Profile of the *tprK* gene in primary syphilis patients based on next-generation sequencing

Dan Liu[1,2¶], Man-Li Tong[1,2¶], Xi Luo[1¶],Li-Li Liu[1,2], Li-Rong Lin[1,2], Hui-Lin Zhang[1], Yong Lin[1], Jian-Jun Niu[1,3], and Tian-Ci Yang[1,2*]

[1] Zhongshan Hospital, Medical College of Xiamen University, Xiamen, China

[2] Institute of Infectious Disease, Medical College of Xiamen University, Xiamen, Fujian Province, China

[3] Zhongshan Hospital, Fujian Medical University, Xiamen, China

[*] Corresponding author: Tian-Ci Yang

E-mail: yangtianci@xmu.edu.cn

[¶]These authors contributed equally to this work.

Short title: Profile of the *tprK* gene

Abstract

Background

The highly variable *tprK* gene of *Treponema pallidum* has been acknowledged to be the cause of persistent infection. Previous studies mainly focused on the heterogeneity in *tprK* in propagated strains using a clone-based Sanger approach. Few studies have investigated *tprK* directly from clinical samples using deep sequencing.

Methods/Principal findings

We conducted a comprehensive analysis of 14 primary syphilis clinical isolates of *T. pallidum* via next-generation sequencing to gain better insight into the profile of *tprK* in primary syphilis patients. Our results based on primary syphilis clinical samples showed that there was a mixture of distinct sequences within each V region of *tprK*. Except for the predominant sequence for each region as previously reported using the clone-based Sanger approach, there were many minor variants of all strains that were mainly observed at a frequency of 1-5%. Interestingly, the identified distinct sequences within the regions were variable in length and differed only by 3 bp or multiples of 3 bp. In addition, amino acid sequence consistency within each region was found between the 14 strains. Among the regions, the sequence IASDGGAIKH in V1 and the sequence DVGHKKENAANVNGTVGA in V4 showed a high stability of inter-strain redundancy.

Conclusions

The seven V regions of the *tprK* gene in primary syphilis infection demonstrated high

37 diversity; they generally contained a high proportion sequence and numerous
38 low-frequency minor variants, most of which are far below the detection limit of
39 Sanger sequencing. The rampant variation in each region was regulated by a strict
40 gene conversion mechanism that maintained the length difference to 3 bp or multiples
41 of 3 bp. The highly stable sequence of inter-strain redundancy may indicate that the
42 sequences play a critical role in *T. pallidum* virulence. These highly stable peptides
43 are also likely to be potential targets for vaccine development.
44

Author summary

Variations in *tprK* have been acknowledged to be the major contributors to persistent *Treponema pallidum* infections. Previous studies were based on the clone-based Sanger approach, and most of them were performed in propagated strains using rabbits, which could not reflect the actual heterogeneous characteristics of *tprK in vivo*. In the present study, we employed next-generation sequencing (NGS) to explore the profile of *tprK* directly from 14 patients with primary syphilis. Our results showed a mixture of distinct sequences within each V region of *tprK* in these clinical samples. First, the length of identified distinct sequences within the region was variable, which differed by only 3 bp or multiples of 3 bp. Then, among the mixtures, a predominant sequence was usually observed for each region, and the remaining minor variants were mainly observed at a frequency of 1-5%. In addition, there was a scenario of amino acid sequence consistency within the regions between the 14 primary syphilis strains. The identification of the profile of *tprK* in the context of human primary syphilis infection contributes to further exploration of the pathogenesis of syphilis.

Introduction

Syphilis, caused by *Treponema pallidum*, is an ancient sexually transmitted disease that dates back to the 15th century and is a public health threat that cannot be neglected [1, 2]. The completion of the first whole genome sequencing of the Nichols strain of *T. pallidum* provides a wealth of information about the characteristics of this pathogen, and since then the sequence of other experimental treponemal strains has also been released [3-8]. These particular achievements have revealed slight variations between different strains in a small genome (~1.1 Mb), and most of the genetic diversity occurs in six genomic regions, including a polymorphic multigene family encoding 12 paralogous proteins (*tpr A* through *tprL*), highlighting most likely a factor in the pathogenesis of *T. pallidum* [2, 6, 9].

Within the *tpr* family, the antigen-coding *tprK* has been found to be the direct target of the human immune response [10]. Several remarkable studies performed in the rabbit model have demonstrated that the *tprK* gene possesses high genetic diversity at both the intra- and inter-strain levels, and the genetic variation in *tprK* is localized to seven variable regions (V1-V7) flanked by highly conserved domains [11-13]. Theoretically, through gene conversion, variations in the V regions would generate millions of chimeric *tprK* variants, resulting in a constant alteration in the *T. pallidum* antigenic profile [14]. Therefore, the *tprK* gene is acknowledged to have a pivotal role in immune evasion and pathogen persistence [15-17].

Previous studies have been mainly based on the clone-based Sanger approach; when using this approach, one would inevitably encounter a bottleneck in clone

82 selection where minor variants, especially at low frequencies, are lost; consequently,
83 the complete mutation profile of *tprK* is not fully understood. In addition, few studies
84 have explored how *tprK* diversifies in the context of human infection, thus reflecting
85 the actual heterogeneous characteristics of *tprK in vivo*, with the exception of one
86 recent publication that reported on whole genome sequencing directly from clinical
87 samples of *T. pallidum* [18]. Research has shown that after being cultured in rabbits,
88 the genes of *T. pallidum* mutate and cannot retain their naïve characteristics, let alone
89 the *tprK* gene, which has rampant potential to vary [19].
90 In the present study, we seek to systematically reveal the profile of *tprK* in *T.*
91 *pallidum* directly from patients with primary syphilis by employing next-generation
92 sequencing (NGS), thus providing important insights into the understanding of the
93 diversity of *tprK* directly from primary syphilis patients and contributing to further
94 explorations of the mechanisms of long-term *T. pallidum* infection.

Methods

Ethics statement

Written consent was obtained with signatures from all patients in accordance with institutional guidelines prior to the study. The study was approved by the Ethics Committee of Zhongshan Hospital, Xiamen University, after a formal hearing and was in conformance with the Declaration of Helsinki.

Sample collection

Swab samples were obtained from the skin lesions of 14 patients (X-1~14) with primary syphilis. The clinical diagnosis of syphilis was based on the US Centers for Disease Control and Prevention (CDC) [20] and the European CDC (ECDC) [21].

