Estimation of Full-Length TprK Diversity in *Treponema pallidum* subspecies *pallidum*

Amin Addetia[1], Michelle Lin[1], Quynh Phung[1], Hong Xie[1], Meei-Li Huang[1], Giulia Ciccarese[2], Ivano Dal Conte[3], Marco Cusini[4], Francesco Drago[2], Lorenzo Giacani[5], Alexander L. Greninger[1*]

[1]Department of Laboratory Medicine, University of Washington, Seattle, Washington, USA

[2]Health Sciences Department, Section of Dermatology, San Martino University Hospital, Genoa, Italy

[3]STI Clinic, Infectious Diseases Unit, University of Turin, Turin, Italy

[4]Fondazione IRCCS Ca' Granda, Ospedale Maggiore Policlinico, Milan, Italy

[5]Department of Medicine, University of Washington, Seattle, Washington, USA

Key words: *Treponema pallidum*, syphilis, TprK, antigenic variation, outer membrane protein, immune evasion.

Short title: *tprK* deep-sequencing profiling

*Address correspondence to Alexander L. Greninger, agrening@uw.edu

Abstract

Immune evasion and disease progression of *Treponema pallidum* subspecies *pallidum* are associated with sequence diversity in the hypervariable, putative outer membrane protein TprK. Previous attempts to study variation within TprK have sequenced at depths insufficient to fully appreciate the hypervariable nature of the protein, failed to establish linkage between the protein's 7 variable regions, or were conducted on strains passed through rabbits. As a consequence, a complete profiling of *tprK* during infection in the human host is still lacking. Furthermore, prior studies examining how *T. pallidum* uses its repertoire of genomic donor sites to generate diversity within the V regions of the *tprK* also yielded a partial understanding of this process, due to the limited number of *tprK* alleles examined. In this study, we used short- and long-read deep sequencing to directly characterize full-length *tprK* alleles from *T. pallidum* collected from early lesions of patients attending two STD clinics in Italy. Our data, combined with recent data available on Chinese *T. pallidum* strains, show the near complete absence of overlap in TprK sequences among the 41 strains profiled to date. Moreover, our data allowed us to redefine the boundaries of *tprK* V regions, identify 55 donor sites, and estimate the total number of TprK variants that *T. pallidum* can potentially generate. Altogether, our results support how *T. pallidum* TprK antigenic variation system is an unsurmountable obstacle for the human immune system to naturally achieve infection eradication, and reiterate the importance of this mechanism for pathogen persistence in the host.

Importance

Syphilis continues to be a significant public health issue in both low- and high-income nations, including the United States, where the number of infectious syphilis cases has increased dramatically over the past five years. *T. pallidum*, the causative agent of syphilis, encodes an outer membrane protein TprK that undergoes segmental gene conversion to constantly create new sequences. We performed deep TprK profiling to understand full-length TprK diversity in *T. pallidum*-positive clinical specimens and compared these to all samples for which TprK deep sequencing is available. We found almost no overlap in TprK sequences between different patients. We further estimate that the total baseline junctional diversity of full-length TprK rivals that of current estimates of the human adaptive immune system. These data underscore the immunoevasive ability of TprK that allows *T. pallidum* to establish lifelong infection.

Introduction

Syphilis, caused by the spirochete *Treponema pallidum* subspecies *pallidum* (*T. pallidum*), is a significant global health problem. Although most syphilis cases occur in low-income countries, where the disease is endemic, rates of syphilis infection have been steadily increasing for the last two decades in high-income nations, particularly in men who have sex with men (MSM) and HIV-infected individuals [1–4]. Syphilis is an acute and chronic sexually transmitted infection marked by distinct early and late stages [5]. These stages are generally distinguished by unique clinical manifestations with symptoms associated with the late stage developing up to several decades after initial infection and following a long period of disease latency [6].

The mechanisms that allow *T. pallidum* to persist for the lifetime of an infected individual are not fully understood. During natural and experimental syphilis infection, a robust host immune response is developed against *T. pallidum* [7]. This suggests immune evasion strategies developed by *T. pallidum* are a key aspect of its pathogenesis [8].

The ability of *T. pallidum* to evade the host immune response is attributed to the organism's scarcity of surface-exposed outer membrane proteins (OMPs), very slow generation time (~33h), and the ability to stochastically and rapidly switch on and off expression of genes encoding putative OMPs through phase variation [9]. Chief among the immune evasion strategies evolved by *T. pallidum* is its ability to generate diversity within the putative OMP TprK [10–12]. TprK harbors seven discrete variable (V) regions, V1-V7. In the putative TprK beta-barrel structure, each variable region is predicted to form a loop exposed at the host-pathogen interface [13]. Generation of variants in these

V regions occurs through non-reciprocal segmental gene conversion, a process in which sections from donor sites flanking the *tprD* gene (*tp0117*) are stitched together to create new sequences [14,15]. Forty-seven putative donor sites have been identified thus far [15], however, the total number of unique TprK sequences that can be generated in a *T. pallidum* strain has yet to be determined.

Gene conversion results in significant intra- and inter-strain diversity of the TprK protein [16,17,14,18,19]. In rabbit models, diversity in TprK actively accumulates over the course of an infection and appears to be driven by the host's immune response [20,21]. At least five of the variable regions, V2 and V4-V7, of TprK elicit an antibody response in rabbit models [22]. These antibodies are specific for a single variable sequence [22], which further supports that generation of new V region sequences allows *T. pallidum* to evade the host response. Furthermore, increased diversity of TprK is directly correlated with more advanced stages of syphilis [19,23]. In both rabbit models and humans, *T. pallidum* strains isolated from cases of secondary/disseminated syphilis contained more TprK diversity than those isolated from cases of primary syphilis [19,23].

Previous studies to evaluate TprK variability within *T. pallidum* strains have sequenced of a limited number of TprK clones, failed to resolve linkage between variable regions, or been conducted on strains passed through rabbits [16,18–20] [24]. As a result, no studies to date have adequately profiled TprK within *T. pallidum* during natural infection in the human host. Furthermore, understanding of how different donor sites contribute to variable region sequences has been hindered by the analysis using a limited number of *tprK* clones [25]. In this study, we used short- and long-read deep

sequencing to directly characterize full-length TprK genes amplified from *T. pallidum* collected from early genital or anal lesions of 13 individuals attending two STD clinics in Milan and Turin, Italy [26]. We then combine our data with recent short-read *tprK* sequencing data from 28 *T. pallidum* strains from collected in China to illustrate the near complete lack of overlap in TprK sequences among all 41 clinical strains directly and deeply profiled to date. Additionally, our data help to redefine the TprK variable regions and to provide an estimate of the number of TprK variants that *T. pallidum* can potentially generate with its repertoire of donor cassettes. Overall, our data reiterate the pivotal importance of the TprK antigenic variation system to allow *T. pallidum* persistence in the host during infection.

Methods

Sample collection

Swabs from genital or anal lesions were collected from syphilis patients attending the Dermatology Clinics of the University of Turin and the Ospedale Maggiore in Milan from approximately December 2016 to March 2017. The only exclusion criterion for sample collection was an existing record of antibiotic therapy initiated within 30 days from the patient visit. For sample collection, whenever possible, the lesion area was gently squeezed to imbibe the swabs with exudate. The swabs were then placed in sterile microcentrifuge tubes containing 1 ml of 1X lysis buffer (10 mM Tris-HCl, 0.1 M ethylenediaminetetraacetic acid, and 0.5% sodium dodecyl-sulfate) suitable for DNA extraction. The swab shafts were then cut to leave the swab in the buffer. Samples were kept frozen at -80°C until DNA extraction. Sample collection was authorized by the

Human Subject Committee of each collecting institution (Protocol code: PR033REG2016 for the University of Turin, and Protocol Code TREPO2016 for the University of Milan) and informed consent was obtained from each patient. Specimens were then sent as de-identified samples in dry ice to the University of Washington for DNA extraction. The University of Washington Institutional Review Board determined this investigation not to be human subject research. Patient demographics were also collected as well as information on sexual orientation, HIV status, syphilis stage and serology results (VDRL/RPR and TPHA/TPPA) at the time of patient visit.

DNA extraction and strain typing

Frozen samples were thawed at room temperature and vortexed before processing. DNA was extracted from 200 µl of sample suspension using the QIAamp DNA mini kit (Qiagen, Valencia, CA) according to the manufacturer's instructions. DNA was resuspended in 100 µl of elution buffer provided with the kit. Successful DNA extraction was checked by amplification of a fragment of the human β-globin gene (Sense primer 5´-CAA CTT CAT CCA CGT TCA CC-3´, Antisense primer 5`- GAA GAG CCA AGG ACA GGT A-3´; expected size: 268 bp). Amplifications were performed in a 50 µl final volume using 5 µl of DNA template and 2.5 units of GoTaq polymerase (Promega, Madison, WI). Final concentrations of $MgCl_2$, dNTPs, and each primer were 1.5 mM, 200 µM, and 0.32 µM, respectively. Cycling conditions were initial denaturation at 95°C for 4 minutes, followed 95°C for 1 min, 60°C for 1 min and 72°C for 1 min for a total of 40 cycles. Final extension was at 72°C for 5 min.

Quantification of Treponemal Load within Patient Samples

The treponemal load of each sample was measured by qPCR as previously described [24]. Briefly, a portion of *tp47* was amplified using 14.33 µL of 2x QuantiTect multiplex PCR mix, 0.65 µL of 2x QuantiTect multiplex PCR mix with ROX, 0.03 unit of UNG and the following primers 5´-CAA GTA CGA GGG GAA CAT CGA T-3´ and 5´-TGA TCG CTG ACA AGC TTA GG-3´. Amplification was monitored with the following probe: 5´-FAM-CGG AGA CTC TGA TGG ATG CTG CAG TT-NFQMGB-3´. The following conditions were used for the qPCR reaction: 50°C for 2 minutes, 95°C for 15 minutes, and 45 cycles of 94°C for 1 minute and 60°C for 1 minute.

Direct from sample amplification and next-generation sequencing of tprK

PCR amplification of *tprK* was conducted using previously described conditions [24] and *tprK*-specific primers appended to 16 bp Pacbio barcodes (Table S1). The resulting 1.7kb product was cleaned using 0.6x volumes of AMPure XP beads (Beckman-Coulter). For long-read sequencing, library construction and sequencing on a Sequel I SMRT Cell 1M with a 10-hour movie were completed by the University of Washington PacBio Sequencing Services. A minimum of 5,224 PacBio reads were obtained for each of the samples. Short-read libraries from the same full-length amplicons were constructed with the Nextera XT kit (Illumina), cleaned with 0.6x volumes of AMPure XP beads (Beckman-Coulter), and sequenced on 1x192 bp Illumina MiSeq runs. A minimum of 101,000 Illumina sequencing reads, corresponding to a minimum mean coverage of 6,672x, were obtained for each sample. Sequencing metadata is available in Table S2.

Sequencing analysis

Analysis of *tprK* was performed using custom python/R scripts available on GitHub (https://github.com/greninger-lab/tprK_diversity). For the Italian samples, because of the tagmentation-based library preparation, we quality- (Q20) and adapter-trimmed Illumina reads using trimmomatic v0.38. PacBio Q20 CCS reads between 1400-1800 bp were trimmed of PCR primers using the dada2 preprocessing pipeline and denoised using RAD [27,28]. Previously published short-read tiling sequencing data for *tprK* from 14 primary and 14 secondary syphilis infections in adults from Xiamen University was downloaded from the NCBI Sequence Read Archive [29,30]. Because of the tiling PCR library design followed by 2x300bp, both paired-end reads were used in analysis of the Xiamen samples after adapter trimming using the same options as above. Variable regions were extracted from all samples using fuzzy regular expression matching using 18bp of neighboring conserved sequence with up to a 3bp mismatch. Because of the slight differences in coverage, we required a minimum of 5 reads of support for a given variable region amino acid sequence from the Xiamen samples, while for the Italian samples we used a minimum of 10 reads. We additionally included short-read sequencing data from 2 *T. pallidum* strains passaged in rabbits, which we profiled in a previous investigation [24], in our analysis. Similar to the Italian strains, we required each unique identified variable region sequence in these strains to be supported by a minimum of 10 sequencing reads and present at or above a frequency of 0.2%.

For full-length TprK phylogenetic analysis, we removed any TprK sequences that contained stop codons or which failed to fuzzy match a 20 amino acid region (allowing 3 mismatches) in any conserved region abutting a variable region, which we found was indicative of two frame shifts in consecutive variable regions in two TprK sequences. We also removed any full-length TprK sequences that comprised <0.2% of sequences present in a given sample for the purposes of display.

We used blastn with exact matching over a word size of 10 – our estimate of the smallest, high-confidence contribution of a donor site – to identify potential donor sites within a 17.2kb locus containing the *tprD* gene. We limited the number of potential contributions of each donor site to a variable region to three by restricting the maximum high scoring pairs (-max_hsps 3). We used the subject_besthit option to force non-overlapping HSPs. In order to generate a list of high-confidence donor sites and reduce putative false positives due to the smaller word size and to control for potential sequencing error, we only used variable regions with greater than 50 reads of support and 0.2% frequency (within-sample) and also required donor sites to be used in recovered *tprK* variable region sequences in at least 2 separate samples.

Shannon diversity scores for each sample were calculated using the R package VEGAN [31]. Differences in the number of variable region sequences and diversity scores for strains stratified by host factors were assessed using the Wilcoxon Rank-Sum test.

Data availability

Illumina and PacBio reads from *tprK* sequencing of the samples used in this study are available under the NCBI BioProject number PRJNA589065.

Results

Italian patient metadata

We selected 13 *T. pallidum* strains collected from syphilis patients, comprising 7 primary and 6 secondary syphilis cases, in Milan and Turin, Italy (Table 1, Table S3). All patients reported to be MSM and the median age of individuals was 39 years (range 20-57 years). Eight of the individuals sampled were HIV positive and, for nine of the patients, this was the first syphilis diagnosis. Seven of the specimens were collected from genital lesions, while the remaining six were collected from anal lesions.

