Title

Phylo-Plex: A phylogenetically informed, low-cost amplicon sequencing platform for deployable high-resolution genomic epidemiology

Authors

Mathew A Beale[1,*], Vignesh Shetty[1], Kirsty E Ambridge[1], George Lacey[1], Sam Dougan[1], William Roberts-Sengier[1], Beth Sampher[1], Florent Lassalle[1], Matthew J Dorman[1,2,3], Mahlape P Mahlangu[4], Johanna ME Venter[4], Bianca Da Costa Dias[4], Martha Chipinduro[5,6], Tendai M. Washaya[5], Luanne Rodgers[5], Beauty Makamure[5], Ethel Dauya[5], Michael Marks[7,8,9], Etienne E. Müller[4], Rashida A Ferrand[5,7], Nicholas R Thomson[1,7]

[1]Parasites and Microbes Programme, Wellcome Sanger Institute, Cambridgeshire, UK

[2] School of Mathematical and Statistical Sciences, College of Science and Engineering, University of Galway, University Road, Galway, Ireland

[3]Department of Microbiology, Moyne Institute of Preventive Medicine, School of Genetics and Microbiology, Trinity College Dublin, Dublin, Ireland

[4]Centre for HIV & STIs, National Institute for Communicable Diseases, Johannesburg, South Africa

[5]Biomedical Research and Training Institute, Harare, Zimbabwe

[6]Faculty of Medicine and Health Sciences, Midlands State University, Gweru, Zimbabwe

[7]Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, UK

[8]Hospital for Tropical Diseases, University College London Hospital, London, UK

[9]Division of Infection and Immunity, University College London, London, UK

*Correspondence to mathew.beale@sanger.ac.uk

Abstract

Genomic pathogen surveillance is a powerful tool for public health and research, but is costly and unachievable in low-resource settings. Most sub-genomic typing methods sacrifice

resolution whilst remaining costly. We developed "Phylo-Plex", a novel approach that identifies information-rich genomic regions to maximise phylogenetic information whilst minimising the number of regions. Applied to *Treponema pallidum* and *Neisseria gonorrhoeae*, we designed a high-resolution multiplex PCR sequencing scheme for lineage tracking pathogens with different extremes of genome variation. For *Treponema pallidum*, we also designed and evaluated the Phylo-Plex scheme in the laboratory and field settings by sequencing 72 clinical samples using MinION Flongle cells. Our *T. pallidum* scheme comprising 59 multiplex amplicons achieved high discrimination of fine-scale sublineages comparable to those defined using whole genomes, and demonstrating a qPCR detection limit ≤Ct 32. Variant calls from MinION amplicon sequencing were highly correlated with Illumina whole genome sequencing. We successfully deployed the method in a low-resource laboratory in Zimbabwe, costed at <£300/24 samples (£12.47/sample). Phylo-Plex enables low-cost tracking of priority pathogenic lineages in low resource settings and at scale.

Introduction

Genomic pathogen surveillance has proven itself a powerful tool for both public health epidemiology and research[1]. Discriminating closely related isolates and deconvoluting complex samples and communities can enable finescale global tracking of pathogen populations, provide insights into epidemiology, transmission and the impact of clinical interventions such as antimicrobial usage or vaccines[2]. However, whole genome sequencing (WGS) pathogens at scale has substantial barriers to implementation including cost, complexity of data analysis, and time to results. This is particularly the case for low-resource settings or where pathogens cannot be readily cultured, necessitating direct or enriched metagenomic sequencing from clinical swabs and complex analytics[3,4]. Currently, the only molecular alternatives to WGS are typing approaches such as single gene or multi-locus sequence typing (MLST). Although designed for very different purposes, compared to WGS, typing is by definition comparatively low resolution and is not responsive to the evolutionary dynamics of the target pathogen[5–8]. A solution to this is multiplex PCR[9–12] approaches such as the ARTIC network protocol (https://artic.network/)[11], which amplifies the whole genome in a series of tiled amplicons, and was used during the SARS-CoV-2 pandemic to generate over 18 million complete viral genome sequences. This approach allowed near real-time tracking and, because of the small size of the genome (30Kb), for comprehensive identification of variants of interest and concern globally. However, this approach

is not cost effective for generating larger complete bacterial genomes which range between 1-8 Mb in length, necessitating a more targeted approach.

Syphilis, caused by the bacterium *Treponema pallidum subspecies pallidum*, is an important sexually transmitted infection. Globally, *T. pallidum* causes 8 million new cases per annum and 700,000 cases of congenital syphilis leading to 220,000 early foetal deaths, stillbirths or neonatal deaths[13]. In most countries diagnosis of syphilis is restricted to either serological testing (in the context of antenatal care) or syndromic management (in the context of symptomatic individuals). *T. pallidum* is extremely difficult to culture[14] and so whole genome-based surveillance of syphilis is not widely used, even in high resource settings. Until now the majority of available genomes have been sequenced with costly sequence capture enriched metagenomic approaches which require detailed bioinformatic analysis and significant computational infrastructure[8,15–17]. *T. pallidum* is subdivided into three phylogenetically distinct subspecies (subspecies *pallidum*, *pertenue* and *endemicum*), and subspecies *pallidum* comprises two main lineages (Nichols and SS14), as well as a number of sublineages[15]. Molecular typing schemes include a variable tandem repeat method[18], a three-gene MLST scheme[7], and a recently described seven-gene MLST[19].

In this study, we developed and employed a novel whole genome informed approach for selecting optimal genome regions which can be targeted for multiplex PCR, dubbed “Phylo-Plex”. Given a representative collection of pathogen genomes, Phylo-Plex clusters single nucleotide polymorphisms (SNPs) along lineage-defining phylogenetic branches into information-rich amplicons which maximise resolution whilst minimising the number of genomic regions required for sequencing. Whilst Phylo-Plex is not intended to replace formal typing schemes, such as MLST, this approach yields high-resolution phylogenies which recapitulate population structures obtained from WGS data, and can be performed quickly and easily at a fraction of the cost. We applied the Phylo-Plex approach to identify optimal regions for PCR sequencing for both a low diversity pathogen (*T. pallidum*) and a high diversity pathogen (*Neisseria gonorrhoeae*). For *T. pallidum*, we then developed a high resolution Phylo-Plex scheme comprising 59 multiplex amplicons. We deployed our method in Harare, Zimbabwe and show how, when paired with low-cost nanopore sequencing, this approach provides a fast and flexible framework for SNP-based molecular sequence typing, that can be performed in a low resource setting at low cost. The value of Phylo-Plex is in its universal applicability to any pathogen where we have existing genomic data and key lineages need to be found or priority pathogens need to be tracked in any setting.

Results

The Phylo-Plex approach we developed here takes WGS data from a representative global population framework, and identifies all discriminating SNPs linked to the major topological features (lineages, sublineages, etc) within the phylogeny. Then, regardless of the sublineage supported, all discriminatory SNPs are positionally clustered along the genome and used in a hierarchical selection algorithm to find the optimal SNP set which maximises the discriminatory power for each lineage, whilst minimising the number of amplicons required. Candidate regions are then used for multiplex PCR primer design, which can be validated *in silico* and *in vitro* before deployment (Figure 1).

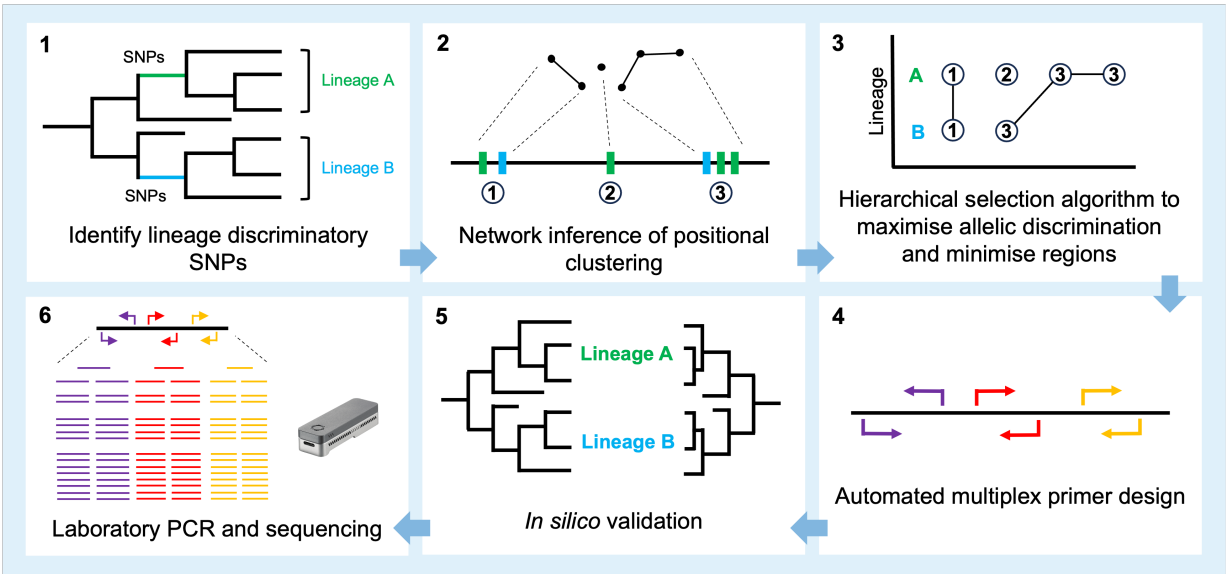


Figure 1. Overview of Phylo-Plex workflow. 1 – Pathogen genome populations are analysed to determine SNPs that delineate individual lineages along ancestral branches. 2 – Discriminatory SNPs are clustered based on their genome position to identify information-rich candidate regions. 3 – Hierarchical selection of candidate regions maximises discrimination for each lineage whilst minimising the number of regions chosen. 4 – Selected regions are used for automated multiplex primer design. 5 – Primer designs are validated *in silico* to ensure the final scheme will amplify efficiently and without contamination, and that genome population structures are accurately reconstructed. 6 – Primers are then optimized and validated in the laboratory using sequencing before deployment.

Identifying population structure defining SNPs

We constructed a globally representative dataset of 607 publicly available *T. pallidum* genomes (as of 2022), including high-quality genomes with ≥95% of reference positions at ≥5X coverage, and spanning *T. pallidum* subspecies *pallidum* (TPA, n=530), *T. pallidum* subspecies *pertenue* (TPE, n=75) and *T. pallidum* subspecies *endemicum* (TEN, n=2) (Supplementary Table 1). Notably,

121 the dataset was enriched for genomes from North America and Europe, and underrepresented
122 for genomes from lower- and middle-income countries, particularly in Africa, reflecting biases in
123 publicly available data[15]. Sequencing reads were mapped to a common reference genome
124 (NC_021508.1) and hypervariable and recombining regions were masked using an initial list of
125 known recombinant genes, followed by *de novo* inference of recombination using Gubbins[20]. We
126 inferred a whole genome phylogeny using IQ-tree[21,22], and designated phylogenetic sublineages
127 using rPinecone[23] v0.1.0 with a threshold of up to 10 SNPs from the common ancestral node. This
128 resulted in 40 sublineages and 33 singletons, including the 17 sublineages previously defined
129 within TPA[15], and 23 newly-defined sublineages within TPE (Supplementary Figure 1).
130
131 Next, within this population structure we identified vertically inherited discriminatory SNPs by
132 applying population genomic statistical tests of allelic segregation to a recombination-masked
133 alignment, using Fixation Index[24] (F_{ST}) to define a list of SNPs that discriminate sublineages. Given
134 that most *T. pallidum* genomic data are metagenomically derived and frequently contain
135 uncertain genomic positions due to low coverage, this approach enabled delineation of
136 informative SNPs whilst accommodating missing data. Retaining SNPs with F_{ST} ≥0.9, a
137 combination of multiple SNPs together enables robust identification of each lineage or
138 sublineage (Supplementary Figure 2). We applied this approach to each of the 40 sublineages
139 defined in our population, as well as to delineate subspecies (TPA and TPE) and TPA major
140 lineages (Nichols and SS14), resulting in a single list comprising 1549 discriminatory SNPs, of
141 which 855 occurred on phylogenetic branches leading to individual sublineages, and thus
142 discriminatory for sublineages, whilst the remaining 694 discriminatory SNPs fell on deeper
143 branches and distinguished *T. pallidum* subspecies and major lineages (Supplementary Figure
144 3).
145
146 Examining the distribution of the 855 SNPs discriminating sublineages, we found that the number
147 of SNPs along discriminatory branches of the phylogeny differed between sublineages (median
148 26 SNPs/sublineage, range 1-153; Supplementary Figure 4). Consistent with previous findings[15],
149 the power to distinguish sublineages within the highly clonal SS14-lineage of TPA was limited
150 (median 4 SNPs/sublineage, range 1-6). In contrast, sublineages within Nichols-lineage TPA
151 (median 20.5 SNPs/sublineage, range 7-60) and in TPE had much higher discrimination (median
152 39 SNPs/sublineage, range 16-153), reflecting more deeply branching sublineages in these parts
153 of the phylogeny.
154

SNP refinement and hierarchical selection of candidate regions

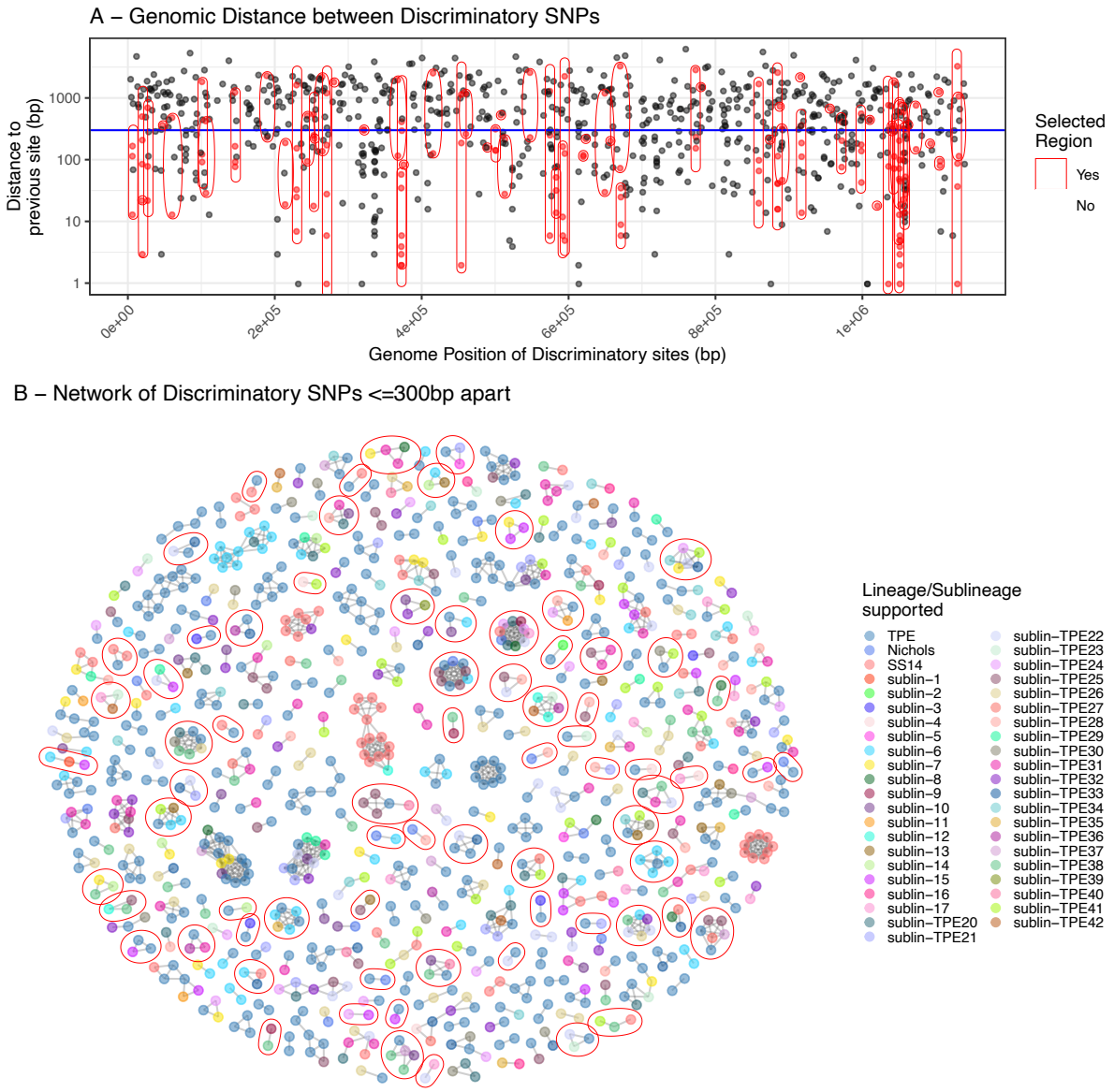
To design an efficient amplicon scheme, we needed to optimise the choice of SNPs to maximise useful information whilst minimising the overall number of amplicons. To ensure multiplex PCR was efficient, we also needed to ensure amplicons were kept small (<1000 bp), and had similar lengths and melting temperatures. Examining the genomic positions of discriminatory SNPs, we found that although these SNPs were widely distributed across the genome for each sublineage (median 13,164 bp between SNPs, Supplementary Figure 5), they could be positionally clustered when analysed as a population (median 694 bp between SNPs). We evaluated the effect of clustering SNPs based on their genome position, forming network edges between SNPs based on different distance thresholds, and selected a distance of 300 bp since this clustered a high number of SNPs whilst keeping the majority of clusters below 1000 bp in length (Supplementary Figure 6). Of our initial 1549 discriminatory SNPs, 988 sites fell within 300 bp of at least one other discriminatory site, suggesting that this was an effective method for identifying localised phylogenetically information-rich regions of the genome (Figure 2A).