Isolation of DNA

Treponemal DNA was extracted from the swab samples using the QIAamp DNA Mini Kit (Qiagen, Inc., Valencia, CA, USA) according to the manufacturer's instructions, and careful precautions were implemented to avoid DNA cross-contamination between isolates [22]. Each sample was quantified by targeting *tp0574* through qPCR using a 96-well reaction plate with a ViiA 7 Real-Time PCR System (Applied Biosystems, USA). A standard curve was constructed using 10-fold serial dilutions of cloned plasmids (for *tp0574*) generated through TOPO TA technology (Invitrogen, Carlsbad, CA, USA) and transformation of DH5α *Escherichia coli* cells [23]. The DNA samples that tested positive were stored at -20°C for further processing.

Segmented amplification of the *tprK* gene

First, the extracted DNA was directly used in the amplification of the *tprK* full open reading frame (ORF). The primers used for the amplification are listed in Table 1. For amplification, KOD FX Neo polymerase (Toyobo, Osaka, Japan) was used. The reaction mixture contained 25 μL of 2× PCR buffer, 0.4 mM deoxynucleoside triphosphates, 0.3 μM of each primer, 1 U of KOD FX Neo polymerase, and 5 μL of genomic DNA in a final volume of 50 μL. The cycling conditions were as follows: 94°C for 2 min, followed by 40 cycles of 98°C for 10 s, 60°C for 30 s, and 68°C for 30 s. Then, the amplicons were gel purified and stored at -20°C for further processing as the next segmented amplification template.

Second, the *tprK* ORF was separated into four fragments overlapping by at least 20 bp (approximately 400~500 bp) for amplification. The primers are listed in Table 1. The purified DNA was diluted 1000 times and used as a template. The amplification mixture was the same as described above except that the primers were 0.15 μM. The cycling conditions were denaturation at 94°C for 2 min, followed by 30 cycles of 98°C for 10 s, 55°C for 30 s, and 68°C for 30 s. The size of all the products was verified by 2% agarose gel electrophoresis, and the products were gel purified. All purified amplicons were stored at -20°C for further processing.

134 **Table 1. The primers for *tprK* amplification and sequencing**

Primer	Purpose	Sequence (5'→3')
tprK-S	Amplification of *tprK* ORF	ACCGGGCATGAATTTTCTTT
tprK-As		GTAGGCCCCATAACAGTGCA
tprK-frag1-S	Amplification of *tprK* fragment1	ATGATTGACCCATCTGCCAC
tprK-frag1-As		GTAGGCCCCATAACAGTGCA
tprK-frag2-S	Amplification of *tprK* fragment2	GGTGGAGCAAAGTTTGACAC
tprK-frag2-As		TTAATGTATTCCTGCACGCC
tprK-frag3-S	Amplification of *tprK* fragment3	GAAGATGGCGTGCAGGAATA
tprK-frag3-As		TCAACACCCAAATCAAGACC
tprK-frag4-S	Amplification of *tprK* fragment4	TATTAAGCTCGAAACCAAGG
tprK-frag4-As		CCAAATCAAGCGACATGCCC
M13forward	Sequencing	CTGGCCGTCGTTTTAC
M13 reverse		CAGGAAACAGCTATGAC

135

Library construction and next-generation sequencing

Library construction and sequencing were performed by the Sangon Biotech Company (Shanghai, China) on the MiSeq platform (Illumina, San Diego, CA, USA) in paired-end bi-directional sequencing (2×300 bp) mode. FastQC (http://www.bioinformatics.babraham.ac.uk/project/fatsqc/) and FASTX (http://hannonlab.cshl.edy/fastx_toolkit) tools were applied to check and improve the quality of the raw sequence data, respectively. The final reads collected from 14 patients were compared with the *tprK* of the Seattle Nichols strain (GenBank accession number AF194369.1) using Bowtie 2 (version 2.1.0).

An in-house Perl script was developed and applied to specifically capture DNA sequences within seven regions of 14 strains from raw data, both forward and reverse, as previously reported [18]. Briefly, the user-defined strings that matched the conserved sequence flanking the variable regions were used to catch the variable sequences. The defined strings referred to the mapping result of the reference and should be as long as necessary to ensure specificity (approximately 12-16 bp). Thus, the exact number of distinct sequences within seven regions across all strains was acquired. The intrastrain heterogeneous sequences were valid if the following conditions were simultaneously supported: 1) the number of the captured sequence was at least fifty reads and 2) the less frequent sequence displayed a frequency above 1%. Then, the relative frequency of the sequences within each variable region was calculated.

***TprK* analysis by clone-based Sanger sequencing**

An aliquot of DNA was also used for the amplification of the *tprK* full ORF according to the procedure described previously [11]. The purified amplicons were cloned into the pCR-2.1 TOPO vector (Invitrogen, Carlsbad, CA, USA) and were used to transform TOP10 competent *Escherichia coli* according to the manufacturer's instructions. Approximately 10 clone plasmids from each sample were randomly selected and sequenced; each clone was

sequenced not only in both directions with the M13 forward and reverse primers but also in the middle with the appropriate primers for a third reaction to ensure accuracy (Table 1). All sequencing was accomplished by the Bioray Biotechnology Company (Xiamen, China). The sequences within each intrastrain variable region were analysed using the BioEdit Sequence Alignment Editor Program (www.mbio.ncsu.edu/BioEdit/bioedit.html).

Results

1. Description of clinical samples and *tprK* sequencing by NGS

The samples (N=14) were collected from patients diagnosed with primary syphilis at Zhongshan Hospital, Xiamen University. The clinical data of patients are shown in Table 2. The qPCR data of *Tp0574* showed that the number of treponemal copies in each clinical sample was eligible for the amplification of the *tprK* full ORF. The median sequencing depth of the *tprK* segment samples ranged from 10568.99 to 56676.38 and the coverage ranged from 99.34% to 99.61%, showing high homogeneity with the *tprK* gene of the Seattle Nichols strain.

Table 2. Description of clinical samples and *tprK* sequencing by NGS

Isolate	Gender	Age (year)	Serum RPR titer	Serum TPPA	Dark field microscopy	*T. pallidum* genome copies by *Tp0574*	Total reads	On-target reads (%)	Mean coverage of depth
X-1	Male	45	1:16	+	Positive	8.2E+03	357382	99.41	51967.28
X-2	Male	27	1:16	+	Positive	8.82E+04	340240	99.47	49660.18
X-3	Male	62	1:16	+	Positive	4.55E+04	398898	99.41	56676.38
X-4	Male	65	1:4	+	Positive	1.15E+04	365060	99.34	52742.09
X-5	Male	76	1:16	+	Positive	5.73E+04	363940	99.61	52960.83
X-6	Male	64	1:32	+	Positive	2.33E+02	106934	99.37	14249.15
X-7	Female	56	1:16	+	Positive	1.26E+04	114012	99.37	15579.12
X-8	Male	46	1:4	+	Positive	1.41E+04	103280	99.43	12951.11
X-9	Male	40	1:4	+	Positive	1.39E+03	119552	99.43	15864.28
X-10	Male	66	1:32	+	Positive	9.17E+03	114064	99.37	14927.08
X-11	Male	44	1:2	+	Positive	2.67E+02	94572	99.50	12935.89
X-12	Male	39	-	+	Positive	6.40E+03	114588	99.43	14944.66
X-13	Male	63	1:16	+	Positive	2.02E+02	118634	99.37	15013.54
X-14	Male	61	1:1	+	Positive	1.16E+03	82812	99.37	10568.99

Abbreviations: NGS, next generation sequencing; RPR, reactive plasma reagin; TPPA, *T. pallidum* particle agglutination; +, positive;-, negative.

