprK* results in antigenic variation of seven extracellular loops of the *Tp* outer membrane protein TprK. Donor cassettes (DCs) from 53 chromosomal sites are recombined into the seven variable (V) regions to generate novel *tprK* alleles, tens or hundreds of which may be found in primary or secondary syphilis lesions[10,11]. Antigenic variation in *tprK* is ongoing during in vitro culture[12] and can theoretically generate 10^{13}-10^{15} unique mosaic alleles[10] that can evade host adaptive immune responses[13,14]. Understanding *tprK* evolution can provide insight into *Tp*-immune dynamics that cause the relapsing clinical course of syphilis[14]. Additionally, *tprK*'s rapid evolution may be used to identify syphilis transmission networks[15] where whole genome sequencing (WGS)-based methods lack resolution due to low diversity of the *Tp* core genome.

Here, we use WGS to genomically characterize contemporary *Tp* circulating in Argentina, Brazil, and Peru. We couple WGS analyses with investigations of *tprK* sequence diversity in the Buenos Aires cohort, which includes samples associated with previously-

146 identified cases of intra-household syphilis transmission[16], enabling us to extend our prior work
147 using *tprK* to elucidate syphilis transmission networks[15] as well as examine the effect of *Tp*
148 genetic background on antigenic variation.
149
150 **Methods**
151 **Ethics Approvals**
152 In Buenos Aires, enrollment of pediatric participants and their caregivers (DI-2018-245-
153 HGNRG) as well of adult participants (DI-2022-155-HGNRG) received approval from by the
154 Committee and Review Board of the Hospital de Niños Ricardo Gutierrez. Sequencing of *Tp*
155 from deidentified specimens was approved by the University of Washington IRB
156 (STUDY00000885) and the University of North Carolina at Chapel Hill IRB (24-2696). Written
157 consent was obtained from all adult participants and caregivers of enrolled children.
158
159 **Enrollment Criteria and Screening**
160 A convenience sampling method was used to recruit patients in two public health services
161 in Buenos Aires, Argentina between October 2018 and January 2023. A total of 94 adult
162 outpatients were assisted in the Dermatology Unit of the Hospital Javier Muñiz, whereas another
163 cohort of 47 patients, including pediatric outpatients (n=28) and their caregivers (n=19) were
164 assisted at the Hospital de Niños Ricardo Gutierrez and studied as previously described[16]. Briefly,
165 serological, epidemiological, and clinical data, including prenatal screenings and treatment data,
166 were collected from the children and their cohabitants. A medical and psychosocial evaluation of
167 the children and their caregivers was performed to evaluate sexual abuse[16]. Pediatric cases
168 where sexual abuse was suspected were excluded. Additional screening criteria is found in
169 **Supplementary Methods**, as are enrollment criteria for Brazil and Peru cohorts, which were
170 included in WGS analyses.
171
172 ***T. pallidum* whole genome sequencing**
173 WGS was performed as previously described[17] on samples with more than 25 *tp0574*
174 copies/μL DNA eluate, equivalent to a minimum of approximately 200 genome copies into
175 library preparation. Briefly, genomic DNA was fragmented using the KAPA HyperPlus kit
176 (Roche) and TruSeq unique dual indices (Illumina) ligated prior to 14-16 cycles of pre-capture
177 PCR. Hybrid capture using our custom DNA probes was performed using the manufacturer's
178 instructions (Integrated DNA Technologies) prior to 14-16 cycles of post-capture amplification.

179 Paired-end 150nt reads were generated on Illumina instruments including the HiSeqX,
180 Novaseq6000, and NextSeq2000.
181
182 ***T. pallidum* genome assembly and phylogenetics**
183 A total of 96 samples from 70 participants from Argentina, 8 samples from 8 participants
184 from Brazil and 3 samples from 3 participants from Peru were newly sequenced for this study
185 and included in a phylogenetic analysis alongside 1,772 publicly available genome sequences
186 from the NCBI Sequence Read Archive (**Supplementary Data**). Genomes were assembled
187 using a custom pipeline available at https://github.com/greninger-lab/tpallidum-variant-calling,
188 outlined in **Supplementary Methods**.
189
190 ***tprK* sequencing and data pre-processing**
191 The *tprK* gene was amplified from *T. pallidum* genomic DNA using 35 cycles of PCR with
192 barcoded primers as previously described[17] in technical duplicate and sequencing performed on
193 the PacBio Revio. Following demultiplexing, HiFi reads were filtered to remove sequences
194 shorter than 1400 bp or longer than 1700 bp or with unexpected primer orientations. Only reads
195 with each base Q30 or higher were included in subsequent analyses. Variable regions and
196 intervening constant regions were translated and reads containing frameshifts or premature stop
197 codons in any region excluded from analysis. Custom analyses are described in
198 **Supplementary Methods**.
199
200 **Statistics and Reproducibility**
201 All calculations and statistics were performed in R v4.2.2. Custom processing scripts are
202 available at https://github.com/greninger-lab/tprk-argentina/.
203
204 **Data Availability**
205 All WGS data generated during this study are available at NCBI BioProject
206 PRJNA723099, with accessions in **Supplementary Data**. *tprK* PacBio raw data are available in
207 NCBI BioProject PRJNA1415869.
208
209 **Results**
210 **Cohort demographics**
211 A total of 141 patients (113 adults and 28 children) were enrolled, including seven
212 individuals associated with three intra-household transmissions. Demographic characteristics of

213 the adult and pediatric cohorts are summarized in **Table 1**. All adults were cisgender, with median
214 age 26 years (IQR 22-33). All 45 men who have sex with men (MSM) were adult males, as were
215 all 27 persons living with HIV (PLWH), including 5 individuals newly diagnosed during enrollment.
216 Twenty-eight patients had prior syphilis. The pediatric cohort included four newborns with
217 congenital syphilis, 12 children (5 males and 7 females) with a median age of 2 (IQR: 1-6) years
218 and 12 adolescents (7 males and 5 females) with a median age of 16 (IQR: 15-17) years.
219 Adolescents with primary and secondary syphilis were sexually active. Children with secondary
220 syphilis have a median age of 3 years (IQRs: 2-6), caused by non-sexual close contact with active
221 lesions from their caregivers[16,18]. Clinical findings are included in **Supplementary Analysis**.
222
223 **Table 1 Demographic characteristics by cohort.** *The p value was derived from a chi-square
224 test comparing the proportion of patients with sequenced genomes to the total patient population.

	Adult patients (n=113) n (%)	**Pediatric patients (n=28) n (%)**	**Total Patients (n=141) n (%)**	**Patients with WGS (n=70) n (%)**	**p value ***
AGE: median (IQRs) years	26 (22-33)	6 (2-16)	24 (19-31)	27 (18-33.5)	
Adults (>18)	113 (100.0)	-	113 (80.1)	51 (72.8)	0.31
Pediatric (0-18)	-	28 (100.0)	28 (19.9)	19 (27.2)	
BIOLOGICAL SEX					
Male	78 (69.0)	13 (46.4)	91 (64.5)	48 (68.6)	0.67
Female	35 (31.0)	15 (53.6)	50 (35.5)	22 (31.4)	
SEXUAL ORIENTATION (among adult and adolescent males)					
MSM	45 (57.7)	0 (0)	45 (52.9)	19 (43.2)	0.39
No MSM	33 (42.3)	7 (100%)	40 (47.1)	25 (56.8)	
HIV					
Positive	26 (23.0)	0 (0)	26 (18.4)	9 (13.8)	0.41
Negative	87 (77.0)	28 (100%)	115 (81.6)	61 (86.2)	
PRIOR SYPHILIS					
Yes	27 (23.9)	1 (3.6)	28 (19.8)	10 (14.2)	0.28
No	85 (75.2)	27 (96.4)	112 (79.4)	70 (85.8)	

Not specified	1 (0.9)	0 (0)	1 (0.7)	0 (0)	
CLINICAL STAGE					
Primary	21 (18.6)	3 (10.7)	24 (17.0)	16 (22.8)	0.24
Secondary	92 (81.4	21 (75.0)	113 (80.1)	54 (77.2)	
Congenital syphilis	0 (0)	4 (14.3)	4 (2.8)	0 (0)	

Yield of molecular testing and genome sequencing

A total of 250 samples were collected from the 141 enrolled patients, primarily oral and anogenital swabs (135/250, 54.0% and 92/250, 36.8% of samples, respectively). Molecular positivity for the single copy *tp0574 Tp* locus was highest among anogenital swabs (81/92, 88.0%), extragenital rash (9/13, 69.2%), and oral swabs (81/135, 60%). All 22 primary chancre swabs were positive. Two of thirteen (15.4%) blind swabs of normal mucosa in primary syphilis were *tp0547* positive. Among secondary syphilis, swabs of oral lesions were positive in 37 of 42 cases (88.1%), while blind swabs of normal oral mucosa were positive in 36 of 66 cases (54.5%) (**Figure S1A**). Primary lesion swabs had the highest interpolated *tp0574* copies per positive swab (median: 48,121 copies, n=16), followed by secondary disseminated anogenital lesions (median: 17,025 copies n=39) (**Figure S1B**). Specimens from patients that had had one or more prior episodes of syphilis (n=44 specimens from 28 patients) were less frequently *tp0574* positive than those who had no prior syphilis episodes (23/44, 52.3% vs 153/204, 75.0%), p=0.0047, chi-square test).

We attempted WGS on samples for which we could input at least 200 genome copies to library preparation. A total of 96 samples from 70 patients had sufficient coverage for analysis, which we defined as having less than 25% ambiguous bases in SNV positions. Among all sample types, the highest yield of near-complete genomes was from primary anogenital lesions (16/22, 72.7%; **Figure S1C**), followed by oral and anogenital lesions of the mucosa. No near-complete genomes were recovered from DNA extracted from whole blood. The proportion of genomes recovered among demographic groups was representative of the enrolled cohort (**Table 1**, all *p* >0.05, chi-square test), although no genomes were recovered from infants with congenital syphilis.

Strains from Nichols and SS14 lineages of *T. pallidum* circulate in Buenos Aires

Prior studies have shown that both SS14- and Nichols-lineage strains of multiple sequence types circulate in Buenos Aires[16,19], however, South America is the least sampled of all continents other than Antarctica, and only a single South American sample was included in a

254 recent comprehensive subclassification scheme[8]. Both lineages of *Tp*, which are named after the
255 Nichols and SS14 strains, were detected in contemporary Colombian samples[9,20]. We built upon
256 these studies by analyzing the genomes recovered from Argentine samples alongside publicly
257 available South American *Tp* sequence data as well as three new samples from Peru and eight
258 from Brazil. First, we contextualized South American strains within a worldwide phylogeny
259 including a total of 1,879 *Tp* samples (**Figure 1A**). Most South American samples are closely
260 related (within 10 SNVs) to previously classified samples in globally distributed sublineages 1
261 (SS14 lineage) and 8 (Nichols lineage)[8] (**Figure 1B**). In Argentina, 32/71 (45.1%) individuals had
262 Nichols-lineage treponemes, similar to Colombia (8/23, 34.5%). Nichols-lineage *Tp* has not been
263 identified in Peru. Although genotypic macrolide resistance, conferred by A2058G mutation in the
264 23S rRNA gene, was relatively low in Argentina compared to other South American countries
265 (39.1%; **Figure 1C**), these data reflect a progressive increase in resistance within Buenos Aires
266 over the past decade [16,19,21].
267
268 **No association between sublineage and demographic characteristics in Buenos Aires**
269 Although co-infection with both Nichols- and SS14-lineage strains has only been reported
270 once before[22], two individuals in the Buenos Aires cohort had evidence for co-infection. A sample
271 from a residual genital chancre in an individual with secondary syphilis who reported multiple sex
272 partners in the preceding three months yielded mixed SS14- and Nichols-like sequence reads at
273 lineage defining positions[8] (**Figure 2A**). In another person, also reporting multiple sex partners,
274 SS14- and Nichols-lineage strains were recovered from different lesions (**Figure 2B**). Co-
275 infections were confirmed by strain-specific digital PCR[22] using DNA re-extracted from the original
276 swab. In one of 20 individuals with WGS from multiple secondary syphilis lesions, we found
277 compartmental SNVs in *tp0006*, a gene of unknown function (**Figure S2, Supplementary**
278 **Analysis**).
279 We constructed a phylogeny of *Tp* genomes from Buenos Aires, including strains from our
280 current study and one sample collected in 2016[15] (n=73 genomes from 71 individuals, **Figure**
281 **2C**). Thirty-one samples formed nine distinct clusters with 0 core-genome SNVs, including
282 clusters of individuals within households as well as clusters with no known epidemiologic link.
283 Among individuals without Nichols-SS14 co-infection and with clinical information (n=68), we did
284 not find any association between patient sex, sexual orientation, or HIV status with *Tp l*ineage.
285 Macrolide resistance was more prevalent in the SS14 lineage than the Nichols (22/38, 57.9% vs
286 5/30, 16.7%, $p<0.01$), and in MSM (11/17, 64.7%) compared to MSW or females (10/29, 34.5%