T. pallidum DNA quantitation from clinical specimens

We first assessed the impact of the amount of treponemal DNA input into our initial PCR reaction on our ability to detect diversity within the 7 variable regions of *tprK*. For 3 strains, we compared inputs of 1,000 copies of treponemal DNA to the maximum possible input for our *tprK* PCR amplification reaction. Additionally, we performed a technical replicate using the same copy number input for another strain. For strains AS10, AS11, and AS12, the maximum input for *tprK* PCR was 5,362, 2,736, 6,663 copies of treponemal DNA. For strain AS18, we repeated the *tprK* PCR with 1,013 copies of treponemal DNA. The number of identified variants and the diversity measures for each variable region were similar despite the varying inputs (Table S4,

239 Figure S1). For our subsequent analyses, we normalized the input for the initial *tprK*
240 amplification to 1,000 treponemal copies for each sample.
241
242 *TprK diversity in T. pallidum strains directly sampled from humans*
243 We used short-read sequencing to examine the diversity within the seven V
244 regions of TprK and required each identified amino acid sequence from an isolate to be
245 supported by a minimum of 10 sequencing reads and present at a relative frequency
246 greater than or equal to 0.2%. We identified a median of 65 (range: 37-162) unique
247 sequences from all seven V regions from our 13 *T. pallidum* strains. Across the 13
248 strains, V1 contained the overall fewest unique sequences (median: 4, range: 1-7) and,
249 as determined by the Shannon diversity index, was the least diverse variable region
250 (median: 0.119, range: 0-0.836). V6 contained the greatest number of unique variants
251 (median: 20; range: 3-65) and was also the most diverse variable region (median:
252 1.603; range: 0.363-2.760) (Table S5).
253 We next examined the diversity of TprK in the context of different clinical
254 characteristics. *T. pallidum* strains collected from cases of secondary syphilis contained
255 significantly more unique variable region sequences (p=0.004) and were significantly
256 more diverse (p=0.002) than those strains collected from cases of primary syphilis
257 (Table 2). The number of unique sequences did not significantly differ (p=0.174)
258 between strains collected from anal or genital lesions. However, strains collected from
259 anal lesions exhibited significantly more diversity (p=0.035) across the seven V regions.
260 No significant differences were observed in the number of unique variants or diversity

when stratified by HIV status of the patient (p=0.187; p=0.171) or history of prior *T. pallidum* infection (p=0.537; p=0.711).

In a previous investigation, we profiled TprK in two *T. pallidum* strains collected from a single patient and amplified by two passages of strains in New Zealand white rabbits [24]. To assess the impact of the additional passage through rabbits on TprK, we compared the number of unique variants and diversity across the seven V regions identified from the 13 Italian *T. pallidum* strains and our two previously profiled strains. Strains passed through rabbits contained a greater number of variable region sequences (median 177.5 vs. 65, p=0.051) and greater diversity across the seven variable regions (median 12.1 vs. 4.2, p=0.076) compared to those directly sequences from clinical samples, though these differences were not significant given the few rabbit strains previously sampled.

To ensure accurate estimation of variable regions in both unlinked and linked analyses, we compared results from both short-read (Illumina) and long-read (PacBio) sequencing of all *tprK* amplicons generated in this study. The variable region allele percentages as measured by short-read and long-read sequencing were highly correlated (median r^2 = 0.995, range 0.974-1.00) (Figure S2), illustrating the high quality of modern long-read sequencing. PacBio sequencing exhibited an overall positive bias in variable region allele percentage compared to Illumina sequencing with an average linear regression slope of 1.029 (range 1.011-1.092), likely due to clustering during read denoising and post-filtering.

Using our long read data, we recovered a total of 634 full-length TprKs across the 13 samples, ranging from 26-95 different full-length TprKs within each sample at ≥

0.2% frequency. The most prevalent TprK in each sample was generally located near the root of the TprK phylogenetic tree for that particular sample (Figure 1). We found that only 3 full-length TprK sequences were shared among the 634 TprK sequences recovered from all 13 patients after removing sequences at a frequency of <0.2% within each sample. Two of these overlapping TprK sequences comprised the most common sequences in at least one of the samples. For example, the most common full-length TprK sequence in AS12 comprised 72.5% of TprK sequences present in that sample and was also present at 3.2% of TprK sequences in MI01. Likewise, a TprK comprising 53.0% of sequences present in MI06 was also present in 0.3% of MI04 TprK sequences.

Comparison of TprK diversity between Italian and Chinese strains

We next examined whether the TprK V region sequences present in our 13 Italian individuals shared any overlap with TprK sequences derived from short-read sequencing of 28 primary and secondary syphilis specimens recently reported from China [29,30]. Given the extraordinary diversity present in this gene, for print display, we filtered out any variable sequences constituting <20% of the species present in a given sample (Figure 3). More complex data filtered with a minimum frequency of 1% is displayed in an interactive figure in Supplemental File 1.

The heatmap shows the impressive diversity present across the TprK variable regions. V1 and V4 were the most conserved (Figure 2). The same two V1 sequences comprised the highest frequency species present in 9/13 (69.3%) Italian specimens and 16/28 (57.1%) Chinese specimens. Only 12 majority V4 sequences were present across

the 41 specimens. However, the most common V4 sequence present in the Chinese samples was only represented once in the Italian cohort and even then it was not the major species present.

V3, V5, V6, and V7 regions demonstrated almost no overlap among the 41 specimens (Figure 2). Only 6 of 39 major V7 sequences were shared between any Italian and any Chinese specimen. No shared V6 sequences were seen among any samples as the majority species present for each sample.

Redefining conserved and variable regions in tprK

The sequences we mined from variable regions were initially based off of prior definitions of the conserved and variable portions of *tprK*, which themselves were based off comparatively few *tprK* sequences [25]. While identifying donor sites, we noticed systematic biases in variable region sequence lengths mined from sequencing reads and the total blastn HSP length (Figure S3A/B and reflected in Figure 2). For instance, >98% V3 region sequences started with the same 23 bp sequence (5´-TCATACTCACCTTAGCCCCGACA-3´), and all other sequences had a Levenshtein edit distance of 1 from this sequence, suggesting this sequence may mark a conserved portion of *tprK.* Similarly, 100% of V5 region sequences started with the same 13 bp sequence (5´-AATATAGGCAGCA-3´) and no V5 sequence had less than 13 bp difference in sequence and blast hit length. For V2, 99.3% of sequences began with the same 14 bp sequence (5´-AGTATGGATTGGGG-3´) and the lone alternative sequence could be explained by low-frequency Illumina sequencing error associated with G-quadruplexes [32]. Removal of these sequences improved the ability to align *tprD* donor

sites across the length of *tprK* variable region sequences, leaving a 4 nucleotide common sequence (5'-CCTA-3') in V4 region sequences that we left based on its short nature (Figure S3C/D).

Contribution of donor sites to variable regions

We next examined how each variable region sequence was generated from different donor sites using data from all 41 samples. We found a total of 55 donor sites, corresponding to 5 for V1, 5 for V2, 13 for V3, 5 for V4, 6 for V5, 14 for V6, and 7 for V7 (Figure 3A). Forty-seven sites were previously reported by Centurion et al. [25]. There was considerable overlap between the two sets, suggesting a finite limit to the number of donor sites for *tprK.* Of note, two donor sites in the *tprD* locus (VS1-15, VS2-21) had single nucleotide variants compared to our reference sequence but exactly matched their previously deposited *tprD* locus (AY587909.1) [25], indicating that chromosomal mutations in donor sites can affect *tprK* variable region sequences. The vast majority of the donor sites found in this analysis, 51/55, were clustered downstream of *tprD*, while the remaining 4 donor sites were located upstream of *tprD*. Notably, all 51 of the donor sites located downstream of *tprD* were in the same orientation as *tprD* and had the highest utilization, while the 4 sites upstream of *tprD* faced in the opposite orientation. Donor sites for specific variable regions were collocated together, such as V1-V4-V5, V2-V7, and V3-V6. V3-V6 donor sites were almost uniformly derived from overlapping sequences (Figure 3B). Donor sites for V1 and V4 were the shortest, measuring an average of 39.2 and 34.8 nucleotides, while V5 and V7 donor sites were the longest at 58.5 and 64.7 nucleotides (Figure 3C).

Estimate of total potential diversity of tprK

Using this new inventory of *tprK* donor sites flanking the *tprD* gene, we next estimated the total coding diversity of TprK. Assuming a simple model in which only one donor site contributes to each variable region sequence, the 55 *tprD* donor sites across 7 variable regions could combine to create a total of 955,500 different full-length TprK sequences assuming no mutation. However, multiple donor sites can contribute material to the same *tprK* variable region to create a mosaic variable region. Our manual review of donor site contributions to variable regions suggested that donor sites were limited to three separate contributions to create mosaic variable regions, so we set a limit of three for the number of high scoring pairs in our blastn analysis of donor sites against each variable region sequence. The plurality of V1 region sequences only had one donor site contribute sequence while no V3 or V7 sequences were generated from only one donor site (Figure 4A). However, all variable regions had the potential for three donor site contributions. Adding up all potential combinations of one, two, and three-segment gene conversions that generate different sequences (assuming no single-segment V3 and V7 sequences) and assuming independence between variable regions leads to a potential diversity of TprK of 2.69×10^{18} full-length protein sequences if donor sites are reused, or 1.11×10^{17} protein sequences without reuse (Table S5).

We next examined whether certain donor sites were not represented in specific sections of a given variable region. Consistent with the segment usage data in Figure 4, we found biases in donor site contribution in every variable region. For instance, every V4 sequence starts with contributions from the same donor site and only two of five total

V1 donor sites contribute to the third segment in V1, when using high-confidence variable region sequences (present in more than 50 reads and >0.2% within sample frequency). In addition, V3 and V6 variable regions make use of almost all of their donor sites in both the second and third segments, but less than half of potential donor sites in the first segment. Taking into account differential use of donor sites by variable region segment reduced total potential total diversity to 1.23×10^{15} full-length TprK sequences with replacement, or 7.95×10^{13} sequences without replacement. Across 1544 individual high-confidence variable region sequences, we found 146 variable region sequences that used the same donor site more than once in the same variable region sequence, indicating that some donor site reuse is allowed in the generation of *tprK* variable regions.

Discussion

Here we combine deep, full-length profiling of TprK from *T. pallidum*-positive patient specimens with data-mining of additional TprK short-read sequencing from 28 Chinese patients to explore the diversity of the consummate *T. pallidum* immunoevasion protein TprK. We find exceedingly little overlap within specific variable regions within and between each patient cohort. Only 3 of 634 high-quality, full-length TprK sequences were shared among any samples in the 13 patients on which we performed long-read sequencing. Consistent with previous reports, we found greater TprK diversity to be associated with secondary syphilis compared to primary syphilis [29,33]. We then used this dataset of TprK diversity to find additional donor sites and to begin to piece together the grammar of variable region generation.

Based on the lexicon of *tprK* donor sites measured using deep sequencing across 41 samples, we estimate a potential full-length TprK combinatorial diversity approaching 10^{14} – 10^{18} proteins, assuming independence across donor sites. These estimates may be overestimates if our assumption of independence between variable region sequences is incorrect. These estimates may also underestimate the total diversity potential of TprK due to varying lengths of donor site contributions to variable regions. Most importantly, this junctional diversity is similar to if not greater than measures of the human adaptive immune system [34–36].

Our data also provide insights into differences in measured diversity among different variable regions. The limited diversity in V1 is associated with higher use of single-segment gene conversions to generate the variable region, while the limited diversity in V4 is associated with biased positional usage of different donor sites. Using the same or similar numbers of overall donor sites, V2 and V5 are able to generate 2-9 times more possible diversity than V1 and V4, which is reflected in direct sequencing measurements. This increase in diversity generation is due to either less positional bias of donor sites or greater proportions of three-segment donor site contributions, or both.

Of note, we measured fewer than 100 full-length TprKs present in any given sample using our filtering criteria, which is substantially less than our theoretical diversity estimates. Though we demonstrated that increasing our PCR template to the maximum allowed per reaction did not greatly affect variable region diversity measurements, these measured estimates could be biased by the limited copy numbers (<10,000 copies) available for *T. pallidum* positive clinical specimens and the limited range of copy numbers tested in our study.

Our work was chiefly limited by the few numbers of clinical samples and *T. pallidum* strains that have been deeply profiled for TprK diversity. Here, we profiled 13 new *T. pallidum* positive clinical specimens and combined them with 28 previously sequenced samples. However, given the considerable coding potential of TprK, 41 specimens is far too little to understand its overall coding diversity. Because of the limited number of total variable regions sampled across these 41 samples (~10^3) versus the potential diversity, we considered ourselves underpowered to examine linkages or epistasis between different variable regions. Future work will have to examine whether certain variable region sequences segregate together within a given TprK. The sampling requirements to determine that association are likely quite considerable and beyond the scope of the work presented here.

In addition to our limited understanding of how TprK variable regions interact with each other, our work here also does not fully inform how TprK interacts with the immune system. As the overall coding diversity of specific variable regions is somewhat limited, it is possible that epistatic interactions between variable regions could influence epitope structure. Certainly, the paucity of variation across the 41 samples in the V4 region is surprising given that anti-V4 antibodies have been detected in humans [37]. We also note the lower number of measured V3 diversity could be associated with a lack of immunological pressure, especially considering its number of potential donor sites and three-segment gene conversions [37]. Also, if there were not epistatic interactions between variable regions, it is not necessarily clear why seven different variable regions with broad but distinct coding potentials would be required in TprK. Alternatively, if there is no or limited epistasis between variable regions and cross-protective antibody is

generated against individual variable regions, the diversity generating potential of individual variable regions combined with the rate of gene conversion could put an upward bound on the time period before *T. pallidum* becomes latent in humans.

In summary, our work provides a basis for one mechanism of how *T. pallidum* maintains lifelong infection, through the constant generation of TprK diversity using a lexicon that approaches that of the baseline human adaptive immune system. Therapeutic interventions that target mechanisms of TprK diversity generation may prove beneficial. We further hypothesize that loss of the TprK diversity generation will be one of the first changes associated with longitudinal passage of *T. pallidum* in the new *in vitro* culture system that provides it respite from constant immune selection.