198

199 **2. Sequence variability of *tprK* directly from primary syphilis samples**

200 **The number and length variation of distinct sequences in seven regions** According to the
201 strategy, we extracted sequences within seven V regions to evaluate the sequence variability
202 of *tprK* directly from primary syphilis samples. Altogether, 335 distinct nucleotide sequences
203 were captured. The number of distinct sequences in the seven regions ranged from 21-76,
204 with the highest number in V6 and the lowest in V1 across all samples (Fig 1). The length of
205 the captured sequences within each region was also found to be variable, particularly in V3,
206 V6 and V7, with 11 or 12 forms. In contrast, the length of the sequence in V5 had only two
207 forms, namely, 84 bp and 90 bp. When the length of all sequences within each sample was
208 calculated, the length of all distinct sequences differed by 3 bp or multiples of 3 bp.
209 Interestingly, although the lengths of V3, V6 and V7 were particularly variable across all
210 populations, these lengths continued to change by 3 bp. In this regard, the lengths of V1, V4
211 and V5 appeared to vary in intervals of 6 bp.

212 **The proportion distribution of distinct sequences in seven regions** The captured
213 sequences were ranked by relative frequency within each V region of each strain. As Fig 2a
214 shows, there was a predominant sequence in each region of all samples directly from primary
215 syphilis patients, and the proportion of this sequence was almost 80%. It is worth noting that
216 the frequency of the predominant sequence in some V regions of 4 samples (X-6, 8, 10, 13)
217 was lower than 60%. In total, the frequency decrease appeared in the V2, V5, V6 and V7
218 regions, and the frequency in V6 of X-6 was even lower at 20.8%.

219 Apart from the detected predominant sequence within seven V regions, there was still a
220 mixture of minor variants in each region. To investigate the relative frequency distribution of
221 minor variants, we used three thresholds to explore the characteristics (Fig 2b). The major
222 proportion of the variants in primary syphilis samples was in the 1-5% (181/237) range, and

the lowest was in the 10-60% (22/237) range. At the two thresholds (5-10% and 10-60%), the observed variants were all mainly in V2, V5, V6 and V7 and from 4 samples (X-6, 8, 10, 13). This corresponded to the lower proportion of their predominant sequences.

Fig 1. The varied length forms of distinct sequences within each region of *tprK*. The varied length forms within each V region are presented as the frequencies in each region and are filled with the gradient colour. All distinct sequences captured for each region are also shown above the V region.

Fig 2. The proportion distribution of distinct sequences within each V region of *tprK*. (A) The dots indicate the relative frequency of identified distinct sequences within each V region of *tprK* across all 14 primary clinical samples, and the colour specifies the strain. (B) The graph shows the number of minor variants within each V region from all strains. The three thresholds (1-5%, 5-10% and 10-60%) are characterized by three different shapes, and the colour specifies the strain.

3. Comparison with the heterogeneity of the clones within the population

Because previous studies were mainly based on the clone-based Sanger approach, we also applied this approach to analyse the *tprK* gene in these 14 strains and then compared the results with those of the NGS. Ten clones of each sample were obtained and identified by Sanger sequencing. Among the ten sequences, the predominant sequence within each V region of the primary syphilis samples was observed. The observed predominant sequence was consistent with the sequences obtained by NGS, such as in the strain of X-2 (Fig 3). However, where the frequency of the predominant sequence declined, especially when the frequency was less than 30%, for example, in the case of V6 in X-8, it became too ambiguous

to distinguish the predominant sequences (S1 Fig). In all clones, the sequence was nearly undetectable for the minor variants with a frequency of 1-5%.

Fig 3. Predominant sequences within seven V regions identified by NGS compared to the results obtained by clone-based Sanger sequencing in this study. The alignment was performed on the X-2 strain as a representative sample. The identical nucleotides are shown on dots, and gaps in the sequence are shown by dashes.

4. Inter-population redundancy of the deducted amino acid sequence

A total of 335 nucleotide sequences were translated into amino acid sequences *in silico*. Ten sequences (10/335) were found to be synonymous, and at least 325 unique amino acid sequences were obtained. Unexpectedly, no sequence yielded a *tprK* frame shift or premature termination. When distinct sequences within each V region of all strains were compared, a scenario of sequence consistency was found. As Fig 4 shows, V1 and V4 presented a strong shared sequence capacity. The sequence IASDGGAIKH in V1 was observed in five strains (5/14) and DVGHKKENAANVNGTVGA in V4 was shared across seven strains (7/14). However, the parallel sequences in V3 and V6 did not seem as significant as in other V regions, especially in V6.

To further explore whether the shared scenario was usually displayed by the predominant sequence across all the strains, we involved only the predominant sequence in the V region of each sample, which was represented by the bold arc in Fig 4, and found that V1 and V4 still presented similar shared sequence abilities despite the decreased redundant sequences. The occurrence of the consistent sequence in V1 and V4 could reach five strains and six strains, respectively. For the V3 and V6 regions, which were rarely consistent with sequences, the shared sequence in V3 occurred only between two strains, and there was no consistent

274 sequence found in V6. Meanwhile, there was also no redundant sequence observed in V7.

275

276 **Fig 4. The scenario of redundant *tprK* amino acid sequences between all 14 primary**
277 **syphilis clinical samples.** The 14 strains are specified by coloured solid circles, and the
278 predominant sequence and minor variants within each V region of one strain are represented
279 by a bold arc and thin arcs, respectively. Each grey circle indicates the occurrence of
280 sequence consistency between the strains.

281

282 ## Discussion

283 Due to the inability to long term culture *T. pallidum in vitro*, research on this pathogen has
284 been greatly hindered. The whole genome sequencing of the Nichols strain of *T. pallidum*
285 provides a new perspective for the study of treponemal genes and proteins. Among these
286 genes, *tprK* has been extensively studied because of its highly variable antigenic profile. It
287 could effectively compromise the immunological function of specific antibodies generated by
288 the host [14, 24-26] and cause immune evasion, thereby further leading to the development of
289 late syphilis, neurosyphilis or serofast. Hence, intensive studies on the heterogeneity of *tprK,*
290 especially in the context of human infection, would contribute to a deep understanding of the
291 pathogenesis of syphilis.

