AMERICAN SOCIETY FOR MICROBIOLOGY
Microbiology Spectrum

RESEARCH ARTICLE

Genetic Profiling of the Full-Length *tprK* Gene in Patients with Primary and Secondary Syphilis

Dan Liu,[a,b] Li-Li Liu,[a,b] Xin-Qi Zheng,[b] Rui Chen,[b] Li-Rong Lin,[a,b] Tian-Ci Yang,[a,b] Man-Li Tong[a,b]

[a]Center of Clinical Laboratory, Zhongshan Hospital, School of Medicine, Xiamen University, Xiamen, China
[b]Institute of Infectious Disease, School of Medicine, Xiamen University, Xiamen, China

Dan Liu, Li-Li Liu, and Xin-Qi Zheng contributed equally to this work. Author order was determined by division of work.

ABSTRACT TprK antigenic variation is acknowledged as an important strategy developed by *Treponema pallidum* to achieve immune evasion. Previous studies applied short-read sequencing to explore *tprK* gene sequence diversity in clinical samples; however, due to the limitations of short-read sequencing, it was difficult to determine the linkage between the seven V regions, and crucial information about full-length *tprK* variants was lost. Although two recent studies explored complete *tprK* gene profiles in natural human syphilis infection, there are still too few profiled full-length *tprK* variants among clinical *T. pallidum* isolates to fully understand the characteristics of TprK coding diversity. Here, Pacific Biosciences (PacBio) long-read sequencing was applied to examine the diversity of full-length *tprK* variants in 21 clinical *T. pallidum* isolates from 11 patients with primary syphilis and 10 patients with secondary syphilis. A total of 398 high-confidence full-length sequences, which presented remarkable sequence heterogeneity, were found. However, these full-length *tprK* variants exhibited limited variation in length and GC content, showing 24 length types and average GC content of 51.5 ± 0.42% and 51.6 ± 0.26% for primary and secondary syphilis samples, respectively. Additionally, the combined patterns of mutated V regions generating new *tprK* variants were obviously different in primary and secondary syphilis samples. The diversity of *tprK* gene sequences in primary syphilis samples may represent the underlying variability of the bacterium; conversely, the variability of the *tprK* gene in secondary syphilis samples may more accurately reflect how *T. pallidum* escapes host immune clearance. These data highlight the *tprK* gene as an important coding gene that shows conflicting genetic characteristics but underlies the persistence of spirochete infection.

IMPORTANCE The resurgence of syphilis in both low- and high-income countries has attracted attention, and persistent infection by the pathogen has long been a research focus. The *tprK* gene, encoding the hypervariable outer membrane protein, is thought to be responsible for pathogen immune evasion and persistent infection. Here, PacBio long-read sequencing was applied to examine the diversity of full-length *tprK* variants in 21 clinical *T. pallidum* isolates from 11 patients with primary syphilis and 10 patients with secondary syphilis. The results showed that the sequences of the *tprK* gene were remarkably heterogeneous; however, the sequences presented limited variation in length and GC content. The investigation of the combined patterns of the V regions allowed us to gain insight into the features of the *tprK* gene generating new variants at different clinical stages. The findings of this study will be helpful for further exploration of the pathogenesis of syphilis.

KEYWORDS *Treponema pallidum*, full-length *tprK* variants, sequence analysis, syphilis

Syphilis, which is caused by the spirochete *Treponema pallidum*, is a chronic sexually transmitted infection (1, 2). Its explosive resurgence in many countries, particularly among men who have sex with men (MSM) and persons living with HIV, has sounded

Editor Mark William Pandori, Nevada State Public Health Laboratory

Address correspondence to Li-Rong Lin, linlirong@xmu.edu.cn, Tian-Ci Yang, yangtianci@xmu.edu.cn, or Man-Li Tong, tongmanli@xmu.edu.cn.

The authors declare no conflict of interest.

Received 30 November 2022
Accepted 17 March 2023
Published 10 April 2023

alarm bells (3, 4). The persistence of the pathogen for the lifetime of an infected individual is the greatest threat to the patient's quality of life (5); however, the mechanism underlying persistent infection is not fully understood. A generally accepted view is that antigen variation allows *T. pallidum* to escape host immune clearance (6, 7). The rare surface-exposed outer membrane proteins of the pathogen have attracted much attention. The paralogous families of *T. pallidum* repeat proteins (Tprs) have been identified as belonging to the repertoire of spirochete outer membrane proteins (8). TprK, a member of the Tpr family, not only is thought to be a principal candidate vaccinogen but also is considered to represent an important immune evasion strategy developed by *T. pallidum* (9–12).

The protein-coding *tprK* gene harbors seven V regions (V1 to V7), each of which is flanked by highly conserved domains (13). Based on a predicted TprK structure, each V region was shown to be exposed as a loop on the surface, and the V regions were found to be antibody targets in experimental syphilis infection (14, 15). The generation of V region diversity was driven by nonreciprocal segmental gene conversion. This gene conversion could create more than 1 million new V sequences and result in significant intra- and interstrain diversity of the *tprK* gene (13). In response to host immune selection, the *tprK* gene showed remarkable accumulation of diversity (16, 17). This finding indicates that understanding the variation of TprK is critical for deciphering the mechanism of lifelong spirochete infection (18, 19).

Recent approaches aimed at understanding TprK antigenic variation have turned to short-read sequencing to push the field forward (19–21). Researchers discovered V sequences on a large scale and gained insight into the diversity of the profiles of the seven V regions of the *tprK* gene. However, due to the limitations imposed by short-read sequencing, these studies had difficulty resolving the linkage between V regions, resulting in the loss of important information about the profiles of full-length *tprK* variants (19–21).