References

1. Patton ME, Su JR, Nelson R, Weinstock H, Centers for Disease Control and Prevention (CDC). Primary and secondary syphilis--United States, 2005-2013. MMWR Morb Mortal Wkly Rep. 2014;63: 402–406.

2. Centers for Disease Control and Prevention. Syphilis Surveillance Supplement 2013–2017. Atlanta, U.S.: Department of Health and Human Services; 2019.

3. Stamm LV. Syphilis: Re-emergence of an old foe. Microb Cell. 2016;3: 363–370. doi:10.15698/mic2016.09.523

4. Doherty L. Syphilis: old problem, new strategy. BMJ. 2002;325: 153–156. doi:10.1136/bmj.325.7356.153

5. Centers for Disease Control and Prevention (CDC). Primary and secondary syphilis--United States, 1997. MMWR Morb Mortal Wkly Rep. 1998;47: 493–497.

6. Singh AE, Romanowski B. Syphilis: review with emphasis on clinical, epidemiologic, and some biologic features. Clin Microbiol Rev. 1999;12: 187–209.

7. Salazar JC, Hazlett KRO, Radolf JD. The immune response to infection with Treponema pallidum, the stealth pathogen. Microbes Infect. 2002;4: 1133–1140. doi:10.1016/s1286-4579(02)01638-6

8. Cruz AR, Ramirez LG, Zuluaga AV, Pillay A, Abreu C, Valencia CA, et al. Immune Evasion and Recognition of the Syphilis Spirochete in Blood and Skin of Secondary Syphilis Patients: Two Immunologically Distinct Compartments. PLoS Negl Trop Dis. 2012;6. doi:10.1371/journal.pntd.0001717

9. Giacani L, Brandt SL, Ke W, Reid TB, Molini BJ, Iverson-Cabral S, et al. Transcription of TP0126, Treponema pallidum putative OmpW homolog, is regulated by the length of a homopolymeric guanosine repeat. Infect Immun. 2015;83: 2275–2289. doi:10.1128/IAI.00360-15

10. LaFond RE, Lukehart SA. Biological Basis for Syphilis. Clinical Microbiology Reviews. 2006;19: 29–49. doi:10.1128/CMR.19.1.29-49.2006

11. Cameron CE, Lukehart SA. Current status of syphilis vaccine development: need, challenges, prospects. Vaccine. 2014;32: 1602–1609. doi:10.1016/j.vaccine.2013.09.053

12. Centurion-Lara A, Castro C, Barrett L, Cameron C, Mostowfi M, Van Voorhis WC, et al. Treponema pallidum major sheath protein homologue Tpr K is a target of opsonic antibody and the protective immune response. J Exp Med. 1999;189: 647–656.

13. Centurion-Lara A, Giacani L, Godornes C, Molini BJ, Brinck Reid T, Lukehart SA. Fine analysis of genetic diversity of the tpr gene family among treponemal species, subspecies and strains. PLoS Negl Trop Dis. 2013;7: e2222. doi:10.1371/journal.pntd.0002222

14. Giacani L, Brandt SL, Puray-Chavez M, Reid TB, Godornes C, Molini BJ, et al. Comparative investigation of the genomic regions involved in antigenic variation of the TprK antigen among treponemal species, subspecies, and strains. J Bacteriol. 2012;194: 4208–4225. doi:10.1128/JB.00863-12

15. Centurion-Lara A, LaFond RE, Hevner K, Godornes C, Molini BJ, Van Voorhis WC, et al. Gene conversion: a mechanism for generation of heterogeneity in the tprK gene of Treponema pallidum during infection: tprK gene conversion. Molecular Microbiology. 2004;52: 1579–1596. doi:10.1111/j.1365-2958.2004.04086.x

16. Stamm LV, Bergen HL. The sequence-variable, single-copy tprK gene of Treponema pallidum Nichols strain UNC and Street strain 14 encodes heterogeneous TprK proteins. Infect Immun. 2000;68: 6482–6486.

17. LaFond RE, Centurion-Lara A, Godornes C, Rompalo AM, Van Voorhis WC, Lukehart SA. Sequence Diversity of Treponema pallidum subsp. pallidum tprK in Human Syphilis Lesions and Rabbit-Propagated Isolates. Journal of Bacteriology. 2003;185: 6262–6268. doi:10.1128/JB.185.21.6262-6268.2003

18. Liu D, Tong M-L, Luo X, Liu L-L, Lin L-R, Zhang H-L, et al. Profile of the tprK gene in primary syphilis patients based on next-generation sequencing. Picardeau M, editor. PLOS Neglected Tropical Diseases. 2019;13: e0006855. doi:10.1371/journal.pntd.0006855

19. Liu D, Tong M-L, Lin Y, Liu L-L, Lin L-R, Yang T-C. Insights into the genetic variation profile of tprK in Treponema pallidum during the development of natural human syphilis infection. Giacani L, editor. PLoS Negl Trop Dis. 2019;13: e0007621. doi:10.1371/journal.pntd.0007621

20. LaFond RE, Centurion-Lara A, Godornes C, Van Voorhis WC, Lukehart SA. TprK Sequence Diversity Accumulates during Infection of Rabbits with Treponema pallidum subsp. pallidum Nichols Strain. Infection and Immunity. 2006;74: 1896–1906. doi:10.1128/IAI.74.3.1896-1906.2006

21. Giacani L, Molini BJ, Kim EY, Godornes BC, Leader BT, Tantalo LC, et al. Antigenic Variation in Treponema pallidum: TprK Sequence Diversity Accumulates in Response to Immune Pressure during Experimental Syphilis. The Journal of Immunology. 2010;184: 3822–3829. doi:10.4049/jimmunol.0902788

22. LaFond RE, Molini BJ, Van Voorhis WC, Lukehart SA. Antigenic Variation of TprK V Regions Abrogates Specific Antibody Binding in Syphilis. Infection and Immunity. 2006;74: 6244–6251. doi:10.1128/IAI.00827-06

528 23. Reid TB, Molini BJ, Fernandez MC, Lukehart SA. Antigenic Variation of TprK
529 Facilitates Development of Secondary Syphilis. Morrison RP, editor. Infect Immun.
530 2014;82: 4959–4967. doi:10.1128/IAI.02236-14

531 24. Addetia A, Tantalo LC, Lin MJ, Xie H, Huang M-L, Marra CM, et al. Comparative
532 Genomics and Full-Length TprK Profiling of Treponema pallidum subsp. pallidum
533 Reinfection. bioRxiv. 2019; 841395. doi:10.1101/841395

534 25. Centurion-Lara A, LaFond RE, Hevner K, Godornes C, Molini BJ, Van Voorhis WC,
535 et al. Gene conversion: a mechanism for generation of heterogeneity in the tprK
536 gene of Treponema pallidum during infection. Mol Microbiol. 2004;52: 1579–96.
537 doi:10.1111/j.1365-2958.2004.04086.x

538 26. Giacani L, Ciccarese G, Puga-Salazar C, Dal Conte I, Colli L, Cusini M, et al.
539 Enhanced Molecular Typing of Treponema pallidum subspecies pallidum Strains
540 From 4 Italian Hospitals Shows Geographical Differences in Strain Type
541 Heterogeneity, Widespread Resistance to Macrolides, and Lack of Mutations
542 Associated With Doxycycline Resistance. Sex Transm Dis. 2018;45: 237–242.
543 doi:10.1097/OLQ.0000000000000741

544 27. Kumar V, Vollbrecht T, Chernyshev M, Mohan S, Hanst B, Bavafa N, et al. Long-
545 read amplicon denoising. Nucleic Acids Res. 2019;47: e104–e104.
546 doi:10.1093/nar/gkz657

547 28. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP.
548 DADA2: High-resolution sample inference from Illumina amplicon data. Nature
549 Methods. 2016;13: 581–583. doi:10.1038/nmeth.3869

550 29. Liu D, Tong M-L, Lin Y, Liu L-L, Lin L-R, Yang T-C. Insights into the genetic
551 variation profile of tprK in Treponema pallidum during the development of natural
552 human syphilis infection. PLOS Neglected Tropical Diseases. 2019;13: e0007621.
553 doi:10.1371/journal.pntd.0007621

554 30. Liu D, Tong M-L, Luo X, Liu L-L, Lin L-R, Zhang H-L, et al. Profile of the tprK gene
555 in primary syphilis patients based on next-generation sequencing. PLOS Neglected
556 Tropical Diseases. 2019;13: e0006855. doi:10.1371/journal.pntd.0006855

557 31. Dixon P. VEGAN, a package of R functions for community ecology. Journal of
558 Vegetation Science. 2003;14: 927–930. doi:10.1111/j.1654-1103.2003.tb02228.x

559 32. Chambers VS, Marsico G, Boutell JM, Di Antonio M, Smith GP, Balasubramanian
560 S. High-throughput sequencing of DNA G-quadruplex structures in the human
561 genome. Nat Biotechnol. 2015;33: 877–881. doi:10.1038/nbt.3295

562 33. Reid TB, Molini BJ, Fernandez MC, Lukehart SA. Antigenic variation of TprK
563 facilitates development of secondary syphilis. Infect Immun. 2014;82: 4959–4967.
564 doi:10.1128/IAI.02236-14

565 34. Charles A Janeway J, Travers P, Walport M, Shlomchik MJ. The generation of
566 diversity in immunoglobulins. Immunobiology: The Immune System in Health and
567 Disease 5th edition. 2001 [cited 2 Jan 2020]. Available:
568 https://www.ncbi.nlm.nih.gov/books/NBK27140/

569 35. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. The Generation of
570 Antibody Diversity. Molecular Biology of the Cell 4th edition. 2002 [cited 2 Jan
571 2020]. Available: https://www.ncbi.nlm.nih.gov/books/NBK26860/

572 36. Mora T, Walczak AM. Quantifying lymphocyte receptor diversity. bioRxiv. 2016;
573 046870. doi:10.1101/046870

574 37. LaFond RE, Molini BJ, Van Voorhis WC, Lukehart SA. Antigenic variation of TprK V
575 regions abrogates specific antibody binding in syphilis. Infect Immun. 2006;74:
576 6244–51. doi:10.1128/IAI.00827-06

577

578

579 **Tables**

580 Table 1. Summary statistics of patient metadata for strains sequenced in this study

Location		% (n=13)
	Milan	30.77
	Turin	69.23
Stage		
	Primary	53.85
	Secondary	46.15
Type		
	14d/g	61.54
	13d/g	23.08
	13d/d	7.69
	6d/f	7.69
Gender		
	Male	100
Sexual Orientation		
	MSM	100
Age		
	Median [Min-Max]	39 [20-57]
HIV Status		
	Positive	61.54
	Negative	38.46
Lesion Location		
	Genital	53.85
	Anal	46.15
Genotypic Antibiotic Resistance		
	Tetracycline Resistance	0
	Macrolide Resistance	100
Infection Status		
	First Time Infected	69.23
	Previous Infection	30.77

581

582

Table 2. Comparison of the total number of variable region sequences and total Shannon diversity scores across the 7 variable regions of TprK in the context of clinical characteristics and passage history.

		Strains n	Total no. of variable region sequences Median (Range)	p-value	Total Diversity Median (Range)	p-value
Stage						
	Primary	7	58 (37-65)	0.004	2.70 (1.49-4.75)	0.002
	Secondary	6	100 (65-162)		7.69 (4.20-9.39)	
HIV Status						
	Positive	8	93.5 (37-162)	0.187	4.74 (2.64-9.39)	0.171
	Negative	5	60 (55-65)		2.70 (1.49-7.06)	
Lesion Location						
	Genital	7	60 (54-96)	0.174	7.69 (3.27-9.39)	0.035
	Anal	6	84.5 (37-162)		2.70 (1.49-5.28)	
Infection Status						
	First Time Infected	9	61 (37-128)	0.537	4.20 (1.40-9.24)	0.711
	Previous Infection	4	80.5 (54-162)		4.69 (2.64-9.39)	
Passaged in Rabbits						
	Yes	2*	177.5 (136-219)	0.051	12.12 (8.70-15.53)	0.076
	No	13	67 (37-162)		4.20 (1.49-9.39)	

*Isolates were previously profiled by Addetia, et. al. [24]

599 **Figure Legends**

600 **Figure 1 – Full-length TprK phylogeny of all protein sequences present at greater**
601 **than 0.2% within each individual from 13 patients from Italy.** Only intact full-length
602 TprK sequences derived from PacBio sequencing were used to generate the
603 phylogenetic tree. Each individual is labeled by a different color and the proportion of
604 sequences is shown by node size. Only three total sequences were shared among the
605 634 TprK sequences present in the 13 *T. pallidum* specimens sequenced in this study.
606
607 **Figure 2 – TprK variable region sequence heatmap.** Heatmap display of all deep
608 sequenced *tprK* from clinical specimens to date, comprising 13 individuals from Italy
609 sequenced here and 28 Chinese individuals from prior work. For print display, only
610 those variable region sequences present at ≥ 20% frequency within a sample are
611 depicted. Any variable frequencies less than 2% in other samples are not shown. The
612 proportion of sequences is illustrated by color for each heatmap for V1 (A), V2 (B), V3
613 (C), V4 (D), V5 (E), V6 (F), and V7 (F). A heatmap filtered at a frequency of 1% is
614 provided as an interactive html as Supplemental File 1.
615
616 **Figure 3 – Map of *tprK* donor sites flanking the *tprD* locus**. Variable region
617 sequences were blastn aligned against a 17.2kb locus that contained putative *tprK*
618 donor sites based on manual review. A) The usage of all 55 donor sites across the *tprD*
619 *locus* by variable region is depicted based on the sum of within-sample percentages
620 across all 41 samples. The entire 17.2kb locus is depicted due to recovery of a V4
621 donor site within the *phnU* gene at 16.9kb. Nucleotide numbering is shown based on

622 the strain UW-148B2 (CP045004.1). B) Zoomed in depiction of the locus immediately
623 downstream of *tprD* containing *tprK* donor sites. Donor sites are in the same orientation
624 as the *tprD* locus. The light brown sites include 45 of the 47 donor sites reported
625 previously by Centurion et al. [25]. The bottom donor sites include 51 of the 55 donor
626 sites found in this study and are colored based on their associated variable region.
627
628 **Figure 4 – Donor site segments and position by V region.** A) The number of donor
629 site contribution segments in each high-confidence variable region sequence was
630 determined in blastn output across the 41 samples. Usage was determined by the sum
631 percentage of variable region sequences by segment. For instance, V1 has the most
632 number of variable region sequences where only one donor site segment is used in a
633 given V region sequence, consistent with its overall lack of diversity. B) The position of
634 donor site contributions within a variable region sequence was also determined for each
635 donor site (i.e. “First” means the donor site was found as the first alignment within a
636 variable region sequence, “Second” as the second portion of the variable region
637 sequence alignment, and “Third” as the third). Within-sample percentages were
638 summed for each variable region in order to adjust for differences in read coverage at
639 each locus between samples. These summed percentages were then adjusted by the
640 total summed percentage to add up to 100% for each variable region.
641

642

643 **Supplemental Figure Legends**

644 **Supplemental Figure 1** – Measurements of diversity are consistent in technical
645 replicates and not significantly affected by template DNA > 1,000 treponemal copies.
646 For strains AS10 (A), AS11 (B), AS12 (C), and AS18 (D) we compared using the
647 maximal template amount allowed by our PCR reaction versus 1,000 treponemal
648 copies. Matched variable regions between the high and normal (1,000 copies) samples
649 are connected by a line and share the same color.