292 In the present study, we performed NGS, a more sensitive and reliable approach, to gain
293 better insight into the profile of *tprK* in primary syphilis patients. Overall, there was a wide
294 sequence mixture focused on seven V regions of *tprK* in primary syphilis clinical samples.
295 Among the seven V regions, V1 and V6 were known to have the lowest and highest
296 variability, respectively [18, 27]. Our results also corroborated this feature in primary syphilis
297 infection, in which the highest distinct number was found in V6 and the lowest distinct
298 number was found in V1 (Figs 1 and 2a). Although *tprK* was revealed to have rampant

genetic diversity within each strain [11, 12, 28], little is known about the exact proportion of these variant sequences within one strain. It is an advantage of NGS to fully discover the variants and determine the frequency [29, 30]. In this study, by using an in-house Perl script, we were able to retrieve the variants within the regions of each strain and calculate the relative frequency of the variants, thus disclosing the proportion of these variant sequences in primary syphilis patients. As shown in Fig 2a, there was a predominant sequence (the proportion above 80%) within each V region across all the strains.

In addition to the predominant sequence within each region, there was also a mixture of minor sequences (Fig 2a). Moreover, these minor variants were found to be mostly distributed at a frequency of 1-5% (Fig 2b), which was extremely below the detection limit for Sanger sequencing [31]. These results demonstrated that although the diversity of the V regions in antigen-coding *tprK* in primary syphilis infection was also presented to be wild, the V regions generally maintained their high proportion pathogenic sequence and numerous low-frequency minor variants. It is worth noting that the proportion of predominant sequences in some V regions of 4 samples (X-6, 8, 10, 13) was apparently lower than in others, and almost all the minor sequences in the 5-10% and 10-60% ranges were from these four samples, more specifically, mostly from the V2, V5, V6 and V7 regions (Fig 2). This result suggested that with the progression of disease or with increasing immunity, some V regions (V2, V5, V6 and V7) began to change. As a result, the frequency of the predominant sequence was decentralized. Instead, the frequency of a minor variant (or a new variant) gradually increased and further promoted the genetic diversity of *tprK* to escape immune clearance. Additionally, among the observed four regions, the frequency of the predominant sequence in V6 was particularly low (Fig 2a), suggesting that V6 may be the first affected region and is involved in immune evasion during the course of infection [14, 18].

We also applied the clone-based Sanger approach to analyse the *tprK* in primary syphilis

patients in comparison with the results of NGS. As described in a previous study [32], the Sanger results generally displayed the predominant sequence within each region, which was consistent with the sequence found during NGS (Fig 3). However, for the lower frequency variants within the region, it became difficult to distinguish the predominant sequence, and we were unable to identify all the minor variants (S1 Fig). An increase in the number of clones selected would potentially alleviate this problem, but it would take more time and money. Additionally, the minor variants at a frequency of 1-5% were nearly undetectable in all selected clones. Therefore, use of the clone-based Sanger approach would lose much information about the complete profile of *tprK*, particularly in primary syphilis clinical samples in which *tprK* contained numerous low-frequency variants.

In this study, except for the distinct variations in *tprK* sequences, we found that the heterogeneity in *tprK* also presented in length (Fig 1). More length forms appeared in V3, V6 and V7, which was similar to the findings of Pinto et al. [18], demonstrating that the variations in these three regions could more easily cause changes in length. Despite this, the length forms were too far away to match with the number of sequences within each region; that is, the variants within each region were of the same length, but the context still had a considerable difference. For example, there were many different sequences observed in V5, but there were only two forms of length which was also observed in previous study [14]. This result indicated that the variants of *tprK* preferred a conversion without changing the initial length of the V region. Additionally, it was interesting that the length of all distinct sequences differed only by 3 bp or multiples of 3 bp, and previous research data also supported this pattern change [14, 18]. A multiple of 3 bp change pattern just matched with the triplet codon in protein coding, which has made us think about this feature probably explain why it is rare to uncover a *tprK* frame shift. It also suggests that the rampant diversity of *tprK* is regulated by a strict gene conversion mechanism.

Another noteworthy finding was the shared amino acid sequences across all the strains from the primary syphilis patients, which has also been observed in previous research [18, 27]. In our study, when all the distinct amino acid sequences within each region were aligned, at least half of the strains had sequences shared by other strains (Fig 4), which was similar to previous findings [18]. However, when only the predominant sequence within each region was analysed, *tprK* inter-population redundancy remained at a high level in only V1 and V4, in contrast to other regions, especially V6 and V7. This result suggested that the redundant sequences in V1 and V4 between strains were the ones that mostly dominated within a single strain. As the same antigenic sequences originated throughout the evolution of *T. pallidum* in different patients, this may reflect that the sequences become the best antigenic profiles to address the immune response of the host. The high stability of inter-population redundancy in V1 and V4 found in primary syphilis may confirm that the shared antigenic sequences in V1 and V4 play a critical role in *T. pallidum* virulence. In previous research [24, 25, 33], the molecular localization in the N-terminal region of *tprK* displayed promising partial protection in a rabbit model. Therefore, the highly stable shared peptide of V1 and V4 across all the strains would also likely be a potential target for vaccine development.

Finally, the limitations of our research should be discussed. First, the findings reported above lacked data about these clinical strains propagated by rabbits and could not directly highlight the difference from the naïve characteristics of *tprK* in human infection. This remains to be confirmed by animal experiments with NGS in the future. In addition, in this study, the number of clones selected for Sanger sequencing might be insufficient, although the current data were enough to verify the accuracy and reliability of our NGS results.

In summary, the characteristic profile of *tprK* in primary syphilis patients was unveiled to generally contain a high proportion sequence and many low-frequency minor variants within each V region. The variations in V regions were regulated by a strict gene conversion

374 mechanism to keep the length differences to 3 bp or multiples of 3 bp. The findings could
375 provide important information for further exploration of the role of *tprK* in immune evasion
376 and persistent infection with syphilis. Furthermore, the peptide of each V region, especially
377 the highly conserved peptide found in this study, could serve as a database of B cell epitopes
378 of TprK for human immunological studies in the future.
379

380 ## Supporting information

381 **S1 Fig. Comparison of the results of NGS and clone-based Sanger sequencing in V6 of**
382 **the X-8 strain.** RF values indicate the relative frequency of each sequence.
383 **S1 Table. The nucleotide sequences within the seven variable regions (V1-V7) of *tprK***
384 **captured directly from 14 primary syphilis clinical samples.**
385 **S2 Table. The amino acid sequences within the seven variable regions (V1-V7) of *tprK***
386 **captured directly from 14 primary syphilis clinical samples.** * indicates synonymous
387 nucleotide sequences within the same strain.
388

389 ## References

390 1. Smolak A, Rowley J, Nagelkerke N, Kassebaum NJ, Chico RM, Korenromp EL, et al.
391 Trends and Predictors of Syphilis Prevalence in the General Population: Global Pooled
392 Analyses of 1103 Prevalence Measures Including 136 Million Syphilis Tests. Clinical
393 infectious diseases : an official publication of the Infectious Diseases Society of America.
394 2018;66(8):1184-91. Epub 2017/11/15. doi: 10.1093/cid/cix975. PubMed PMID: 29136161;
395 PubMed Central PMCID: PMCPmc5888928.
396 2. Everall I, Sanchez-Buso L. Bringing Treponema into the spotlight. Nature reviews
397 Microbiology. 2017;15(4):196. Epub 2017/03/14. doi: 10.1038/nrmicro.2017.23. PubMed
398 PMID: 28286342.