Two recent studies by Addetia et al. applied third-generation (Pacific Biosciences [PacBio]) sequencing to directly profile full-length *tprK* variants (22, 23). In contrast to the redundancy of the V sequences found at the interstrain level (19), none of the full-length *tprK* variants was shared between the analyzed samples. Given the extraordinary diversity present in the *tprK* gene, there are still too few profiled full-length *tprK* variants among clinical *T. pallidum* isolates to fully understand the characteristics of TprK coding diversity (23). Furthermore, the two studies did not investigate the combinations of the seven V regions in single full-length *tprK* variants, which could provide important information about the generation of a new *tprK* variant.

In this study, we applied PacBio long-read sequencing to explore the diversity of full-length *tprK* variants in 21 clinical *T. pallidum* isolates from 11 patients with primary syphilis and 10 patients with secondary syphilis. Deep profiling of the diversity of the full-length *tprK* gene and exploration of the linkage between V regions would help us to better understand the characteristic variation of the *tprK* gene at different clinical stages. Obtaining high-confidence full-length *tprK* variants among clinical *T. pallidum* isolates would provide a critical foundation for developing further serological studies to characterize the anti-TprK immune response.

RESULTS

Long-read sequencing data for the *tprK* gene from 28 syphilis samples. Long-read sequencing was performed to analyze the full-length *tprK* genes of 28 syphilis samples. This analysis resulted in an average of 4,986 amplicon sequences for the 28 samples (range, 1,086 to 7,602 reads); the average read length was 1,587 bp, with a median read length of 1,590 bp. The detailed read data are shown in Table S2 in the supplemental material. Next, we applied the quality-filtering strategy to obtain high-confidence full-length *tprK* sequences for the 28 samples (Fig. 1A). Four samples with poor sequencing quality (S-11, S-14, X-6, and X-13) were excluded because of excessive nonspecific amplified fragments. Additionally, to evaluate the quality of our long-read sequencing results,

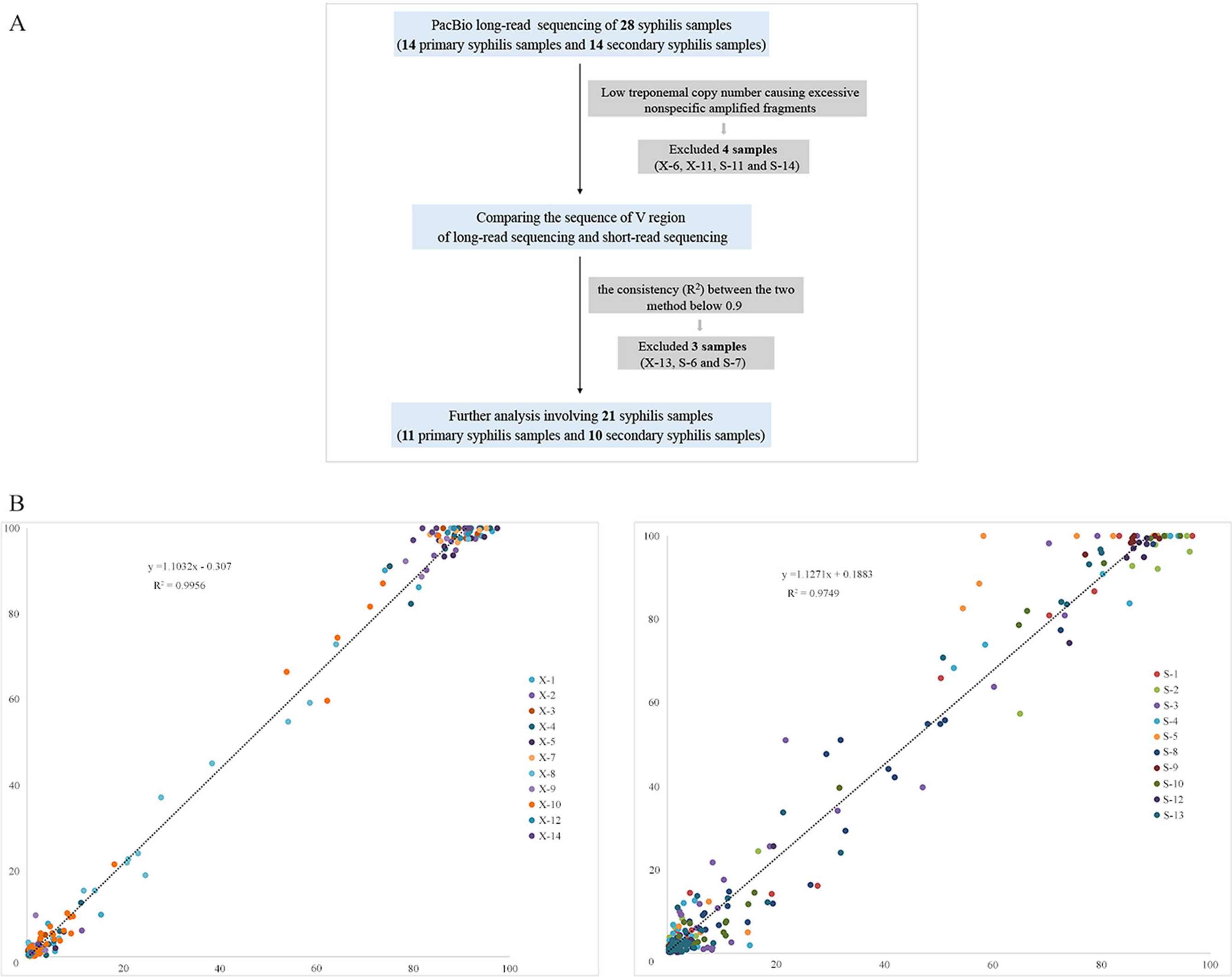

FIG 1 Sample filtering strategy. (A) Flowchart of the strategy. (B) Comparison of long-read and short-read sequencing of the V sequences from 11 primary syphilis samples and 10 secondary syphilis samples.