650

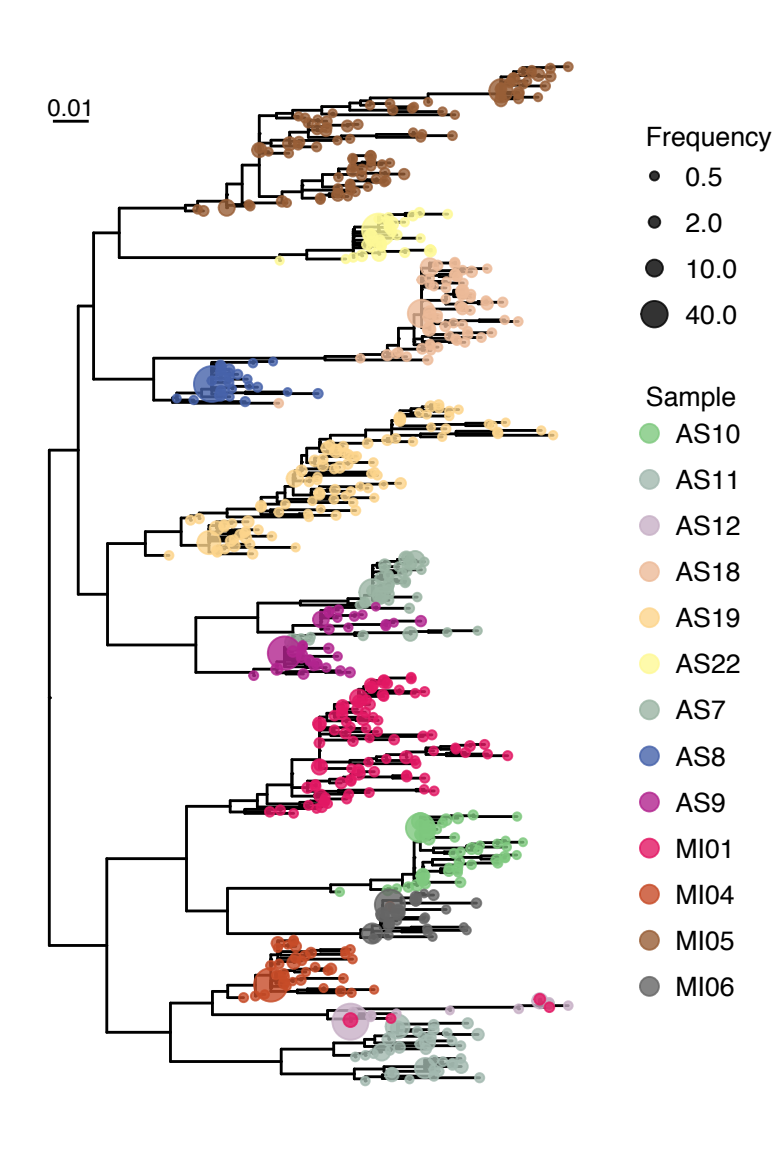
651 **Supplemental Figure 2** – Illumina versus PacBio variable sequence allele frequencies
652 scatterplot for each sample. Each data point is a specific variable region sequence and
653 different regions are labeled by color.

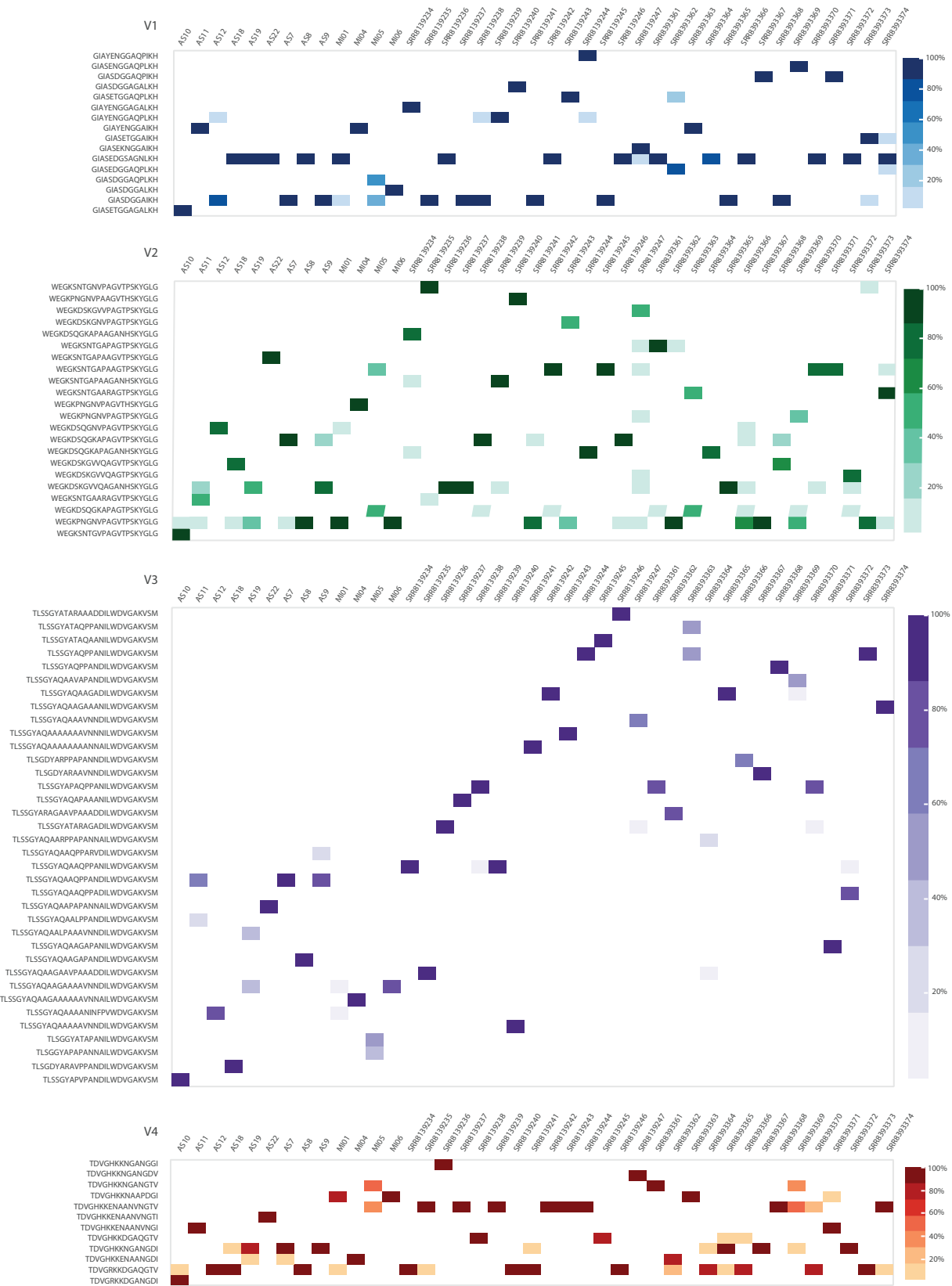
654

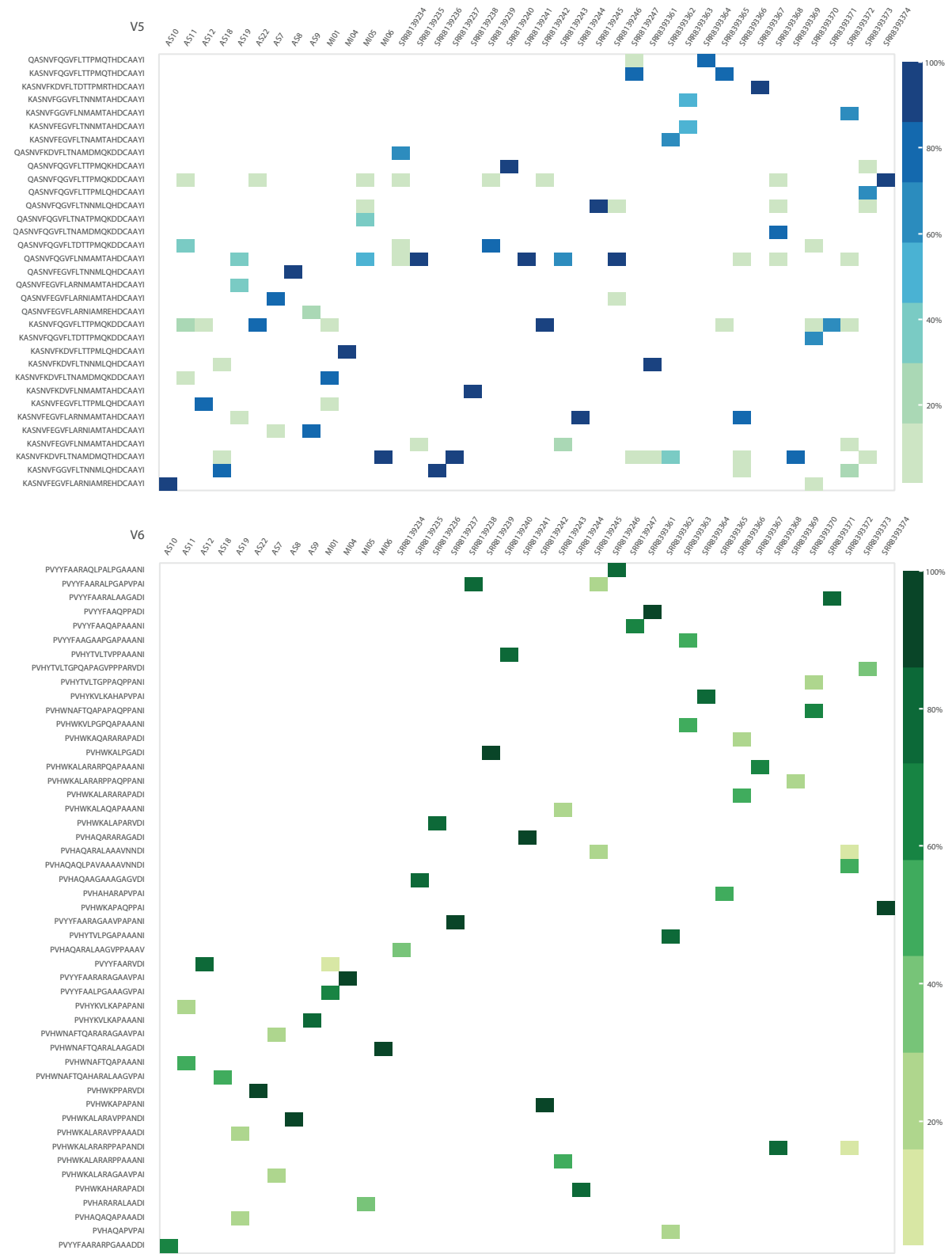
655 **Supplemental Figure 3** – Blast sequence length alignment versus variable region
656 sequence length plots before and after variable region sequence filtering. A) Scatterplot
657 of total sequence length alignment versus variable region sequence length after filtering
658 of V2, V3, and V5 variable region sequences of likely conserved region sequences. B)
659 Corresponding histogram of differences in total sequence length and alignment length
660 after filtering. C) Scatterplot of total sequence length alignment versus variable region
661 sequence length based on prior definitions of variable region sequences. D)
662 Corresponding histogram of differences in total sequence and alignment length without
663 filtering. Counts are absolute sequencing read counts across all 41 samples.