We constructed a network of clustered discriminatory SNPs using pairwise positional genome distances, allowing edges where SNPs were separated by ≤300 bp, and identifying 311 distinct network components, each containing >1 discriminatory sites (median 2, range 2-15) - henceforth referred to as 'candidate regions' (Figure 2B). For the remaining 561 discriminatory SNPs not included within a network component (i.e. singleton sites not positionally linked to another), we identified these as additional singleton regions, leaving a total of 872 candidate regions comprising 1549 individual SNPs. Of these, we excluded 50/872 larger candidate regions with total length ≥385 bp to minimise the PCR amplicon length, leaving 822 candidate regions.

Given the varying levels of SNP support for the different sublineages of TPA and TPE, we needed to optimise the number of potential amplicons whilst maintaining support for each sublineage. We designed an algorithm to iteratively select candidate regions to maximise the number of discriminatory sites for each sublineage, whilst minimising the total number of regions included in the final scheme (Supplementary Figure 7). We considered the importance of maximising SNPs supporting each sublineage, whilst incorporating redundancy for PCR failures - aiming to avoid any single region being the sole supporter of a sublineage (which could lead to biases driven by hypervariable regions or residual recombination, as well as the risk of losing all support should PCR for that region fail).

189 Briefly, we constructed a ranked list of candidate regions based on the number of distinct
190 sublineages supported by the SNPs within each region. Using an iterative loop, where initial
191 support for each sublineage was considered to be zero regions, each candidate region was
192 evaluated for its SNP support of specific sublineages. If a region contained SNPs that supported
193 a sublineage, the counts for each supported sublineage were updated and the region was added
194 to the scheme, unless the sublineage had already reached a minimum threshold of regions. For
195 the *T. pallidum* scheme, we selected a minimum threshold of three regions, representing a
196 compromise to provide robust discrimination whilst keeping the total number of amplicons
197 small. Once a sublineage reached this threshold, new regions were not evaluated against it
198 unless they contributed to another sublineage. To minimise potential biases driven by
199 hypervariable regions or residual recombination, counts were increased by one for each
200 sublineage supported by a candidate region even if that region contained multiple discriminatory
201 sites to that sublineage. This process was repeated iteratively for all sublineages and regions.
202 Consequently, our algorithm ensures all sublineages contain at least the minimum threshold
203 discriminatory sites, unless there were fewer than this threshold initially available, whilst some
204 sublineages (typically those representing more deeply branching lineages) contain more than the
205 minimum (as subsequent sublineage picks may enrich discriminatory sites for other
206 sublineages).
207
208 Our Phylo-Plex algorithm prioritised 76/822 candidate genomic regions that define the 40
209 sublineages within the dual global phylogenetic framework of TPA and TPE (Supplementary Figure
210 4). We re-examined the discriminatory power for each sublineage using this revised list and found
211 that support for each sublineage was maintained, whilst the variability in terms of sites per
212 sublineage was greatly reduced (mean 5.5 SNPs, range 1-48) (Supplementary Figure 4). To ensure
213 similar PCR amplification dynamics and to make room for flanking primer design, we extended
214 the length of each Phylo-Plex amplicon (we termed "Phy-cons") to 700 bp by adding equal length
215 flanking sequences to each end. We used these candidate regions for automated design of 20-
216 24 bp primers with 59–63°C melting temperatures with PrimalScheme[10,11] (Supplementary Table
217 2), which produced 72 primer pairs with an amplicon size distribution of 511-628 bp (4/76
218 candidate regions were unsuited to multiplex primer design and were excluded). We also
219 designed and integrated a primer set for recovering the ribosomal 23S region (important for
220 inferring antimicrobial resistance to macrolides in *T. pallidum* – in different pathogens, gene-
221 specific markers for other resistance or virulence determinants could be used) to complete the

222 final amplicon scheme of 73 Phy-cons. All 73 primer pair designs were initially evaluated for
223 specificity using primer BLAST and *in silico* PCR (see Methods).
224
225



226
227 **Figure 2. Positional clustering of discriminatory SNPs enables identification of information**
228 **rich regions for PCR.** A – Genomic distance between discriminatory SNPs for 40 sublineages.
229 Blue line indicates 300 bp distance between sites. Black and Red points represent all
230 discriminatory SNP-sites. Red points represent the clustered location-optimised SNPs
231 represented in final amplicon panel design. B – Network showing the discriminatory SNPs
232 coloured by the lineage they define and clustered by genome position. Nodes indicate individual
233 SNPs and are coloured according to the sublineage supported. Edges indicate SNPs ≤ 300bp from
234 each other, and form clusters of information-rich genomic regions. Red rings indicate clusters
235 included in the final design.
236

Evaluation of laboratory sequencing and bioinformatics

Next, we evaluated the robustness of the laboratory and bioinformatic methods. We performed initial validation using ten-fold dilution series of genomic DNA extracted from four rabbit-passaged reference strains. We then selected 72 clinical swab samples from PCR-confirmed syphilis patients with a *Treponema* qPCR Ct between 23.9-32.2 from South Africa, and generated multiplex PCR amplicons in batches of 24 using the *Treponema* Phylo-Plex scheme (TP-Phylo-Plex). Because all Phy-cons were of a similar size, we used agarose gels to confirm overall pooled PCR performance, and sequencing coverage metrics to evaluate individual amplicon performance.

To ensure our scheme was easily and affordably deployable in low resource settings, we opted for MinION sequencing (Oxford Nanopore). We ligated individual sample barcodes to multiplex-PCR products once PCR success had been confirmed by gel electrophoresis, before pooling the barcoded PCR products and generating sequencing libraries. We used multiplex PCR products pooled from 24 individual samples as input for sequencing on MinION (Oxford Nanopore Technologies) Flongle flow cells, yielding 172,494-303,104 reads per run (after filtering of low quality <Q9, unmapped or reads outside of the accepted size range of 450-800bp) mapping to the targeted regions of the reference genome (NC_021508.1). This amounted to a median of 8,858 reads per sample (range 1,878-20,414), distributed across 73 individual amplicons, with median coverage per amplicon of 104 (range 0-612) across the three sequencing runs and 72 samples.

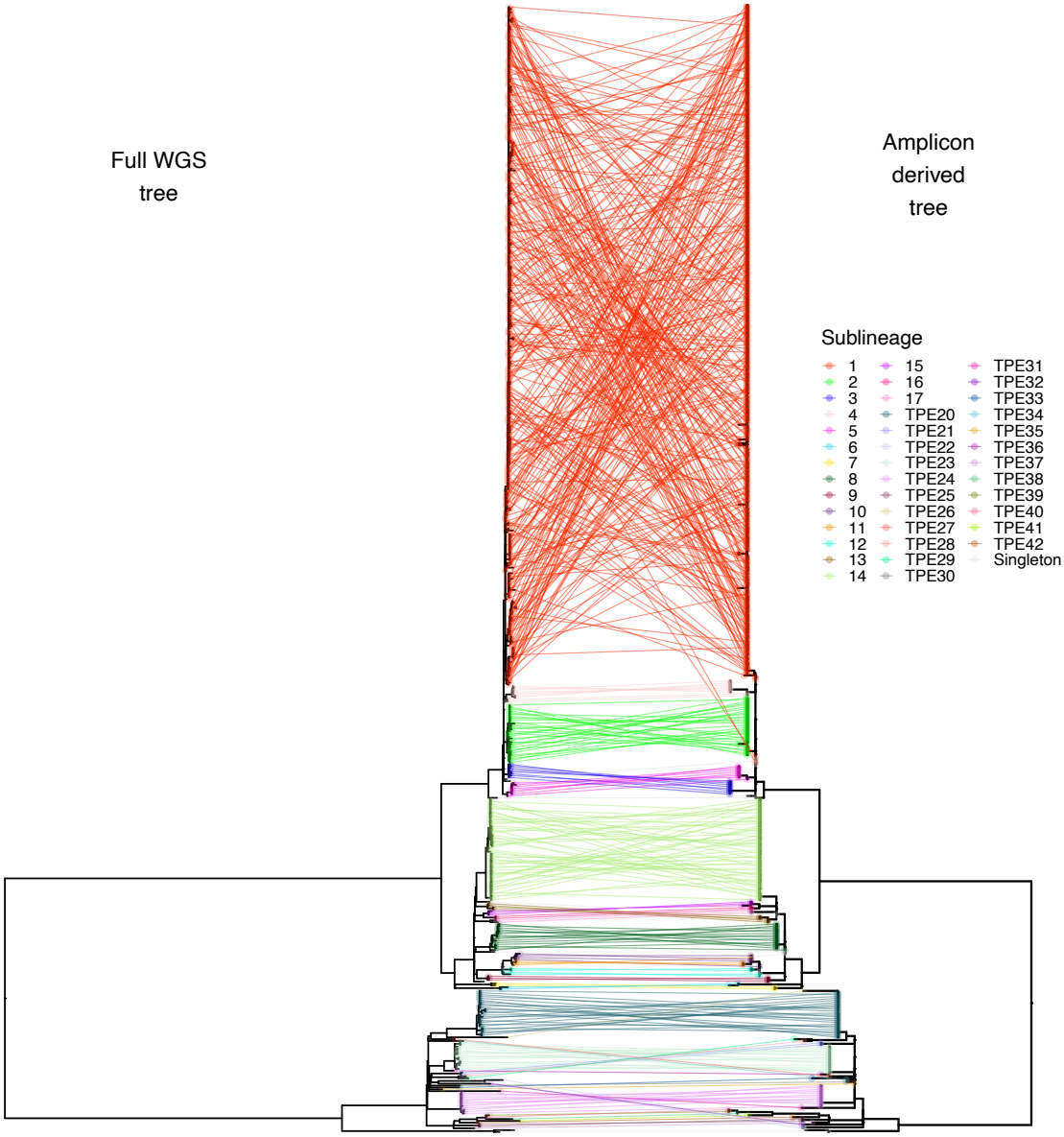
Using read mapping coverage for each Phy-con, we found that 4/72 samples performed poorly with a mean coverage of <50X across all amplicons, and these were excluded from further analysis (Supplementary Figure 8). Notably, these four samples were not outliers in terms of Treponema qPCR - indeed the majority of South African samples performed well, suggesting the upper sensitivity limit for successful sequencing of the TP-Phylo-Plex assay was at least Ct 32 (Supplementary Figure 8).

Analysing coverage for individual Phy-cons, as expected amplicon recovery was improved for samples with increased pathogen load (determined by qPCR Ct; Supplementary Figure 9). However, 15/74 Phy-cons in the scheme consistently yielded reduced coverage in multiplex PCR (but performed well as single-plex PCR assays). For the purposes of the scheme development, we removed these amplicons and retrospectively evaluated the impact on discriminatory power for sublineages (Supplementary Figure 4), henceforth focussing on the 59 remaining Phy-cons

271 that worked consistently in multiplex. Note that this stringent approach was primarily for the
272 purposes of scheme evaluation in the field, and further validation and primer balancing might
273 have enabled retention of some Phy-cons. This would not have changed the way TP-Phylo-Plex
274 was seen to work under a range of different conditions.
275
276 We next evaluated the sequencing accuracy of the Phy-cons generated for TP-Phylo-Plex on
277 MinION. For this, we performed whole genome sequencing of the same 72 South African samples
278 using the pooled sequence capture assay[3,15] on Illumina NovaSeq to provide a gold standard
279 comparator (Supplementary Table 3). We obtained high quality (≥75% reference sites at ≥5X
280 coverage) Illumina consensus genomes for 59 of the 68/72 samples which also had high Phy-con
281 sequence coverage using MinION. Focusing on the regions of the whole genome represented by
282 the Phy-cons, we directly compared the variant calls generated from these 59 samples by the two
283 orthogonal sequencing methods, examining the 181 variable sites contained within the 59 Phy-
284 cons for discordant results. Of these, we found perfect concordance for 179/181 variable sites
285 across all samples. Of the two discrepant sites, the first (NC_021508.1 position 7002) affected
286 1/59 samples and was due to a deletion detected in the Illumina sequence (7003delG) compared
287 to the reference and MinION sequences (7003G). On closer inspection, this was found to be a
288 novel deletion present in both sets of sequencing reads, but was missed in the MinION
289 sequencing because our pipeline does not call INDELS. The second discrepancy (NC_021508.1
290 position 916096) occurred in 2/59 samples, and this site was found to be variable across the
291 dataset (reference C allele, samples were either C or T). The two discrepant samples contained
292 the 'T' variant allele in the Illumina data, but both had very low sequencing coverage in the
293 relevant MinION sequencing amplicon and SNP loci (one mixed SNP with 2/3 reads supporting
294 'T'; one clear 'T' allele supported by 6 reads). In both samples, this discrepancy with the Illumina
295 data was caused by insufficient read coverage from the MinION sequencing, leading to the variant
296 caller failing to correctly call the SNP. Given that we would not want a SNP to be inferred without
297 sufficient coverage, this is intended behaviour. To improve reproducibility and accessibility, we
298 reimplemented the bioinformatics pipeline using NextFlow[25], and this included introducing
299 additional quality steps around coverage that were designed to account for these issues in future,
300 including ensuring reference sites with insufficient coverage in MinION data were called as 'N' in
301 the final inferred sequence.
302

Recapitulation of *T. pallidum* population structures by TP-Phylo-Plex

To assess the accuracy of phylogenetic delineation, we extracted the SNPs present within the final 59-amplicon scheme from a whole genome alignment and compared the resulting phylogeny to the one derived from the whole genomes (Figure 3). The distance matrices were correlated (Mantel statistic r=0.363, p<0.0001, 10,000 permutations), and we used Treespace[26] to demonstrate the trees were highly concordant with sublineage (concordance=0.900) compared to whole genome tree bootstraps (median concordance=0.337, 100 bootstraps) and whole genome tip-randomised trees (median concordance=0.188, 100 permutations) (Supplementary Figure 10).



314 **Figure 3. Recapitulation of whole genome population structure using Phylo-Plex.** Tanglegram
315 comparing whole genome phylogeny with phylogeny calculated from the *in silico* predicted
316 amplicons. Broad sublineage clustering is replicated in the vast majority of cases. Note, tree
317 scales are not identical, since the branch lengths in the amplicon-derived tree were extended to
318 illustrate differences; the underlying topologies were not changed.
319

320 Next, we evaluated TP-Phylo-Plex's ability to discriminate the 40 sublineages used in the original
321 design. *In silico* evaluation of the original 73 amplicon design successfully delineated all 40
322 sublineages. However, the removal of 15 poorly performing amplicons resulted in a 59- Phy-con
323 scheme which did not include highly discriminatory SNPs for 8 sublineages. Since Phy-cons can
324 contain additional SNPs which have F_{ST} <0.95 that may still discriminate profiles, we examined
325 the unique amplicon profiles predicted for each sublineage from the original WGS dataset. We
326 found two instances where two sublineages (sublineages 10 and 11; sublineages 39 and 40)
327 could not be discriminated from each other and thus TP-Phylo-Plex could not distinguish within
328 these two groups using the 59- Phy-con scheme (whilst still being clearly delineated from other
329 sublineages) (Supplementary Figure 11). Notably, those sublineages occurred in clonal lineages
330 such as TPA SS14-Lineage, which had the smallest number of available discriminatory sites even
331 at the WGS level, and were thus most affected by removal of amplicons. Because the level of
332 discrimination can be 'tuned' in Phylo-Plex, users can make decisions about the epidemiological
333 value of discriminating between those sublineages in their analyses.
334

335 We further evaluated the ability of TP-Phylo-Plex to discriminate previously described population
336 structures, using a previously published whole genome phylogeny of 237 TPA genomes from the
337 UK[27] (Supplementary Figure 12). The dataset used for designing the scheme incorporated many
338 of the same genomes and targeted the same sublineages, and TP-Phylo-Plex effectively
339 recapitulated the population structure of UK syphilis.
340

341 Application to more complex pathogen populations

342 For initial development, we used *T. pallidum* genomes which are highly genetically conserved,
343 with sparsely distributed SNPs across the core genome, limiting the available genomic sites
344 which can be used for discrimination. In populations of pathogens with more diverse genomes,
345 there would be more available SNPs and these may be more densely distributed, enabling
346 simpler design with shorter clustering distances. To evaluate the method with a more diverse
347 population structure we selected the highly diverse sexually transmitted bacteria, *Neisseria*
348 *gonorrohoeae*, using a large and recent genomic epidemiology study[28] comprising 5578 genomes

349 collected in Victoria, Australia. Taouk and colleagues used core gene MLST to identify 1262
350 transmission pairs or clusters[28], and we selected all clusters comprising 10 or more genomes
351 (n=79) for tracking. We reanalysed the sequencing data, inferring discriminatory SNPs for the 79
352 transmission clusters (8,692/88,6771 SNPs had F_{ST} ≥0.95). We examined the optimal positional
353 distance between SNPs for clustering, and found that this was lower than for *T. pallidum*
354 (Supplementary Figure 13). However, for the purposes of this evaluation we used the same
355 settings as previously, clustering SNPs within 300 bp to form a positional network, resulting in 849
356 multi-SNP Phy-cons and 965 singleton SNPs. We excluded Phy-cons smaller than 700 bp, and
357 performed hierarchical Phy-con selection to find the optimal combination of Phy-cons as
358 previously, resulting in a final scheme comprising 169 Phy-cons and 785 discriminatory SNPs
359 (Supplementary Figure 14, Supplementary Table 6). Notably, the computational resources
360 required to repeat the analysis with the more complex dataset were substantially increased
361 (requiring up to 128 Gb memory for some processes). This *in silico* scheme showed
362 discriminatory support for 76/79 transmission clusters (Supplementary Figure 14) with clear
363 phylogenetic discrimination (Supplementary Figure 15).
364

365 Application in a low-resource setting

366 To determine the practical and operational feasibility of setting-up and using Phylo-Plex in a low
367 resource setting, we set-up TP-Phylo-Plex at the Biomedical Research and Training Institute (BRTI;
368 Harare, Zimbabwe), a small, ISO 15189 accredited service laboratory based in a converted
369 townhouse. Equipment and molecular biology and sequencing reagents were shipped to
370 Zimbabwe from the UK; key learning points here were around delays due to obtaining import
371 permits. We implemented DNA extraction, qPCR diagnostics, multiplex PCR, library prep and
372 sequencing on the MinION Flongle device (Oxford Nanopore Technologies) (Figure 4). We also
373 performed full bioinformatic analysis using a Nextflow pipeline developed for processing Phy-con
374 data, and real-time interpretation using an easy-to-use custom web app developed in Shiny (see
375 Methods). We established realistic timelines for performing the assay in this setting, and were
376 able to generate amplicons, perform sequencing and bioinformatics in under two days (Figure 4,
377 Supplementary Table 4). As part of this process, two local PhD students were trained in all
378 laboratory processes. Notably, power cuts occurred daily across Harare, but the BRTI laboratory
379 had backup solar power cells. The backup system did malfunction, preventing PCR for two days.
380 This highlights the need for robust power supply systems or qPCR machines that can run from
381 batteries when using the method in a low resource setting.