we compared the sequences of the V regions obtained from these high-confidence sequences to those obtained from short-read sequencing in previous studies (19). This comparison showed that the R^2 values for consistency of the V sequences between the short reads and the filtered long reads ranged from 0.8821 to 0.993 for primary syphilis samples and from 0.8657 to 0.9976 for secondary syphilis samples. We included only the 21 samples (11 primary syphilis samples and 10 secondary syphilis samples) with consistency values above 0.9 in the final analysis. The comparison of the two sets of sequencing data showed high consistency (R^2 values of 0.9956 and 0.9749 for primary syphilis samples and secondary syphilis samples, respectively) (Fig. 1B).

Remarkable diversity of full-length *tprK* variants in primary syphilis samples and secondary syphilis samples. Using the long-read data, we identified a total of 193 *tprK* variants across 11 primary syphilis samples and 205 variants across 10 secondary syphilis samples (see Table S3). We next found that the distribution of the variants in most primary syphilis samples was dominated by one variant with a frequency of >60% (Fig. 2A). In contrast, the distribution of the variants in most of the secondary syphilis samples was dominated by one variant with a frequency of <60% (Fig. 2B). As shown in Fig. 2, the frequencies of other *tprK* variants found across primary syphilis samples were generally as low as 5% or lower. While this corresponded to the lower frequency of the dominant variant across secondary syphilis samples, some of the minor

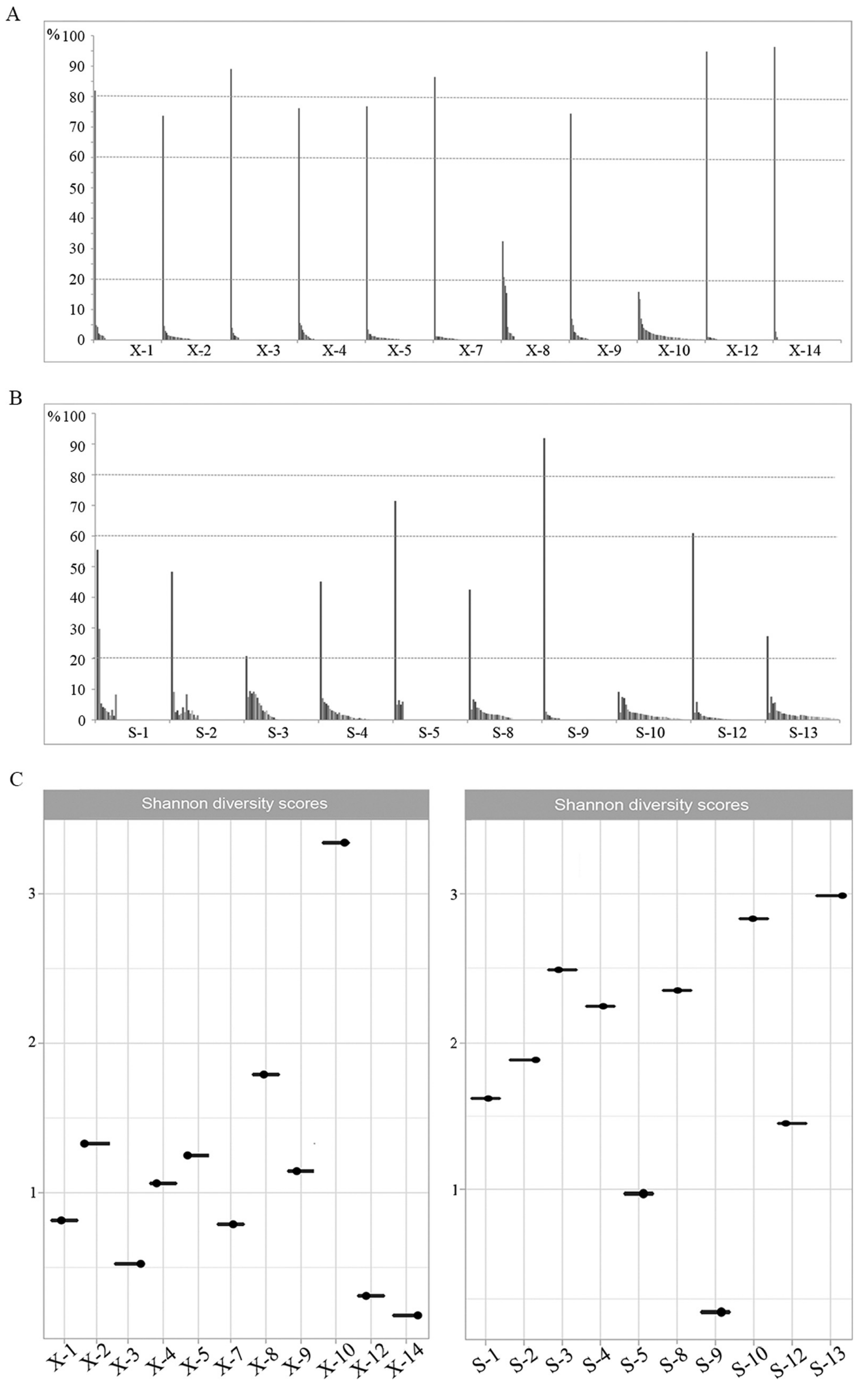

FIG 2 Diversity of full-length *tprK* variants in primary syphilis samples and secondary syphilis samples. (A) Proportional distribution of the distinct full-length *tprK* variants obtained from primary syphilis samples. (B) Proportional distribution of the distinct full-length *tprK* variants obtained from secondary syphilis samples. (C) Shannon diversity measures of the full-length *tprK* variants from primary and secondary syphilis samples.