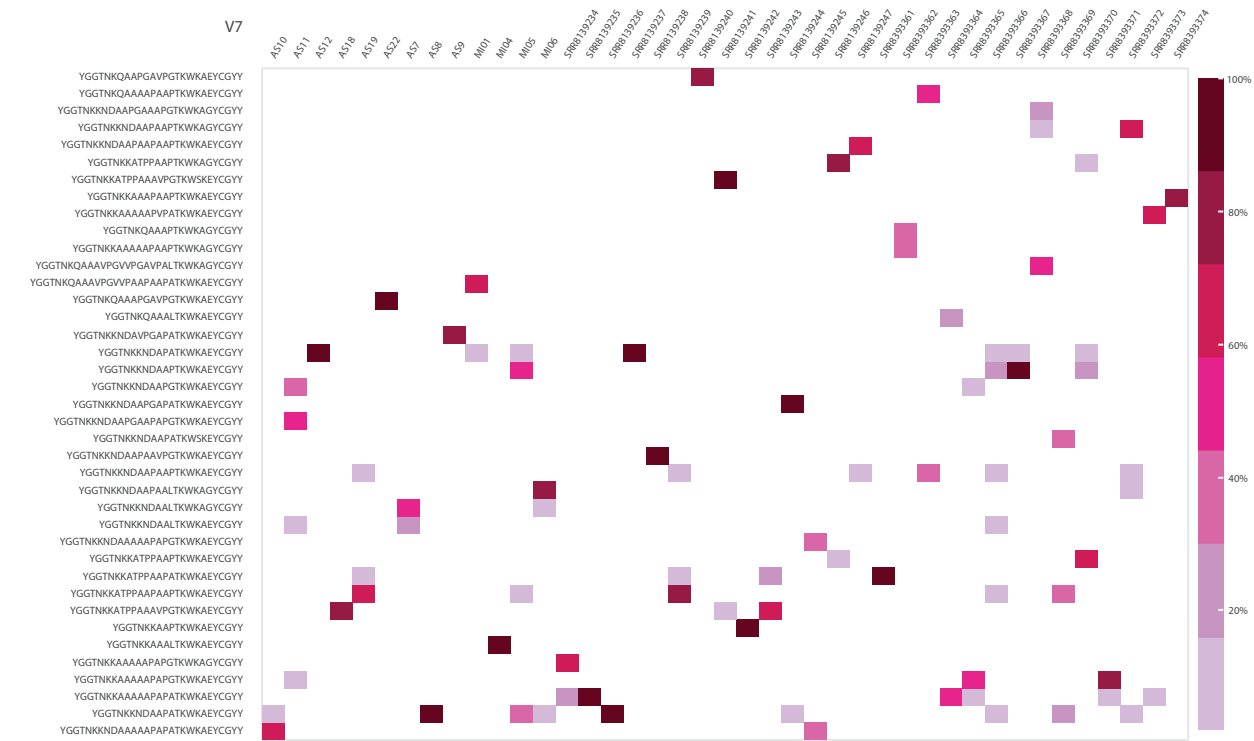
664

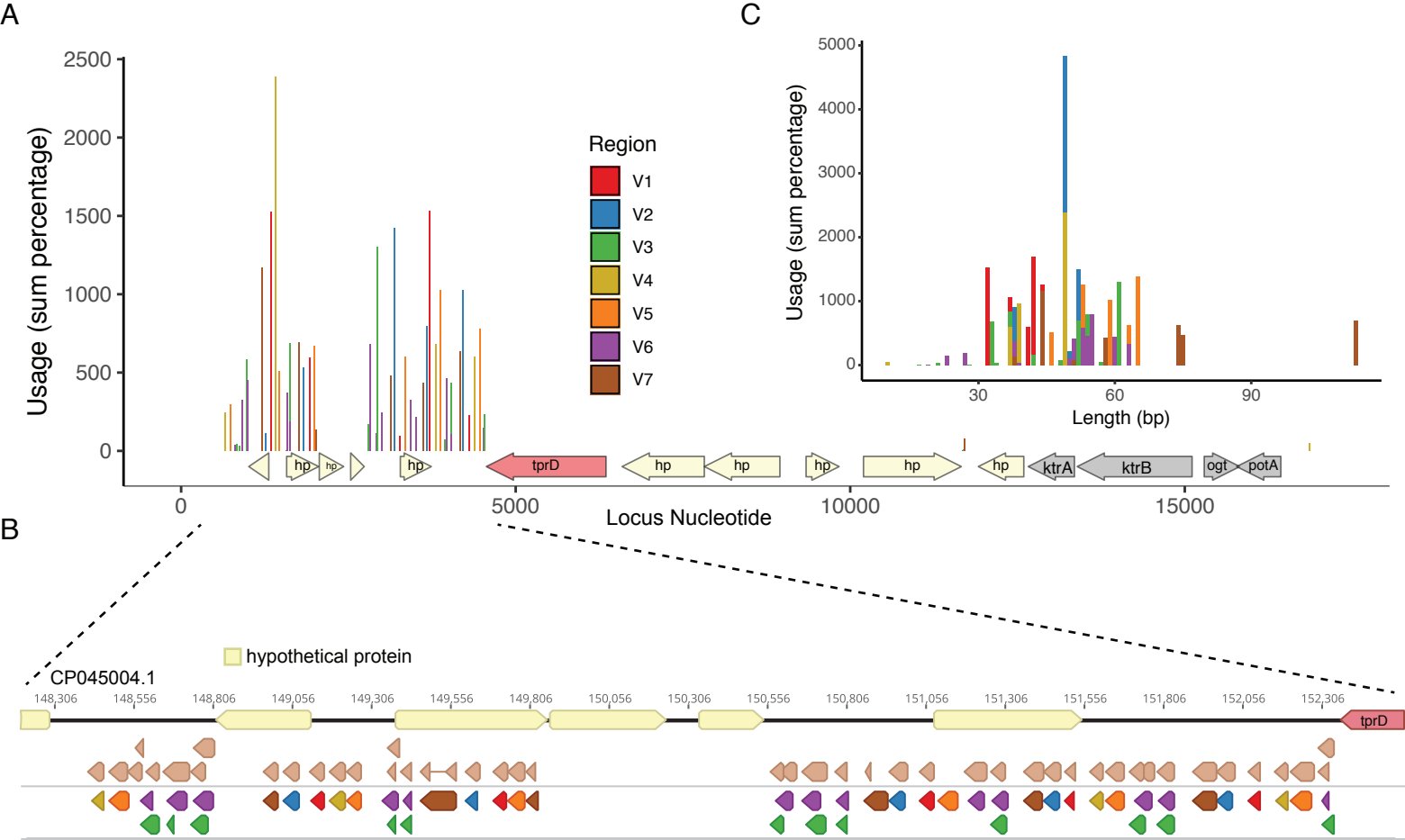
665 **Supplemental File 1 –** Interactive HTML heatmap of TprK V region frequencies across
666 13 Italian individuals and 28 Chinese individuals. The file contains any variable region
667 sequence present in at least one strain at a frequency greater than 1%.
668
669 **Supplemental File 2** – CSV file containing TprK V region sequences extracted from 13
670 Italian individuals and 28 Chinese individuals previously deep sequenced in *tprK*. Only
671 variable regions present in at least one sample with a minimum of 1% frequency are
672 displayed, as in Supplemental File 1.
673
674 **Supplemental File 3** – GFF file of *tprD* locus used in this study with previous donor
675 sites and newly annotated donor sites.