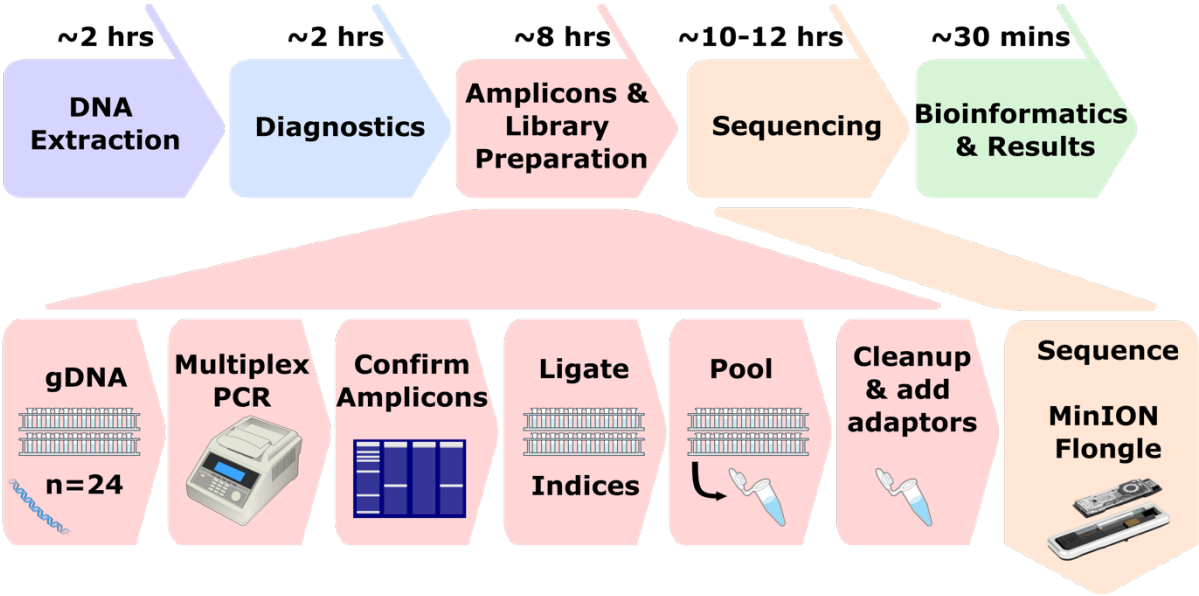


Figure 4. Workflow and timelines for Phylo-Plex protocol. DNA is extracted from clinical swabs, before molecular diagnostics (multiplex qPCR). Confirmed PCR positive samples proceed to Phylo-Plex multiplex PCR and library preparation, before sequencing on MinION Flongle flowcells. Bioinformatic analysis is automated using a single command Nextflow pipeline and interpretation is performed using a web-based Shiny App.

For this field evaluation, we prospectively recruited 100 individuals with genital ulcers in Harare, Zimbabwe, as part of an ongoing multi-country research study investigating the aetiology of genital ulcers. Of 100 genital ulcer swabs collected, 14 were qPCR positive for *T. pallidum* (qPCR Cts range 21.9 – 35.3). We used the 59 Phy-con TP-Phylo-Plex assay to sequence the 14 samples (including different types of technical replicates: independent DNA extractions of swabs taken from the same patient lesions, samples 454a and 454b, samples mz446a/mz446c and mz446b; independent PCRs performed on the same DNA extraction, samples mz446a and mz446c, samples mz476a and mz476b; Total 18), and recovered 12 complete TP-Phylo-Plex profiles. The technical replicates for sample mz476 had identical SNP calls, whilst although both of the replicates for sample mz454 successfully sequenced, one did not pass analytical quality checks. Three samples (5 replicates) with very high qPCR Ct >34 all failed to amplify at sufficient depth) (Supplementary Figure 16). This is consistent with our earlier observation that samples from South Africa performed well up to a qPCR Ct of 32, and suggests an approximate limit for successful sequencing between Ct 32 and 34. These samples yielded novel SNP profiles, reflecting that *T. pallidum* sequencing from Zimbabwe (as with the rest of Africa) was sparse prior to this work.

407 ## Phylo-Plex offers a low-cost alternative for genomic surveillance

408 We evaluated the cost of implementing the 59-amplicon TP-Phylo-Plex, including all reagents
409 (purchased in the UK using publicly displayed list prices) and consumables required for
410 sequencing of PCR-positive DNA extracts. Excluding one-time setup costs (such as purchase of
411 a sequencer and high-performance laptop), and assuming samples were sequenced in batches
412 of 24, we estimated the cost to be £12.47/sample (Table 1), in addition to existing costs for DNA
413 extraction and quality control. Of this cost, the most expensive components were the sequencing
414 (£64.32/Flongle flow cell, £2.68/sample) and the native sample barcoding (£108.00/24 samples,
415 £4.50/sample) reagents.
416
417 Future replacement of native barcoding by ligation with custom primers introduced during PCR
418 would enable further cost savings, whilst sequencing in larger batches on higher capacity MinION
419 flowcells (e.g. 96-384 plex) could reduce sequencing costs. Moreover, whilst the scheme used
420 here implemented amplicons of approximately 700 bp, designing schemes with smaller
421 amplicons (e.g. 300-400 bp) would allow for higher multiplexing of hundreds to thousands of
422 samples on Illumina short-read sequencing platforms.

Stage	Reagents	per 24	per sample
PCR Primers	PCR Primers	£4.32	£0.18
PCR	Q5 Hot Start High-Fidelity 2X Master Mix (100)	£11.76	£0.49
PCR	Agarose / loading dye / EtBr / ladder	£1.56	£0.07
Library	NEBNEXT Ultra II End Repair/dA-tailing module (96)	£35.28	£1.47
Library	NEB Blunt/TA Ligase Master Mix (250)	£44.40	£1.85
Library	NEBNext Quick Ligation Module (20)	£20.40	£0.85
Library	ONT Native Barcoding Kit 24 V14	£108.00	£4.50
Library	Qubit dsDNA HS Assay Kit (500 rxns)	£1.48	£0.06
Sequencing	ONT Flongle Flow Cell	£64.32	£2.68
	Reagent Total (Assuming List Prices)	**£291.52**	**£12.15**
	Plasticware		
PCR	PCR Tubes	£0.52	£0.02
Library	Qubit Assay Tubes	£1.96	£0.08
Library	Microfuge Tubes	£0.30	£0.01
All	Filter Tips	£5.00	£0.21
	Plasticware Total (Assuming List Prices)	**£7.78**	**£0.32**
	Total Assay Cost	**£299.30**	**£12.47**

Table 1. Cost estimate for Phylo-Plex assay. Costs estimated using UK list pricing (institutional discounting may reduce this), and based on reagent volumes and usage. Initial setup costs may require larger purchases including MinION sequencer, High-performance laptop, Qubit Fluorometer, Thermocycler. Microcentrifuge, although many laboratories equipped for molecular biology, including those in Lower- and Middle Income Countries, may already have much of the required infrastructure (sometimes as a legacy of SARS-C0V-2 sequencing).

We compared the costs of our Phylo-Plex assay to that of traditional MLST, which typically uses 7 amplicons and requires 14 Sanger sequencing reads per sample, costing commercially £28–£112/sample (£2-£8/reaction). MLST can also be adapted for high-throughput sequencing, such as nanopore. One approach is to adapt the Phylo-Plex methods by using existing MLST primers in discrete PCRs for each gene which are pooled before library preparation and sequencing (Supplementary Table 5), resulting in an increased per-sample cost of PCR (7 PCR reactions at £3.89 compared to a single PCR reaction at £0.56). MLST would also require less sequencing depth (7 vs. 59 regions), allowing higher multiplexing per run. Since the MinION Flongle run cost is fixed at £64.32, sequencing cost per sample decreases with higher throughput. However, because library costs remain fixed and additional PCRs must be performed, applying an unaltered 7-gene MLST schemes to 24-plex samples would be more expensive (£15.32/sample) than TP-Phylo-Plex (£12.15/sample), and 7-gene MLST remains more costly even at 192-plex. Alternatively, if MLST PCR primers are redesigned to enable multiplex PCR, overall costs of a 7-gene MLST can be reduced to £11.99/sample at 24-plex and £9.64 at 192-plex.

Future proofing the Phylo-Plex and the Phy-con design

Although Phylo-Plex was specifically optimised to recover known population structures and sublineages, our approach of retaining and using SNP information means that detection of novel SNPs and sublineages is theoretically possible, provided the genomic variants occur within the targeted amplicon regions. We originally used rPinecone to cluster *T. pallidum* sublineages on the basis of being within 10-SNPs from a common ancestral node[23], meaning that individual genomes within the same sublineage can be separated by up to 20 SNPs. Within the TP-Phylo-Plex amplicons, representing approximately 3.6% of the genome, most samples from the same sublineage were identical or differed by a single SNP (Supplementary Figure 11). Detection of a novel sublineage using TP-Phylo-Plex would therefore require at least two novel SNPs occurring within the available Phy-cons. Notably, *T. pallidum* accumulates approximately one substitution

458 per genome every 7 years[15], and novel sublineages with sufficient SNPs divergence for detection
459 using TP-Phylo-Plex would therefore also be epidemiologically distinct.
460
461 Next, to quantify the sensitivity of TP-Phylo-Plex to novel variation, we simulated vertically
462 inherited (non-recombining) SNPs accumulating in the genome by introducing random *in silico*
463 mutations to the 1.139 Mb *T. pallidum* genome at different frequencies (0.1%, 0.05%, 0.01%,
464 0.005%, 0.001%, 0.0005%), corresponding to defined SNPs/genome (1140, 570, 114, 57, 11, 6).
465 We counted how many WGS mutations occurred within TP-Phylo-Plex amplicons, evaluating
466 each mutational frequency over 250 simulations, and found that at 0.001% genomic sites
467 mutated (11 SNPs/genome – similar to the discriminatory level within our sublineages), only 5.2%
468 of simulations had ≥2 SNPs occurring within the 59-amplicon TP-Phylo-Plex scheme. However,
469 this increased to 45.6% of simulations at 0.005% genome sites mutated (57 SNPs) and to 87.6%
470 of simulations at 0.01% genomic sites (114 SNPs) (Supplementary Figure 17). Therefore, the
471 probability of distinguishing a novel sublineage depends on the level of genetic divergence –
472 substantially divergent novel sublineages separated by >100 SNPs have a high probability of being
473 identified, whilst the probability is reduced with less divergent sublineages. This is similar to how
474 a standard MLST scheme would detect a new sequence type, except Phylo-Plex captures a larger
475 proportion of the genome, and therefore has greater opportunity to detect novelty.
476
477 Should novel diversity be identified (either through Phylo-Plex, or through more conventional
478 WGS surveillance), a further consideration is how straightforward it would be to enhance
479 detection of novel diversity through adding additional amplicons. Notably, PCR primers represent
480 a very small fraction of total assay costs – including a small number of additional primers would
481 have minimal cost implications. However, since the primers are designed to work together in a
482 multiplex, new primers would need to be validated *in silico* to ensure they were compatible with
483 existing primers, as well as *in vitro* to ensure the multiplex PCR still functioned correctly. A simpler
484 and more flexible solution would be to add a second more locally or regionally targeted multiplex
485 PCR including new Phy-cons. This same approach could also be used for tracking non-
486 chromosomal elements such as plasmids or AMR genes. The PCR products of both multiplexes
487 could be combined before library preparation, at an estimated additional cost of £0.67/sample
488 (Table 1), largely driven by the reagent costs of performing a second PCR. Alternatively, these
489 secondary primer sets could simply be deployed on samples showing variation in the more
490 general Phylo-Plex amplicon set.

Discussion

As genomic surveillance becomes routine in high income countries, there is a need to maximise the utility of global microbial whole genome data we generate as a community so it can be used for epidemiological tracking in all settings. As a consequence of the international SARS-CoV-2 response, multiplex amplicon sequencing is now a proven technology widely used by laboratories around the world[11]. We sought to build on this capacity by developing a novel approach which enables low-cost and adaptable sequence-based pathogen surveillance in all settings, including those that are resource poor[29].

Genome sequencing of 'difficult to cultivate pathogens' such as *Treponema pallidum* and some other STIs represents an ongoing technical challenge. Approaches for direct sequencing from clinical specimens include metagenomics[30], and this can be supplemented with methods for host depletion through differential lysis or differential methylation patterns[31], enrichment of on-target reads through adaptive sequencing[32], selective multiple displacement amplification[33], and sequence capture using baits[34,35]. Each of these offers varying degrees of enrichment, with sequence capture often being preferred for enriching low abundance target organisms[35].

In the UK, Illumina sequencing of bacterial genomes is offered commercially in the region of £65/sample, and typically makes use of high throughput sequencers to reduce per-sample costs, requiring large batches of samples to make them affordable for pathogen sequencing. This can impact the timeliness of results whilst awaiting sufficient samples to complete a batch. Typing methods such as MLST are sometimes perceived as a cheaper alternative, but a typical MLST scheme uses 7 amplicons and requires 14 Sanger sequencing reads/sample, available at £2-£8/reaction commercially. Therefore, traditional MLST costs £28-£112/sample. By comparison, our TP-Phylo-Plex protocol costs £12.47/sample, and still recapitulates the current global population structure derived from whole genomes with a high degree of precision. Phylo-Plex achieves laboratory cost savings through both multiplexing and by using high-throughput sequencing, and existing MLST schemes (whilst designed for different use cases) could similarly reduce costs in this manner, particularly if redesigned to work in multi-plex PCR. Moreover, Phylo-Plex can be run in small batches (<£300/24 sample run) and give results in 24-48 hours, but because it can use multiple sequencing technologies, it can be scaled to enable very high throughput (hundreds of samples/run) sequencing. This makes it ideal for deployment to investigate emerging outbreaks *in situ* as well as more long-term surveillance studies.