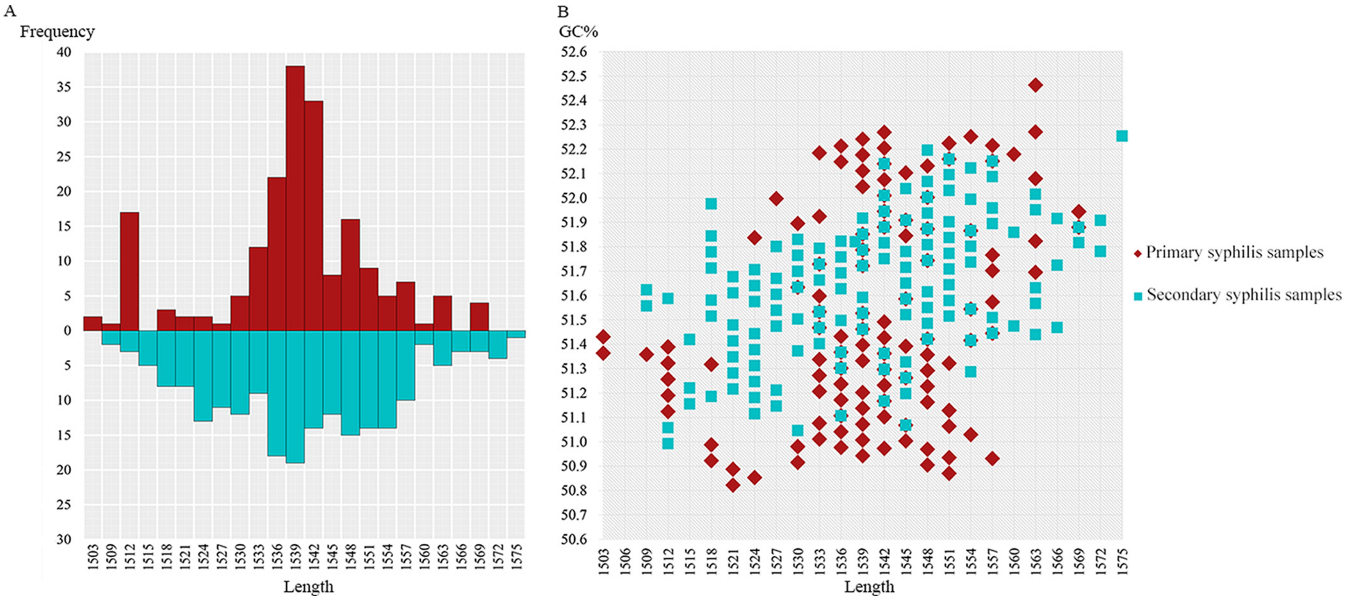

FIG 3 Variation in the length and GC content of full-length *tprK* variants. (A) Distribution of the length of *tprK* variants obtained from primary and secondary syphilis samples. (B) Scatterplot of the length and GC content of *tprK* variants from primary and secondary syphilis samples.

variants were distributed at frequencies of 5% to 20%. Then, we applied Shannon diversity scores to compare the diversity of the *tprK* variants at the two different stages. The scores for each sample are shown in Fig. 2C. The full-length *tprK* variants from secondary syphilis samples were more diverse than those from primary syphilis samples ($P = 0.035$). Additionally, when we aligned all 398 full-length *tprK* variants, it was surprisingly found that none of them was shared across these 21 samples. The variants in each sample appeared to form an independent population.

Limited variation in the length and GC content of full-length *tprK* variants. Although the *tprK* variants identified across these 21 samples presented remarkable diversity, the length range of these variants was clearly limited. The lengths of the variants ranged from 1,503 bp to 1,575 bp, including only 24 length types. The distribution of these lengths exhibited a nearly normal distribution, regardless of whether the samples were from primary syphilis or secondary syphilis (Fig. 3A). However, it is worth noting that the length types of the variants in primary syphilis samples presented a more concentrated distribution, centered around a length of 1,539 bp. In contrast, the distribution of secondary syphilis samples was relatively scattered. Subsequently, we calculated the GC content of each full-length *tprK* variant and found that the GC content of the *tprK* variants from primary and secondary syphilis samples showed only slight differences, with average GC content of 51.5 ± 0.42% and 51.6 ± 0.26%, respectively (Fig. 3B). The relationship between GC content and length across the two groups was also analyzed. There was a weak positive correlation between GC content and length in primary syphilis samples ($r = 0.242$, $P < 0.05$) and a moderate positive correlation in secondary syphilis samples ($r = 0.496$, $P < 0.05$).