287 and 6/22, 27.3%, respectively, p<0.05). Mutations to 16S rRNA sites homologous to those
288 conferring tetracycline resistance were not observed[23,24].
289
290 **The number and distribution of *tprK* alleles depends on stage, site, and *T. pallidum* lineage**
291 We next interrogated the hypervariable *tprK* gene using high fidelity PacBio sequencing
292 coupled with our customized analyses (**Figure S3**). TprK undergoes antigenic variation within
293 seven of its extracellular loops to evade host adaptive immunity[13,14,25–27] by replacing fragments
294 of *tprK* with donor cassettes (DCs) from 53 chromosomal regions arranged in two loci (**Figure**
295 **S4**). Variable regions were defined essentially as described previously[10,25], however, for ease of
296 interpretation we split variable and constant regions between codons. Sequences between
297 variable loops exhibited minimal variation, with no changes to a known B cell epitope[28] (**Figure**
298 **S5, Supplementary Analysis**).
299 We examined the relationship between clinical and demographic factors in the number of
300 unique *tprK* alleles, V regions, and Shannon diversity index in each sample, including a total of
301 56 samples from 44 individuals with sufficient read depth and quality (**Figure S3**). V regions were
302 only counted if present in both technical replicates. We did not limit inclusion of full-length *tprK*
303 alleles to only those present in both technical replicates because it had a different effect on
304 primary and secondary samples (**Figure S6A-B**), consistent with differences in the underlying
305 distribution of alleles between primary and secondary syphilis[11]. No full-length *tprK* alleles
306 recovered were shared between patients. Among ten individuals with *tprK* sequencing from more
307 than one body site, six paired samples shared one or two full-length *tprK* alleles. The number of
308 *tprK* alleles exceeding 0.2% abundance ranged between 7-141 per sample. Consistent with prior
309 reports[10], secondary syphilis samples of all types yielded a significantly higher number of unique
310 full-length *tprK* alleles (median 63 vs 37, p <0.001, **Figure 3A**) as well as more V regions and
311 higher Shannon diversity index. When lesion type and location was considered, oral secondary
312 lesions had significantly higher *tprK* diversity than anogenital primary lesions, though all
313 secondary lesion types except rash had a higher median number of alleles and greater diversity
314 than primary lesions (**Figure 3B**). Among the five individuals from whom paired oral and
315 anogenital lesions of any type were collected, oral lesions consistently had higher *tprK* diversity
316 than anogenital lesions (**Figure 3C**). Among 55 samples from 43 patients with paired WGS and
317 full-length *tprK* profiling, SS14-lineage samples had more unique *tprK* alleles than those from the
318 Nichols lineage (**Figure 3D**), driven primarily by increased diversity in V6 and V7 (**Figure 3E,**
319 **Figure S7**). Among samples with paired WGS, a total of 3,142 unique *tprK* alleles were found,
320 including 1,191 in Nichols-lineage samples (n=25) and 1,951 in SS14-lineage samples (n=30).

321 Patient sex, sexual orientation, prior syphilis, HIV coinfection, and age were not associated with
322 differences in *tprK* diversity, though prior syphilis and PLHV were very small groups (**Figure S8**).
323
324 **SS14 lineage strains generate *tprK* diversity with an expanded repertoire of *tprK* donor**
325 **cassettes**
326 To further explore the observation of increased *tprK* diversity in the SS14 lineage, we
327 applied a k-mer-based analytical approach designed to account for its extensive rearrangement.
328 *tprK* evolves via recombinogenic changes that replace segments of tens of nucleotides in each
329 variable region[25] (**Figure S3B**), precluding the use of maximum-likelihood methods to determine
330 sequence relatedness. Therefore, we examined differences in *tprK* sequence composition
331 between lineages by first breaking each sequence into short overlapping subsequences (k-
332 mers) of lengths ranging between 9-27nt (**Figure S9**), allowing comparison of sequence
333 similarity despite extensive recombination and rearrangement. We focused particularly on k-
334 mers of length 21 nt since the seven-amino acid fragments (7AA-mers) they encode
335 approximate the minimal length of a linear B cell epitope[29], yet are short enough to enable
336 identification of *tprK* sequence fragments that that are exact matches to chromosomal donor
337 cassettes[12,25]. Although some could theoretically have been generated via gene conversion to
338 stitch donor sequences together, we will define *tprK* 21-mers that perfectly match one or more
339 donor cassettes as "donor cassette (DC)-derived". K-mers and AA-mers found only in *tprK*
340 sequences from strains of the Nichols or SS14 lineage are defined as "lineage-specific".
341 To ensure balanced representation of 21-mer diversity, SS14 samples (n=30) were first
342 down-sampled to the same number (25) as Nichols, then 1,000 *tprK* alleles from each lineage
343 randomly selected from the 25 samples. SS14 lineage samples had more V6 and V7 21-mers
344 and 7AA-mers than Nichols (**Figure 4A-B**), with a considerably greater proportion of unique
345 SS14 7AA sequences derived from V7 than in Nichols (**Figure 4C**). While SS14 and Nichols
346 strains had similar proportions of 7AA-mers encoded by 21-mers that were exact matches to
347 those encoded by donor cassettes (DCs), SS14 strains had significantly more lineage-specific
348 DC-derived 7AA fragments than Nichols (44 vs 9, $p<0.01$), half of which were found in V7
349 (**Figure 4D**).
350 We subsequently determined whether the increased diversity in the SS14 lineage strains
351 could be accounted for by differences in the loci that contribute to recombination. We previously
352 showed how differences in donor repertoire can affect *tprK* diversity both by genetically
353 knocking out the major DC locus[27] and detecting a clinical strain missing nine sets of DCs that
354 had a reduced number of *tprK* variable regions relative to wild type strains[15]. Among strains in

355 this analysis, the predominant DC alleles were lineage-specific, with unique mutations between
356 SS14 and Nichols lineages. Differences between Nichols and SS14 loci were confined to SNVs
357 in donor cassettes specific for V3, V6, and V7, as well as SNVs between DCs (**Figure S10**).
358 When the DC-derived, lineage-specific 21-mers were mapped to their respective donor cassette
359 locus, a majority of strains harbored at least one *tprK* allele with a lineage-specific 7AA-mer
360 derived from the V3/V6 overlapping donor cassettes 26 and 25 (**Figure 4E**). The largest
361 difference between Nichols and SS14 donor cassette alleles is a 51bp indel in V7-DC31 (**Figure**
362 **S10**), which directly contributes novel sequences to *tprK* alleles in SS14-lineage samples. When
363 assessing both the type (DC-derived or not) and lineage specificity of unique *tprK* k-mers, we
364 found that SS14 has significantly greater use of lineage-specific V7 7AA fragments than Nichols
365 (**Figure 4G**). Interestingly, the lineage-specific DC sequences found in Nichols V7-DC31 and
366 V7-DC2 are not found in Nichols *tprK* alleles, while they are readily found in SS14 (**Figure 4E-**
367 **F**), thus accounting for the reduced diversity of the V7 *tprK* allele seen in Nichols strains.
368 Together, these data suggest that the increased *tprK* sequence diversity observed in SS14-
369 lineage strains is due in part to an enhanced repertoire of V7 donor cassettes.
370

371 ***tprK* is more similar between individuals within known transmission chains than between**
372 **unrelated individuals**

373 Finally, we extended our previous findings showing that *tprK* sequences shared by two
374 individuals are more similar in patients with a suspected epidemiologic link[15], enabling outbreak
375 tracing. First, we recapitulated our prior results showing that the Pearson correlation between
376 the proportion of V regions between samples is greater within known intra-household
377 transmission events than in epidemiologically unrelated samples with identical core genomes
378 (**Figure 5A,** $p<0.05$, t-test)[15]. Next, we increased the specificity of the comparison by
379 determining the inter-sample Bray-Curtis dissimilarity for each V region and length of k
380 (**Supplementary Methods**). Each variable region had a differential ability to discriminate
381 between samples from the same patient, same household, or unrelated individuals (**Figure**
382 **S11**). V1 and V4 have lower Bray-Curtis dissimilarity between samples from the same
383 household than unrelated individuals, while V6 only has low k-mer dissimilarity in samples from
384 the same individual. These results are consistent with prior observations that V1 and V4 evolve
385 the most slowly and V6 evolves the most rapidly, resulting in the highest diversity among V
386 regions[12]. We calculated a composite Bray-Curtis dissimilarity score that included k-mers from
387 all seven V regions, summed across all values of *k* to account for both the high sensitivity of
388 short kmers and high specificity of longer kmers (**Figure 5B**). We also calculated the median

389 inter-allele global alignment distance between *tprK* alleles within and between each sample
390 (**Figure 5C**) as well as the median inter-allele composite Bray-Curtis score (**Figure 5D**). All
391 methods showed a significant difference in *tprK* similarity between known transmission chains
392 (intra-household) versus epidemiologically unrelated samples. Analysis of full-length alleles
393 (**Figure 5C-D**) did not further resolve distinctions between known transmission events and
394 unrelated samples relative to methods relying only on total V region content (**Figure 5A-B**).
395 Interestingly, among samples separated by 0 SNVs, all methods identified unexpectedly
396 low Bray-Curtis dissimilarity between the oral and genital samples from one patient and the
397 genital sample from a second patient (**Figure 5,** red boxes). Notably, outside of the V regions,
398 strains from both patients also had a W516C mutation in the C-terminus of TprK (**Figure S5B**).
399 Upon further review of patient clinical data, we found that the samples came from a male with
400 secondary syphilis and a female with primary syphilis. Both patients were in their mid-20s,
401 heterosexual, denied stable sexual partners, and visited the clinic on the same day. Together,
402 these findings support the utility of *tprK* sequencing in identifying syphilis transmission chains for
403 epidemiological investigation.
404
405 **Discussion**
406 As the syphilis epidemic in South America continues to accelerate, sequencing of *Tp*
407 provides considerable insight into syphilis epidemiology, revealing that most strains isolated in
408 South America are closely related to those circulating throughout the world[8]. We also extend our
409 previous methods using *tprK* similarity to elucidate transmission events, which can augment
410 public health responses by detecting network links not captured by traditional contact tracing
411 approaches[3]. Using multiple independent methods to profile *tprK* relatedness in a unique cohort
412 of samples that included household linkage, we show *tprK* sequences from individuals within a
413 transmission chain are more closely related than those from unrelated samples with identical
414 core genomes. Moreover, the method revealed a possible transmission event that had not been
415 recorded clinically, showing the power of this technique to identify closely related cases that
416 merit follow-up public health investigation. Analysis of full-length *tprK* alleles did not
417 demonstrate a significantly enhanced ability to resolve inter-sample distances relative to
418 analysis of V regions alone, suggesting that *tprK* profiling could be deployed using short-read
419 sequencing technologies that have been widely adopted in public health labs. The sensitivity of
420 *tprK* profiling to unmask possible transmission links has important implications for patient
421 privacy, a consideration for future study design and use by public health departments.