524 We chose *T. pallidum* as a model pathogen because it is extremely expensive and complicated to
525 culture, syphilis is a major global health concern and there is still limited data on its global
526 diversity and spread. Moreover, Phylo-Plex is particularly attractive for *T. pallidum* because of the
527 need to use costly metagenomic methods to recover genomes, as well as low genetic diversity
528 and minimal recombination which made design against a reference genome straightforward. To
529 show that this approach was suitable for a range of pathogens, we also designed an *in silico*
530 scheme for a large and genetically diverse collection of *N. gonorrhoeae*, enabling discrimination
531 of fine scale transmission clusters. Phylo-Plex is also appropriate for low-cost genomic
532 surveillance of viral, bacterial and fungal pathogens, and because of the robust and portable
533 nature of the method, as well as the relative ease of laboratory training, is ideally suited and
534 proven for use in low resource settings, enabling tracking of diseases of public health importance
535 such as mpox, chlamydia, typhoid, and cholera.
536
537 Unlike traditional molecular typing methods, Phylo-Plex uses population genomics-informed,
538 phylogenetically robust grouping of taxa within specie(s). Inspired by computational hierarchical
539 SNP typing methods such as GenoTyphi[36], we developed an approach that focuses on key
540 ancestral SNPs for clustering genomes, but refined the concept using network approaches to
541 reduce operational complexity and design a low-cost laboratory method suitable for all settings.
542 This is a similar concept to that of Tau-typing[37], which also identifies phylogenetically
543 discriminatory marker regions, but here we focus on maximising resolution with the minimum
544 amount of regions. Despite recovering only 3.6% of the total genome length, our TP-Phylo-Plex
545 scheme showed high concordance with the whole genome phylogeny, and indeed showed higher
546 concordance than bootstrapped datasets derived from the whole genome itself. Moreover, our
547 laboratory validation showed high accuracy in SNP detection, supporting our use of nanopore
548 sequencing for field deployment.
549
550 Phylo-Plex can be used both as a standalone surveillance tool, and also to supplement and
551 extend whole genome sequencing approaches; for a given population, a subset of samples could
552 undergo WGS to identify key lineages, followed by design of a Phylo-Plex scheme to enable
553 scaleup, and we describe important steps for how new schemes can be developed and evaluated
554 before deployment. Existing Phylo-Plex schemes could also be rapidly adapted to a new outbreak
555 simply by adding a separate pool of amplicons. Importantly, the sensitivity of a Phylo-Plex
556 scheme can be adjusted – used a very stringent threshold to remove 15/74 amplicons from the
557 TP-Phylo-Plex for operational reasons, yet this still enabled full delineation of 31/40 known

558 sublineages. As with any multi-plex PCR, careful laboratory optimisation (including balancing
559 primer concentrations, cycling conditions and modifying primers for compatibility) is required to
560 achieve optimal results. Further optimisation might have allowed inclusion of these additional
561 amplicons, thus enabling delineation of the remaining sublineages for the TPA scheme.
562
563 Phylo-Plex is capable of detecting novel lineages, albeit this is largely dependent on the level of
564 nucleotide divergence between lineages. Adding more amplicons offers greater resolution and
565 can enable antimicrobial resistance or virulence genotyping, at minimal additional cost, but this
566 must be balanced against the technical complexity and time to optimise. The sensitivity required
567 can be judged according to whether a particular lineage or sublineage, detectable by WGS and
568 circulating in relevant populations, is of sufficient clinical relevance to require discrimination
569 from other lineages. A recent example would be the current global health emergency associated
570 with mpox lineage 1b which may have distinct disease outcomes[38], where being able to rapidly
571 and cheaply discriminate the causal lineage in an outbreak can inform both public health
572 management and patient treatment. Indeed, we argue this is the real value of Phylo-Plex – a
573 method for phylogenetically delineating strains which can be tuned to any pathogen, population,
574 or epidemiological question.
575

576 Methods

577 *In silico* methods and design

578 Dataset preparation

579 We constructed a globally representative dataset of 607 publicly available *Treponema pallidum*
580 genomes, including high quality genomes with ≥95% of reference positions at ≥5X coverage. The
581 vast majority of these had been sequenced using the pooled sequence capture method on an
582 Illumina sequencer[3] (Supplementary Table 1). Sequencing reads were mapped to a common
583 reference genome (SS14, NC_021508.1), and masked for previously described hypervariable,
584 repetitive and recombining genes, as well as using Gubbins[20] v2.4.1 to further identify and filter
585 recombining regions as previously described[15]. From the resulting multiple sequence alignment,
586 we then constructed a maximum likelihood phylogeny using IQ-Tree[21,22] v1.6.12 with a GTR+R
587 model and 1000 ultrafast bootstraps[39]. We also used IQ-Tree to generate 100 standard non-
588 parametric bootstrap trees for evaluating tree concordance. We inferred phylogenetic

589 sublineages using rPinecone[23] v0.1.0 with a threshold of up to 10 SNPs from the common
590 ancestral node.

591 Identifying and refining discriminatory sites

592 We converted whole genome sequence alignments into Variant Call Format (VCF) using snp-
593 sites[40] v2.5.1, before importing the data into R using vcfR[41] v1.15.0. We created binary matrices
594 for each genome indicating membership of each sublineage. We also included comparisons for
595 *T. pallidum* subspecies (*pallidum* and *pertenue*), as well as the subspecies *pallidum* major
596 lineages Nichols and SS14. For each sublineage, we then tested each variable SNP for population
597 segregation using Weir and Cockerham's F_{ST}, using the `wc` command in hierfstat[42] v0.5.11. We
598 considered sites with F_{ST} ≥0.90 sufficient for downstream analysis. Since our analysis included
599 sites masked to 'N', we identified and excluded sites where the 'N' could potentially be
600 misrepresented as a discriminatory SNP.
601
602 To identify information-rich regions of the genome, we calculated pairwise positional distances
603 between discriminatory SNPs. We then subset data to include distances below individual
604 thresholds, testing 50 bp intervals from 50-1000 bp. For each set of distances, we formed edge
605 networks using the network package v1.18.2 in R, and extracted network components using
606 iGraph v2.0.3, counting the number of network components (genomic regions), as well as the
607 number of discriminatory SNPs per region and size range of the regions. Note that this approach
608 of linking individual SNPs by distance can result in chains of connected SNPs that span many
609 thousands of bases. Increasing the size of the linkage interval results in more SNPs within a
610 network component, but increasingly leads to longer amplicons. We selected 300 bp for linkage
611 as optimal for *T. pallidum*, since this balanced network components and SNP count against
612 region size.

613 Primer design

614 For multiplex primer design, we used the command line version of PrimalScheme[10,11] v1.3.2,
615 using the option to supply multiple distinct alignments for combined primer design. Using the
616 SS14 reference genome (NC_021508.1), we generated a representative pseudomolecule, where
617 genomic positions in our genome collection that were variable in ≥0.5% of genomes were masked
618 to 'N'. We then extracted target regions into individual fasta files, with the target region itself
619 masked to 'N', and flanked by 150bp unmasked sequence. We specified a target amplicon size of
620 684bp (selected due to the typical target region size and allowing for flanking regions) in
621 PrimalScheme to ensure multiplex primer design was forced to occur around the masked target

622 region. We used the standard GC content and specified predicted melting temperatures of 59–
623 63°C [10]. All primers were evaluated *in silico* using primer BLAST[43] (implemented in primerTree[44]
624 v1.0.6), as well as using *in silico* PCR (available at https://github.com/sanger-
625 pathogens/sh16_scripts/blob/master/legacy/in_silico_pcr.py) against SPAdes[45] assemblies of
626 the genome data.

627 Macrolide resistance primer design

628 Macrolide resistance in *T. pallidum* is mediated by point mutations (A2058G, A2059G) in the
629 ribosomal 23S sequences. Although published primers for *Treponema* 23S exist[46,47], these were
630 not optimised for compatibility with the multiplex scheme, necessitating design of new primers.
631 To identify highly conserved *Treponema* specific regions in the 23S gene, which occurs universally
632 across bacteria, we performed Compact Bit-Sliced Signature Index[48] (COBS) search against a
633 collection of ~661,000 uniformly assembled bacterial genomes[49]. We used the *T. pallidum*
634 ribosomal operon as a template, and generated 71 bp sliding windows along the *Treponema*
635 operon as queries for COBS, examining identical (100%) search hits from the full bacterial
636 dataset. Identical search hits included to representatives of other *Treponema* species (the
637 dataset contained 174 assemblies classified as *Treponema* of which were 125 classified as *T.*
638 *pallidum*) as well as a broad range of bacterial genera, with one window resulting in 229,791
639 identical hits (of which 51,460 were to *Streptococcus pneumoniae* and 48,124 were to
640 *Staphylococcus aureus*). The median genome count for each window was 129 (reflecting hits to
641 species in the Treponema genus), and regions of the *T. pallidum* 23S operon identified as
642 matching ≥5% above the median (135.5) were initially flagged as potentially cross-genus
643 matches, and this was confirmed by examining the species identities of the hits. Windows flagged
644 for likely contamination were masked in the fasta file used as input to PrimalScheme for primer
645 design using the conditions described for the full discriminatory scheme, resulting in two
646 alternative primer designs which were screened using primer BLAST. After laboratory validation,
647 the best performing of these two primer pairs was selected for use within the scheme.

648 *Application to Neisseria gonorrhoeae* dataset

649 To evaluate the performance of the Phylo-Plex approach using a more complex dataset, we
650 selected a large *Neisseria gonorrhoeae* dataset comprising 5880 genomes from Victoria,
651 Australia[28]. In the original analysis, the authors defined 1177 putative transmission clusters
652 based on core gene MLST alleles. For this evaluation, we aimed to design a Phylo-Plex scheme
653 that could recover any transmission cluster comprising ≥10 genomes (n=79). We downloaded
654 fastq files for 5878 publicly available sequences from the European Nucleotide Archive (ENA),

655 and all reads were mapped to the *Neisseria gonorrhoeae* reference genome (FA1090,
656 NC_002946.2) using snippy v4.6.0 (available at https://github.com/tseemann/snippy) with
657 default settings to construct a reference mapped alignment. We extracted all SNPs using snp-
658 sites and followed the workflow described above for identifying discriminatory SNPs for the 79
659 transmission clusters. Due to the complexity of the dataset, it was not possible to complete the
660 analysis on a laptop due to excessive memory usage, and individual functions were reformatted
661 to run as independent processes on a high-performance computer cluster. It was also not
662 possible to run gubbins for exclusion of recombining sites on this dataset due to computational
663 complexity. However, recombining regions that are exclusively discriminatory for a sublineage are
664 potentially still useful within the Phylo-Plex framework. Moreover, recombination blocks are
665 typically represented in our network as large components where nearly all SNPs support the
666 same sublineage (Supplementary Figure 14), and the Phylo-Plex algorithm is designed to
667 prioritise regions that support multiple sublineages. For phylogenetic analysis of such a large and
668 complex dataset, we used IQ-Tree as described above with a GTR-Gamma model, but we
669 excluded bootstrapping as this took over 30 days of CPU wall time.

670 *In silico* validation

671 To evaluate the schemes, we initially simulated amplicons from Illumina-derived whole
672 genomes. Based on the target regions, we extracted SNPs and constructed multiple sequence
673 alignments using the vcfR[41] v1.15.0, ape[50] v5.8 and seqinr[51] v4.2.36 packages, from which we
674 inferred phylogenies using IQ-Tree[21] v1.6.12. We compared the whole genome phylogenies to the
675 inferred amplicon phylogenies for phylogenetic consistency using tanglegrams produced using
676 ggtree[52] v3.12.0, distance matrix comparison using Mantel tests in the vegan[53] v2.6.6.1, as well
677 as correlating whole genome derived sublineages with amplicon derived phylogenies using the
678 treeConcordance function in Treespace[26] v1.1.4.3. We examined distinct sublineage clustering
679 *in silico* by constructing networks of identical amplicon profiles, and correlating these with WGS-
680 derived sublineage assignments in R. Since *T. pallidum* genomes often contain uncertain sites,
681 for this aspect we performed ancestral reconstruction of SNPs using pyjar v0.1 (available at
682 https://github.com/simonrharris/pyjar) and extracted the pairwise cophenetic SNP distances
683 from the ancestrally reconstructed phylogeny using ape. We subsequently performed direct
684 comparisons on Illumina whole genome and MinION amplicon variant calls and phylogenetic
685 placement derived from the same samples by integrating the variant calls from both methods
686 (described below) using bcftools merge[54] v1.19.
687

688 To evaluate the ability of Phylo-Plex to detect novel variation, we used a custom python script
689 (randomly_mutate_genome.py, available at https://github.com/matbeale/TP-Phylo-
690 Plex_paper_2025) to introduce random mutations at defined frequencies (0.1%, 0.05%, 0.01%,
691 0.005%, 0.001%, 0.0005%) into the SS14 reference genome (NC_021508.1). We simulated 250
692 different mutated genomes at each frequency, and converted the resulting fasta files to multi-
693 sample variant call files (VCF) using snp-sites. We imported VCFs into R using vcfR and
694 partitioned the SNPs for each simulated genome. For each set of variants, we calculated how
695 many intersected with the amplicon positions using iRanges[55] v2.38.0 and created summary
696 statistics using tidyverse[56] v2.0.0.
697

698 # Laboratory workflow

699 ## Validation samples

700 For initial validation of PCRs, we used ten-fold dilution series of high copy number rabbit
701 passaged strains (SEA-83-1, BAL-2, BAL-6, Haiti-B). We also used a clinical dataset of 72 syphilis
702 swab DNA samples derived from patients with genital ulcers in South Africa, which were
703 sequenced using both Illumina (pooled sequence capture[3]) and the Phylo-Plex method. We also
704 prospectively recruited 100 patients with genital ulcers in Zimbabwe as part of an ongoing study
705 ("Multi-country Aetiology of Genital Ulcer Survey" – MAGUS); 14 *T. pallidum* PCR-positive swab
706 DNA samples were sequenced using the protocol in a low-resource laboratory at the Biomedical
707 Research and Training Institute in Harare, Zimbabwe. This included a combination of replicates
708 for three samples; one sample (mz454) was independently extracted from two separate aliquots
709 of swab lysate as technical replicates; one samples with high total DNA and Treponema qPCR
710 load (mz476) was sequenced in duplicate from the same extract as technical replicates; one
711 sample with low total DNA and Treponema qPCR load (mz446) was sequenced as both types of
712 technical replicates..

713 ## Sample Preparation

714 For the validation in Zimbabwe, ulcers were swabbed using flocked swabs in transport medium
715 (Corning), with 1 ml lysate realiquoted into cryovials for storage at -80°C ahead of processing.
716 DNA extraction was performed on 200µl swab lysate using the QIAamp DNA Mini Kit (Qiagen)
717 according to the manufacturer's instructions.

718 ## PCR amplification

719 We modified the ARTIC network (https://artic.network/) SARS-CoV-2 sequencing protocol
720 (available at https://www.protocols.io/view/artic-sars-cov-2-sequencing-protocol-v4-lsk114-
721 bp2l6n26rgqe/v4) for multiplex PCR and library preparation, adjusting reaction volumes to reduce
722 cost and increasing cycling times to improve PCR efficiency of 700 bp amplicons. Briefly, for each
723 sample we prepared a PCR reaction mix containing 6.25µl Q5 Hot Start High-Fidelity 2X Master
724 Mix (New England Biolabs), 2µl pooled multiplex primers (10 mM; Integrated DNA Technologies),
725 1.75µl nuclease-free water, and 2.5µl template DNA (total reaction volume of 12.5µl). Samples
726 were initially denatured for 30 seconds at 98°C, followed by 35 cycles of 15 seconds at 98°C, 4 ½
727 minutes at 62°C and 30 seconds at 72°C. Performance of bulk PCR products was assessed using
728 agarose gels, ensuring a band was present at 700-800 bp.

729 ## Library preparation and sequencing

730 PCR products were used directly without cleanup, and diluted 1:5 (or less, if fewer than 24
731 samples/run) in nuclease-free water. We performed end repair using the NEBNext Ultra II End
732 Prep module (New England Biolabs), combining 1.2 µl Ultra II End Prep buffer with 0.5µl End Prep
733 Enzyme Mix, and adding 8.3µl of diluted PCR product, before incubation for 15 minutes at room
734 temperature (~20°C), then inactivation for 15 minutes at 65°C, then 1 minute on ice. PCR
735 products were individually barcoded using 5µl Blunt/TA Ligase Master Mix (New England Biolabs),
736 1.25µl DNA barcode (SQK-NBD114.24, or SQK-NBD114.96; Oxford Nanopore Technologies), 3µl
737 nuclease-free water, and 0.75µl end-repaired product from the previous step; ligation reactions
738 were incubated for 20 minutes at room temperature, followed by 10 minutes at 65°C, then 1
739 minute on ice. All barcoded PCR products were then pooled into a single tube for onward
740 preparation. We performed two rounds of SPRI bead cleanup (AmpureXP, Beckman-Coulter) at
741 0.4X, using the Short Fragment Buffer for washing (Oxford Nanopore Technologies). After cleanup
742 and quantification, we ligated sequencing adaptors (Oxford Nanopore Technologies) by
743 combining 10µl NEBNext Quick Ligation Reaction Buffer (5x), Quick T4 DNA Ligase (both New
744 England Biolabs), 5µl Native Adaptor (Oxford Nanopore Technologies), and 30µl pooled library,
745 followed by incubation for 20 minutes at room temperature. We then performed two further
746 rounds of SPRI bead cleanup, before final quality control and library preparation.
747
748 Pooled multiplex amplicon libraries were sequenced in groups of 24 on a single run of the MinION
749 Flongle device (Oxford Nanopore Technologies) using either R9.4.1 (initial validation and testing
750 of South African samples) or R10.4.0 (field deployment and Zimbabwe samples) chemistry, with

751 the change motivated by discontinuation of R9.4.1 reagents and flowcells. Where fewer than 24
752 samples were available (e.g. in Zimbabwe), we reduced the dilution factor of initial PCR products
753 from 1:5 up to 1:3 to ensure sufficient total DNA was available for subsequent steps.