399 3. Fraser CM, Norris SJ, Weinstock GM, White O, Sutton GG, Dodson R, et al. Complete
400 genome sequence of Treponema pallidum, the syphilis spirochete. Science (New York, NY).
401 1998;281(5375):375-88. Epub 1998/07/17. PubMed PMID: 9665876.
402 4. Matejkova P, Strouhal M, Smajs D, Norris SJ, Palzkill T, Petrosino JF, et al. Complete
403 genome sequence of Treponema pallidum ssp pallidum strain SS14 determined with
404 oligonucleotide arrays. Bmc Microbiology. 2008;8. PubMed PMID:
405 WOS:000256297900001.
406 5. Smajs D, Zobanikova M, Strouhal M, Cejkova D, Dugan-Rocha S, Pospisilova P, et al.
407 Complete Genome Sequence of Treponema paraluiscuniculi, Strain Cuniculi A: The Loss of
408 Infectivity to Humans Is Associated with Genome Decay. PloS one. 2011;6(5). PubMed
409 PMID: WOS:000291097600063.
410 6. Cejkova D, Zobanikova M, Chen L, Pospisilova P, Strouhal M, Qin X, et al. Whole
411 Genome Sequences of Three Treponema pallidum ssp pertenue Strains: Yaws and Syphilis
412 Treponemes Differ in Less than 0.2% of the Genome Sequence. Plos Neglected Tropical
413 Diseases. 2012;6(1). PubMed PMID: WOS:000300416100021.
414 7. Petrosova H, Zobanikova M, Cejkova D, Mikalova L, Pospisilova P, Strouhal M, et al.
415 Whole genome sequence of Treponema pallidum ssp. pallidum, strain Mexico A, suggests
416 recombination between yaws and syphilis strains. PLoS Negl Trop Dis. 2012;6(9):e1832.
417 Epub 2012/10/03. doi: 10.1371/journal.pntd.0001832. PubMed PMID: 23029591; PubMed
418 Central PMCID: PMCPmc3447947.
419 8. Zobanikova M, Mikolka P, Cejkova D, Pospisilova P, Chen L, Strouhal M, et al.
420 Complete genome sequence of Treponema pallidum strain DAL-1. Standards in Genomic
421 Sciences. 2012;7(1):12-21. PubMed PMID: WOS:000314405900002.
422 9. Mikalova L, Strouhal M, Cejkova D, Zobanikova M, Pospisilova P, Norris SJ, et al.
423 Genome Analysis of Treponema pallidum subsp. pallidum and subsp. pertenue Strains: Most

424 of the Genetic Differences Are Localized in Six Regions. PloS one. 2010;5(12). PubMed
425 PMID: WOS:000285793200041.
426 10. Morgan CA, Molini BJ, Lukehart SA, Van Voorhis WC. Segregation of B and T cell
427 epitopes of Treponema pallidum repeat protein K to variable and conserved regions during
428 experimental syphilis infection. J Immunol. 2002;169(2):952-7. PubMed PMID:
429 WOS:000176753500040.
430 11. Centurion-Lara A, Godornes C, Castro C, Van Voorhis WC, Lukehart SA. The tprK
431 gene is heterogeneous among Treponema pallidum strains and has multiple alleles. Infect
432 Immun. 2000;68(2):824-31. Epub 2000/01/20. PubMed PMID: 10639452; PubMed Central
433 PMCID: PMCPmc97211.
434 12. Stamm LV, Bergen HL. The sequence-variable, single-copy tprK gene of Treponema
435 pallidum Nichols strain UNC and Street strain 14 encodes heterogeneous TprK proteins.
436 Infection and Immunity. 2000;68(11):6482-6. PubMed PMID: WOS:000090007000055.
437 13. Giacani L, Brandt SL, Puray-Chavez M, Reid TB, Godornes C, Molini BJ, et al.
438 Comparative Investigation of the Genomic Regions Involved in Antigenic Variation of the
439 TprK Antigen among Treponemal Species, Subspecies, and Strains. Journal of Bacteriology.
440 2012;194(16):4208-25. PubMed PMID: WOS:000307198100007.
441 14. Centurion-Lara A, LaFond RE, Hevner K, Godornes C, Molini BJ, Van Voorhis WC, et
442 al. Gene conversion: a mechanism for generation of heterogeneity in the tprK gene of
443 Treponema pallidum during infection. Molecular Microbiology. 2004;52(6):1579-96.
444 PubMed PMID: WOS:000221866300005.
445 15. Cox DL, Luthra A, Dunham-Ems S, Desrosiers DC, Salazar JC, Caimano MJ, et al.
446 Surface immunolabeling and consensus computational framework to identify candidate rare
447 outer membrane proteins of Treponema pallidum. Infect Immun. 2010;78(12):5178-94. doi:
448 10.1128/IAI.00834-10. PubMed PMID: 20876295; PubMed Central PMCID:

449 PMCPMC2981305.

450 16. Reid TB, Molini BJ, Fernandez MC, Lukehart SA. Antigenic Variation of TprK
451 Facilitates Development of Secondary Syphilis. Infection and Immunity.
452 2014;82(12):4959-67. PubMed PMID: WOS:000346958400006.

453 17. Radolf JD, Deka RK, Anand A, Smajs D, Norgard MV, Yang XF. Treponema pallidum,
454 the syphilis spirochete: making a living as a stealth pathogen. Nature Reviews Microbiology.
455 2016;14(12):744-59. PubMed PMID: WOS:000388217400008.

456 18. Pinto M, Borges V, Antelo M, Pinheiro M, Nunes A, Azevedo J, et al. Genome-scale
457 analysis of the non-cultivable Treponema pallidum reveals extensive within-patient genetic
458 variation. Nature Microbiology. 2017;2(1). doi: 10.1038/nmicrobiol.2016.190. PubMed
459 PMID: WOS:000396366300010.