422 An important limitation of the current study of *tprK* is the use of convenience sampling
423 from a single location, Buenos Aires, resulting in a relatively small number of participants for a
424 very diverse metric like *tprK*. In particular, the number of known transmission events (three) was
425 very low. Furthermore, the household transmission samples – which are suspected to be
426 acquired syphilis by nonsexual contact[16,18] - may not be fully representative of *tprK* evolution
427 during sexual transmission. Nonetheless, the work described here replicates and significantly
428 extends prior *tprK* profiling in Seattle, Washington[15].
429 In addition to the key public health questions that can be answered using *Tp* WGS and
430 targeted *tprK* sequencing, genomic studies of *Tp* evolution in patient samples play an important
431 role in revealing aspects of pathogenesis and evolution that are challenging to study in vitro
432 given limited experimental tractability and technically challenging cultivation. Unexpectedly, we
433 showed that *tprK* recovered from oral lesions is more diverse than from anogenital lesions, even
434 from patient-matched specimens, which may reflect different immune surveillance or other
435 selection pressures at different mucosal sites. Future investigations of subtle differences in *tprK*
436 diversity will require incorporating unique molecular identifiers or other techniques to better
437 resolve rare alleles, approaches we are currently developing. Additionally, we demonstrate for
438 the first time increased *tprK* diversity in the SS14 lineage relative to the Nichols lineage and
439 attribute it to an increased repertoire of unique donor sequences resulting from a 51 bp insertion
440 in a V7-specific donor cassette. As this increased diversity in the SS14 lineage results in
441 multiple novel potentially antigenic variants not seen in the Nichols lineage, enhanced *tprK*
442 diversity may provide a fitness advantage to SS14 strains by enabling more variants to escape
443 adaptive immune responses to *tprK*, potentially accounting for SS14's enhanced prevalence
444 globally.
445 In conclusion, we find WGS coupled with *tprK* profiling to provide extensive insight into
446 *Tp* pathogenesis. Furthermore, the use of inter-sample Bray's dissimilarity scores of *tprK* V
447 regions is a promising tool to infer syphilis transmission networks. Future work involving a wider
448 range of sampling, including more known transmission events, will be necessary to confirm
449 these findings and further refine methods to infer transmission events from *tprK* profiling.
450 Critically, studies that connect *tprK* variation with host immune responses will be essential to
451 define how antigenic variation shapes immune evasion by SS14 and Nichols lineage *Tp*.
452
453 **Acknowledgements**
454 The authors are grateful for the work of the UW Long Read Sequencing Center.
455

Competing interests

JBP reports past research support from Gilead Sciences and consulting for Zymeron Corporation, and non-financial support from Abbott Laboratories, all outside the scope of this manuscript. ALG reports contract testing from Abbott, Novavax, Pfizer, Janssen, and Hologic, as well as research support from Gilead, all outside the scope of this manuscript.

REFERENCES

1 Syphilis cases increase in the Americas - PAHO/WHO | Pan American Health Organization. 2024; published online May 22. https://www.paho.org/en/news/22-5-2024-syphilis-cases-increase-americas (accessed Dec 13, 2025).

2 Data on syphilis. https://www.who.int/data/gho/data/themes/topics/data-on-syphilis (accessed Dec 13, 2025).

3 Jennings JM, Tilchin C, Meza B, *et al.* Overlapping Transmission Networks of Early Syphilis and/or Newly HIV Diagnosed Gay, Bisexual and Other Men Who Have Sex with Men (MSM): Opportunities for Optimizing Public Health Interventions. *AIDS Behav* 2020; **24**: 2895–905.

4 CDC. National Overview of STIs in 2023. STI Statistics. 2024; published online Nov 7.

5 Edmondson DG, Hu B, Norris SJ. Long-Term In Vitro Culture of the Syphilis Spirochete Treponema pallidum subsp. pallidum. *mBio* 2018; **9**. DOI:10.1128/mBio.01153-18.

6 Romeis E, Tantalo L, Lieberman N, Phung Q, Greninger A, Giacani L. Genetic Engineering of *Treponema pallidum* subsp. *pallidum* , the Syphilis Spirochete. Microbiology, 2021 DOI:10.1101/2021.05.07.443079.

7 Fraser CM, Norris SJ, Weinstock GM, *et al.* Complete genome sequence of Treponema pallidum, the syphilis spirochete. *Science* 1998; **281**: 375–88.

8 Beale MA, Marks M, Cole MJ, *et al.* Global phylogeny of Treponema pallidum lineages reveals recent expansion and spread of contemporary syphilis. *Nat Microbiol* 2021; **6**: 1549–60.

9 Seña AC, Matoga MM, Yang L, *et al.* Clinical and genomic diversity of *Treponema pallidum subsp. pallidum:* A global, multi-center study of early syphilis to inform vaccine research. Infectious Diseases (except HIV/AIDS), 2023 DOI:10.1101/2023.07.19.23291250.

10 Addetia A, Lin MJ, Phung Q, *et al.* Estimation of Full-Length TprK Diversity in Treponema pallidum subsp. *pallidum*. *mBio* 2020; **11**. DOI:10.1128/mBio.02726-20.

11 Liu D, Liu L-L, Zheng X-Q, *et al.* Genetic Profiling of the Full-Length *tprK* Gene in Patients with Primary and Secondary Syphilis. *Microbiol Spectr* 2023; **11**: e04931-22.

12 Lin MJ, Haynes AM, Addetia A, *et al.* Longitudinal TprK profiling of in vivo and in vitro-propagated Treponema pallidum subsp. pallidum reveals accumulation of antigenic variants in absence of immune pressure. *PLoS Negl Trop Dis* 2021; **15**: e0009753.

13 Giacani L, Molini BJ, Kim EY, *et al.* Antigenic variation in Treponema pallidum: TprK sequence diversity accumulates in response to immune pressure during experimental syphilis. *J Immunol* 2010; **184**: 3822–9.

14 Reid TB, Molini BJ, Fernandez MC, Lukehart SA. Antigenic variation of TprK facilitates development of secondary syphilis. *Infect Immun* 2014; **82**: 4959–67.

15 Lieberman NAP, Avendaño CC, Bakhash SAKM, *et al.* Genomic Epidemiology of *Treponema pallidum* and Circulation of Strains With Diminished *tprK* Antigen Variation Capability in Seattle, 2021–2022. *The Journal of Infectious Diseases* 2023; : jiad368.

16 Garcia LN, Morando N, Otero AV, *et al.* Multilocus sequence typing of Treponema pallidum pallidum in children with acquired syphilis by nonsexual contact. *Future Microbiol* 2022; **17**: 1295–305.

17 Lieberman NAP, Lin MJ, Xie H, *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.

18 Moscatelli G, Moroni S, García Bournissen F, *et al.* Acquired Syphilis by Nonsexual Contact in Childhood. *Pediatric Infectious Disease Journal* 2021; **40**: 892–8.

19 Morando N, Vrbová E, Melgar A, Rabinovich RD, Šmajs D, Pando MA. High frequency of Nichols-like strains and increased levels of macrolide resistance in Treponema pallidum in clinical samples from Buenos Aires, Argentina. *Sci Rep* 2022; **12**: 16339.

20 Salazar JC, Vargas-Cely F, García-Luna JA, *et al.* Treponema pallidum genetic diversity and its implications for targeted vaccine development: A cross-sectional study of early syphilis cases in Southwestern Colombia. *PLoS One* 2024; **19**: e0307600.

21 Gallo Vaulet L, Grillová L, Mikalová L, *et al.* Molecular typing of Treponema pallidum isolates from Buenos Aires, Argentina: Frequent Nichols-like isolates and low levels of macrolide resistance. *PLoS One* 2017; **12**: e0172905.

22 Lieberman NAP, Mohamed Bakhash SAK, Nunley BE, *et al.* Disseminated Syphilis Caused by Two Recombining *Treponema pallidum* Strains. *N Engl J Med* 2025; **392**: 1551–3.

23 Pringle M, Fellström C, Johansson K-E. Decreased susceptibility to doxycycline associated with a 16S rRNA gene mutation in Brachyspira hyodysenteriae. *Vet Microbiol* 2007; **123**: 245–8.

24 Trieber CA, Taylor DE. Mutations in the 16S rRNA genes of Helicobacter pylori mediate resistance to tetracycline. *J Bacteriol* 2002; **184**: 2131–40.

25 Centurion-Lara A, LaFond RE, Hevner K, *et al.* Gene conversion: a mechanism for generation of heterogeneity in the tprK gene of Treponema pallidum during infection. *Mol Microbiol* 2004; **52**: 1579–96.

26 LaFond RE, Molini BJ, Van Voorhis WC, Lukehart SA. Antigenic variation of TprK V regions abrogates specific antibody binding in syphilis. *Infect Immun* 2006; **74**: 6244–51.

529 27 Romeis E, Lieberman NAP, Molini B, *et al.* Treponema pallidum subsp. pallidum with an
530 Artificially impaired TprK antigenic variation system is attenuated in the Rabbit model of
531 syphilis. *PLoS Pathog* 2023; **19**: e1011259.

532 28 Giacani L, Romeis E, Haynes A, *et al.* Immunization with full-length TprC variants induces a
533 broad response to surface-exposed epitopes of the Treponema pallidum repeat protein family
534 and is partially protective in the rabbit model of syphilis. *Vaccine* 2025; **61**: 127406.

535 29 Buus S, Rockberg J, Forsström B, Nilsson P, Uhlen M, Schafer-Nielsen C. High-resolution
536 mapping of linear antibody epitopes using ultrahigh-density peptide microarrays. *Mol Cell*
537 *Proteomics* 2012; **11**: 1790–800.

538

539

FIGURE 1

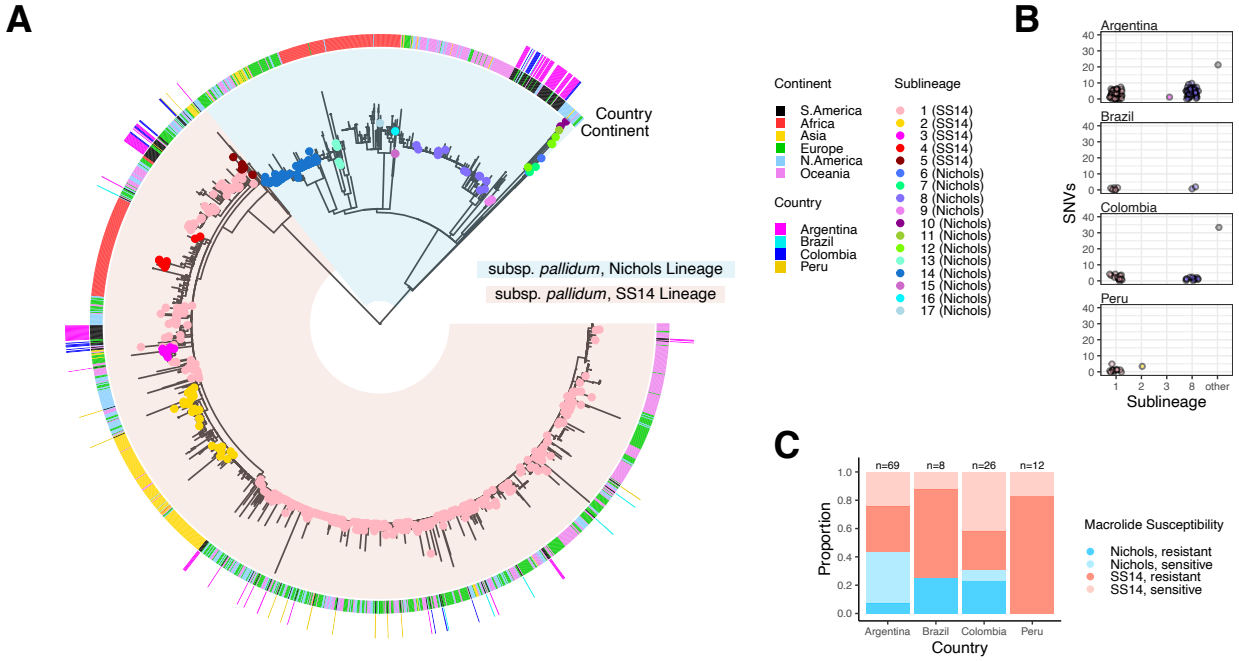
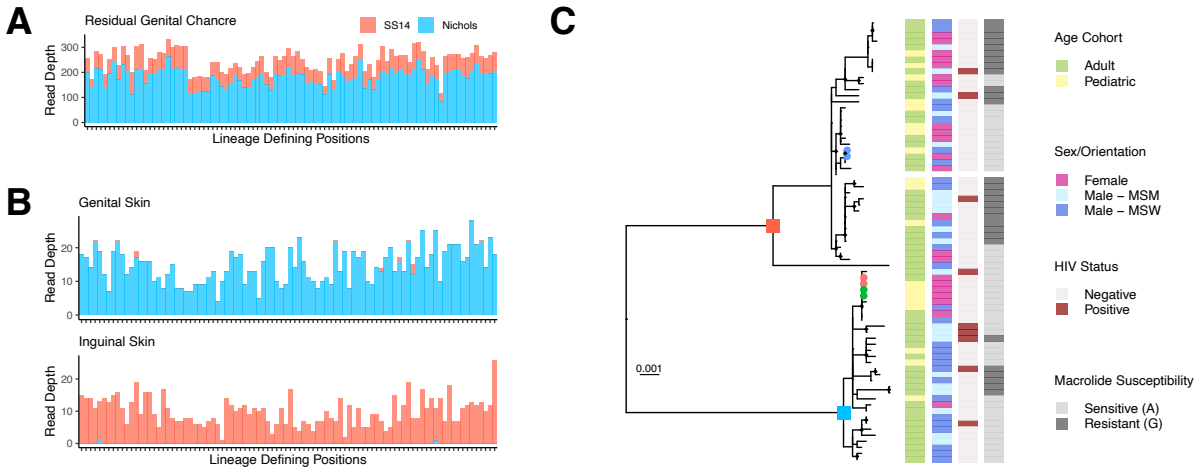