754 Bioinformatic analyses

755 Basecalling, QC, Variant calling

756 Illumina data was processed and analysed as previously described[15]. Briefly, enriched
757 metagenomic sequencing reads were screened using Kraken2[57] v2.1.2 and *Treponema* reads
758 were extracted using seqtk v1.3 (available at https://github.com/lh3/seqtk). We mapped
759 sequencing reads to the SS14 reference (NC_021508.1) after masking previously described
760 recombinogenic regions, called variants using bcftools[54] v1.19 to construct a multiple sequence
761 alignment, and used Gubbins[20] v2.4.1 to identify and mask additional recombinant sites.
762
763 Initial validation of MinION amplicon data was conducted using a custom pipeline. Sequencing
764 runs were basecalled and demultiplexed using the high accuracy model in Guppy v6.4.2.
765 Sequencing reads with >Q9 were mapped to the SS14 reference (NC_021508.1) using
766 minimap2[58] v2.16, and quality metrics were produced using samtools[59]. In particular, we
767 calculated coverage metrics for each individual amplicon by parsing the outputs of samtools
768 depth v1.6. After mapping, we trimmed reads to remove primers using Cutadapt[60] v1.15, before
769 performing SNP calling using Clair3[61] v1.0.1 (available at https://github.com/HKU-BAL/Clair3)
770 using the "res_dna_r9.4.1_e8.1_hac_v033" model. We compared variants inferred from
771 Nanopore data to Illumina data by merging filtered VCF files using bcftools[54] v1.19.
772 After initial validation of methods, we subsequently reimplemented this workflow in a
773 streamlined NextFlow[25] pipeline (available at https://github.com/sanger-pathogens/long-read-
774 ampliseq; workflow shown in Supplementary Figure 18). This replaced the basecaller with
775 Dorado v0.5.1 (available at https://github.com/nanoporetech/dorado), incorporated additional
776 quality steps, and refined variant calling and pseudosequence generation. Raw sequencing
777 reads (fast5/pod5) were base-called using Dorado with the
778 "dna_r10.4.1_e8.2_400bps_hac@v4.3.0" model and minimum quality score set to 9, and
779 demultiplexed using the appropriate barcode kit name specified (e.g. 'SQK-NBD114-24').
780 samtools[59] v1.19.2 was used to convert the resulting bams to fastqs. Primers and low quality
781 (below Q15 at each end) bases were trimmed from fastqs using Cutadapt[60] v4.7, with lower read
782 length cutoff set to 450 and upper read length cutoff set to 800 to ensure only reads of the
783 expected length were included. Remaining bases with quality <15 were also masked using seqtk

784 v1.4 (available at https://github.com/lh3/seqtk). Mapping of the reads against the masked SS14
785 reference was performed using Minimap2[58] v2.26. Unmapped reads and alignments with
786 mapping quality below 50 were filtered using samtools[59]. As each read should map to one of the
787 target regions, alignments not overlapping with these regions or that are secondary or
788 supplementary were also filtered. Variant calling was performed using Clair3[61] v1.0.9 in haploid
789 mode, using the "r1041_e82_400bps_hac_v430" model and specifying to only call SNPs with a
790 minimum coverage of 5. Bcftools[54] v1.20 was used to merge the genome variant call files (gvcfs)
791 from all samples to a multi-sample vcf and also to convert to tab-delimited format for simple
792 parsing. A custom script run in Python 3.10.12 (available in the pipeline) was used to create a
793 consensus fasta file for each sample with individual target regions, as well as all the target regions
794 concatenated, including variants and reference calls with minimum genotype quality of 1 and
795 masking everything else. Individual target region fastas were also concatenated to make a
796 multiple sequence alignment, from which SNPs were extracted with snp-sites[40] v2.5.1.
797 Throughout the pipeline, various tools are used for QC reporting at different stages, including
798 PycoQC v2.5.2 (available at https://github.com/a-slide/pycoQC), FastQC[62] v0.12.1 (available at
799 https://github.com/s-andrews/FastQC), samtools, bedtools v2.31.1, as well as custom Python
800 scripts. MultiQC[63] v1.22.2 is used to create a summary report combining output from PycoQC,
801 FastQC and samtools stats/flagstat for all samples. Typical run time for this pipeline on a high-
802 performance laptop for a TP-Phylo-Plex MinION Flongle sequencing run was 20-40 minutes
803 (depending on the number of reads generated by the sequencing). Fieldwork in Zimbabwe used
804 https://github.com/sanger-pathogens/ONTAP v1.0.0 for analysis.
805
806 We also developed a Shiny web interface ("AmpliSeq_QC_Frontend.R") in R which can be loaded
807 using Rstudio, included with the pipeline code at https://github.com/sanger-pathogens/ONTAP ,
808 to enable straightforward guided interpretation of bioinformatics outputs for use by non-
809 bioinformaticians in the field. This takes the directory path of a completed NextFlow pipeline run
810 as input, and provides a point-and-click interface to assess quality and coverage metrics,
811 allowing the user to select and deselect samples, and reanalysing results (including phylogenies
812 and plots) in real-time.
813

814 ## Evaluation of *in vitro* amplicon performance

815 To identify poorly performing amplicons amongst the South African validation set, we first
816 identified and excluded samples which performed poorly across all amplicons based on an
817 overall mean coverage of <50X. By doing this, we ensured the analysis was not substantially
818 biased by sample. Amongst the samples which performed well overall, we then identified poorly
819 performing amplicons by asking what proportion of amplicons met a minimum coverage
820 threshold of 25X (using the quantile function in R). We then excluded amplicons where >10% of
821 the samples had a mean coverage below 25X. For the TP-Phylo-Plex assay, this led to exclusion
822 of 15/74 amplicons.

823 ## Statistics and plotting

824 Phylogenies and tanglegrams were plotted using ggtree[52] v3.12.0, and all other plots were
825 produced using ggplot2[64] v3.5.1 in R[65] v4.4.2. Networks were plotted using ggnetwork[66] v0.5.13
826 and iGraph[67] v2.0.3, and all statistics were calculated in R. Code workflows were made in draw.io
827 and Lucidchart. Additional diagrams were made in Inkscape. ChatGPT v3.5 was used for initial
828 code scaffolding and to assist with troubleshooting during the development of the Shiny
829 application for Nextflow data interpretation.

830 ## Ethical approvals

831 Clinical samples from South Africa were collected as part of routine public health surveillance,
832 and ethical approval for genome sequencing was granted by the University of the Witwatersrand
833 Human Research Ethics Committee (Medical) (Ethics clearance certificate no. M230157).
834 Clinical swabs from Zimbabwe were collected as part of the Multi-Country Aetiology of Genital
835 Ulcer Study (MAGUS), and ethical approvals were granted for patient recruitment, diagnostics
836 and genomics from the Medical Research Council of Zimbabwe (MRCZ/A/2878), the Biomedical
837 Research and Training Institute Institutional Review Board (Ap/175/2022) and the London School
838 of Hygiene and Tropical Medicine Ethics Committee (26731).

839 ## Data and Code Availability

840 European Nucleotide Archive (ENA) Accessions for Illumina sequencing reads from the initial
841 dataset used for primer design are listed in Supplementary Table 1. Accessions for the Illumina
842 reads from South African syphilis are listed in Supplementary Table 3 and are available at the ENA
843 under project accession PRJEB60271. Oxford Nanopore sequencing reads are available from the

ENA under project accession PRJEB85457, and are listed in Supplementary Table 3 (for South Africa) and Supplementary Table 4 (for Zimbabwe).

The code used in this paper (including for identifying discriminatory sites and the hierarchical selection algorithm) are available at https://github.com/matbeale/TP-Phylo-Plex_paper_2025 and at https://doi.org/10.5281/zenodo.14894111. The Nextflow pipeline for automated data processing of Phylo-Plex data is available at https://github.com/sanger-pathogens/ONTAP.

Author Contributions

Conceptualisation: MAB; Methodology: MAB, VS, KEA, GL, SD, WRS, BS, FL, MJD; Formal Analysis: MAB; Investigation: MAB, VS, KEA, GL, MC, TMW; Resources: MAB, SD, WRS, BS, FL, MPM, JMEV, BDCD, EM, LR, BM, MM, NRT; Data Curation: MAB, VS; Writing – Original Draft: MAB; Visualisation: MAB; Supervision: LR, RAF, NRT; Project Administration: MAB, MM, BM, ED, EM, RAF, NRT.

Acknowledgements

We thank the teams and administrators at the National Institute for Communicable Disease in South Africa and at the Zvitambo and Biomedical Research and Training Institute labs in Harare, Zimbabwe, as well as all partners involved in the MAGUS study. We acknowledge the Core Sequencing, Pathogen Informatics and Samples teams at the Wellcome Sanger Institute.

This work was part-funded by a Gates Foundation grant (INV-035896) to MAB, MM, RAF and NRT. MAB, VS, KEA, GL, SD, WRS, BS, FL, MJD and NRT were supported by Wellcome funding to the Sanger Institute (206545/Z/17/Z).

This work was supported, in whole or in part, by the Gates Foundation [INV-035896] and the Wellcome Trust [206545/Z/17/Z]. The conclusions and opinions expressed in this work are those of the authors alone and shall not be attributed to the Gates Foundation or the Wellcome Trust. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 License has already been assigned to the Author Accepted Manuscript version that might arise from this submission. Please note works submitted as a preprint have not undergone a peer review process.

References

1. Gardy, J. L. & Loman, N. J. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat. Rev. Genet.* nrg.2017.88 (2017) doi:10.1038/nrg.2017.88.
2. Struelens, M. J. *et al.* Real-time genomic surveillance for enhanced control of infectious diseases and antimicrobial resistance. *Front. Sci.* **2**, (2024).
3. Beale, M. A. *et al.* Genomic epidemiology of syphilis reveals independent emergence of macrolide resistance across multiple circulating lineages. *Nat. Commun.* **10**, 3255 (2019).
4. Hadfield, J. *et al.* Comprehensive global genome dynamics of *Chlamydia trachomatis* show ancient diversification followed by contemporary mixing and recent lineage expansion. *Genome Res.* **27**, 1220–1229 (2017).
5. Uelze, L. *et al.* Typing methods based on whole genome sequencing data. *One Health Outlook* **2**, 3 (2020).
6. Luo, Y. *et al.* US Gulf-like toxigenic O1 *Vibrio cholerae* causing sporadic cholera outbreaks in China. *J. Infect.* **72**, 564–572 (2016).
7. Grillová, L. *et al.* Molecular characterization of *Treponema pallidum* subsp. *pallidum* in Switzerland and France with a new multilocus sequence typing scheme. *PloS One* **13**, e0200773 (2018).
8. Lieberman, N. A. P. *et al.* *Treponema pallidum* genome sequencing from six continents reveals variability in vaccine candidate genes and dominance of Nichols clade strains in Madagascar. *PLoS Negl. Trop. Dis.* **15**, e0010063 (2021).
9. Quick, J. nCoV-2019 sequencing protocol. https://www.protocols.io/view/ncov-2019-sequencing-protocol-bbmuik6w (2020).
10. Quick, J. *et al.* Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat. Protoc.* **12**, 1261–1276 (2017).

11. Kent, C. *et al*. PrimalScheme: open-source community resources for low-cost viral genome sequencing. 2024.12.20.629611 Preprint at https://doi.org/10.1101/2024.12.20.629611 (2024).
12. Plante, J. A. *et al*. The variant gambit: COVID-19's next move. *Cell Host Microbe* **29**, 508–515 (2021).
13. WHO. Syphilis. https://www.who.int/news-room/fact-sheets/detail/syphilis.
14. Edmondson, D. G., Hu, B. & Norris, S. J. Long-Term In Vitro Culture of the Syphilis Spirochete *Treponema pallidum* subsp. *pallidum*. *mBio* **9**, e01153-18 (2018).
15. Beale, M. A. *et al*. Global phylogeny of *Treponema pallidum* lineages reveals recent expansion and spread of contemporary syphilis. *Nat. Microbiol.* **6**, 1549–1560 (2021).
16. Taouk, M. L. *et al*. Characterisation of *Treponema pallidum* lineages within the contemporary syphilis outbreak in Australia: a genomic epidemiological analysis. *Lancet Microbe* **3**, e417–e426 (2022).
17. Seña, A. C. *et al*. Clinical and genomic diversity of *Treponema pallidum* subspecies *pallidum* to inform vaccine research: an international, molecular epidemiology study. *Lancet Microbe* **0**, (2024).
18. Marra, C. M. *et al*. Enhanced Molecular Typing of *Treponema pallidum*: Geographical Distribution of Strain Types and Association with Neurosyphilis. *J. Infect. Dis.* **202**, 1380–1388 (2010).
19. Pla-Diaz, M. *et al*. A new typing scheme demonstrates high discriminatory power for *Treponema pallidum* subspecies. *BioRxiv Prepr. Serv. Biol.* 2025.07.10.664125 (2025) doi:10.1101/2025.07.10.664125.
20. Croucher, N. J. *et al*. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* gku1196 (2014) doi:10.1093/nar/gku1196.

21. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
22. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
23. Wailan, A. M. *et al.* rPinecone: Define sub-lineages of a clonal expansion via a phylogenetic tree. *Microb. Genomics* https://doi.org/10.1099/mgen.0.000264 (2019) doi:10.1099/mgen.0.000264.
24. Weir, B. S. & Cockerham, C. C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**, 1358–1370 (1984).
25. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
26. Jombart, T., Kendall, M., Almagro-Garcia, J. & Colijn, C. treespace: Statistical exploration of landscapes of phylogenetic trees. *Mol. Ecol. Resour.* **17**, 1385–1392 (2017).
27. Beale, M. A. *et al.* Genomic epidemiology of syphilis in England: a population-based study. *Lancet Microbe* **4**, e770–e780 (2023).
28. Taouk, M. L. *et al.* Longitudinal genomic analysis of *Neisseria gonorrhoeae* transmission dynamics in Australia. *Nat. Commun.* **15**, 8076 (2024).
29. Wilson, C. N., Musicha, P. & Beale, M. A. Genomic epidemiology on the move. *Nat. Rev. Microbiol.* 1 (2022) doi:10.1038/s41579-022-00836-4.
30. Houldcroft, C. J., Beale, M. A. & Breuer, J. Clinical and biological insights from viral genome sequencing. *Nat. Rev. Microbiol.* **15**, 183–192 (2017).
31. Wang, C. *et al.* Benefits and challenges of host depletion methods in profiling the upper and lower respiratory microbiome. *Npj Biofilms Microbiomes* **11**, 130 (2025).
32. Martin, S. *et al.* Nanopore adaptive sampling: a tool for enrichment of low abundance species in metagenomic samples. *Genome Biol.* **23**, 11 (2022).

33. Thurlow, C. M. *et al.* Selective Whole-Genome Amplification as a Tool to Enrich Specimens with Low *Treponema pallidum* Genomic DNA Copies for Whole-Genome Sequencing. *mSphere* **7**, e00009-22 (2022).
34. Pinto, M. *et al.* Genome-scale analysis of the non-cultivable *Treponema pallidum* reveals extensive within-patient genetic variation. *Nat. Microbiol.* **2**, 16190 (2016).
35. Büttner, K. A. *et al.* Evaluating methods for genome sequencing of *Chlamydia trachomatis* and other sexually transmitted bacteria directly from clinical swabs. *Microb. Genomics* **11**, 001353 (2025).
36. Wong, V. K. *et al.* An extended genotyping framework for *Salmonella enterica* serovar Typhi, the cause of human typhoid. *Nat. Commun.* **7**, 12827 (2016).
37. Seabolt, M. H., Boddapati, A. K., Forstedt, J. J. & Konstantinidis, K. T. Tau-typing: a Nextflow pipeline for finding the best phylogenetic markers in the genome for molecular typing of microbial species. *Bioinformatics* **39**, btad425 (2023).
38. Beiras, C. G. *et al.* Concurrent outbreaks of mpox in Africa—an update. *The Lancet* https://doi.org/10.1016/S0140-6736(24)02353-5 (2024) doi:10.1016/S0140-6736(24)02353-5.
39. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
40. Page, A. J. *et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genomics* **2**, (2016).
41. Knaus, B. J. & Grünwald, N. J. vcfr: a package to manipulate and visualize variant call format data in R. *Mol. Ecol. Resour.* **17**, 44–53 (2017).
42. Goudet, J. & Jombart, T. hierfstat: Estimation and Tests of Hierarchical F-Statistics. (2022).
43. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **215**, 403–410 (1990).

44. Hestor, J. & Cannon, M. primerTree: Visually Assessing the Specificity and Informativeness of Primer Pairs. (2022).
45. Bankevich, A. *et al*. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
46. Grimes, M. *et al*. Two Mutations associated with Macrolide Resistance in *Treponema pallidum*: Increasing Prevalence and Correlation with Molecular Strain Type in Seattle, Washington. *Sex. Transm. Dis.* **39**, 954–958 (2012).
47. Beale, M. A. *et al*. Yaws re-emergence and bacterial drug resistance selection after mass administration of azithromycin: a genomic epidemiology investigation. *Lancet Microbe* **1**, e263–e271 (2020).
48. Bingmann, T., Bradley, P., Gauger, F. & Iqbal, Z. COBS: A Compact Bit-Sliced Signature Index. in *String Processing and Information Retrieval* (eds Brisaboa, N. R. & Puglisi, S. J.) 285–303 (Springer International Publishing, Cham, 2019). doi:10.1007/978-3-030-32686-9_21.
49. Blackwell, G. A. *et al*. Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. *PLOS Biol.* **19**, e3001421 (2021).
50. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
51. Charif, D. & Lobry, J. Seqin{R} 1.0-2: a contributed package to the {R} project for statistical computing devoted to biological sequences retrieval and analysis. in *Structural approaches to sequence evolution: Molecules, networks, populations* (eds Bastolla, U., Porto, M., Roman, H. & Vendruscolo, M.) 207–232 (Springer-Verlag, New York, 2007).
52. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
53. Oksanen, J. *et al*. vegan: Community Ecology Package. R Package. (2015).

54. Danecek, P. & McCarthy, S. A. BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics* **33**, 2037–2039 (2017).
55. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLOS Comput. Biol.* **9**, e1003118 (2013).
56. Wickham, H. *et al.* Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
57. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
58. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
59. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
60. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
61. Zheng, Z. *et al.* Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nat. Comput. Sci.* **2**, 797–803 (2022).
62. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
63. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
64. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer International Publishing, 2016). doi:10.1007/978-3-319-24277-4.
65. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing (2014).
66. Briatte, F. ggnetwork: Geometries to Plot Networks with 'ggplot2'. (2021).
67. Csárdi, G. *et al.* igraph for R: R interface of the igraph library for graph theory and network analysis. https://doi.org/10.5281/zenodo.8240644 (2023) doi:10.5281/zenodo.8240644.