Combined patterns of the seven V regions in full-length *tprK* variants. Due to the limitations of short-read sequencing, the V regions obtained by this approach could not be assembled into a full-length *tprK* variant. In this study, high-confidence full-length *tprK* variants were obtained using long-read sequencing, and the features of the seven V regions in the full-length patterns could be further explored. The V sequences of the predominant full-length *tprK* variants of most samples were the sequences with the greatest proportions in the short-sequencing data, except for X-10 and S-3. For sample X-10, the V3 region sequence of the predominant full-length *tprK* variant was not consistent with the greatest proportion of V3 sequences in the short-read sequencing data. This phenomenon occurred in both the V3 and V7 regions in S-3. Further inspection revealed that, in these two samples, the frequency of the predominant full-length

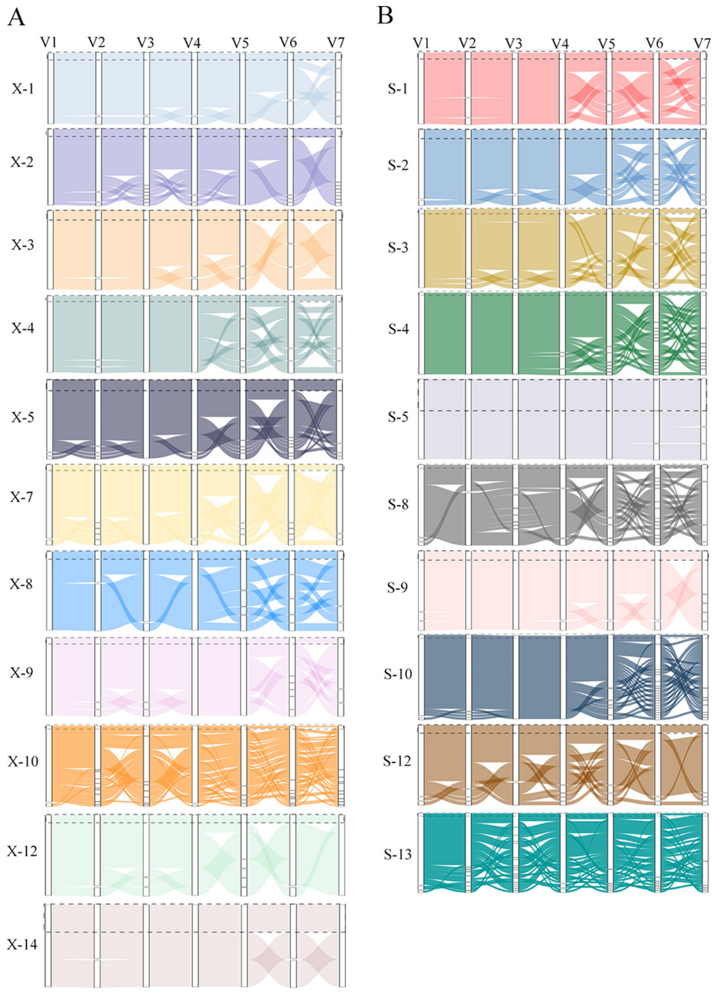

FIG 4 Combined patterns of the seven V regions in full-length *tprK* variants. (A) Primary syphilis samples. (B) Secondary syphilis samples. The sequence usage of the seven V regions in full-length *tprK* variants is represented by the area of the block in each column. The predominant full-length *tprK* variant in each sample is enclosed in a dashed box.

tprK variant was less than approximately 20%, which was lower than that in the other samples (Fig. 4).

Next, how each V variant assembles into a full-length *tprK* variant, creating the rich diversity of the *tprK* gene, was investigated. In the plots of full-length *tprK* variants, we found that most full-length *tprK* variants in each primary syphilis sample differed from the predominant sequence by only one V region, especially in samples containing a high-frequency full-length *tprK* variant (Fig. 4A). It was rare to find more than two mutant V regions concurrently in the minor full-length *tprK* variant, compared to the predominant sequence. Consequently, sequences within individual samples showed a high degree of similarity. However, as shown in Fig. 4A, the combined pattern of the seven V regions in samples X-8 and X-10 was distinctly different from that described above. More than two mutant V regions in a single full-length *tprK* sequence were frequently observed in these two samples. This scenario became more common across secondary syphilis samples. An increasing number of mutant V regions simultaneously appeared in one minor *tprK* sequence, compared to the predominant sequence (Fig. 4B). Therefore, this increased the diversity of the sequences in secondary syphilis samples. Of course, there were also exceptions, such as samples S-5 and S-9, in which the frequency of the predominant *tprK* variants was relatively high and most of the full-length *tprK* variants had only a single variable V region.

DISCUSSION

Our previous studies provided a snapshot of the diversity of *tprK* gene sequences in patients with primary and secondary syphilis, based on short-read sequencing (19, 20). Here, we applied long-read sequencing to further investigate the diversity of the full-length profiles of *tprK*. With our filtering criteria, a total of 398 full-length *tprK* sequences were obtained. Based on the lexicon of the seven V regions of the *tprK* gene obtained in our previous studies (19), we found that the V sequences obtained via the two methods showed high consistency, suggesting that the full-length *tprK* variants obtained in this study were of high confidence. When the distribution of the full-length *tprK* variants in each given sample was analyzed, the distribution trend was similar to that we found previously (19), which was also seen in samples from Italy in the study by Addetia et al. (23). This result indicated that the distribution of the *tprK* variants in the primary and secondary syphilis samples was not affected by the geographic area; instead, it might be a genetic characteristic of primary and secondary syphilis.