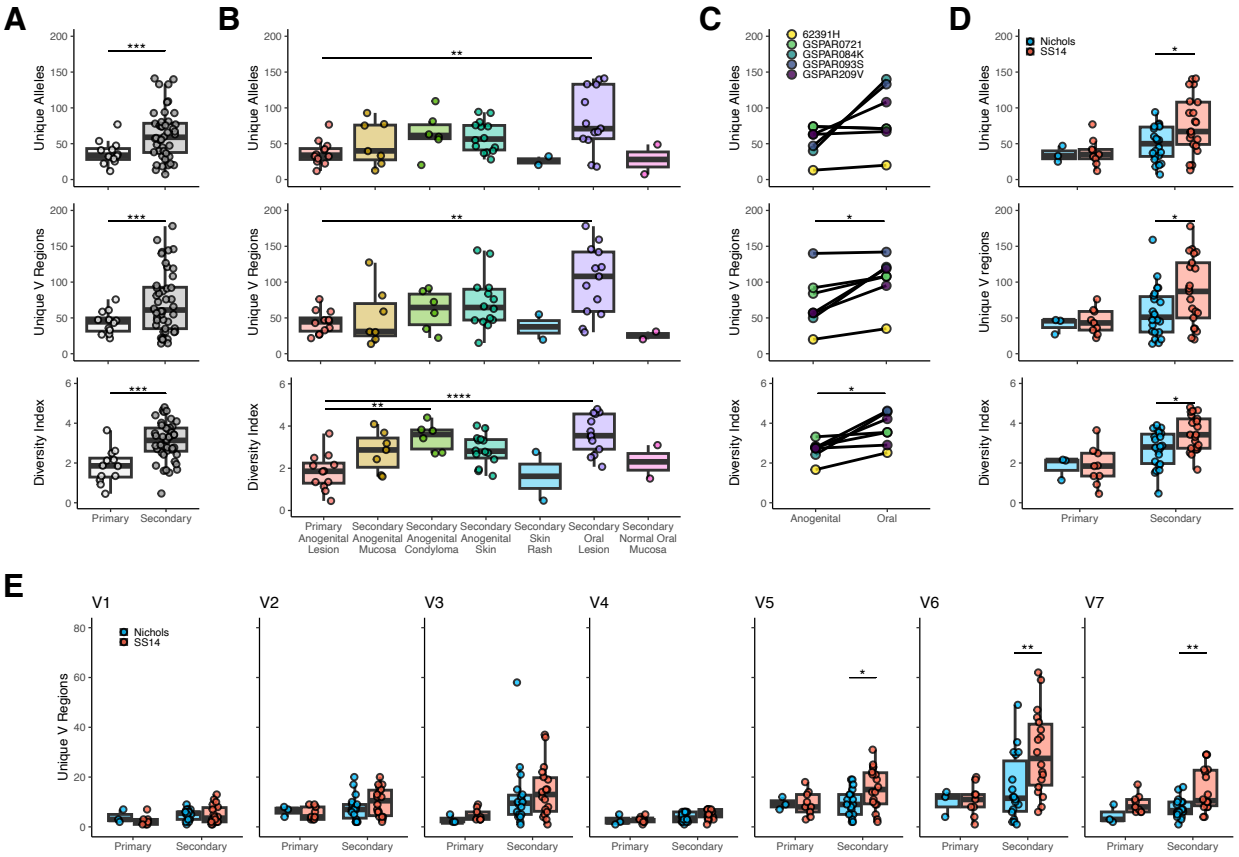
Figure 1: South American *T. pallidum* is closely related to globally dominant sublineages. A) Recombination-masked, maximum likelihood phylogeny of 1,879 *T. pallidum* strains, including 107 newly sequenced for this study. Blue background represents the Nichols lineage and red represents the SS14 lineage. Strains included in a study by Beale *et al* (2021) have tips colored by sublineage. The inner ring denotes continent of collection and outer ring denotes country for South American samples only. B) Number of single nucleotide differences between each South American sample and its most closely related worldwide sample, plotted by sublineage of the worldwide sample. C) Presence of A2058G mutation conferring macrolide susceptibility, by country and lineage.

555 **FIGURE 2**
556



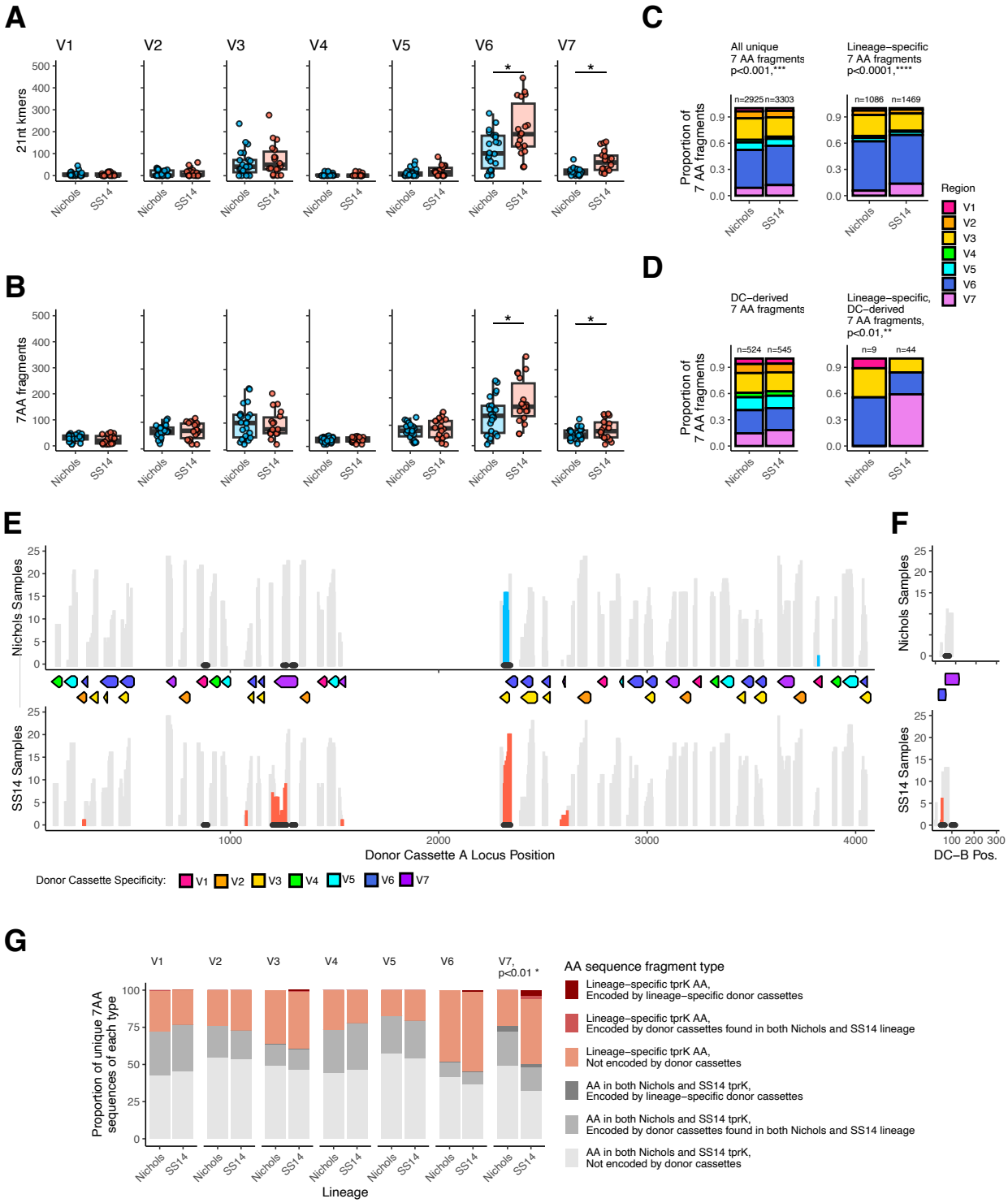
557
558
559 **Figure 2: Co-infection and *Treponema pallidum* genomic epidemiology within Buenos**
560 **Aires, Argentina, 2018-2023.** A) Whole genome sequencing read depth shows the presence of
561 ~80% Nichols- and ~20% SS14- variants at lineage-defining positions, supporting a co-infection
562 in the residual genital chancre of an individual (GSPAR041) with secondary syphilis. B)
563 Secondary syphilis lesions from genital and inguinal skin of another individual (GSPAR083)
564 belong to Nichols and SS14 clades, respectively. Individuals in both A and B had a history of
565 multiple sex partners. C) Phylogeny of strains collected in Buenos Aires, 2018-2023. Nichols
566 and SS14 lineage ancestral nodes are labeled with blue and red squares, respectively. Tips
567 from samples of each patient within the three household transmission clusters are colored red,
568 green, and blue, respectively. Tiles correspond to patient age, sex and sexual orientation, HIV
569 status, and macrolide susceptibility (A2058G).
570
571
572

573 **FIGURE 3**
574



575
576
577 **Figure 3: Number and distribution of *tprK* alleles.** A) Differences in total alleles (top), unique
578 variable regions (middle), and Shannon diversity index (bottom) by syphilis stage. T-test,
579 ***p<0.001. B) Differences in *tprK* by lesion type and location. ANOVA with Tukey's post hoc,
580 **p<0.01, ****p<0.0001. C) Samples from individuals with paired samples from anogenital and
581 oral sites show a higher number of V regions and diversity index in the oral lesion. Paired T-test,
582 *p<0.05. D) *tprK* diversity in Nichols vs SS14 lineage samples, t-test, *p<0.05. E) Variation by
583 clinical stage and lineage in the number of unique V regions. T-test, *p<0.05, **p<0.01.
584
585