1028 **Supplementary Figure 1. Population structure and characteristics of initial *Treponema***
1029 **dataset used for Phylo-Plex design.** A – Maximum Likelihood phylogeny of 607 *Treponema*
1030 *pallidum* genomes, with coloured tracks showing Subspecies, Lineage (within subspecies
1031 pallidum), and Sublineage (across all subspecies). B - Treemap plot showing distribution of
1032 genomes by *T. pallidum* subspecies: TPA – subspecies *pallidum*, TPE – subspecies *pertenue*, TEN
1033 – subspecies *endemicum*. C - Treemap plot showing distribution of genomes by T. pallidum
1034 subspecies (clustered as samples sharing a common ancestral node ≤10 SNPs. D - Treemap plot
1035 showing distribution of genomes by country.
1036
1037 **Supplementary Figure 2. Identifying discriminatory SNPs for a single lineage.**
1038 A – Population fixation analysis (F_{ST}) of variable sites. Sites in red discriminate *T. pallidum* Nichols
1039 Lineage from all other lineages at F_{ST} ≥0.9. B – Maximum likelihood whole genome phylogeny
1040 showing allelic identity at each discriminatory site identified in A. Colours indicate allele detected
1041 (pale grey indicates 'N', where data was missing – common in metagenomic data).
1042
1043 **Supplementary Figure 3. Identifying discriminatory SNPs for 40 sublineages.**
1044 Population fixation analysis (F_{ST}) of variable sites for each sublineage - plots show only sites with
1045 F_{ST} ≥0.9.
1046
1047 **Supplementary Figure 4. Discriminatory site support for each sublineage.**
1048 Plot shows number of discriminatory sites specifically supporting each lineage and sublineage
1049 when considering (i) all discriminatory sites identified in the genome dataset, (ii) sites included in
1050 74 regions selected by the Phylo-Plex selection algorithm, (iii) sites remaining after full
1051 optimization of the 59-amplicon multiplex PCR. The final PCR used for evaluation removed 15
1052 amplicons and therefore removed direct support for 8 sublineages ('*'), but future iterations
1053 would modify the primer designs and balance to reinclude these.
1054
1055 **Supplementary Figure 5. Genomic distance between individual sites is reduced when**
1056 **considered as a total population.**
1057 Genetic distance between adjacent discriminatory SNPs according to sublineage and as a total
1058 population. Blue line highlights 300 bp.
1059
1060 **Supplementary Figure 6. Changing distance used to link SNPs as clusters impacts cluster**
1061 **count, SNPs per amplicon and minimum amplicon size.**

1062 The selection of an appropriate genomic distance for positional clustering can be tuned and
1063 impacts the number of SNPs that merge into clusters. However, increasing cluster distance
1064 between individual SNPs can also substantially impact the total length of SNP networks, resulting
1065 in candidate amplicons too long to reasonably amplify in multiplex PCR. A – Number of positional
1066 clusters produced with different genomic distance between sites. B – Number of SNPs present in
1067 each cluster (candidate amplicon) produced with different genomic distance between sites. C –
1068 Genomic distance between furthest SNPs in cluster (i.e. minimum length of candidate amplicon)
1069 produced with different genomic distance between sites.
1070
1071 **Supplementary Figure 7. Hierarchical selection algorithm for maximising discriminatory**
1072 **power whilst minimising total number of amplicons.**
1073 Each candidate region is evaluated based on the sublineages the SNPs within it support, and
1074 regions are added until support for a sublineage meets a minimum threshold (3 SNPs), after
1075 which, support for that sublineage is no longer considered.
1076
1077 **Supplementary Figure 8. Sequencing performance of 72 South African clinical syphilis**
1078 **samples.**
1079 Points show mean coverage for 74 amplicons from each of 72 samples, ordered by input
1080 Treponema qPCR Ct. Samples were sequenced on three separate runs of MinION Flongle cells.
1081 Blue line indicates 50X coverage. Red points indicate 15 amplicons which consistently performed
1082 poorly (see Supplementary Figure 9). Cyan lines indicate mean coverage of all amplicons in that
1083 sample.
1084
1085 **Supplementary Figure 9. Sequencing performance of 74 multiplex amplicons in clinical**
1086 **syphilis samples.**
1087 Points show mean coverage for 74 amplicons from each of 72 South African clinical syphilis
1088 samples compared to input Treponema qPCR Ct. Blue line indicates 50X coverage. Red points
1089 indicate 15 amplicons where ≥10% of samples had <25X coverage (after excluding samples
1090 which performed poorly overall).
1091
1092 **Supplementary Figure 10. Quantitative comparison between WGS derived phylogeny and**
1093 **Amplicon derived phylogeny demonstrates high concordance.**

A – Collapsed WGS phylogeny (one tip per sublineage), B – Tree concordance between collapsed WGS tree and (i) full WGS tree, (ii) Amplicon derived tree (simulated), (iii) 100 bootstraps derived from full WGS tree, (iv) 100 tip-randomised full WGS trees.

Supplementary Figure 11. *In silico* sublineage concordance between WGS and amplicons. A – Collapsed WGS phylogeny (one tip per sublineage), against Amplicon profiles (identical sequences). Each column represents a cluster of identical amplicon profiles, coloured by the whole genome sublineage. Amplicon profiles which include multiple sublineages (indicated by arrows) are not fully resolved. Some sublineages consist of multiple amplicon profiles, reflecting expected SNP diversity in the amplicons. B – Distribution of samples within the amplicon profiles according to sublineage. Bars are coloured by sublineage proportion, and numbers indicate the genome count in the analysis.

Supplementary Figure 12. *In silico* reconstruction of UK syphilis population structure using genomes and amplicons.
A – Minimum spanning tree of UK syphilis genomes coloured by sublineages (Beale 2023). Branch lengths indicate SNPs, and dotted lines indicate truncated branches ≥20 SNPs. B – Minimum spanning tree of UK syphilis dataset simulated using final 59-amplicon TP-Phylo-Plex scheme.

Supplementary Figure 13. Effect of changing distance used to link SNPs for *Neisseria gonorrhoea*. Plots show cluster count, SNPs per amplicon and minimum amplicon size.

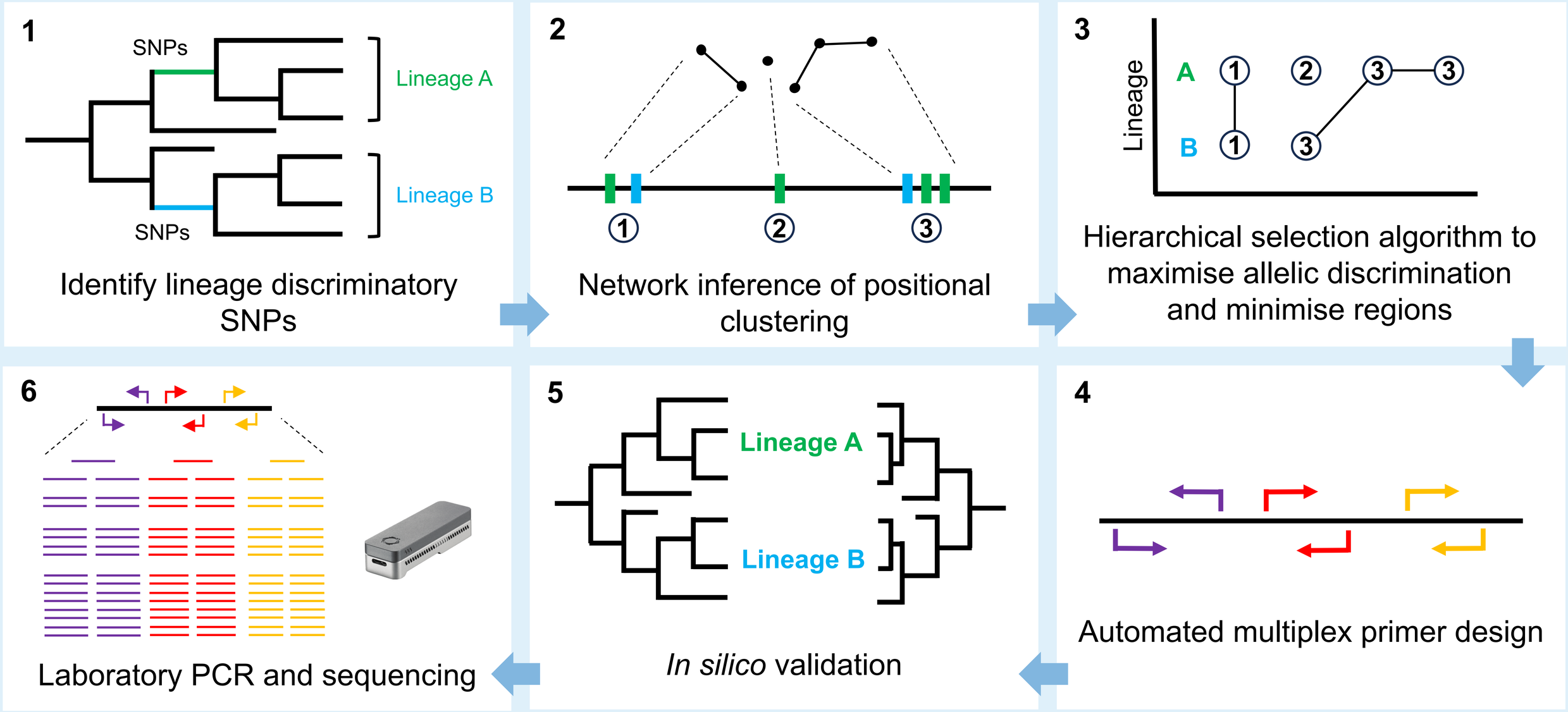
Supplementary Figure 14. Selection of optimal amplicons (Phy-cons) enables design of an efficient scheme to recover transmission clusters in *Neisseria gonorrhoea*. A - Network showing the discriminatory SNPs coloured by the lineage they define and clustered by genome position. Nodes indicate individual SNPs and are coloured according to the sublineage supported. Edges indicate SNPs ≤ 300bp from each other, and form clusters of information-rich genomic regions. Red rings indicate clusters included in the final design. Large blocks of linked SNPs associated with the same sublineage likely indicate regions of recombination. B - Discriminatory site support for each sublineage, showing number of discriminatory sites specifically supporting each lineage and sublineage when considering (i) all discriminatory sites identified in the genome dataset, (ii) sites included in 169 regions selected by the Phylo-Plex selection algorithm.

1128 **Supplementary Figure 15. Phylogenetic recapitulation of *Neisseria gonorrhoeae* whole**
1129 **genome population structure and transmission clusters using Phylo-Plex.** Tanglegram
1130 comparing whole genome phylogeny with phylogeny calculated from the *in silico* predicted
1131 amplicons. Broad sublineage clustering is replicated in the vast majority of cases. Note, tree
1132 scales are not identical, since the branch lengths in the amplicon-derived tree were extended to
1133 illustrate differences; the underlying topologies were not changed.
1134
1135 **Supplementary Figure 16. Amplicon recovery for 17 samples at Zimbabwe field site.**
1136 Points show mean coverage for 59 amplicons from each of 14 samples collected, processed and
1137 sequenced in Zimbabwe, ordered by input Treponema qPCR Ct. Blue line indicates 25X coverage
1138 threshold. Includes technical replicates (mz476a/b, mz446a/b/c).
1139
1140 **Supplementary Figure 17. Phylo-Plex can detect *de novo* mutations and novel sublineages**
1141 **but is insensitive to minor changes.**
1142 *In silico* mutations in whole *T. pallidum* genomes contained within TP-Phylo-Plex amplicons (250
1143 simulations). With 0.001% genomic sites mutated (11 SNPs – similar to discriminatory level
1144 within most *T. pallidum* sub-lineages), only 5.2% of replicates had ≥2 SNPs occurring within Phy-
1145 cons. However, at 0.005% genome sites mutated (56 SNPs), this rose to 45.6% of replicates, and
1146 at 0.01% genomic sites (113 SNPs), it rose to 87.6% of replicates. Blue line indicates 2 SNPs
1147 (minimum SNPs needed to theoretically detect a sublineage as different in amplicon data).
1148
1149 **Supplementary Figure 18. Workflow of NextFlow pipeline for automated processing of**
1150 **Phylo-Plex amplicon sequence data.** Complete workflow for processing raw ONT sequence
1151 files (POD5/Fast5), including basecalling, quality assessment, trimming and filtering, mapping,
1152 coverage assessment, variant calling and pseudosequence generation.
1153
1154
1155 **Supplementary Table 1. Metadata and list of genomes used for designing the TP-Phylo-Plex**
1156 **scheme.**
1157
1158 **Supplementary Table 2. Primers designed for TP-Phylo-Plex scheme.**
1159
1160 **Supplementary Table 3. Description of sequences and metadata for validation samples**
1161 **from South Africa.**

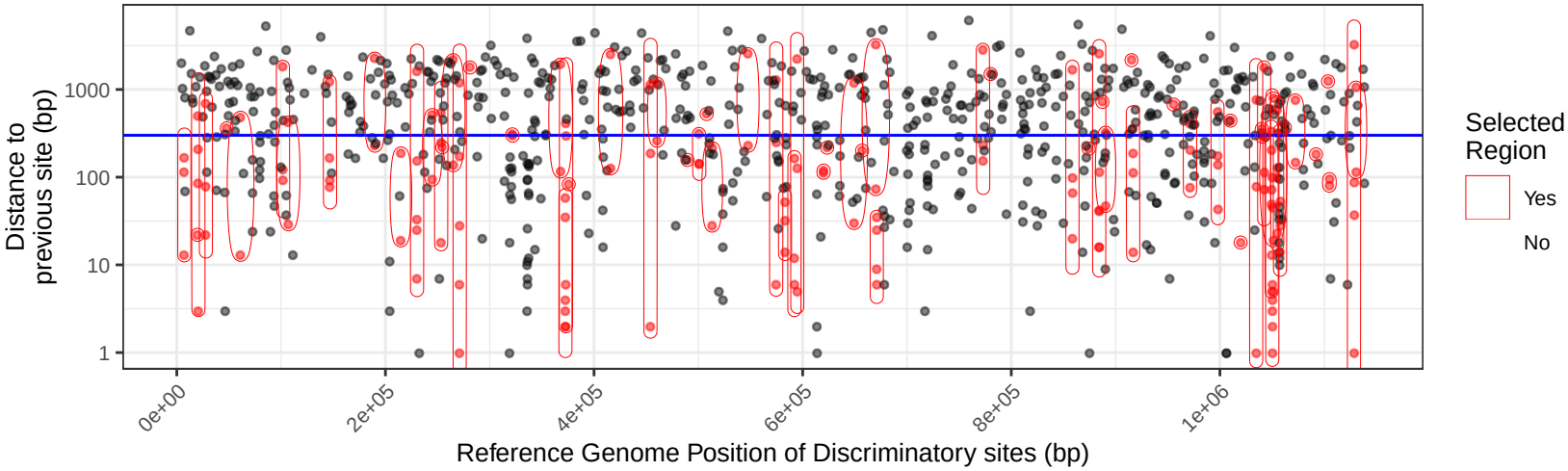
Supplementary Table 4. Description of sequences and metadata for field work samples from Zimbabwe.

Supplementary Table 5. Cost breakdown of applying the Phylo-Plex method to MLST.