The lengths of coding sequences are under both functional and structural constraints (24, 25). Low diversity was found among the lengths of the *tprK* variants, and the distribution of these lengths presented a nearly normal distribution between the samples at two different stages, which might indicate that, although TprK is highly variable, the variation is regulated by a strict gene conversion mechanism to maintain structural stability for protein function. Of course, future research is needed to confirm this hypothesis. In addition, the GC content is a factor that constrains the length of the coding sequence in most organisms (26). Therefore, we further explored the relationship between GC content and length among the *tprK* variants and found a positive correlation between GC content and length, indicating that, in addition to functional and structural constraints, GC content is another factor that constrains the length of *tprK* variants. Moreover, the GC content has been reported to be the most prominent property showing strong correlations with recombination in bacteria (27, 28). In this study, the average GC content of the *tprK* variants in secondary syphilis samples was only slightly higher than that in primary syphilis samples, suggesting that *tprK* gene conversion might not be G/C biased. This finding supported the results of the study by Lin et al., which showed no significant differences in GC content in highly variable V6 sequences (21). Generally, the limited variations in the lengths and GC contents of the full-length *tprK* variants seemed to conflict with the highly variable features of the *tprK* gene but were not unreasonable.

The advantages provided by long reads made the large-scale exploration of the combinations of the seven V regions in a single *tprK* variant possible. By aligning the full-length *tprK* variants of each sample and examining the mutant V regions (taking the predominant sequence as a reference), we found that most full-length *tprK* variants of primary syphilis samples contained only one mutant V region. This pattern of variant emergence should greatly reduce the potential diversity of the *tprK* gene, causing the *tprK* variants within a sample to be highly similar. Conversely, the pattern of variant emergence in secondary syphilis samples was more in line with our expectation that different mutant V regions would appear together in a single full-length *tprK* variant. Different combinations of mutant V regions would significantly increase the diversity of the *tprK* gene, in turn reducing the similarity between sequences and resulting in a greater likelihood of inducing TprK antigenic variation to facilitate secondary syphilis development (18). As reported in a recent study in which the *tprK* gene was shown to maintain a low but detectable basal rate of variation in the complete absence of immune pressure (21), the profile of the *tprK* variants in primary syphilis samples seemed more similar to that observed in the absence of immune pressure, showing basal variability. The *tprK* gene then accumulates more diversity on top of its own basal variation to facilitate disease development, as shown by the profile of the *tprK* gene in secondary syphilis samples. The variability of the *tprK* gene in secondary syphilis samples would more accurately reflect how *T. pallidum* escapes host immune clearance (18).

Notably, the full-length *tprK* variants were not found to be shared among the 21 samples, as reported by Addetia et al. (23). A previous study also demonstrated that an initially infecting isolate and a reinfecting isolate did not share any full-length *tprK* variants (22). Even full-length *tprK* variants were not shared in recurrent cases (29). Instead, the full-length *tprK* variants within the strain were more likely to cluster together (22, 23, 29). This result may indicate that each *tprK* variant is unique, although they were generated from the same repertoire of genome donor sites by nonreciprocal segmental gene conversion (13, 23). Therefore, it has become more important to clarify the function of TprK. It would be helpful to understand the reasons for the high variability of the *tprK* gene.

However, it must be noted that the total number of full-length *tprK* variants discovered in our study was substantially lower than that reported in studies by Addetia et al. (22, 23). The difference in counts between the two studies was mainly due to a large difference in the number of *tprK* sequences in secondary syphilis samples. This could be the consequence of the fewer PacBio reads obtained in our study. Because primary syphilis samples with more *tprK* variants were present at lower frequencies ($<$0.2%), PacBio reads from both studies could not confidently call these *tprK* variants. However, secondary syphilis samples had more *tprK* variants present at frequencies of 5 to 10%, and fewer PacBio reads may increase the chances of losing information about these *tprK* variants. Nevertheless, our results indicated that long-read sequencing was remarkably consistent with short-read sequencing for detecting TprK variants. In addition, we did not further explore the structures of the full-length *tprK* variants to determine whether there were structural changes in the *tprK* variants. Future work should attempt to follow a "structure-to-pathogenesis" approach to map the surface topology of TprK within the context of syphilitic infection, based on a more complete profile of the *tprK* gene.

Overall, the results showed that the *tprK* gene not only presented remarkable sequence heterogeneity but also had relatively conserved properties. The combined patterns of mutated V regions generating new *tprK* variants in primary and secondary syphilis samples were obviously different. The findings of this study will contribute to further exploration of the pathogenesis of syphilis.

MATERIALS AND METHODS

Ethics statement. The subjects included in this study were adults, and all of the subjects provided written informed consent in accordance with the institutional guidelines prior to the study. This study was approved by the Institutional Ethics Committee of Zhongshan Hospital, School of Medicine, Xiamen University, and complied with national legislation and the Declaration of Helsinki guidelines.

Amplification and sequencing of *tprK*. DNA samples from previous studies that had been extracted from 14 primary syphilis samples (genital swab samples) and 14 secondary syphilis samples (lesion biopsy samples) (19, 20) were directly used for the PCR amplification of *tprK*. The detailed characteristics of the samples were described in the previous study (19). High-fidelity PrimeSTAR GXL DNA polymerase (TaKaRa Bio, Inc., Beijing, China) was used in the PCR system to ensure the accuracy of amplification. PCR was conducted with *tprK*-specific primers appended to 16-bp PacBio barcodes (see Table S1 in the supplemental material) under the following conditions: 98°C for 2 min, 35 cycles of 98°C for 10 s, 62°C for 15 s, and 68°C for 2 min, and a final elongation at 68°C for 5 min. The resulting *tprK* amplicon was purified using 0.6 volume of AMPure XP beads (Beckman Coulter, Inc., Brea, CA, USA). Library construction and sequencing on a Sequel I single-molecule real-time (SMRT) Cell 1M system, with a 10-h movie, were completed by Beijing Novogene Bioinformatics Technology.