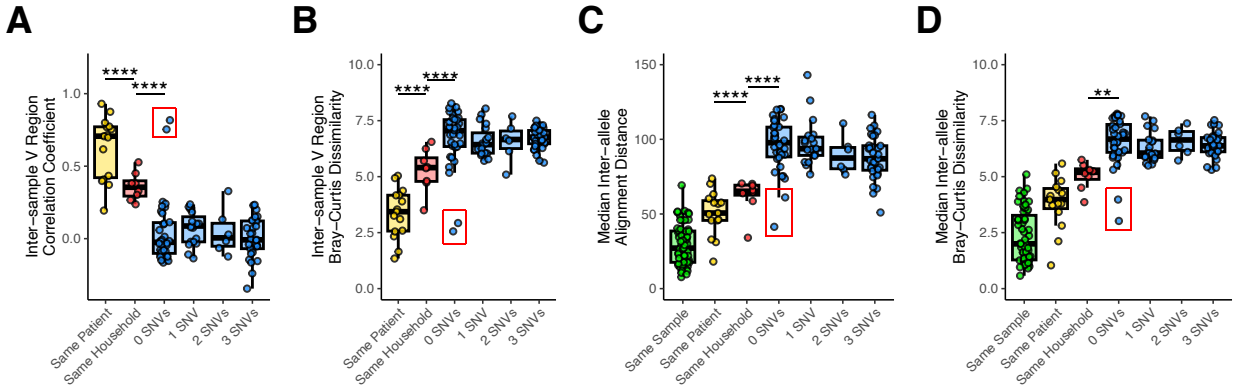
586 **FIGURE 4**
587



588
589 **Figure 4: Lineage-specific differences in *tprK* sequence content.** All analyses in this figure
590 were performed by sampling an equal number of samples from each lineage (25) followed by
591 random subsampling of 1,000 unique alleles per lineage. Significance is defined as a mean

592 p<0.05 across 1,000 independent subsampled replicates. A single representative replicate is
593 visualized. The total number of unique V6 and V7 21 nt k-mers (A) and 7AA-mers (B) are
594 significantly higher in SS14 than Nichols lineage. T-test, *p<0.05. C) The proportion of unique
595 7AA-mers derived from each region differs by lineage among all 7AA-mers as well as those found
596 in only the SS14 or Nichols lineage. ***p<0.001 and ****p<0.0001, chi squared. D) The proportion
597 of 7AA-mers derived from each region and encoded directly by donor cassettes is not different
598 between SS14 and Nichols lineages, but each lineage contains a different proportion of lineage-
599 specific 7AA-mer fragments directly encoded by donor cassettes, **p<0.01, chi squared. SS14
600 uses more lineage-specific donor cassette (DC)-derived sequences than Nichols (44 vs 9). E)
601 Histogram of the number of Nichols (top) or SS14 (bottom) samples containing at least one *tprK*
602 allele using DC-derived 7AA-mer from each donor site in DC-A and DC-B. Locations of donor
603 cassettes colored by V region specificity are shown. DC-derived AA-mers found in *tprK*
604 sequences from both Nichols and SS14 lineages are shown in grey. Lineage-specific 7AA-mers
605 are shown in blue (Nichols) and red (SS14). Locations of chromosomal sequence differences
606 between Nichols and SS14 donor cassettes are marked with black points. G) Proportion of unique
607 7AA-mers of each type – present in both lineages, Nichols only, or SS14 only, and directly
608 encoded by DCs or a mosaic sequence not found in the donor cassettes – by variable region.
609 SS14 lineage strains contain significantly more lineage-specific V7 7AA-mers than Nichols,
610 including both DC-derived and mosaic sequences generated by gene conversion, **p<0.01, chi
611 squared.
612

613 **FIGURE 5**
614



615
616
617 **Figure 5: Inter-sample *tprK* relatedness by pairwise sample relationship.** Pairwise
618 comparisons of *tprK* profiling data were performed using different statistical methods were
619 performed. Samples were grouped based on whether they were taken from different samples
620 from the same individual, individuals from the same household constituting likely transmission,
621 and based on core genome SNV distance. A) Pearson coefficient of relative proportions of
622 variable regions. B) Composite Bray-Curtis Dissimilarity for DNA kmers in variable regions. C)
623 Median pairwise alignment distance between *tprK* alleles in each sample pair type. D) Median
624 composite inter-allele composite Bray-Curtis dissimilarity in each sample pair type. ANOVA with
625 Tukey's post hoc, **p<0.01, ****p<0.0001. Red boxes enclose samples from a pair of patients
626 that may be epidemiologically linked based on patient age, sex and orientation, and
627 presentation to the clinic on the same day,
628

Lineage-specific *tprK* diversification and *Treponema pallidum* transmission dynamics in Buenos Aires, Argentina

Lieberman, Garcia, et al

SUPPLEMENT

Supplementary Methods

Ethics approvals and enrollment criteria (Brazil, Peru)

Enrollment of adults in Brazil was approved by the Brazil: National Research Ethics Commission protocol 56591822.9.0000.5243 and screened according to Brazilian National Guidelines[1]. Enrollment of adults in Peru was performed under the University of Southern California IRB HS-21-00353, approved by the ethical review board at Universidad Peruana Cayetano Heredia in Lima on December 3rd, 2019 (UPCH IRB protocol 103093). Enrollment criteria and screening of the Peruvian cohort have been previously described[2].

DNA Extraction and qPCR

In the Buenos Aires cohort, DNA from 300 µL whole blood samples or dry swabs suspended in 400 µL of PBS was extracted using QIAamp DNA Blood Mini Kit (Qiagen, Germany) according to the manufacturer's instructions. Then, the samples were eluted in Qiagen AE buffer to a final volume of 200 µl for whole blood samples and 150 µl for swabs. Swabs from Brazil were collected in 2 ml of DNA/RNA shield stabilization solution and extracted using the Quick-DNA 96 Kit (Zymo Research; Irvine, CA). Peruvian swabs were collected in DNA lysis buffer (10 mM Tris-HCl pH 8.0, 0.1 mM EDTA, 0.5% SDS) and extracted using the QIAamp DNA Mini Kit. *T. pallidum* DNA was detected via semi-quantitative PCR for the single copy locus *tp0574* as previously described[2]. Samples were considered positive when *tp0574* cycle threshold was less than 40.

T. pallidum variant calling

Raw fastqs were newly generated or downloaded from the NCBI Sequence Read Archive. Reads were preprocessed by removal of human reads with Kraken2 v2.1.3[3] and quality and adapter trimmed with Trimmomatic v0.39[4], then filtered using BBDuk v39.10[5] at low stringency against the SS14 reference genome (NC_021508) to identify reads containing *Tp* sequences. From these, *Tp* rRNA and tRNA were further filtered at high stringency, requiring 99% identity to the conserved *Tp* references. Filtered *Tp* reads were then mapped against the SS14 reference genome using bowtie2 v2.5.2[6] with default parameters, deduplicated with Picard v2.26.3[7] and variants called using GATK v4.6.1[8] HaplotypeCaller assuming ploidy of 1 and maximum indel size of 130. Joint genotyping was performed using GATK and filtered with bcftools v1.2[9] to a minimum depth of 3 reads and allele frequency >0.8. Masking of previously identified

recombinant loci[10–12] and regions with poor assembly quality, such as homopolymers, was performed prior to SNV-only phylogenetic inference.

Bespoke *tprK* analyses

The number of alleles in each sample was defined as the number of unique full-length *tprK* sequences exceeding 0.2% abundance in the technical replicate of each sample with the highest sequencing quality score. K-mers of *tprK* alleles and donor cassette loci were generated using Jellyfish v1.2[13]. High-confidence *tprK* alleles used to evalutate intra-and inter-sample distance (**Figure 5** and **Figure S11**) are defined as those present in both technical replicates. Inter-sample Pearson correlations between V regions were calculated as previously described[14]. Inter-sample Bray's distance was calculated for lengths of k=9, 11, 13, 15, 17, 19, 21, 23, 25, and 27 using V region-derived k-mers aggregated per sample (ie, without consideration for full-length *tprK* alleles). Inter-allele Bray's distance was calculated using k-mers of lengths 9-27 for each pair of alleles and the median inter-allele distance for each sample pair used for statistics and visualization. Inter-sample V region and allele-derived Bray's distances were summed across all values of k. Pairwise global sequence alignments were performed with mafft v7.41[15] between individual V regions and inter-allele alignment distance evaluated as the sum of V1-V7. The median inter-allele alignment distance for each sample pair was used for statistics and visualization.

Supplementary Table 2: Demographic characteristics of participants from Brazil and Peru with WGS newly included in this study

	Brazil (n=8) n (%)	**Peru (n=3) n (%)**
AGE: median (range) years	28 (24-31)	37 (30-38)
STAGE		
Adults (>18)	7 (87.5)	3 (100.0)
Pediatric (0-18)	1 (12.5)	0 (0)
BIOLOGICAL SEX		
Male	7 (87.5)	3 (100.0)
Female	1 (12.5)	0 (0)
SEXUAL ORIENTATION (among adult and adolescent male)		

MSM	4 (57.1)	3 (100.0)
No MSM	3 (4.9)	0 (0%)
HIV		
Positive	1 (12.5)	1 (33.3%)
Negative	7 (87.5)	2 (66.7%)
PRIOR SYPHILIS		
Yes	0 (0)	0 (0)
No	8 (100)	3 (100)
CLINICAL STAGE		
Primary	4 (50)	3 (100)
Secondary	3 (37.5)	0 (0)
Congenital syphilis	1 (12.5)	0 (0)

Supplementary Analysis

Extended clinical findings (Buenos Aires)

All primary syphilis cases in adults (18 male and 3 female), except one, had a chancre at sites of sexual exposure while two cases presented with Folman's Balanitis (one patient had both a penis chancre and balanitis). Three primary syphilis cases had VDRL negative results, including one person living with HIV that presented with a genital chancre and darkfield positive result. Predominant symptoms among secondary syphilis cases (60 male and 32 female) were lymphadenopathy (65/92, 70.7%) and rash (52/92, 56.5%). Rash types included macular (16 cases), maculopapular (16), papular (15), and papulosquamous (5). Some patients also expressed oral or tongue plaque lesions (33/92, 35.9%), genital or perianal condyloma lata (30/92; 32.6%), and genital erosions (17/92, 18.5%). Five patients also had alopecia (5/92, 5.4%).

Congenital syphilis cases (one male and three females) were newborns and three had early signs of congenital syphilis including pemphigus-like lesions, hydrops fetalis, and hepatosplenomegaly. All primary syphilis cases were adolescents (two males and one female) with genital chancre. Secondary syphilis cases included 12 children (five males and seven females) and nine adolescents (five males and four females). Among them, common symptoms included macular or maculo-papular rash (7/21, 33.3%), genital or perianal condyloma (13/21,

61.9%) and palate erosion (5/21, 23.8%) patients. Few patients expressed lymphadenopathy (4/21, 19.0%), alopecia (2/21, 9.5%), and genital erosions (2/21, 9.5%).

Lesion-specific SNVs in *tp0006*

We recovered near-complete genomes from multiple specimens from 20 individuals and found no evidence of intra-host evolution in 19 of the matched sets. In one sample pair, taken from perianal condyloma and skin rash from an individual (GSPAR095M) with secondary syphilis, two clonal SNVs (D136G in the perianal lesion and M273I in the axilla skin sample) were seen in *tp0006,* encoding an accessory protein of unknown function[16]. Analysis of the predicted structure using Foldseek[17] revealed TP0006 to have structural homology to substrate binding domains of ATP-binding cassette transporters (**Figure S2;** Dr. Chad Brautigam, personal communication). Of note, the patient reported treatment with amoxicillin for an unrelated infection approximately six weeks prior to his presentation with secondary syphilis.

Limited diversity in *tprK* constant regions:

Among 79 samples from 54 patients with matched WGS, *tprK* constant regions showed limited sequence variation, with only six positions out of approximately 1,600 nucleotides different between Nichols and SS14 lineages and none in two previously identified long-lived T cell epitopes[18]. In samples from one patient (GSPAR031R), a 59-year-old cisgender heterosexual man newly diagnosed with HIV, *tprK* alleles from both genital mucosa and interdigital skin had a novel complex mutation resulting in conservative amino acid change T271A in the conserved region between V2 and V3, corroborated by WGS in four separate site samples from the same patient. A 13 nt sequence containing the altered sequence (**Figure S5A**) was identical to a motif within the conserved N terminus of *tpr* subfamily I, providing a donor sequence for a putative recombination event. Additionally, two patients had an A-to-C transversion that conferred a W516C mutation in the C-terminal tail of TprK (**Figure S5B**). In patient GSPAR093S the mutation was present in the resolving genital chancre at approximately 53% abundance and 90% in the oral lesion. In the matched WGS sample from the genital lesion, no other genomic positions were mixed, suggesting *de novo* mutation rather than co-infection. Using WGS data, the W516C variant is found between 10-100% allele frequency in 124/1,879 (6.6%) samples (**Figure S5C**).