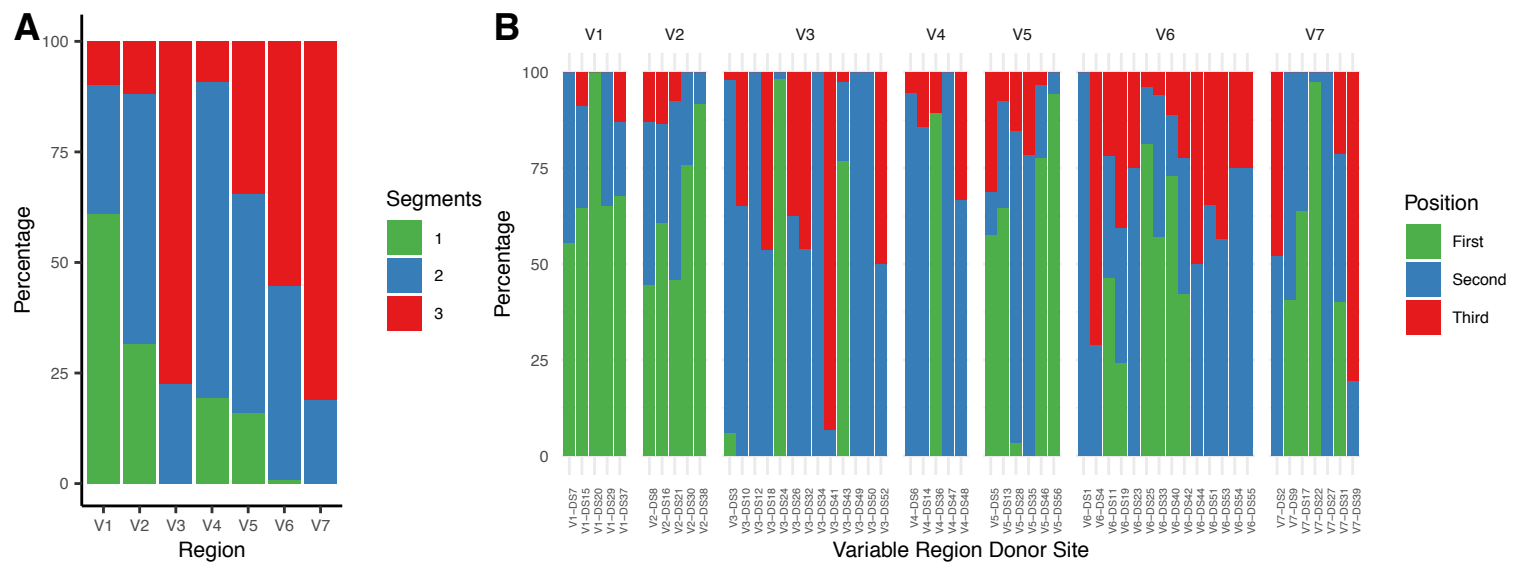

0.01
Frequency
0.5
2.0
10.0
40.0
Sample
AS10
AS11
AS12
AS18
AS19
AS22
AS7
AS8
AS9
MI01
MI04
MI05
MI06

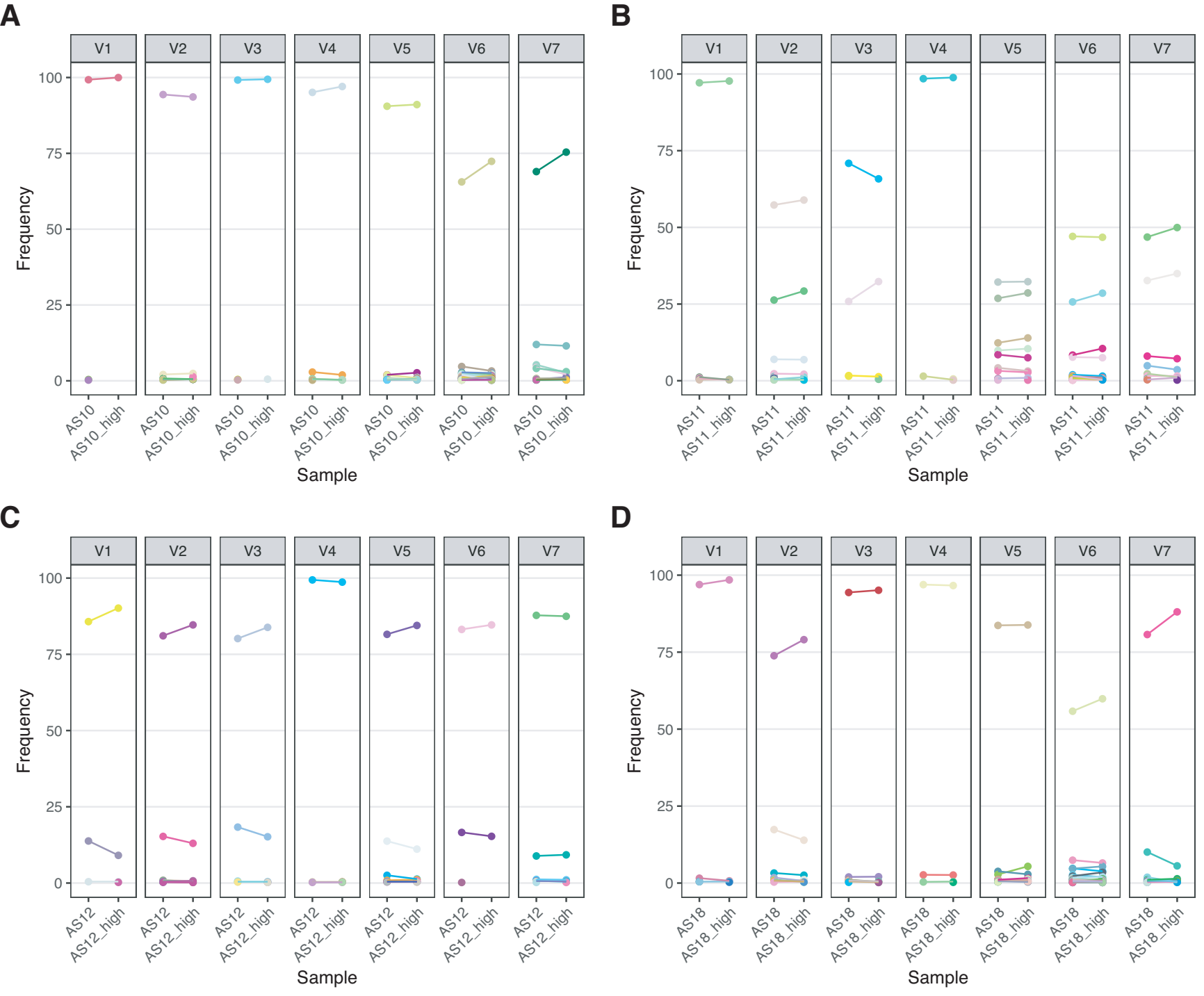

V1
AS10 AS11 AS12 AS18 AS19 AS22 AS7 AS8 AS9 MI01 MI04 MI05 MI06 SRR8139234 SRR8139235 SRR8139236 SRR8139237 SRR8139238 SRR8139239 SRR8139240 SRR8139241 SRR8139242 SRR8139243 SRR8139244 SRR8139245 SRR8139246 SRR8139247 SRR8393361 SRR8393362 SRR8393363 SRR8393364 SRR8393365 SRR8393366 SRR8393367 SRR8393368 SRR8393369 SRR8393370 SRR8393371 SRR8393372 SRR8393373 SRR8393374
GIAYENGGAQPIKH
GIASENGGAQPLKH
GIASDGGAQPIKH
GIASDGGAGALKH
GIASETGGAQPLKH
GIAYENGGAGALKH
GIAYENGGAQPLKH
GIAYENGGAIKH
GIASETGGAIKH
GIASEKNGGAIKH
GIASEDGSAGNLKH
GIASEDGGAQPLKH
GIASDGGAQPLKH
GIASDGGALKH
GIASDGGAIKH
GIASETGGAGALKH
100%
80%
60%
40%
20%
V2
AS10 AS11 AS12 AS18 AS19 AS22 AS7 AS8 AS9 MI01 MI04 MI05 MI06 SRR8139234 SRR8139235 SRR8139236 SRR8139237 SRR8139238 SRR8139239 SRR8139240 SRR8139241 SRR8139242 SRR8139243 SRR8139244 SRR8139245 SRR8139246 SRR8139247 SRR8393361 SRR8393362 SRR8393363 SRR8393364 SRR8393365 SRR8393366 SRR8393367 SRR8393368 SRR8393369 SRR8393370 SRR8393371 SRR8393372 SRR8393373 SRR8393374
WEGKSNTGNVPAGVTPSKYGLG
WEGKPNGNVPAAGVTHSKYGLG
WEGKDSKGVVPAGTPSKYGLG
WEGKDSKGNVPAGTPSKYGLG
WEGKDSQGKAPAAGANHSKYGLG
WEGKSNTGAPAGTPSKYGLG
WEGKSNTGAPAAGVTPSKYGLG
WEGKSNTGAPAAGTPSKYGLG
WEGKSNTGAPAAGANHSKYGLG
WEGKSNTGAARAGTPSKYGLG
WEGKPNGNVPAGVTHSKYGLG
WEGKPNGNVPAGTPSKYGLG
WEGKDSQGNVPAGVTPSKYGLG
WEGKDSQGKAPAGVTPSKYGLG
WEGKDSQGKAPAGANHSKYGLG
WEGKDSKGVVQAGVTPSKYGLG
WEGKDSKGVVQAGTPSKYGLG
WEGKDSKGVVQAGANHSKYGLG
WEGKSNTGAARAGVTPSKYGLG
WEGKDSQGKAPAGTPSKYGLG
WEGKPNGNVPAGVTPSKYGLG
WEGKSNTGVPAGVTPSKYGLG
100%
80%
60%
40%
20%
V3
AS10 AS11 AS12 AS18 AS19 AS22 AS7 AS8 AS9 MI01 MI04 MI05 MI06 SRR8139234 SRR8139235 SRR8139236 SRR8139237 SRR8139238 SRR8139239 SRR8139240 SRR8139241 SRR8139242 SRR8139243 SRR8139244 SRR8139245 SRR8139246 SRR8139247 SRR8393361 SRR8393362 SRR8393363 SRR8393364 SRR8393365 SRR8393366 SRR8393367 SRR8393368 SRR8393369 SRR8393370 SRR8393371 SRR8393372 SRR8393373 SRR8393374
TLSSGYATARAAADDILWDVGAKVSM
TLSSGYATAQPPANILWDVGAKVSM
TLSSGYATAQAANILWDVGAKVSM
TLSSGYAQPPANILWDVGAKVSM
TLSSGYAQPPANDILWDVGAKVSM
TLSSGYAQAAVAPANDILWDVGAKVSM
TLSSGYAQAAGADILWDVGAKVSM
TLSSGYAQAAGAAANILWDVGAKVSM
TLSSGYAQAAAAVNNDILWDVGAKVSM
TLSSGYAQAAAAAAAVNNNILWDVGAKVSM
TLSSGYAQAAAAAAAANNAILWDVGAKVSM
TLSGDYARPPAPANNDILWDVGAKVSM
TLSGDYARAAVNNDILWDVGAKVSM
TLSSGYAPAQPPANILWDVGAKVSM
TLSSGYAQAPAAANILWDVGAKVSM
TLSSGYARAGAAVPAAADDILWDVGAKVSM
TLSSGYATARAGADILWDVGAKVSM
TLSSGYAQAARPPAPANNAILWDVGAKVSM
TLSSGYAQAAQPPARVDILWDVGAKVSM
TLSSGYAQAAQPPANILWDVGAKVSM
TLSSGYAQAAQPPANDILWDVGAKVSM
TLSSGYAQAAQPPADILWDVGAKVSM
TLSSGYAQAAPAPANNAILWDVGAKVSM
TLSSGYAQAALPPANDILWDVGAKVSM
TLSSGYAQAALPAAAVNNDILWDVGAKVSM
TLSSGYAQAAGAPANILWDVGAKVSM
TLSSGYAQAAGAPANDILWDVGAKVSM
TLSSGYAQAAGAAAVPAAADDILWDVGAKVSM
TLSSGYAQAAGAAAAAVNNDILWDVGAKVSM
TLSSGYAQAAGAAAAAAVNNAILWDVGAKVSM
TLSSGYAQAAAAANINFPVWDVGAKVSM
TLSSGYAQAAAAAVNNDILWDVGAKVSM
TLSGGYATAPANILWDVGAKVSM
TLSGGYAPAPANNAILWDVGAKVSM
TLSGDYARAVPPANDILWDVGAKVSM
TLSSGYAPVPANDILWDVGAKVSM
100%
80%
60%
40%
20%
V4
AS10 AS11 AS12 AS18 AS19 AS22 AS7 AS8 AS9 MI01 MI04 MI05 MI06 SRR8139234 SRR8139235 SRR8139236 SRR8139237 SRR8139238 SRR8139239 SRR8139240 SRR8139241 SRR8139242 SRR8139243 SRR8139244 SRR8139245 SRR8139246 SRR8139247 SRR8393361 SRR8393362 SRR8393363 SRR8393364 SRR8393365 SRR8393366 SRR8393367 SRR8393368 SRR8393369 SRR8393370 SRR8393371 SRR8393372 SRR8393373 SRR8393374
TDVGHKKNGANGGI
TDVGHKKNGANGDV
TDVGHKKNGANGTV
TDVGHKKNAAPDGI
TDVGHKKENAANVNGTV
TDVGHKKENAANVNGTI
TDVGHKKENAANVNGI
TDVGHKKDGAQGTV
TDVGHKKNGANGDI
TDVGHKKENAANGDI
TDVGRKKDGAQGTV
TDVGRKKDGANGDI
100%
80%
60%
40%
20%

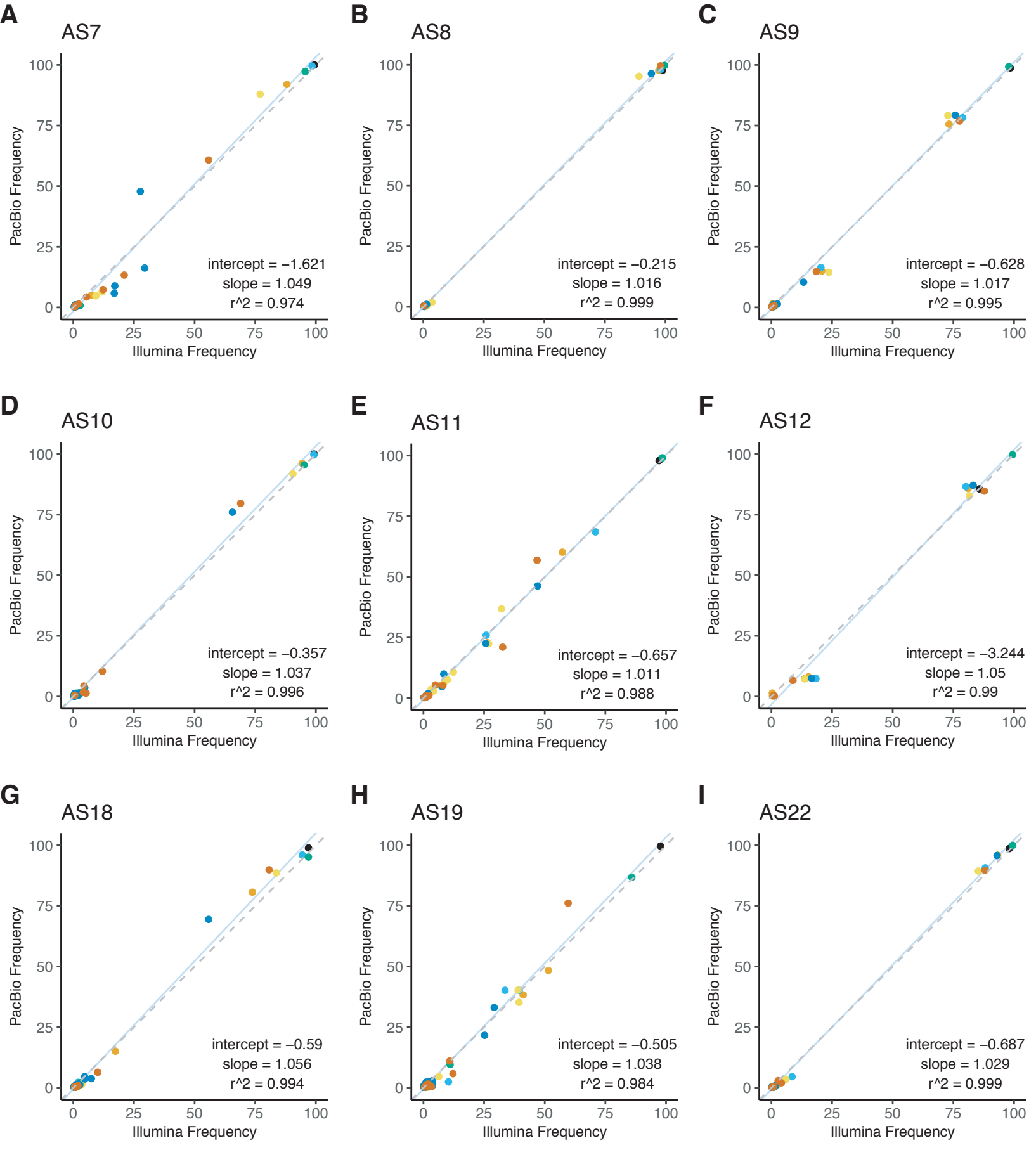

V5
A510
A511
A512
A518
A519
A522
A57
A58
A59
M01
M04
M05
M06
SRR1392234
SRR1392235
SRR1392236
SRR1392237
SRR1392238
SRR1392239
SRR1392240
SRR1392241
SRR1392242
SRR1392243
SRR1392244
SRR1392245
SRR1392246
SRR1392247
SRR1393361
SRR1393362
SRR1393363
SRR1393364
SRR1393365
SRR1393366
SRR1393367
SRR1393368
SRR1393369
SRR1393370
SRR1393371
SRR1393372
SRR1393373
SRR1393374
QASNVFQGVFLTTPMQTHDCAAYI
KASNVFQGVFLTTPMQTHDCAAYI
KASNVFKDVFLTDTTPMRTHDCAAYI
KASNVFGGVFLTNNMTAHDCAAYI
KASNVFGGVFLNMAMTAHDCAAYI
KASNVFEGVFLTNNMTAHDCAAYI
KASNVFEGVFLTNAMTAHDCAAYI
QASNVFKDVFLTNAMDMQKDDCAAYI
QASNVFQGVFLTTPMQKHDCAAYI
QASNVFQGVFLTTPMQKDDCAAYI
QASNVFQGVFLTTPMLQHDCAAYI
QASNVFQGVFLTNNMLQHDCAAYI
QASNVFQGVFLTNATPMQKDDCAAYI
QASNVFQGVFLTNAMDMQKDDCAAYI
QASNVFQGVFLTDTTPMQKDDCAAYI
QASNVFQGVFLNMAMTAHDCAAYI
QASNVFEGVFLTNNMLQHDCAAYI
QASNVFEGVFLARNMAMTAHDCAAYI
QASNVFEGVFLARNIAMTAHDCAAYI
QASNVFEGVFLARNIAMREHDCAAYI
KASNVFQGVFLTTPMQKDDCAAYI
KASNVFQGVFLTDTTPMQKDDCAAYI
KASNVFKDVFLTTPMLQHDCAAYI
KASNVFKDVFLTNNMLQHDCAAYI
KASNVFKDVFLTNAMDMQKDDCAAYI
KASNVFKDVFLNMAMTAHDCAAYI
KASNVFEGVFLTTPMLQHDCAAYI
KASNVFEGVFLARNMAMTAHDCAAYI
KASNVFEGVFLARNIAMTAHDCAAYI
KASNVFEGVFLNMAMTAHDCAAYI
KASNVFKDVFLNAMDMQTHDCAAYI
KASNVFGGVFLTNNMLQHDCAAYI
KASNVFEGVFLARNIAMREHDCAAYI
100%
80%
60%
40%
20%
V6
A510
A511
A512
A518
A519
A522
A57
A58
A59
M01
M04
M05
M06
SRR1392234
SRR1392235
SRR1392236
SRR1392237
SRR1392238
SRR1392239
SRR1392240
SRR1392241
SRR1392242
SRR1392243
SRR1392244
SRR1392245
SRR1392246
SRR1392247
SRR1393361
SRR1393362
SRR1393363
SRR1393364
SRR1393365
SRR1393366
SRR1393367
SRR1393368
SRR1393369
SRR1393370
SRR1393371
SRR1393372
SRR1393373
SRR1393374
PVYYFAARAQLPALPGAAANI
PVYYFAARALPGAPVPAI
PVYYFAARALAAGADI
PVYYFAAQPPADI
PVYYFAAQAPAAANI
PVYYFAAGAAPGAPAAANI
PVHYTVLTVPPAAANI
PVHYTVLTGPQAPAGVPPPARVDI
PVHYTVLTGPPAQPPANI
PVHYKVLKAHAPVPAI
PVHWNAFTQAPAPAQPPANI
PVHWKVLPGPQAPAAANI
PVHWKAQARARAPADI
PVHWKALPGADI
PVHWKALARARPQAPAAANI
PVHWKALARARPPAQPPANI
PVHWKALARARAPADI
PVHWKALAQAPAAANI
PVHWKALAPARVDI
PVHAQARARAGADI
PVHAQARALAAAVNNDI
PVHAQAQLPAVAAAAVNNDI
PVHAQAAGAAAGAGVDI
PVHAHARAPVPAI
PVHWKAPAQPPAI
PVYYFAARAGAAVPAPANI
PVHYTVLPGAPAAANI
PVHAQARALAAGVPPAAAV
PVYYFAARVDI
PVYYFAARARAGAAVPAI
PVYYFAALPGAAAGVPAI
PVHYKVLKAPAPANI
PVHYKVLKAPAAANI
PVHWNAFTQARARAGAAVPAI
PVHWNAFTQARALAAGADI
PVHWNAFTQAPAAANI
PVHWNAFTQAHARALAAGVPAI
PVHWKPPARVDI
PVHWKAPAPANI
PVHWKALARAVPPANDI
PVHWKALARAVPPAAADI
PVHWKALARARPPAPANDI
PVHWKALARARPPAAANI
PVHWKALARAGAAVPAI
PVHWKAHARAPADI
PVHARARALAADI
PVHAQAQAPAAADI
PVHAQAPVPAI
PVYYFAARARPGAAADDI
100%
80%
60%
40%
20%