460 19. Giacani L, Jeffrey BM, Molini BJ, Le HT, Lukehart SA, Centurion-Lara A, et al.
461 Complete genome sequence and annotation of the Treponema pallidum subsp. pallidum
462 Chicago strain. J Bacteriol. 2010;192(10):2645-6. Epub 2010/03/30. doi:
463 10.1128/jb.00159-10. PubMed PMID: 20348263; PubMed Central PMCID:
464 PMCPmc2863575.

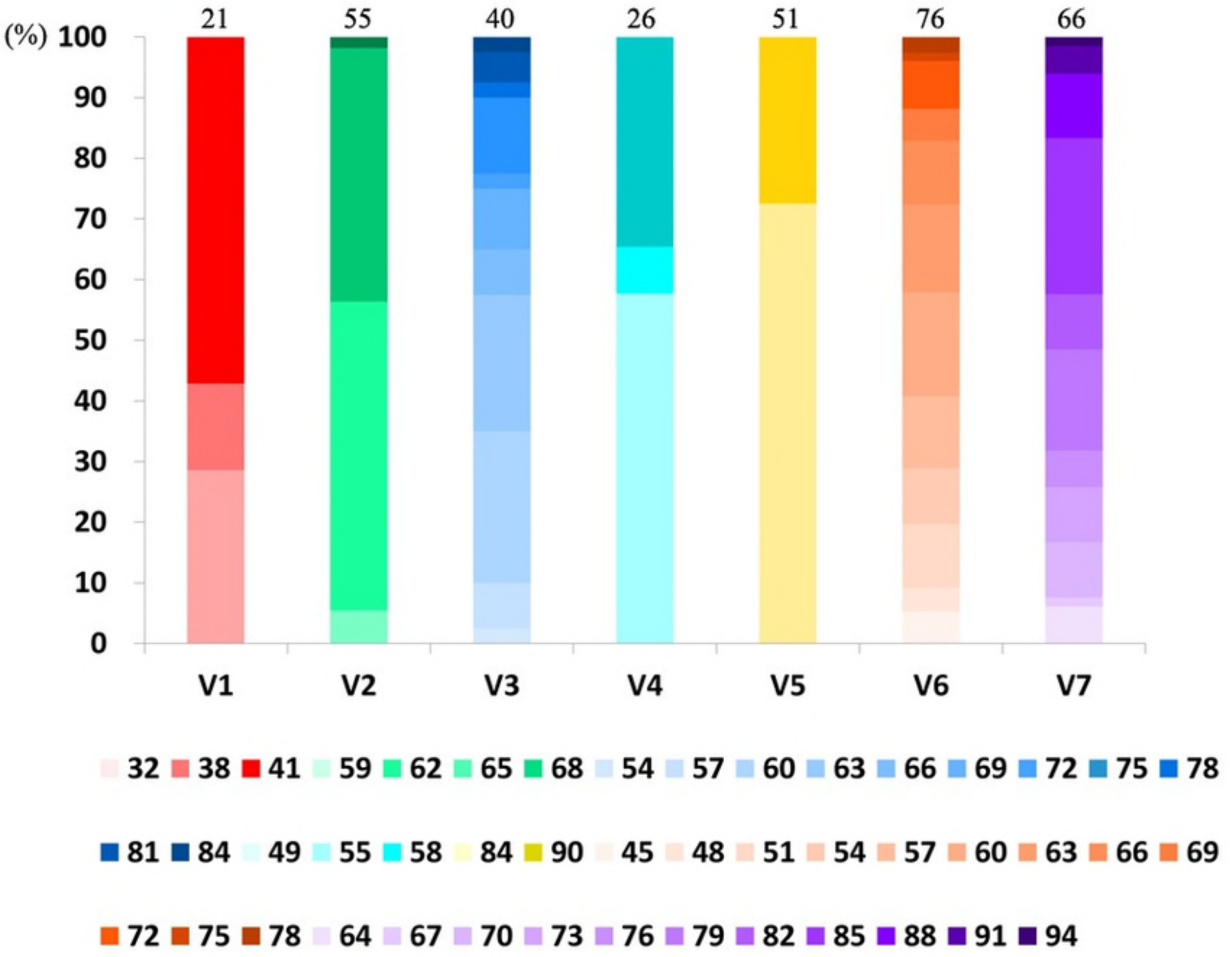
465 20. Workowski KA, Bolan GA. Sexually Transmitted Diseases Treatment Guidelines, 2015.
466 MMWR Recommendations and reports : Morbidity and mortality weekly report
467 Recommendations and reports. 2015;64(RR-03):1-137. PubMed PMID: PMC5885289.

468 21. Janier M, Hegyi V, Dupin N, Unemo M, Tiplica GS, Potočnik M, et al. 2014 European
469 guideline on the management of syphilis. Journal of the European Academy of Dermatology
470 and Venereology. 2014;28(12):1581-93. doi: doi:10.1111/jdv.12734.

471 22. Tong ML, Zhao Q, Liu LL, Zhu XZ, Gao K, Zhang HL, et al. Whole genome sequence
472 of the Treponema pallidum subsp. pallidum strain Amoy: An Asian isolate highly similar to
473 SS14. 2017;12(8):e0182768. doi: 10.1371/journal.pone.0182768. PubMed PMID: 28787460.