Supplementary References

1. Freitas FLS, Benzaken AS, Passos MRL de, Coelho ICB, Miranda AE. Brazilian Protocol for Sexually Transmitted Infections 2020: acquired syphilis. Rev Soc Bras Med Trop 2021;54(suppl 1):e2020616.
2. Kojima N, Park H, Konda KA, et al. The PICASSO Cohort: baseline characteristics of a cohort of men who have sex with men and male-to-female transgender women at high risk for syphilis infection in Lima, Peru. BMC Infect Dis 2017;17(1):255.
3. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol [Internet] 2019 [cited 2025 July 9];20(1). Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1891-0
4. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30(15):2114–20.
5. Bushnell B. BBMap short read aligner, and other bioinformatic tools.
6. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9(4):357–9.
7. Broad Institute. Picard [Internet]. Available from: http://broadinstitute.github.io/picard.
8. Auwera G van der, O'Connor BD. Genomics in the cloud: using Docker, GATK, and WDL in Terra. First edition. Sebastopol, CA: O'Reilly Media; 2020.
9. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. GigaScience [Internet] 2021 [cited 2025 July 9];10(2). Available from: https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giab008/6137722
10. Beale MA, Marks M, Cole MJ, et al. Global phylogeny of Treponema pallidum lineages reveals recent expansion and spread of contemporary syphilis. Nat Microbiol 2021;6(12):1549–60.
11. Grillová L, Oppelt J, Mikalová L, 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.
12. Pla-Díaz M, Sánchez-Busó L, Giacani L, et al. Evolutionary Processes in the Emergence and Recent Spread of the Syphilis Agent, Treponema pallidum. Mol Biol Evol 2022;39(1):msab318.
13. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of *k* -mers. Bioinformatics 2011;27(6):764–70.

14. Lieberman NAP, Avendaño CC, Bakhash SAKM, et al. Genomic Epidemiology of *Treponema pallidum* and Circulation of Strains With Diminished *tprK* Antigen Variation Capability in Seattle, 2021–2022. The Journal of Infectious Diseases 2023;jiad368.
15. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution 2013;30(4):772–80.
16. Edmondson DG, De Lay BD, Hanson BM, Kowis LE, Norris SJ. Clonal isolates of Treponema pallidum subsp. pallidum Nichols provide evidence for the occurrence of microevolution during experimental rabbit infection and in vitro culture. PLoS One 2023;18(3):e0281187.
17. van Kempen M, Kim SS, Tumescheit C, et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol 2024;42(2):243–6.
18. Morgan CA, Molini BJ, Lukehart SA, Van Voorhis WC. Segregation of B and T cell epitopes of Treponema pallidum repeat protein K to variable and conserved regions during experimental syphilis infection. J Immunol 2002;169(2):952–7.

FIGURE S1

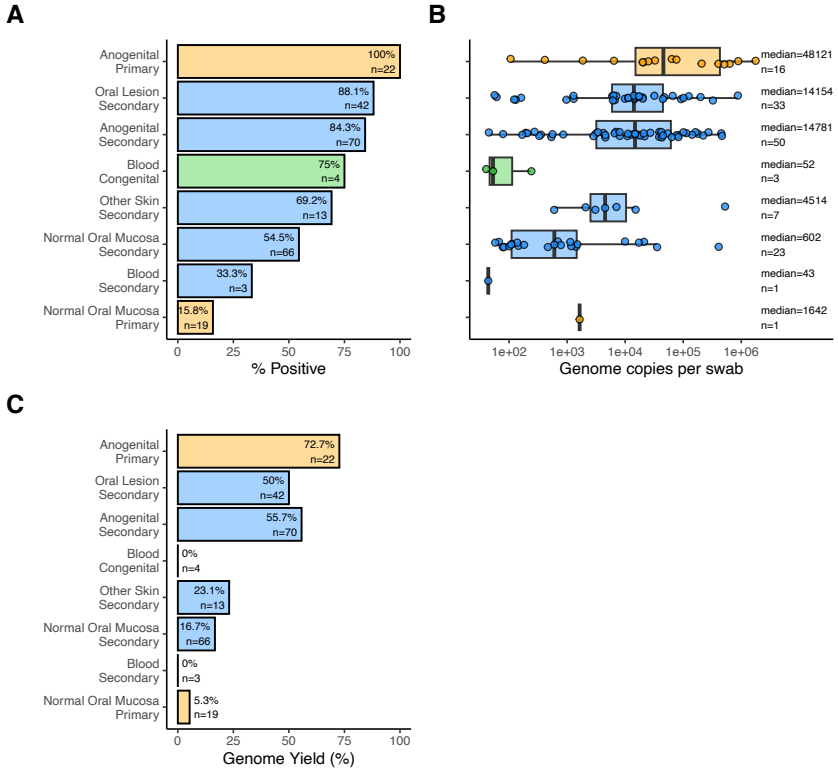


Figure S1: *T. pallidum* molecular positivity and treponemal burden by sample anatomic location and patient's clinical stage. A) Percent of samples positive for the single copy locus *tp0574*. B) Extrapolated total *Tp* genome copies from a single swab or 300 μL whole blood, C) Percent of samples that yielded a complete genome.

FIGURE S2

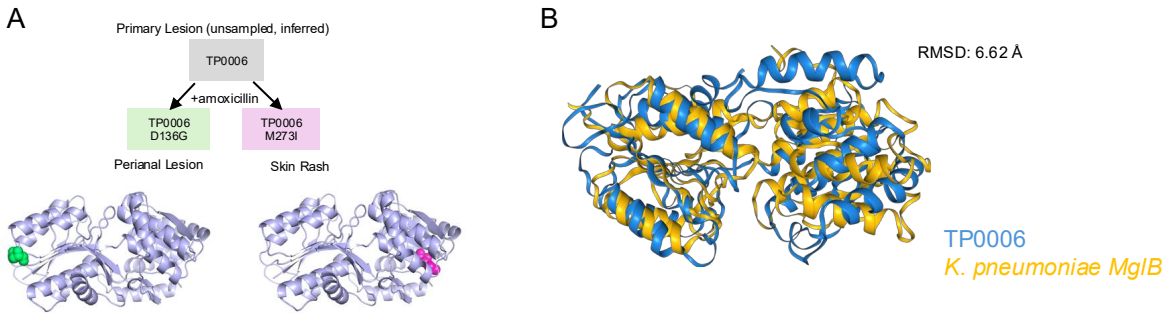


Figure S2: Mutations to *tp0006*, encoding a protein of unknown function. A) Two separate mutations were found in *T. pallidum* DNA recovered from a male patient's perianal and skin lesions relative to the inferred genotype of the infecting strain, which was not sampled. Mutations are shown on the AlphaFold3-predicted structure of TP0006. B) TP0006 has structural homology to the substrate binding proteins of ABC transporters such as *MglB*. RMSD, root mean squared deviation.

FIGURE S3

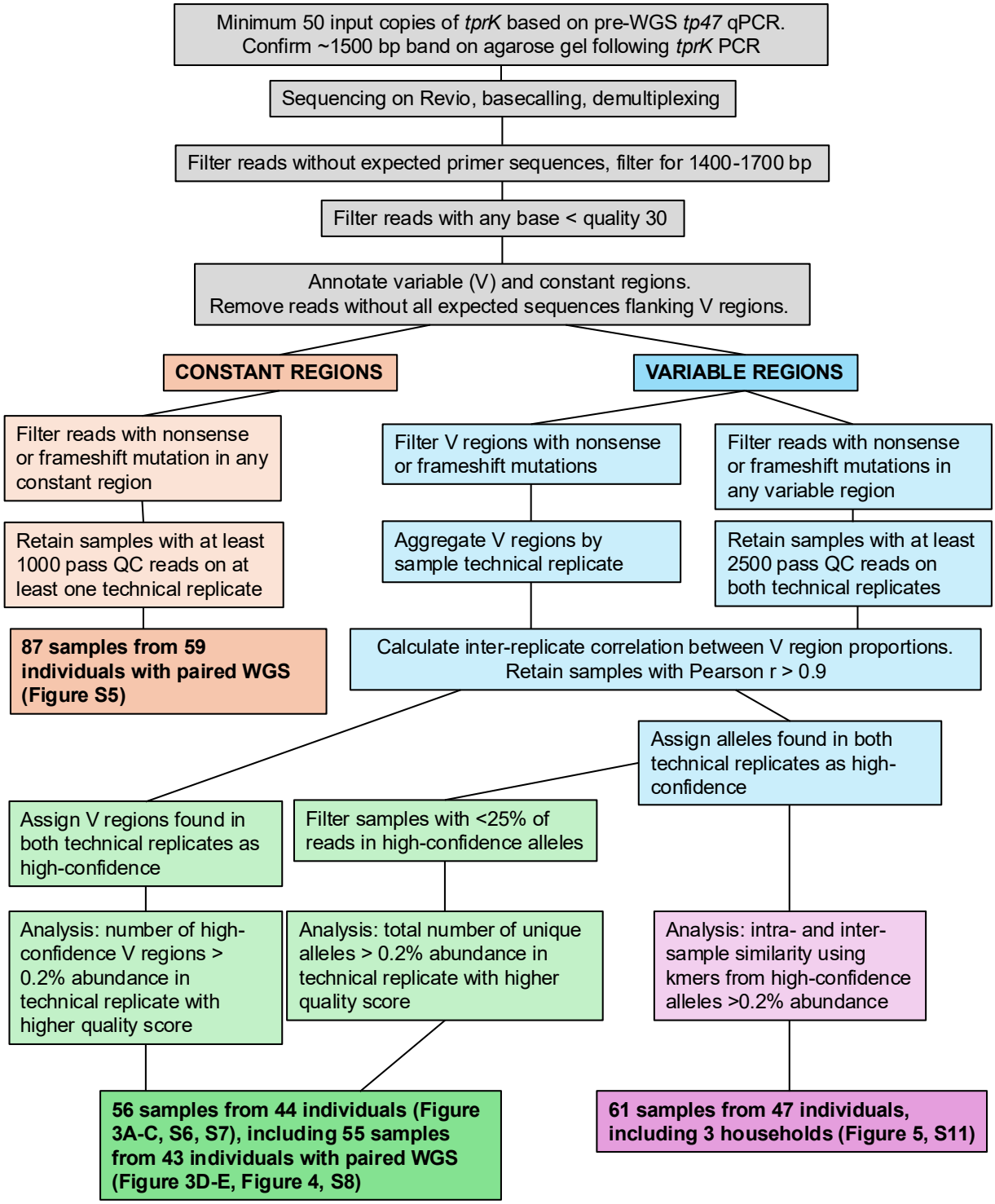


Figure S3: Quality control and filtering processes for each *tprK* analysis.