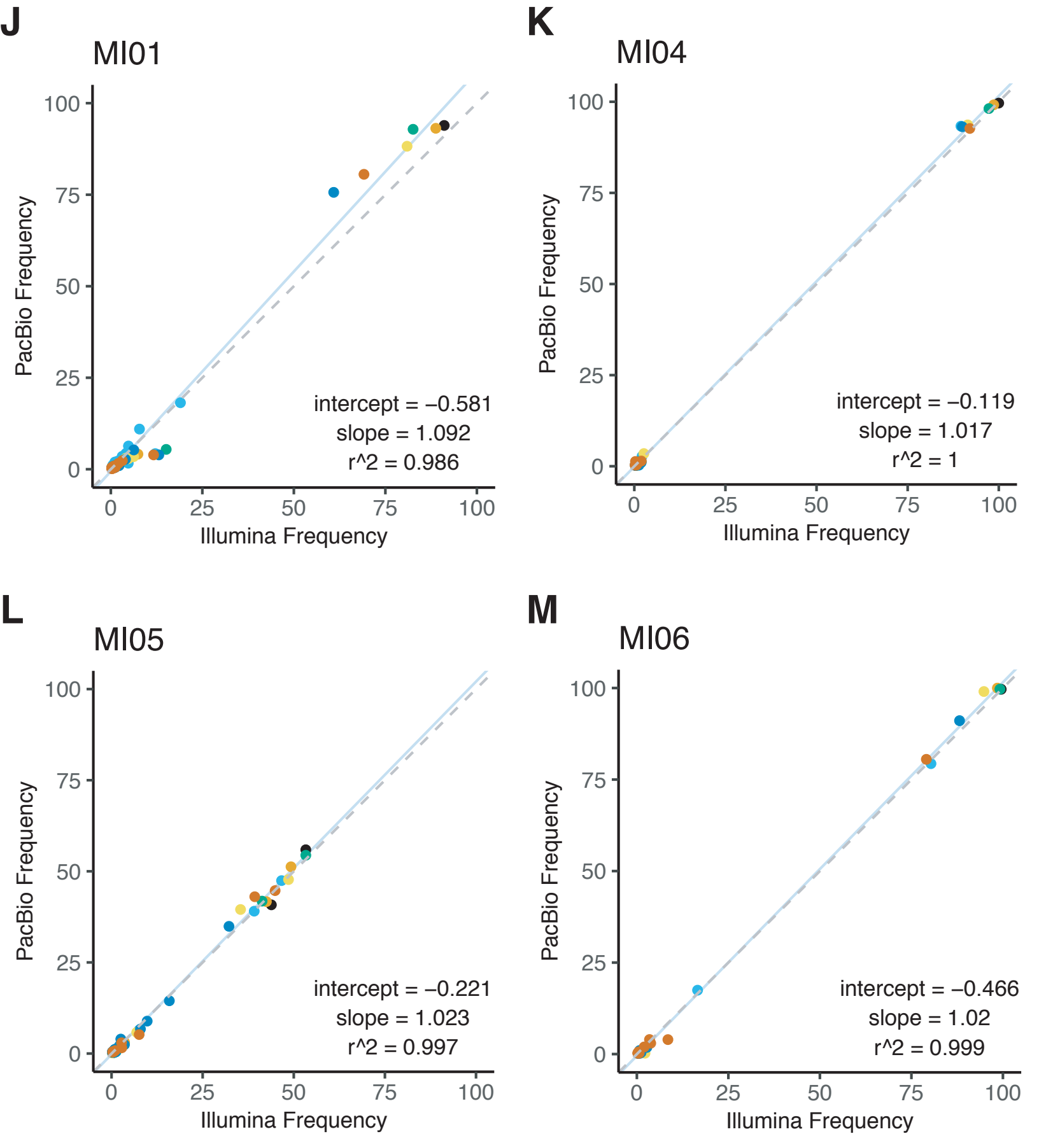

V7
A510
A511
A512
A518
A519
A522
A57
A58
A59
M01
M04
M05
M06
SRR8139234
SRR8139235
SRR8139236
SRR8139237
SRR8139238
SRR8139239
SRR8139240
SRR8139241
SRR8139242
SRR8139243
SRR8139244
SRR8139245
SRR8139246
SRR8139247
SRR8393361
SRR8393362
SRR8393363
SRR8393364
SRR8393365
SRR8393366
SRR8393367
SRR8393368
SRR8393369
SRR8393370
SRR8393371
SRR8393372
SRR8393373
SRR8393374
YGGTNKQAAPGAVPGTKWKAEYCGYY
YGGTNKQAAAAPAAPTKWKAEYCGYY
YGGTNKKNDAAPGAAAPGTKWKAGYCGYY
YGGTNKKNDAAPAAPTKWKAGYCGYY
YGGTNKKNDAAPAAPAAPTKWKAEYCGYY
YGGTNKKATPPAAPTKWKAGYCGYY
YGGTNKKATPPAAAVPGTKWSKEYCGYY
YGGTNKKAAAPAAPTKWKAEYCGYY
YGGTNKKAAAAAPVPATKWKAEYCGYY
YGGTNKQAAAPTKWKAGYCGYY
YGGTNKKAAAAAPAAPTKWKAGYCGYY
YGGTNKQAAAVPGVVPGAVPALTKWKAGYCGYY
YGGTNKQAAAVPGVVPAAPAAPATKWKAEYCGYY
YGGTNKQAAPGAVPGTKWKAEYCGYY
YGGTNKQAAALTKWKAEYCGYY
YGGTNKKNDAVPGAPATKWKAEYCGYY
YGGTNKKNDAPATKWKAEYCGYY
YGGTNKKNDAAPTKWKAEYCGYY
YGGTNKKNDAAPGTKWKAEYCGYY
YGGTNKKNDAAPGAPATKWKAEYCGYY
YGGTNKKNDAAPGAAPAPGTKWKAEYCGYY
YGGTNKKNDAPATKWSKEYCGYY
YGGTNKKNDAAPAAVPGTKWKAEYCGYY
YGGTNKKNDAAPAAPTKWKAEYCGYY
YGGTNKKNDAAPAALTKWKAGYCGYY
YGGTNKKNDAALTKWKAGYCGYY
YGGTNKKNDAALTKWKAGYCGYY
YGGTNKKNDAAAAAPAPGTKWKAEYCGYY
YGGTNKKATPPAAPTKWKAEYCGYY
YGGTNKKATPPAAPATKWKAEYCGYY
YGGTNKKATPPAAAPAAPTKWKAEYCGYY
YGGTNKKATPPAAAVPGTKWKAEYCGYY
YGGTNKKAAPTKWKAEYCGYY
YGGTNKKAAALTKWKAEYCGYY
YGGTNKKAAAAAPAPGTKWKAGYCGYY
YGGTNKKAAAAAPAPGTKWKAGYCGYY
YGGTNKKAAAAAPAPATKWKAEYCGYY
YGGTNKKNDAAPATKWKAEYCGYY
YGGTNKKNDAAAAAPAPATKWKAEYCGYY
100%
80%
60%
40%
20%


A
C
B
Usage (sum percentage)
Region
V1
V2
V3
V4
V5
V6
V7
Locus Nucleotide
Length (bp)
hp
tprD
ktrA
ktrB
ogt
potA
hypothetical protein
CP045004.1


A
100
75
50
25
0
Percentage
V1
V2
V3
V4
V5
V6
V7
Region
Segments
1
2
3
B
V1
V2
V3
V4
V5
V6
V7
100
75
50
25
0
Percentage
V1-DS7
V1-DS15
V1-DS20
V1-DS29
V1-DS37
V2-DS8
V2-DS16
V2-DS21
V2-DS30
V2-DS38
V3-DS3
V3-DS10
V3-DS12
V3-DS18
V3-DS24
V3-DS26
V3-DS32
V3-DS34
V3-DS41
V3-DS43
V3-DS49
V3-DS50
V3-DS52
V4-DS6
V4-DS14
V4-DS36
V4-DS47
V4-DS48
V5-DS5
V5-DS13
V5-DS28
V5-DS35
V5-DS46
V5-DS56
V6-DS1
V6-DS4
V6-DS11
V6-DS19
V6-DS23
V6-DS25
V6-DS33
V6-DS40
V6-DS42
V6-DS44
V6-DS51
V6-DS53
V6-DS54
V6-DS55
V7-DS2
V7-DS9
V7-DS17
V7-DS22
V7-DS27
V7-DS31
V7-DS39
Variable Region Donor Site
Position
First
Second
Third


A
B
C
D
V1
V2
V3
V4
V5
V6
V7
Frequency
0
25
50
75
100
AS10
AS10_high
AS11
AS11_high
AS12
AS12_high
AS18
AS18_high
Sample


A
AS7
intercept = −1.621
slope = 1.049
r^2 = 0.974
B
AS8
intercept = −0.215
slope = 1.016
r^2 = 0.999
C
AS9
intercept = −0.628
slope = 1.017
r^2 = 0.995
D
AS10
intercept = −0.357
slope = 1.037
r^2 = 0.996
E
AS11
intercept = −0.657
slope = 1.011
r^2 = 0.988
F
AS12
intercept = −3.244
slope = 1.05
r^2 = 0.99
G
AS18
intercept = −0.59
slope = 1.056
r^2 = 0.994
H
AS19
intercept = −0.505
slope = 1.038
r^2 = 0.984
I
AS22
intercept = −0.687
slope = 1.029
r^2 = 0.999
PacBio Frequency
Illumina Frequency
0
25
50
75
100


J
MI01
PacBio Frequency
Illumina Frequency
intercept = −0.581
slope = 1.092
r^2 = 0.986
K
MI04
PacBio Frequency
Illumina Frequency
intercept = −0.119
slope = 1.017
r^2 = 1
L
MI05
PacBio Frequency
Illumina Frequency
intercept = −0.221
slope = 1.023
r^2 = 0.997
M
MI06
PacBio Frequency
Illumina Frequency
intercept = −0.466
slope = 1.02
r^2 = 0.999

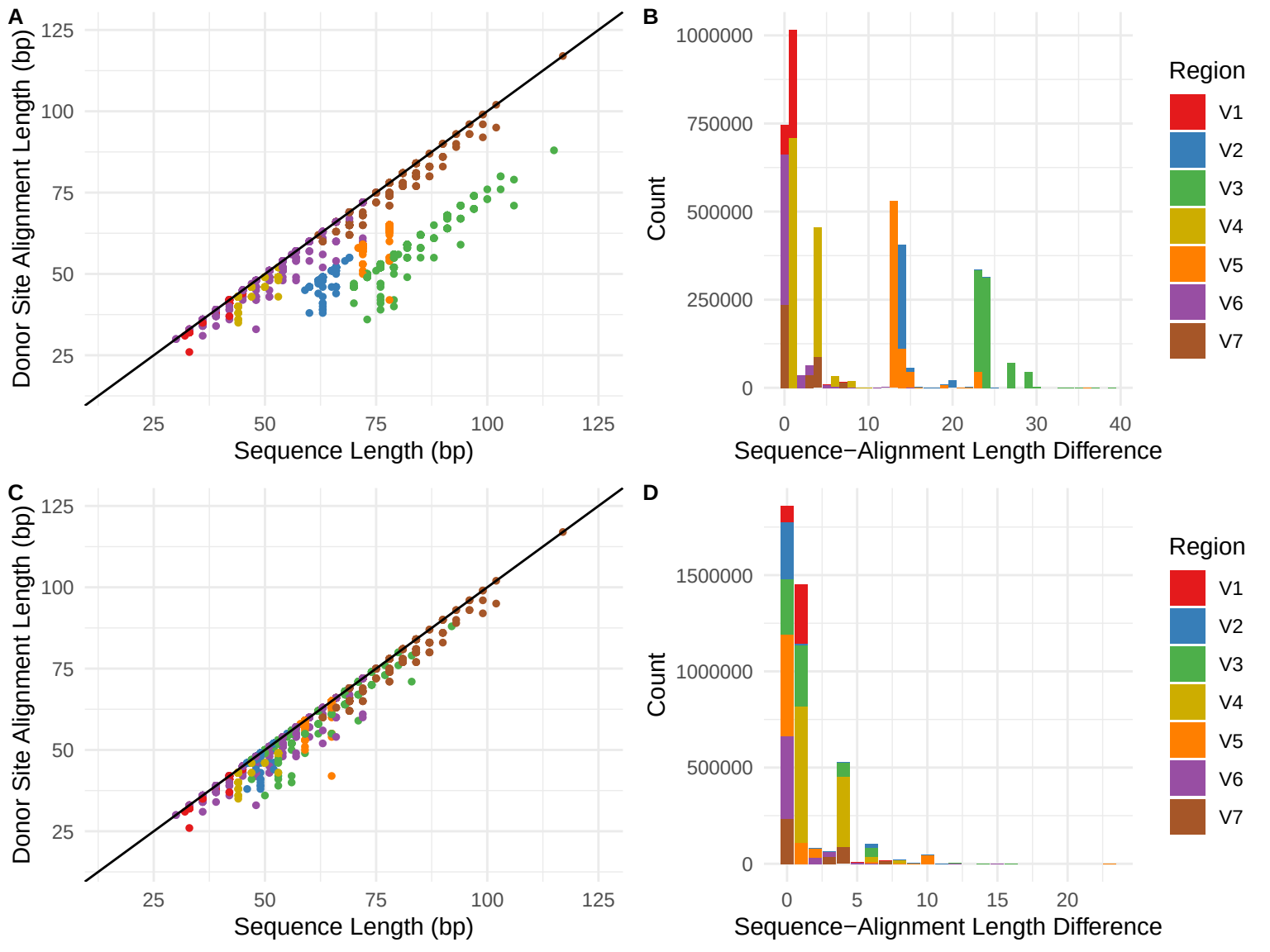

A
Donor Site Alignment Length (bp)
125
100
75
50
25
25
50
75
100
125
Sequence Length (bp)
B
Count
1000000
750000
500000
250000
0
0
10
20
30
40
Sequence−Alignment Length Difference
Region
V1
V2
V3
V4
V5
V6
V7
C
Donor Site Alignment Length (bp)
125
100
75
50
25
25
50
75
100
125
Sequence Length (bp)
D
Count
1500000
1000000
500000
0
0
5
10
15
20
Sequence−Alignment Length Difference
Region
V1
V2
V3
V4
V5
V6
V7

Table S1. PacBio barcoded *tprK* primers used in this study.