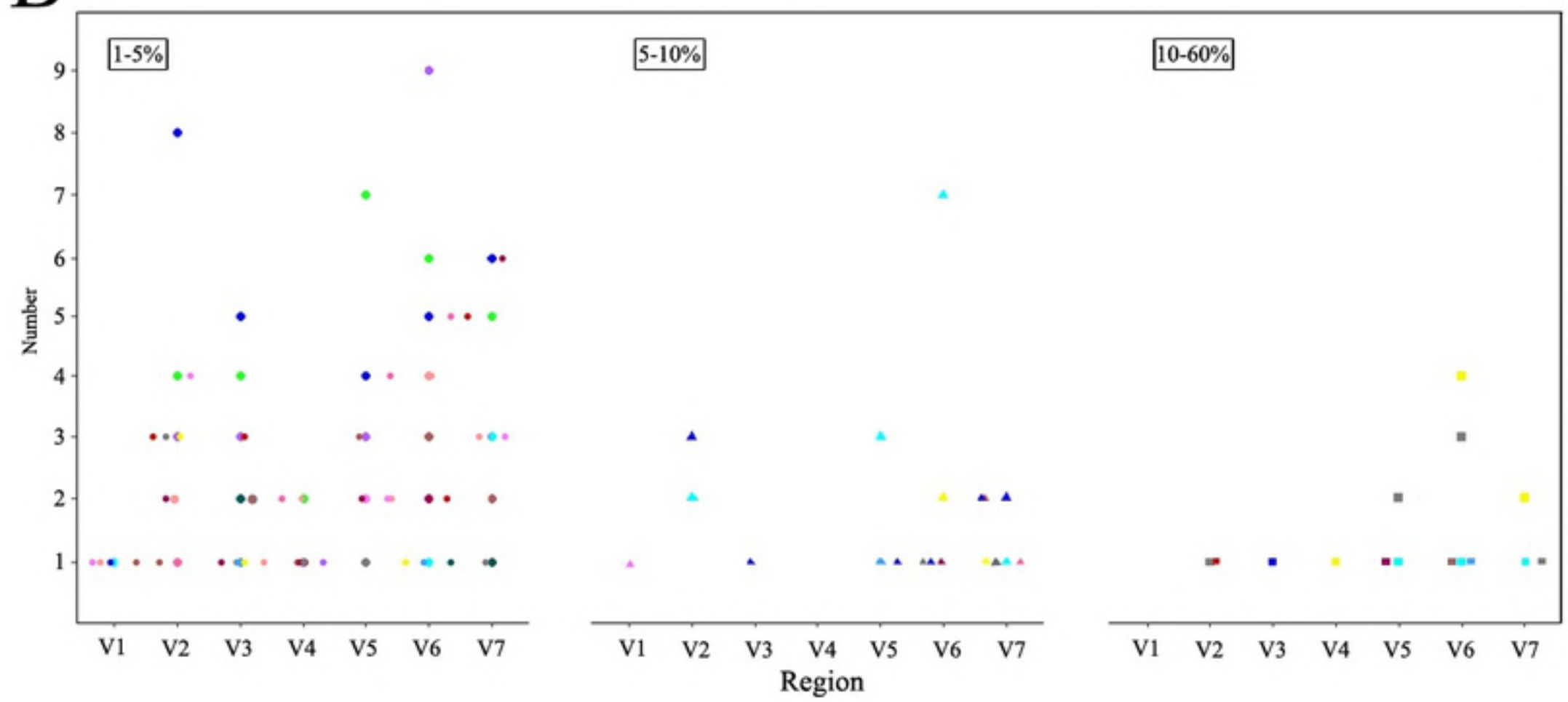
474 23. Zhu XZ, Fan JY, Liu D, Gao ZX, Gao K, Lin Y, et al. Assessing effects of different
475 processing procedures on the yield of Treponema pallidum DNA from blood. Analytical
476 biochemistry. 2018;557:91-6. Epub 2018/07/25. doi: 10.1016/j.ab.2018.07.019. PubMed
477 PMID: 30040912.
478 24. Centurion-Lara A, Castro C, Barrett L, Cameron C, Mostowfi M, Van Voorhis WC, et al.
479 Treponema pallidum major sheath protein homologue Tpr K is a target of opsonic antibody
480 and the protective immune response. J Exp Med. 1999;189(4):647-56. Epub 1999/02/17.
481 PubMed PMID: 9989979; PubMed Central PMCID: PMCPmc2192927.
482 25. Morgan CA, Lukehart SA, Van Voorhis WC. Protection against syphilis correlates with
483 specificity of antibodies to the variable regions of Treponema pallidum repeat protein K.
484 Infect Immun. 2003;71(10):5605-12. doi: 10.1128/iai.71.10.5605-5612.2003. PubMed PMID:
485 WOS:000185551200020.
486 26. LaFond RE, Molini BJ, Van Voorhis WC, Lukehart SA. Antigenic variation of TprK V
487 regions abrogates specific antibody binding in syphilis. Infection and Immunity.
488 2006;74(11):6244-51. PubMed PMID: WOS:000241600500025.
489 27. LaFond RE, Centurion-Lara A, Godornes C, Rompalo AM, Van Voorhis WC, Lukehart
490 SA. Sequence diversity of Treponema pallidum subsp pallidum tprK in human syphilis
491 lesions and rabbit-propagated isolates. Journal of Bacteriology. 2003;185(21):6262-8. doi:
492 10.1128/jb.185.21.6262-6268.2003. PubMed PMID: WOS:000186037600005.
493 28. LaFond RE, Centurion-Lara A, Godornes C, Van Voorhis WC, Lukehart SA. TprK
494 sequence diversity accumulates during infection of rabbits with Treponema pallidum subsp
495 pallidum Nichols straint. Infect Immun. 2006;74(3):1896-906. doi:
496 10.1128/iai.74.3.1896-1906.2006. PubMed PMID: WOS:000235817500052.
497 29. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, et al. DNA sequencing of
498 a cytogenetically normal acute myeloid leukaemia genome. Nature. 2008;456(7218):66-72.

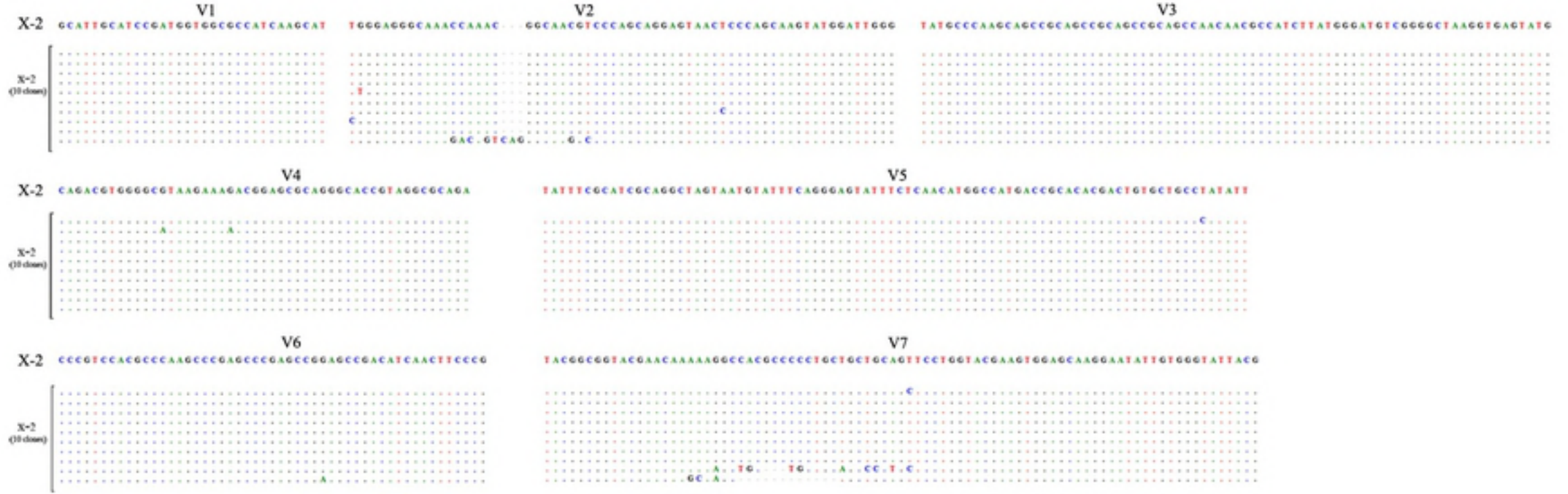
499 Epub 2008/11/07. doi: 10.1038/nature07485. PubMed PMID: 18987736; PubMed Central
500 PMCID: PMCPmc2603574.
501 30. Solmone M, Vincenti D, Prosperi MC, Bruselles A, Ippolito G, Capobianchi MR. Use of
502 massively parallel ultradeep pyrosequencing to characterize the genetic diversity of hepatitis
503 B virus in drug-resistant and drug-naive patients and to detect minor variants in reverse
504 transcriptase and hepatitis B S antigen. Journal of virology. 2009;83(4):1718-26. Epub
505 2008/12/17. doi: 10.1128/jvi.02011-08. PubMed PMID: 19073746; PubMed Central PMCID:
506 PMCPmc2643754.
507 31. Palmer S, Kearney M, Maldarelli F, Halvas EK, Bixby CJ, Bazmi H, et al. Multiple,
508 linked human immunodeficiency virus type 1 drug resistance mutations in
509 treatment-experienced patients are missed by standard genotype analysis. Journal of clinical
510 microbiology. 2005;43(1):406-13. Epub 2005/01/07. doi: 10.1128/jcm.43.1.406-413.2005.
511 PubMed PMID: 15635002; PubMed Central PMCID: PMCPmc540111.
512 32. Pinto M, Borges V, Antelo M, Pinheiro M, Nunes A, Azevedo J, et al. Genome-scale
513 analysis of the non-cultivable Treponema pallidum reveals extensive within-patient genetic
514 variation. Nat Microbiol. 2016;2:16190. Epub 2016/10/18. doi:
515 10.1038/nmicrobiol.2016.190. PubMed PMID: 27748767.
516 33. Morgan CA, Lukehart SA, Van Voorhis WC. Immunization with the N-terminal portion
517 of Treponema pallidum repeat protein K attenuates syphilitic lesion development in the rabbit
518 model. Infect Immun. 2002;70(12):6811-6. doi: 10.1128/iai.70.12.6811-6816.2002. PubMed
519 PMID: WOS:000179377600039.
520