FIGURE S4

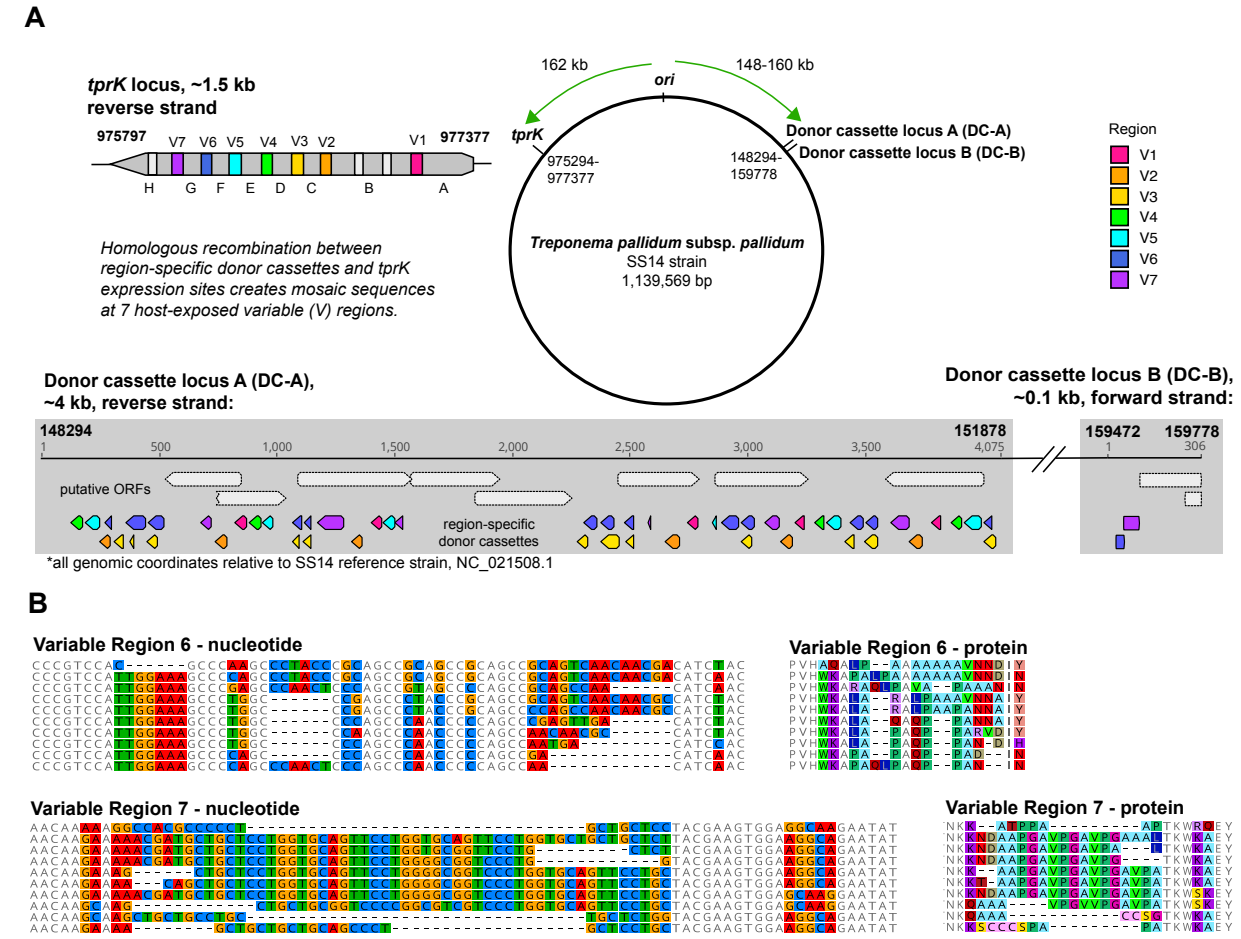


Figure S4: Gene conversion between donor cassettes and expression sites results in antigenic variation in seven variable loops of TprK. A) The *tprK* locus is found 162kb from the origin of replication on the reverse strand of the *T. pallidum* chromosome. *tprK* encodes a ~500 residue predicted outer membrane protein. Seven of its ten variable loops undergo antigenic variation (colored). Constant regions A:H are found between variable regions. Two loci encoding donor cassettes (DCs) are 148 and 159 kb from the origin of replication, respectively. DC-A contains 51 discrete sites with donor cassettes on the reverse strand. DC-B contains 2 sites with donor cassettes on the forward strand. All DCs are V region-specific, with regions V3 (yellow) and V6 (blue) having overlap. DCs overlap with several transcribed open reading frames (ORFs). B) Examples of ten V6 and V7 sequences (out of 1020 and 313, respectively). Both nucleotide and amino acid variability are shown.

FIGURE S5

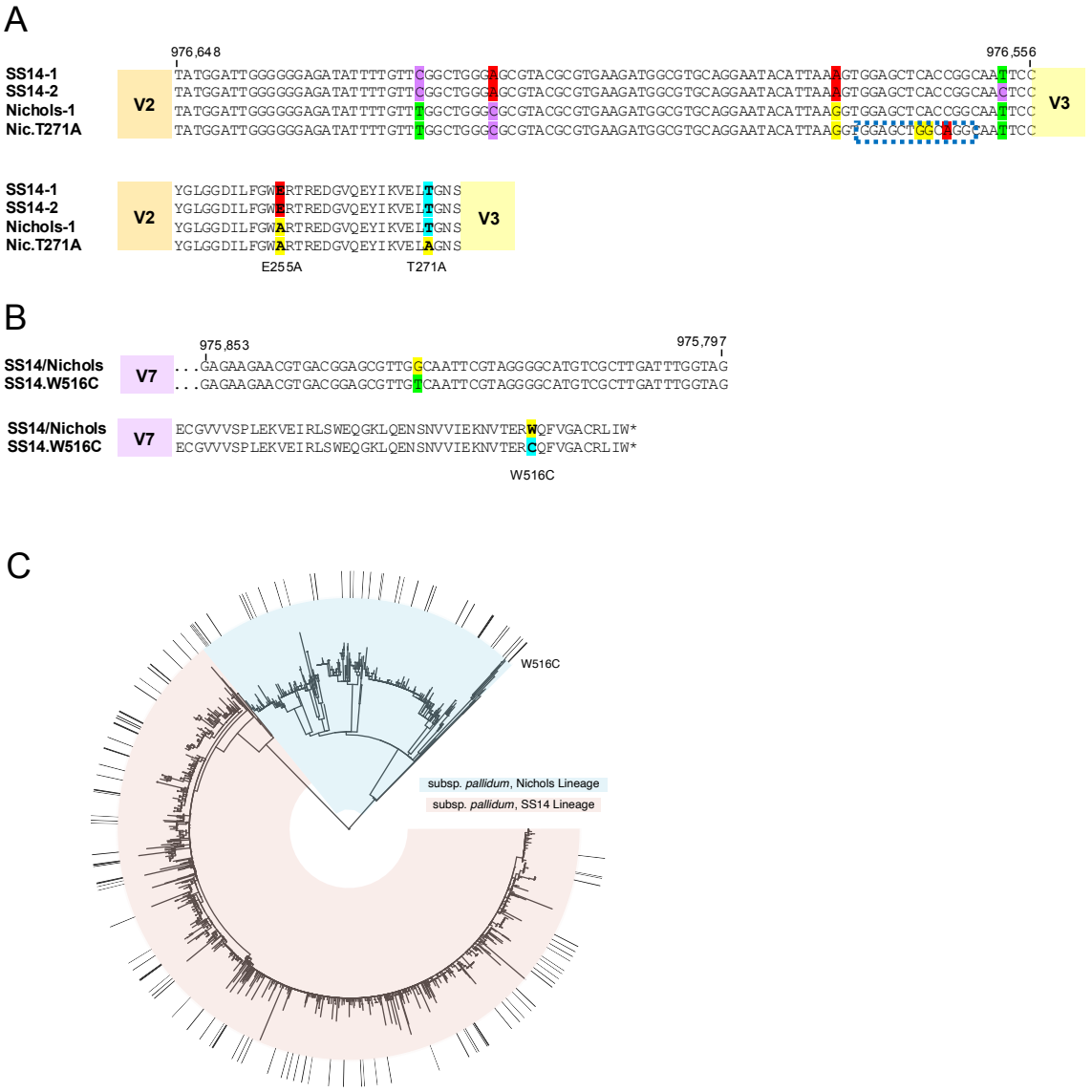


Figure S5: Nucleotide and amino acid differences found in *tprK* constant regions. A) Lineage-defining mutations between V2 and V3. In one patient, a recombination event from a donor site in a subfamily I *tpr* (blue box) was found in samples from four unique body sites, resulting in T271A. B) A SNV resulting in W516C is fully fixed in one patient and incompletely fixed in another. C) Global phylogeny of 1879 strains from Figure 1A. Nichols lineage samples have a blue background and SS14 red. The outer ring denotes strains with W516C fully or partially fixed.

FIGURE S6

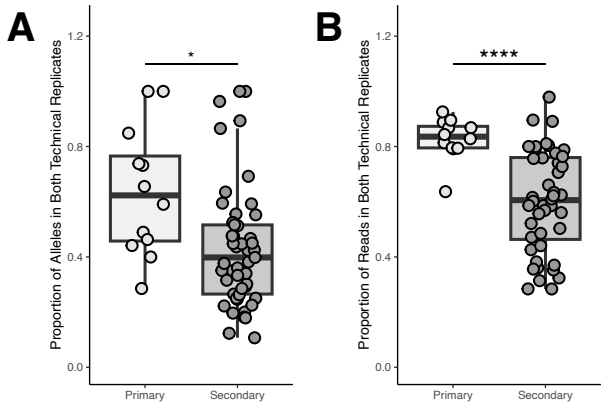


Figure S6: The proportion of unique *tprK* alleles (A) and the percent of total reads (B) in high confidence alleles varies by stage. T-test, *p<0.05, ****p<0.0001

FIGURE S7

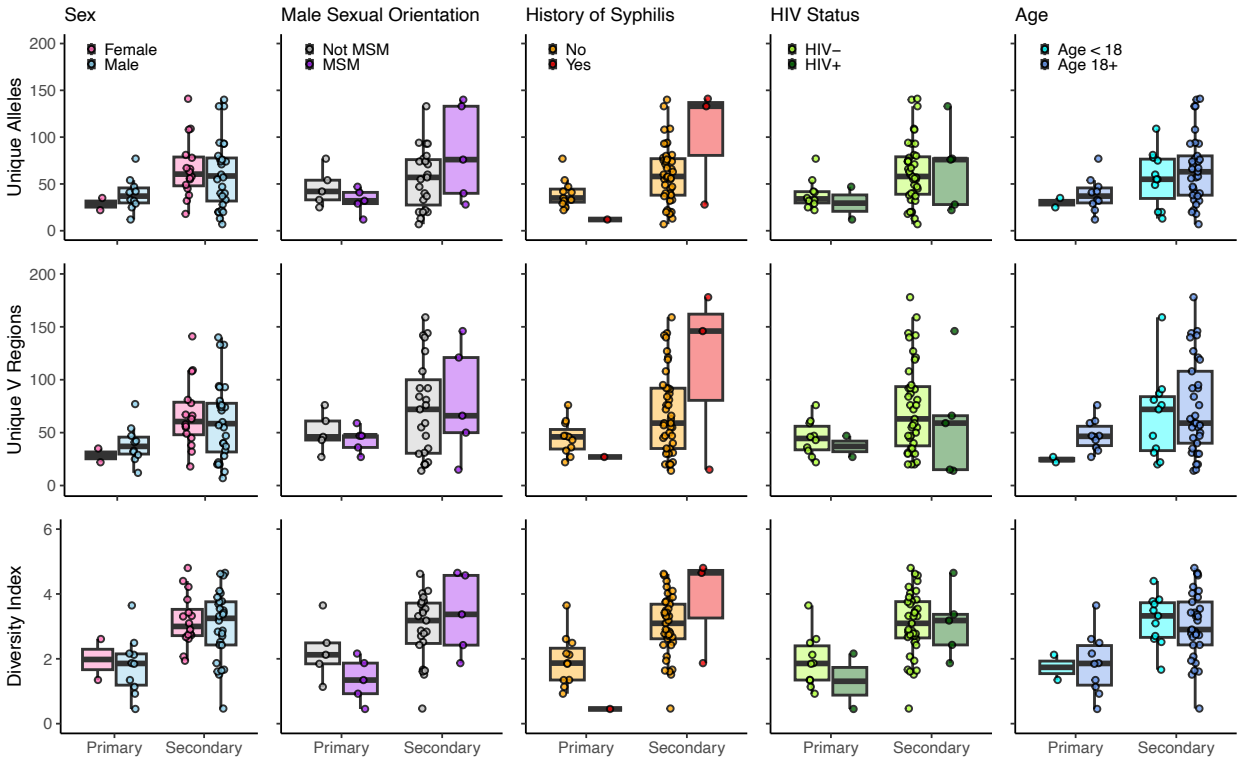


Figure S7: No differences were seen in the number of *tprK* alleles, V regions, or diversity index by patient sex, sexual orientation, reinfection, HIV status, or age.