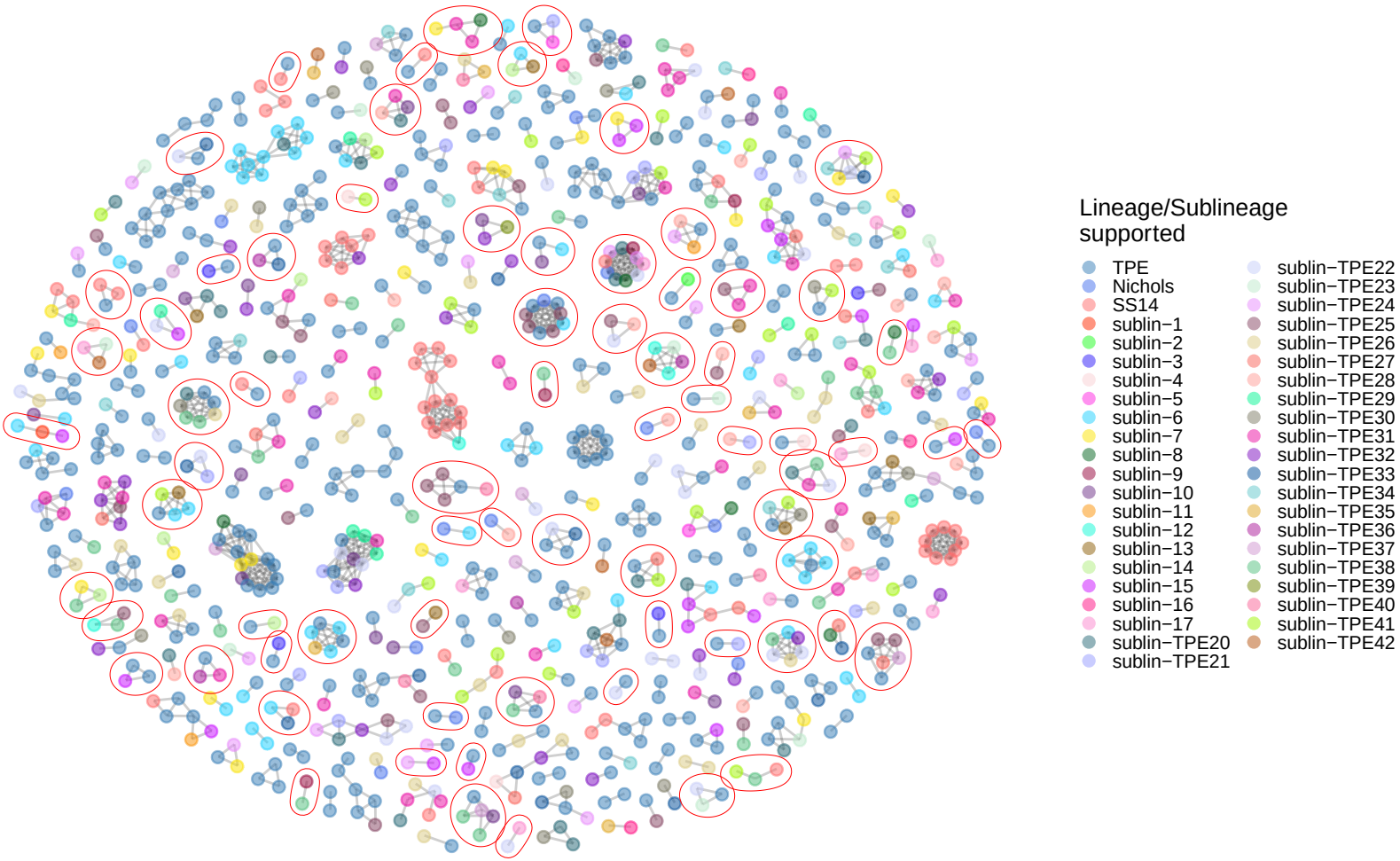
Supplementary Table 6. Genomic regions inferred for efficient reconstruction of *Neisseria gonorrhoeae* transmission clusters.

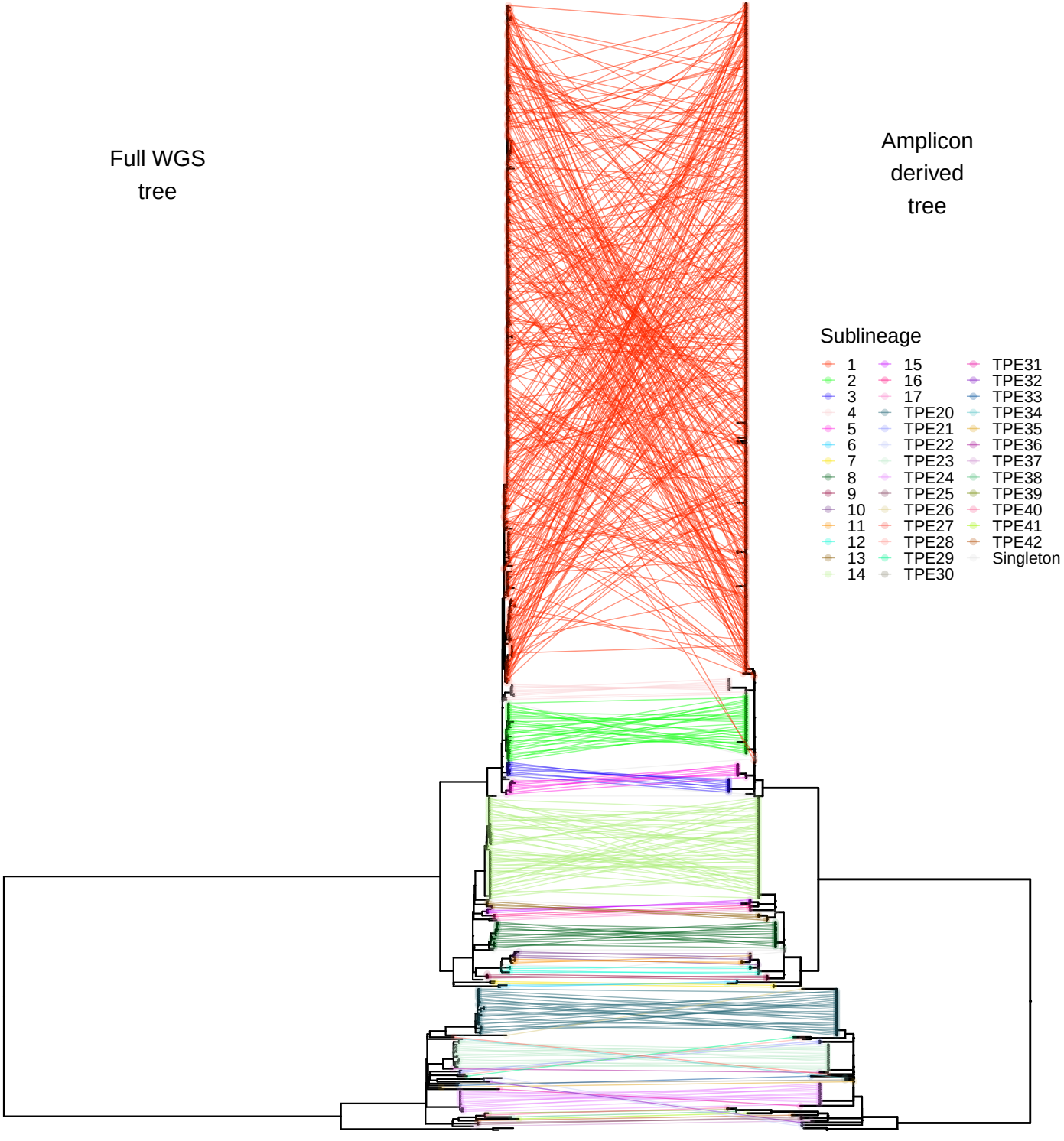

1
SNPs
Lineage A
Lineage B
SNPs
Identify lineage discriminatory SNPs
2
1
2
3
Network inference of positional clustering
3
Lineage
A
B
1
2
3
3
1
3
Hierarchical selection algorithm to maximise allelic discrimination and minimise regions
4
Automated multiplex primer design
5
Lineage A
Lineage B
In silico validation
6
Laboratory PCR and sequencing

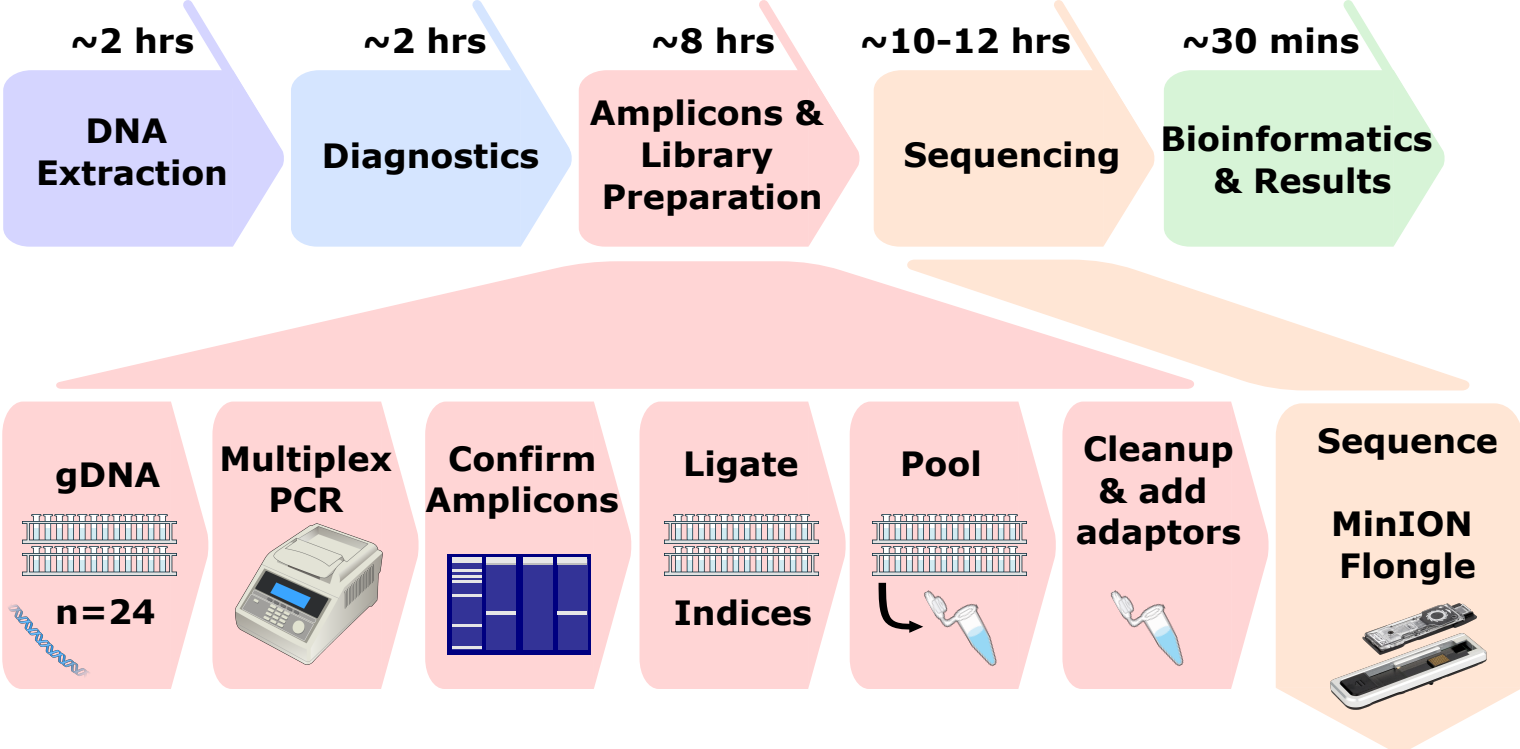
A − Genomic Distance between Discriminatory SNPs



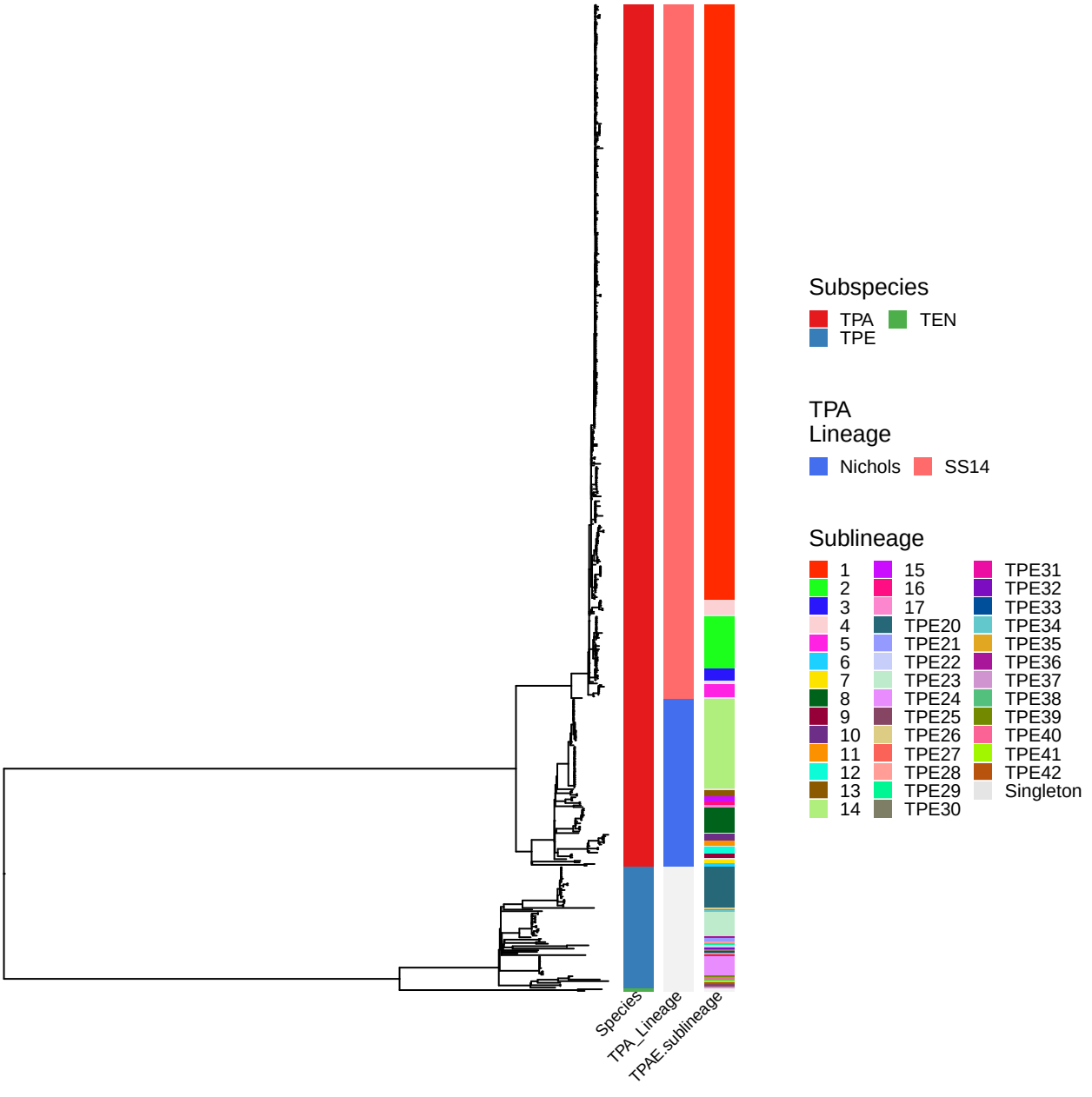
B − Network of Discriminatory SNPs <=300bp apart



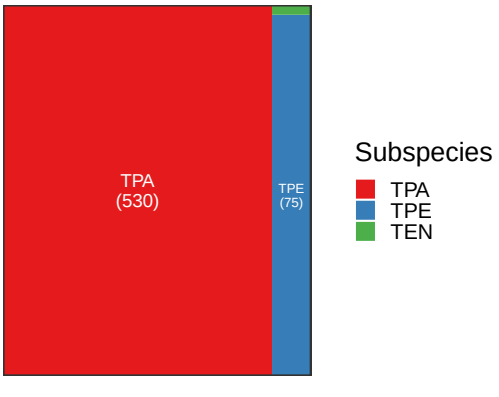

Full WGS tree
Amplicon derived tree
Sublineage
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
TPE20
TPE21
TPE22
TPE23
TPE24
TPE25
TPE26
TPE27
TPE28
TPE29
TPE30
TPE31
TPE32
TPE33
TPE34
TPE35
TPE36
TPE37
TPE38
TPE39
TPE40
TPE41
TPE42
Singleton


~2 hrs
DNA Extraction
~2 hrs
Diagnostics
~8 hrs
Amplicons & Library Preparation
~10-12 hrs
Sequencing
~30 mins
Bioinformatics & Results
gDNA
n=24
Multiplex PCR
Confirm Amplicons
Ligate
Indices
Pool
Cleanup & add adaptors
Sequence
MinION
Flongle