Processing and filtering of PacBio reads. The output data were filtered and processed using SMRT Link v4.0, and PacBio subreads generated for each amplicon were converted to circular consensus (CCS) reads using CCS v3.0.0. Reads with a consensus base call confidence of $<$99% were excluded. Based on the published custom Python/R scripts available on GitHub (22, 23), scripts were modified to run on our data in-house. Briefly, the retained PacBio reads with lengths between 1,400 and 1,800 bp were trimmed of PCR primers by using the DADA2 preprocessing pipeline and were further denoised by using RAD. The V region sequences previously obtained by short-read sequencing were used for quality checking of the full-length *tprK* sequences (19). A receivable full-length *tprK* sequence had to satisfy the following conditions: (i) each V region must be supported by $\geq$5 reads from short-read sequencing, and (ii) it must not contain a stop codon or frameshift. The detailed bioinformatic pipeline used in the study is also available on GitHub (https://github.com/lauisld/Amoy-research-lab). Shannon diversity scores for each sample were calculated as described previously (23). Diversity scores for strains stratified by clinical stage were assessed using the Wilcoxon rank-sum test. Two-sided *P* values of $<$0.05 were considered statistically significant.

Data availability. PacBio reads obtained from the *tprK* sequencing of the samples in this study are available under NCBI BioProject accession number PRJNA902089.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 0.2 MB.

SUPPLEMENTAL FILE 2, XLSX file, 0.05 MB.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (grants 82272370, 82271387, 82172331, 82001292, 81971147, 81972028, 81973104, and 81871729) and the Natural Science Foundation of Fujian Province, China (grants 2022J02055, 2022J011346, 2021J02055, 2021J01073, and 2020D017). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author roles were as follows: conceptualization: D.L. and M.-L.T.; data curation: D.L., L.-L.L., and X.-Q.Z.; formal analysis: L.-R.L. and T.-C.Y.; funding acquisition: D.L., L.-L.L., L.-R.L., T.-C.Y., and M.-L.T.; investigation: X.-Q.Z. and R.C.; methodology: D.L.; project administration: T.-C.Y. and M.-L.T.; software and visualization: X.-Q.Z. and M.-L.T.; supervision: L.-L.L. and M.-L.T.; validation: D.L. and M.-L.T.; writing (original draft): D.L. and M.-L.T.; writing (reviewing and editing): L.-L.L. and L.-R.L.