Primer Name	Sequence (5' -> 3')
tprK-F_bc1001	CACATATCAGAGTGCGGGAAAGAAAAGAACCATACATCC
tprK-F_bc1002	ACACACAGACTGTGAGGGAAAGAAAAGAACCATACATCC
tprK-F_bc1003	ACACATCTCGTGAGAGGGAAAGAAAAGAACCATACATCC
tprK-F_bc1004	CACGCACACACGCGCGGGAAAGAAAAGAACCATACATCC
tprK-F_bc1005	CACTCGACTCTCGCGTGGAAAGAAAAGAACCATACATCC
tprK-F_bc1006	CATATATATCAGCTGTGGAAAGAAAAGAACCATACATCC
tprK-F_bc1007	TCTGTATCTCTATGTGGGAAAGAAAAGAACCATACATCC
tprK-F_bc1008	ACAGTCGAGCGCTGCGGGAAAGAAAAGAACCATACATCC
tprK-R_bc1009	ACACACGCGAGACAGACGCAGTTCCGGATTCTG
tprK-R_bc1010	ACGCGCTATCTCAGAGCGCAGTTCCGGATTCTG
tprK-R_bc1011	CTATACGTATATCTATCGCAGTTCCGGATTCTG
tprK-R_bc1012	ACACTAGATCGCGTGTCGCAGTTCCGGATTCTG
tprK-R_bc1013	CTCTCGCATACGCGAGCGCAGTTCCGGATTCTG
tprK-R_bc1014	CTCACTACGCGCGCGTCGCAGTTCCGGATTCTG
tprK-R_bc1015	CGCATGACACGTGTGTCGCAGTTCCGGATTCTG
tprK-R_bc1016	CATAGAGAGATAGTATCGCAGTTCCGGATTCTG

Table S2. SRA Accessions for Sequencing Libraries

SRA Accession	Library
SRR10953926	AS7_tprk_nextera
SRR10953925	AS7_tprk_pacbio
SRR10953914	AS8_tprk_nextera
SRR10953903	AS8_tprk_pacbio
SRR10953902	AS9_tprk_nextera
SRR10953901	AS9_tprk_pacbio
SRR10953900	AS10_tprk_nextera
SRR10953899	AS10_tprk_nextera_highcopy
SRR10953898	AS10_tprk_pacbio
SRR10953897	AS11_tprk_nextera
SRR10953924	AS11_tprk_nextera_highcopy
SRR10953923	AS11_tprk_pacbio
SRR10953921	AS12_tprk_nextera_highcopy
SRR10953922	AS12_tprk_nextera
SRR10953920	AS12_tprk_pacbio
SRR10953919	AS18_tprk_nextera
SRR10953918	AS18_tprk_nextera_highcopy
SRR10953917	AS18_tprk_pacbio
SRR10953916	AS19_tprk_nextera
SRR10953915	AS19_tprk_pacbio
SRR10953913	AS22_tprk_nextera
SRR10953912	AS22_tprk_pacbio
SRR10953911	MI01_tprk_nextera
SRR10953910	MI01_tprk_pacbio
SRR10953909	MI04_tprk_nextera
SRR10953908	MI04_tprk_pacbio
SRR10953907	MI05_tprk_nextera
SRR10953906	MI05_tprk_pacbio
SRR10953905	MI06_tprk_nextera
SRR10953904	MI06_tprk_pacbio

Table S3. Individual metadata for strains sequenced in this study

Sample ID	Tp0548 sequ	TprEGJ type	ARP type	23S rRNA PCI	23S A2058G	23S A2059G	Strain Type (r	Macrolide re	16S Tetracyc	Gender	Age	MSM	HIV status	Syphilis Stag	Lesion Locati	Non-treponer	Treponemal test performed at sample collection	Sexual networking/travel history data?	First syphilis diagnosis of previously diagnosed?
AS7	G	D	14	POSITIVE	PRESENT	ABSENT	14D/G	YES	Negative	M	26	yes	neg	primary	genital	RPR neg	TPPA1:80	new sex partner with suspected syphilis-4 prtrs last 6 mo	first time diagnosed -
AS8	G	D	14	POSITIVE	PRESENT	ABSENT	14D/G	YES	Negative	M	22	yes	neg	primary	genital	RPR neg	TPPA1:640	4 prtrs last 6 mo	first time diagnosed -
AS9	G	D	14	POSITIVE	PRESENT	ABSENT	14D/G	YES	Negative	M	51	yes	pos	primary	anal	RPR 1:8	TPPA not done	HIV positive partner plus 8 ptrs last six mo.	syphilis in 2009 -no others episodes till now
AS10	G	D	14	POSITIVE	PRESENT	ABSENT	14D/G	YES	Negative	M	41	yes	pos	secondary	genital	RPR 1:16	TPPA1:20480	8 ptrs last six mo.	first time diagnosed -
AS11	G	D	14	POSITIVE	PRESENT	ABSENT	14D/G	YES	Negative	M	32	yes	neg	secondary	anal	RPR 1:16	TPPA1:20480	Sex contacts in north-east Italy; 13 ptrs last six mo.	first time diagnosed -
AS12	G	D	14	POSITIVE	PRESENT	ABSENT	14D/G	YES	Negative	M	54	yes	pos	primary	anal	RPR neg	TPPA not done	frequent sex contacts in Milan in saunas ; 60 ptrs last six mo.	first time diagnosed -
AS18	G	D	14	POSITIVE	PRESENT	ABSENT	14D/G	YES	Negative	M	33	yes	pos	secondary	genital	RPR 1:32	TPPA not done	habitual cruising ;saunas; regular gay chat; 20 ptrs last six mo.	secondary syphilis in 2014 no other episodes till now
AS19	G	D	14	POSITIVE	PRESENT	ABSENT	14D/G	YES	Negative	M	42	yes	pos	secondary	anal	RPR 1:16	TPPA not done	habitual cruising ;saunas; regular gay chat 35 ptrs last six mo.	syphilis in 1996 no others episodes till now
AS22	G	D	13	POSITIVE	PRESENT	ABSENT	13D/G	YES	Negative	M	35	yes	neg	primary	genital	RPR neg	TPPA1:20480	only one regular partner last 6 mo.	first time diagnosed
MI01	F	D	6	POSITIVE	PRESENT	ABSENT	6D/F	YES	Negative	M	39	yes	pos	secondary	anal	VDRL 1:32	TPHA 1:80	milano	no
MI04	G	D	13	POSITIVE	PRESENT	ABSENT	13D/G	YES	Negative	M	20	yes	neg	primary	genital	VDRL 1:4	TPHA 1:80	milano	no
MI05	D	D	13	POSITIVE	PRESENT	ABSENT	13D/D	YES	Negative	M	54	yes	pos	secondary	anal	VDRL 1:32	TPHA 1:80	canarie	no
MI06	G	D	13	POSITIVE	PRESENT	ABSENT	13D/G	YES	Negative	M	57	yes	pos	primary	genital	VDRL 1:16	TPHA 1:80	milano	si

Table S4. Comparison of the number of variable region sequences and diversity measures identified the 7 variable regions based on t

		V1			V2			V3			V4			V5			V6			V7			Total		
	Copies of treponemal DNA input into *tprK* PCR	No. of variable region sequences	Eveness	Shannon Div	No. of variable region sequences	Eveness	Shannon Div	No. of variable region sequences	Eveness	Shannon Div	No. of variable region sequences	Eveness	Shannon Div	No. of variable region sequences	Eveness	Shannon Div	No. of variable region sequences	Eveness	Shannon Div	No. of variable region sequences	Eveness	Shannon Div	No. of variable region sequences	Eveness	Shannon Div
AS10	1000	3	0.042	0.047	10	0.141	0.324	3	0.047	0.052	6	0.142	0.254	14	0.202	0.532	39	0.485	1.776	16	0.437	1.210	91	1.495	4.196
AS10_high	5362	1	0	0	8	0.167	0.347	2	0.050	0.035	5	0.101	0.163	12	0.198	0.492	33	0.415	1.453	15	0.375	1.016	76	1.306	3.504
AS11	1000	5	0.103	0.165	13	0.498	1.276	4	0.524	0.726	2	0.113	0.078	12	0.725	1.801	16	0.578	1.603	13	0.551	1.414	65	3.091	7.064
AS11_high	2736	7	0.076	0.148	10	0.470	1.082	4	0.522	0.724	4	0.054	0.074	9	0.771	1.695	16	0.533	1.477	9	0.554	1.218	59	2.980	6.417
AS12	1000	3	0.391	0.430	9	0.293	0.643	5	0.351	0.565	3	0.036	0.040	7	0.330	0.642	3	0.422	0.464	7	0.253	0.491	37	2.076	3.276
AS12_high	6663	4	0.254	0.352	7	0.271	0.527	5	0.303	0.488	5	0.054	0.087	8	0.287	0.596	2	0.618	0.428	8	0.242	0.503	39	2.029	2.983
AS18	1000	5	0.107	0.173	9	0.406	0.891	7	0.158	0.308	3	0.135	0.148	16	0.303	0.841	40	0.561	2.069	16	0.307	0.850	96	1.706	4.593
AS18_high	1013	4	0.068	0.095	10	0.336	0.774	10	0.124	0.285	4	0.122	0.169	13	0.304	0.780	40	0.517	1.907	12	0.234	0.582	93	1.977	5.280

Table S5. Number of variable region sequences and diversity measures for the 7 variable regions of TprK for the 13 strains profiled in this study.

	V1			V2			V3			V4			V5			V6			V7			Total		
	No. of variable region sequences	Eveness	Shannon Diversity	No. of variable region sequences	Eveness	Shannon Diversity	No. of variable region sequences	Eveness	Shannon Diversity	No. of variable region sequences	Eveness	Shannon Diversity	No. of variable region sequences	Eveness	Shannon Diversity	No. of variable region sequences	Eveness	Shannon Diversity	No. of variable region sequences	Eveness	Shannon Diversity	No. of variable region sequences	Eveness	Shannon Diversity
AS7	2	0.052	0.036	9	0.241	0.529	4	0.07	0.097	5	0.141	0.226	7	0.396	0.771	20	0.59	1.767	11	0.55	1.319	58	2.04	4.745
AS8	3	0.069	0.076	7	0.088	0.172	7	0.067	0.131	2	0.034	0.023	15	0.218	0.591	15	0.134	0.363	6	0.072	0.13	55	0.682	1.486
AS9	4	0.068	0.094	11	0.356	0.853	3	0.499	0.548	5	0.082	0.131	8	0.359	0.746	23	0.325	1.019	11	0.298	0.714	65	1.987	4.105
AS10	3	0.042	0.047	10	0.141	0.324	3	0.047	0.052	6	0.142	0.254	14	0.202	0.532	39	0.485	1.776	16	0.437	1.21	91	1.496	4.195
AS11	5	0.103	0.165	13	0.498	1.276	4	0.524	0.726	2	0.113	0.078	12	0.725	1.801	16	0.578	1.603	13	0.551	1.414	65	3.092	7.063
AS12	3	0.391	0.43	9	0.293	0.643	5	0.351	0.565	3	0.036	0.04	7	0.33	0.642	3	0.422	0.464	7	0.253	0.491	37	2.076	3.275
AS18	5	0.107	0.173	9	0.406	0.891	7	0.158	0.308	3	0.135	0.148	16	0.303	0.841	40	0.561	2.069	16	0.307	0.85	96	1.977	5.28
AS19	4	0.095	0.132	14	0.406	1.072	26	0.53	1.727	7	0.268	0.521	22	0.519	1.603	65	0.661	2.76	24	0.497	1.578	162	2.976	9.393
AS22	5	0.074	0.119	10	0.159	0.367	7	0.249	0.484	3	0.033	0.036	12	0.278	0.692	11	0.167	0.401	13	0.236	0.604	61	1.196	2.703
MI01	7	0.207	0.402	6	0.259	0.465	43	0.813	3.059	6	0.311	0.557	11	0.342	0.821	25	0.5	1.608	30	0.412	1.401	128	2.844	8.313
MI04	1	0	0	3	0.074	0.081	12	0.23	0.571	3	0.128	0.14	8	0.212	0.441	20	0.2	0.599	13	0.178	0.458	60	1.022	2.29
MI05	6	0.467	0.836	10	0.458	1.055	12	0.493	1.224	9	0.435	0.956	14	0.494	1.303	43	0.694	2.608	10	0.546	1.257	104	3.587	9.239
MI06	2	0.042	0.029	5	0.061	0.098	8	0.305	0.634	3	0.048	0.052	6	0.156	0.279	16	0.238	0.659	14	0.337	0.89	54	1.187	2.641