FIGURE S8

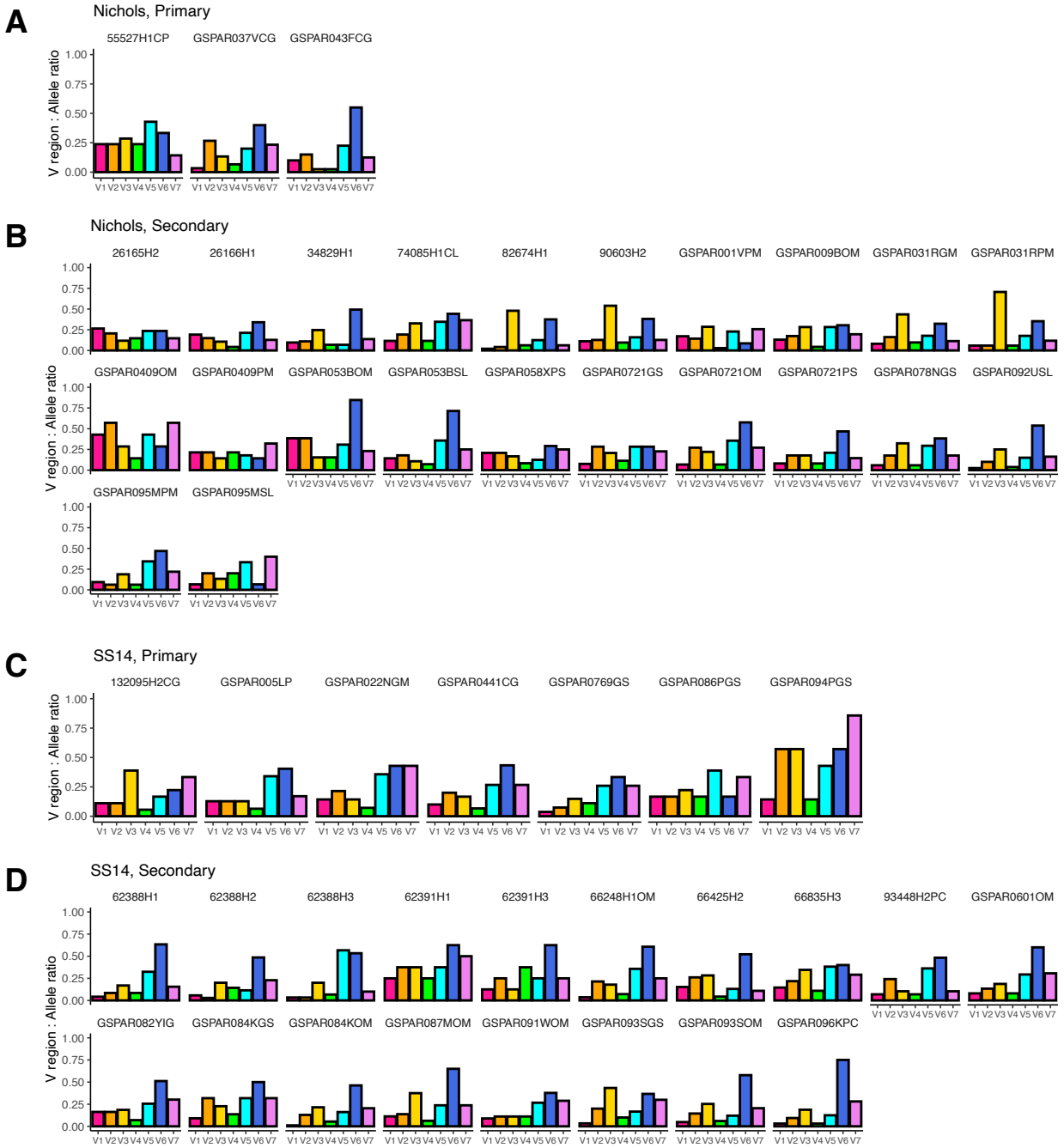


Figure S8: Distribution of unique V regions. The proportion of unique V regions to the number of unique *tprK* alleles is shown for Nichols lineage primary and secondary (A, B) and SS14 lineage primary and secondary (C, D). For example, a sample with 20 unique *tprK* alleles, 2 unique V1 sequences, and 10 unique V6 sequences would have a V1 V:Allele ratio of 0.1 and a V6 V:Allele ratio of 0.5.

FIGURE S9

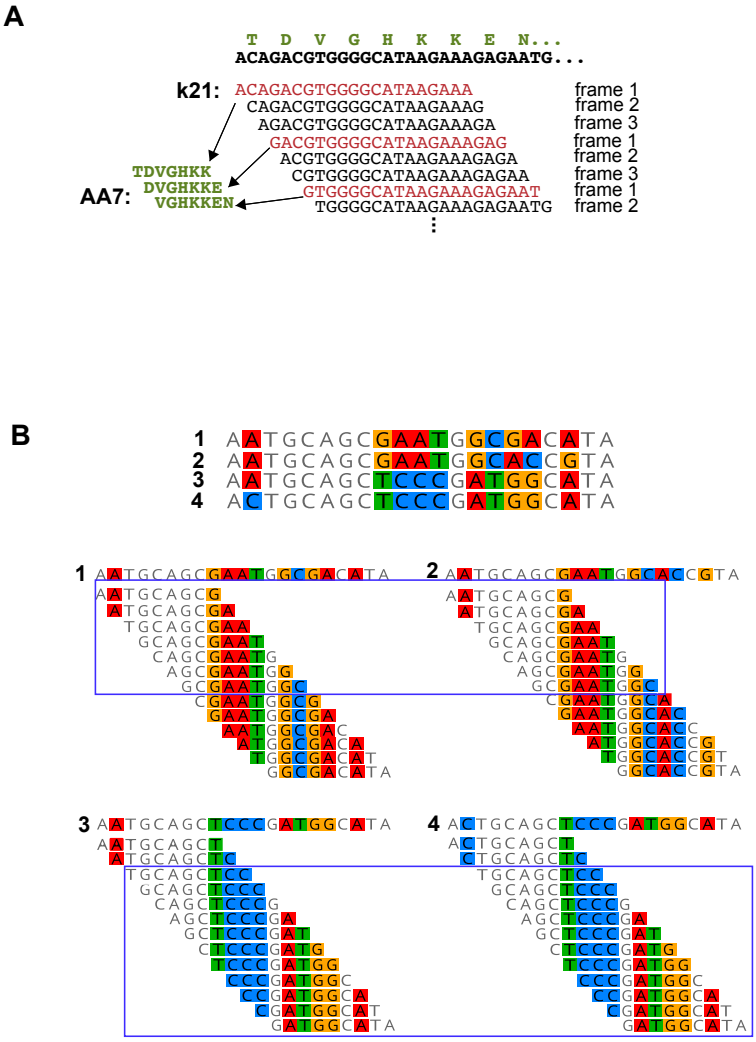


Figure S9: *tprK* analysis using kmers. A) Each *tprK* sequence fragment is broken into overlapping sequences of length k (k=21 is shown). Kmers in frame 1 encode amino acid fragments (AA-mers) of length k/3. B) Similarity between *tprK* alleles can be evaluated using the number of shared kmers/AA-mers. Although all four sequences shown derive from V4, those in the top row share no kmers with the bottom row.

FIGURE S10

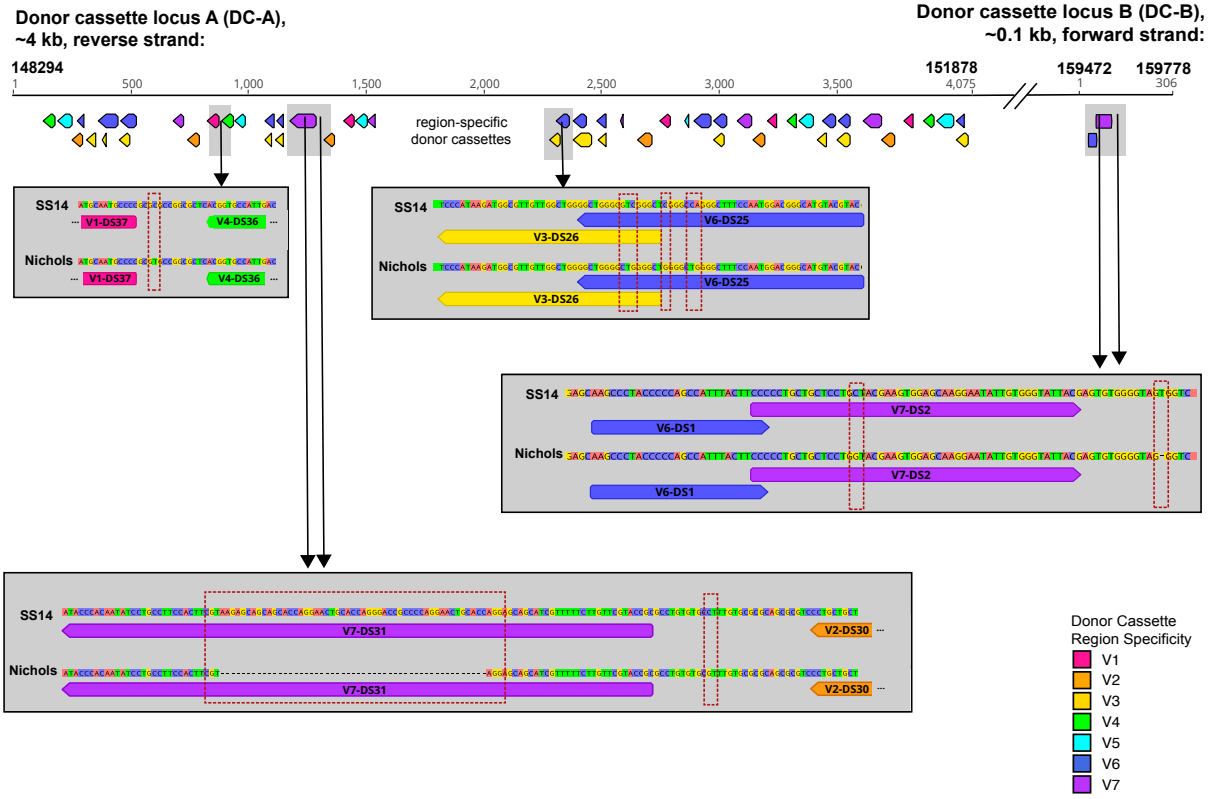


Figure S10: Differences in donor cassette loci between the Nichols and SS14 lineages (grey background) are confined to donor cassettes in regions V3-DC25/V6-DC26, V7-DC31, and V7-DC2, as well as sequences between cassettes.

FIGURE S11

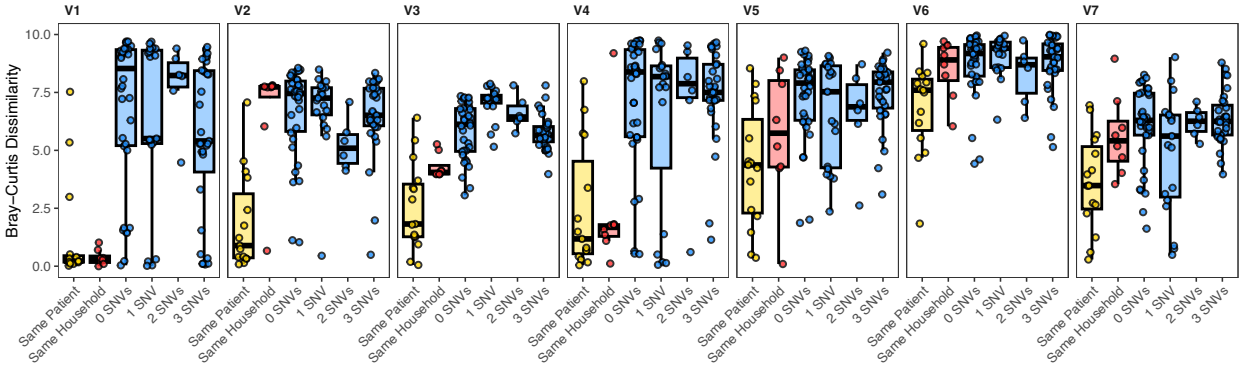


Figure S11: Inter-sample dissimilarity between *tprK* V regions. Dissimilarity of k-mers of length 21 at each sample relationship type varies by V region, with V1 and V4 generally similar within a single patient or household, while V6 is highly variable