REFERENCES

1. Li W, Li QL, Xu QY, Wang XT, Yang TC. 2022. Tp47 promoted the phagocytosis of HMC3 cells though autophagy induced by endoplasmic reticulum stress. J Eur Acad Dermatol Venereol 36:2224–2234. https://doi.org/10.1111/jdv.18295.
2. Xu QY, Wang YJ, Lin LR, Liu LL, Yang TC. 2022. The outer membrane lipoprotein Tp0136 stimulates human platelet activation and aggregation through PAR1 to enhance G_q/G_i signaling. Front Immunol 13:818151. https://doi.org/10.3389/fimmu.2022.818151.
3. Liu D, He SM, Zhu XZ, Liu LL, Lin LR, Niu JJ, Yang TC. 2020. Molecular characterization based on MLST and ECDC typing schemes and antibiotic resistance analyses of *Treponema pallidum* subsp. *pallidum* in Xiamen, China. Front Cell Infect Microbiol 10:618747. https://doi.org/10.3389/fcimb.2020.618747.
4. Mendez-Lopez A, Stuckler D, Marcus U, Hickson F, Noori T, Whittaker RN, Jansen K, Diaz A, Henszel L, Velter A, Semenza JC, Schmidt AJ. 2022. Social and behavioural determinants of syphilis: modelling based on repeated cross-sectional surveys from 2010 and 2017 among 278,256 men who have sex with men in 31 European countries. Lancet Reg Health Eur 22:100483. https://doi.org/10.1016/j.lanepe.2022.100483.
5. Ghanem KG, Ram S, Rice PA. 2020. The modern epidemic of syphilis. N Engl J Med 382:845–854. https://doi.org/10.1056/NEJMra1901593.
6. Kumar S, Caimano MJ, Anand A, Dey A, Hawley KL, LeDoyt ME, La Vake CJ, Cruz AR, Ramirez LG, Paštěková L, Bezsonova I, Šmajs D, Salazar JC, Radolf JD. 2018. Sequence variation of rare outer membrane protein β-barrel domains in clinical strains provides insights into the evolution of *Treponema pallidum* subsp. *pallidum*, the syphilis spirochete. mBio 9:e01006-18. https://doi.org/10.1128/mBio.01006-18.
7. Radolf JD, Deka RK, Anand A, Smajs D, Norgard MV, Yang XF. 2016. *Treponema pallidum*, the syphilis spirochete: making a living as a stealth pathogen. Nat Rev Microbiol 14:744–759. https://doi.org/10.1038/nrmicro.2016.141.
8. Radolf JD, Kumar S. 2018. The *Treponema pallidum* outer membrane. Curr Top Microbiol Immunol 415:1–38. https://doi.org/10.1007/82_2017_44.
9. Centurion-Lara A, Castro C, Barrett L, Cameron C, Mostowfi M, Van Voorhis WC, Lukehart SA. 1999. *Treponema pallidum* major sheath protein homologue Tpr K is a target of opsonic antibody and the protective immune response. J Exp Med 189:647–656. https://doi.org/10.1084/jem.189.4.647.
10. LaFond RE, Molini BJ, Van Voorhis WC, Lukehart SA. 2006. Antigenic variation of TprK V regions abrogates specific antibody binding in syphilis. Infect Immun 74:6244–6251. https://doi.org/10.1128/IAI.00827-06.
11. Morgan CA, Lukehart SA, Van Voorhis WC. 2002. Immunization with the *N*-terminal portion of *Treponema pallidum* repeat protein K attenuates syphilitic lesion development in the rabbit model. Infect Immun 70:6811–6816. https://doi.org/10.1128/IAI.70.12.6811-6816.2002.
12. Morgan CA, Lukehart SA, Van Voorhis WC. 2003. Protection against syphilis correlates with specificity of antibodies to the variable regions of *Treponema pallidum* repeat protein K. Infect Immun 71:5605–5612. https://doi.org/10.1128/IAI.71.10.5605-5612.2003.
13. Centurion-Lara A, LaFond RE, Hevner K, Godornes C, Molini BJ, Van Voorhis WC, Lukehart SA. 2004. Gene conversion: a mechanism for generation of heterogeneity in the *tprK* gene of *Treponema pallidum* during infection. Mol Microbiol 52:1579–1596. https://doi.org/10.1111/j.1365-2958.2004.04086.x.
14. Centurion-Lara A, Giacani L, Godornes C, Molini BJ, Brinck Reid T, Lukehart SA. 2013. Fine analysis of genetic diversity of the *tpr* gene family among treponemal species, subspecies and strains. PLoS Negl Trop Dis 7:e2222. https://doi.org/10.1371/journal.pntd.0002222.
15. Morgan CA, Molini BJ, Lukehart SA, Van Voorhis WC. 2002. Segregation of B and T cell epitopes of *Treponema pallidum* repeat protein K to variable and conserved regions during experimental syphilis infection. J Immunol 169:952–957. https://doi.org/10.4049/jimmunol.169.2.952.
16. LaFond RE, Centurion-Lara A, Godornes C, Rompalo AM, Van Voorhis WC, Lukehart SA. 2003. Sequence diversity of *Treponema pallidum* subsp *pallidum tprK* in human syphilis lesions and rabbit-propagated isolates. J Bacteriol 185:6262–6268. https://doi.org/10.1128/JB.185.21.6262-6268.2003.
17. Giacani L, Molini BJ, Kim EY, Godornes BC, Leader BT, Tantalo LC, Centurion-Lara A, Lukehart SA. 2010. Antigenic variation in *Treponema pallidum*: TprK sequence diversity accumulates in response to immune pressure during experimental syphilis. J Immunol 184:3822–3829. https://doi.org/10.4049/jimmunol.0902788.
18. Reid TB, Molini BJ, Fernandez MC, Lukehart SA. 2014. Antigenic variation of TprK facilitates development of secondary syphilis. Infect Immun 82:4959–4967. https://doi.org/10.1128/IAI.02236-14.
19. Liu D, Tong M-L, Lin Y, Liu L-L, Lin L-R, Yang T-C. 2019. Insights into the genetic variation profile of *tprK* in *Treponema pallidum* during the development of natural human syphilis infection. PLoS Negl Trop Dis 13:e0007621. https://doi.org/10.1371/journal.pntd.0007621.
20. Liu D, Tong ML, Luo X, Liu LL, Lin LR, Zhang HL, Lin Y, Niu JJ, Yang TC. 2019. Profile of the *tprK* gene in primary syphilis patients based on next-generation sequencing. PLoS Negl Trop Dis 13:e0006855. https://doi.org/10.1371/journal.pntd.0006855.
21. Lin MJ, Haynes AM, Addetia A, Lieberman NAP, Phung Q, Xie H, Nguyen TV, Molini BJ, Lukehart SA, Giacani L, Greninger AL. 2021. 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 15:e0009753. https://doi.org/10.1371/journal.pntd.0009753.
22. Addetia A, Tantalo LC, Lin MJ, Xie H, Huang ML, Marra CM, Greninger AL. 2020. Comparative genomics and full-length Tprk profiling of *Treponema*

pallidum subsp. pallidum reinfection. PLoS Negl Trop Dis 14:e0007921. https://doi.org/10.1371/journal.pntd.0007921.

23. Addetia A, Lin MJ, Phung Q, Xie H, Huang M-L, Ciccarese G, Dal Conte I, Cusini M, Drago F, Giacani L, Greninger Alexander L. 2020. Estimation of full-length TprK diversity in *Treponema pallidum* subsp. *pallidum*. mBio 11:e02726-20. https://doi.org/10.1128/mBio.02726-20.
24. Hawkins JD. 1988. A survey on intron and exon lengths. Nucleic Acids Res 16:9893–9908. https://doi.org/10.1093/nar/16.21.9893.
25. Traut TW. 1988. Do exons code for structural or functional units in proteins. Proc Natl Acad Sci U S A 85:2944–2948. https://doi.org/10.1073/pnas.85.9.2944.
26. Oliver JL, Marín A. 1996. A relationship between GC content and coding-sequence length. J Mol Evol 43:216–223. https://doi.org/10.1007/BF02338829.
27. Bobay LM, Ochman H. 2017. Impact of recombination on the base composition of bacteria and archaea. Mol Biol Evol 34:2627–2636. https://doi.org/10.1093/molbev/msx189.
28. Marsolier-Kergoat MC, Yeramian E. 2009. GC content and recombination: reassessing the causal effects for the *Saccharomyces cerevisiae* genome. Genetics 183:31–38. https://doi.org/10.1534/genetics.109.105049.
29. Myint M, Bashiri H, Harrington RD, Marra CM. 2004. Relapse of secondary syphilis after benzathine penicillin G. Sex Transm Dis 31:196–199. https://doi.org/10.1097/01.OLQ.0000114941.37942.4C.