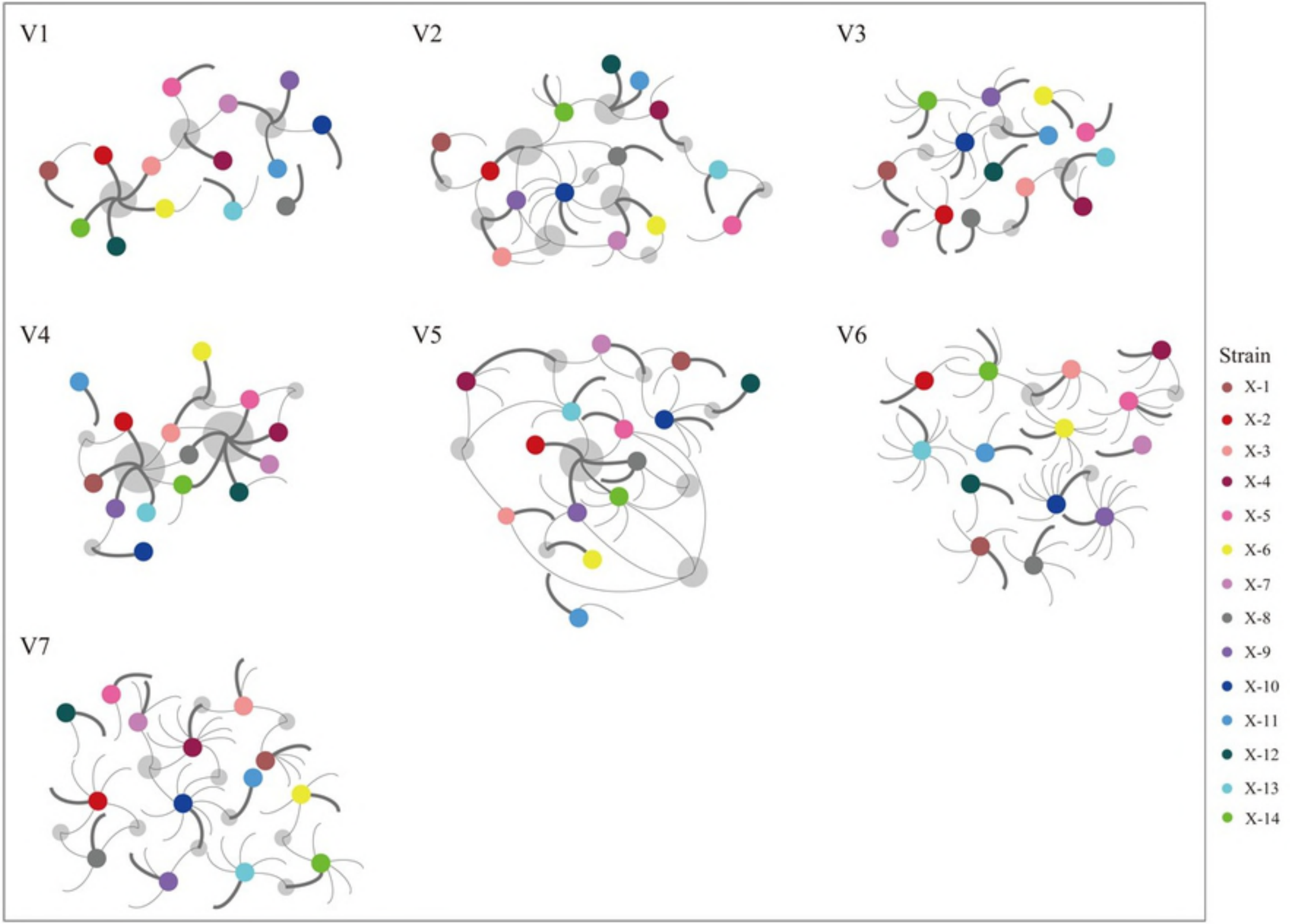

(%) 100
90
80
70
60
50
40
30
20
10
0
21
55
40
26
51
76
66
V1
V2
V3
V4
V5
V6
V7
32 38 41 59 62 65 68 54 57 60 63 66 69 72 75 78
81 84 49 55 58 84 90 45 48 51 54 57 60 63 66 69
72 75 78 64 67 70 73 76 79 82 85 88 91 94

A

B




V1
V2
V3
X-2
X-2
(10 clones)
V4
V5
V6
V7


V1
V2
V3
V4
V5
V6
V7
Strain
X-1
X-2
X-3
X-4
X-5
X-6
X-7
X-8
X-9
X-10
X-11
X-12
X-13
X-14