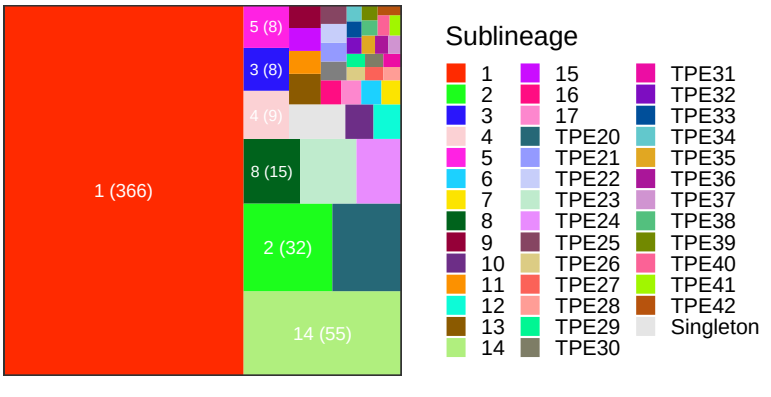
A − Whole Genome Phylogeny



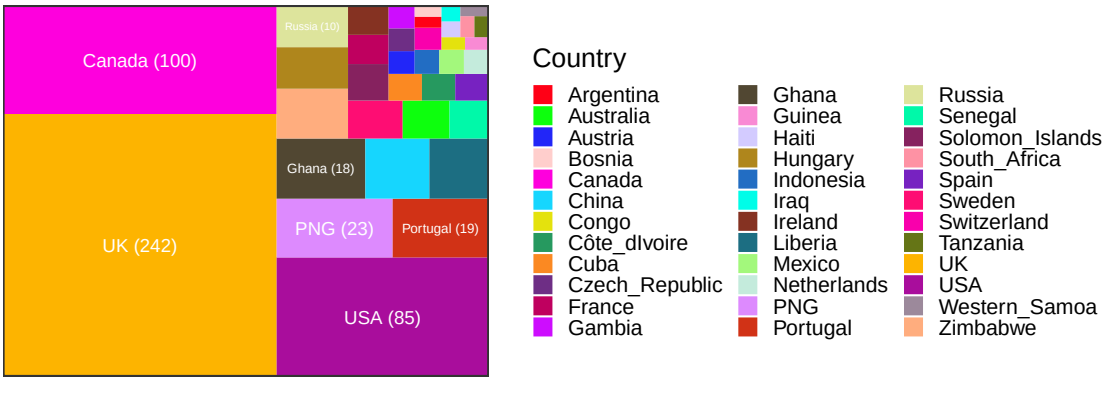
B − Subspecies distributions



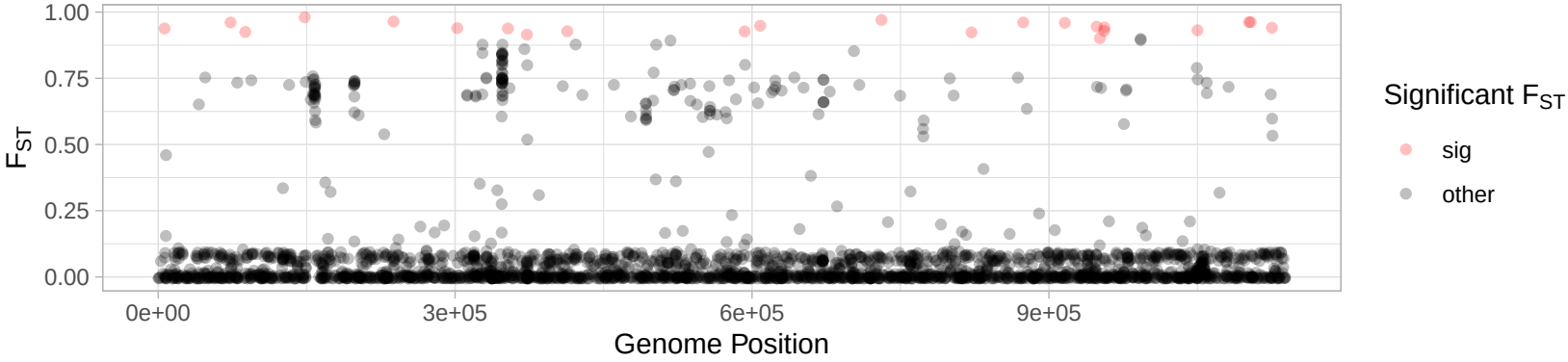
C − Sublineage distributions



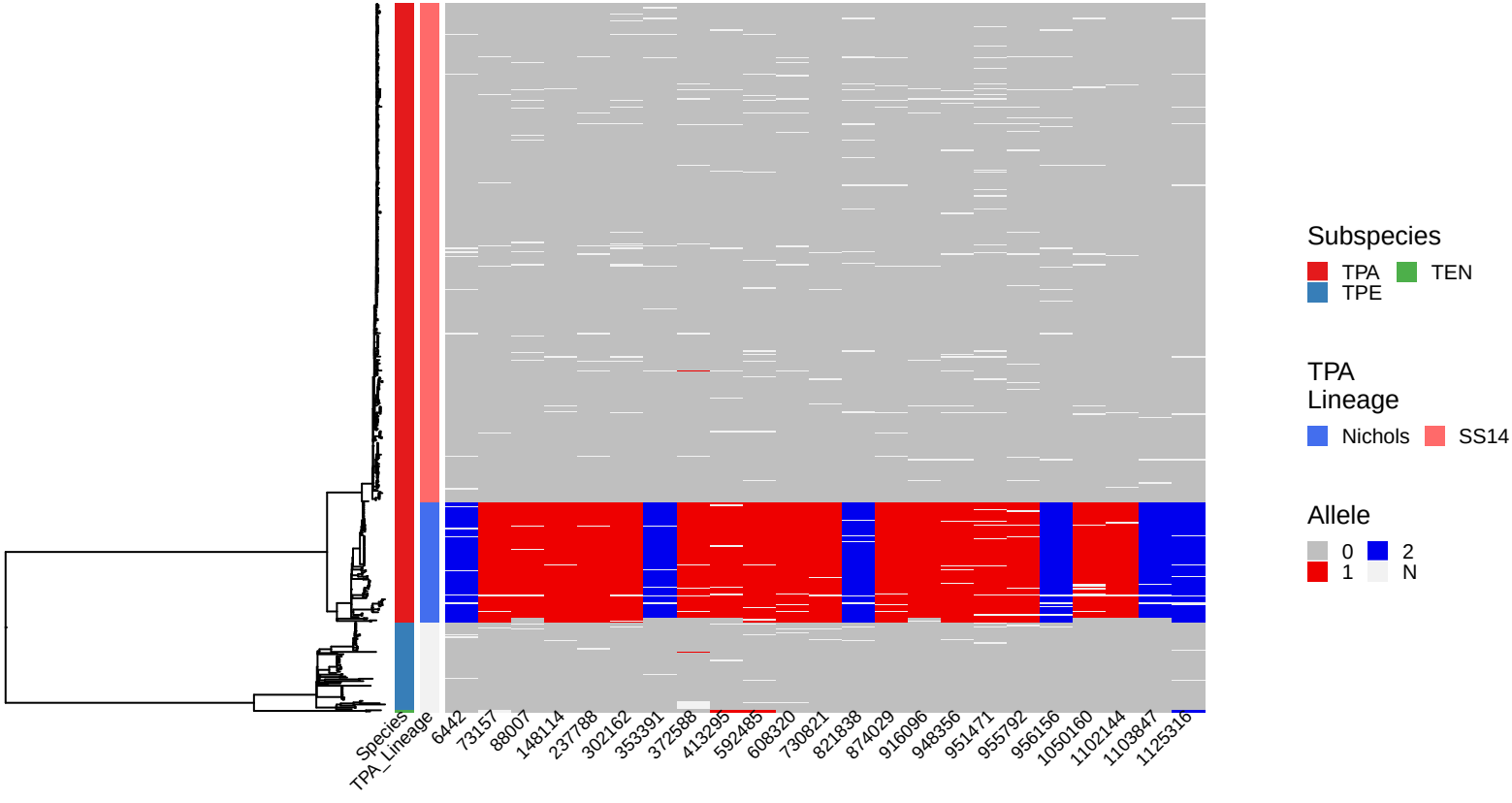
D − Geographical distributions



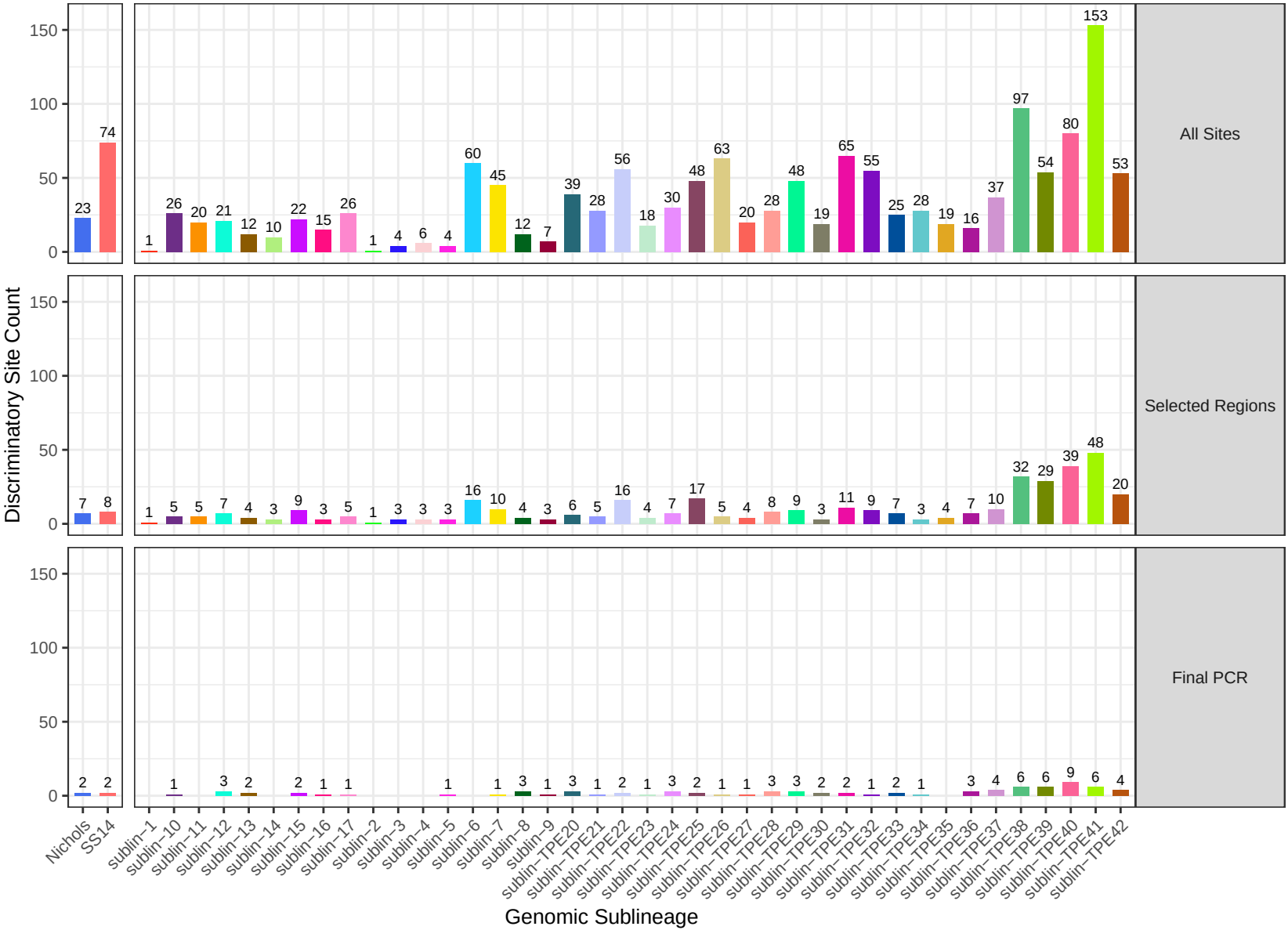
A − Site−by−site F_{ST} discriminating Nichols Lineage

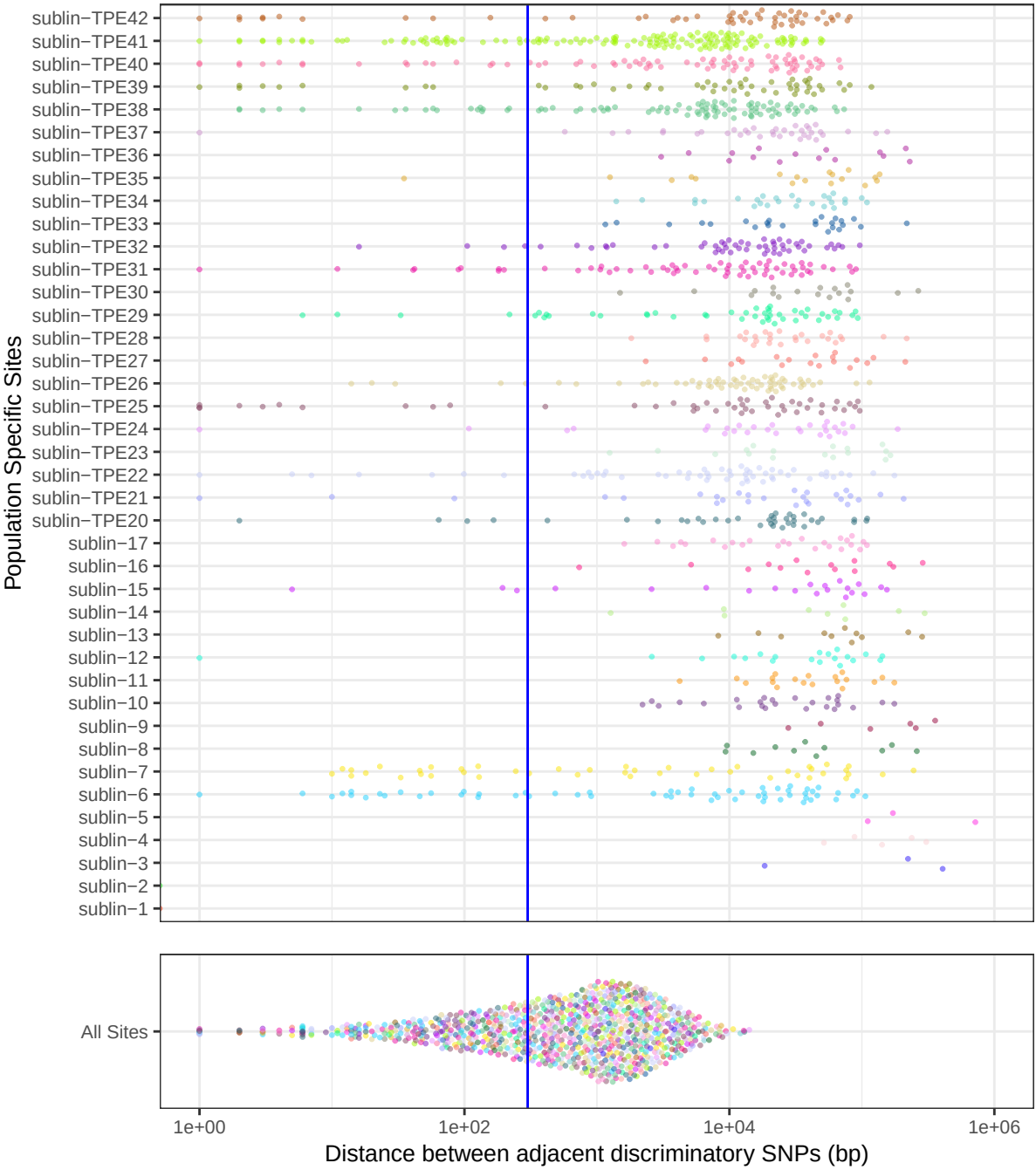


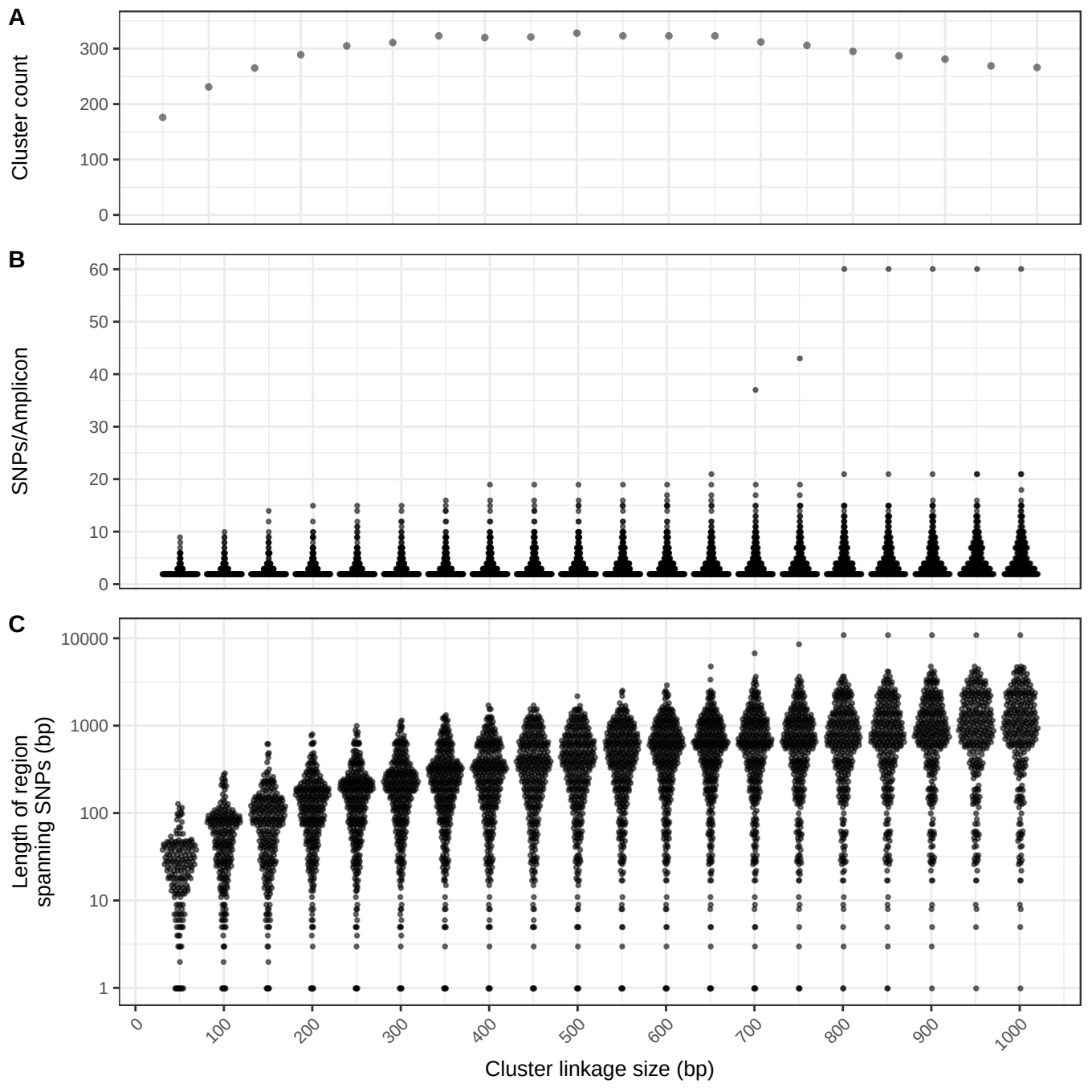
B − Discriminatory SNPs across the phylogeny

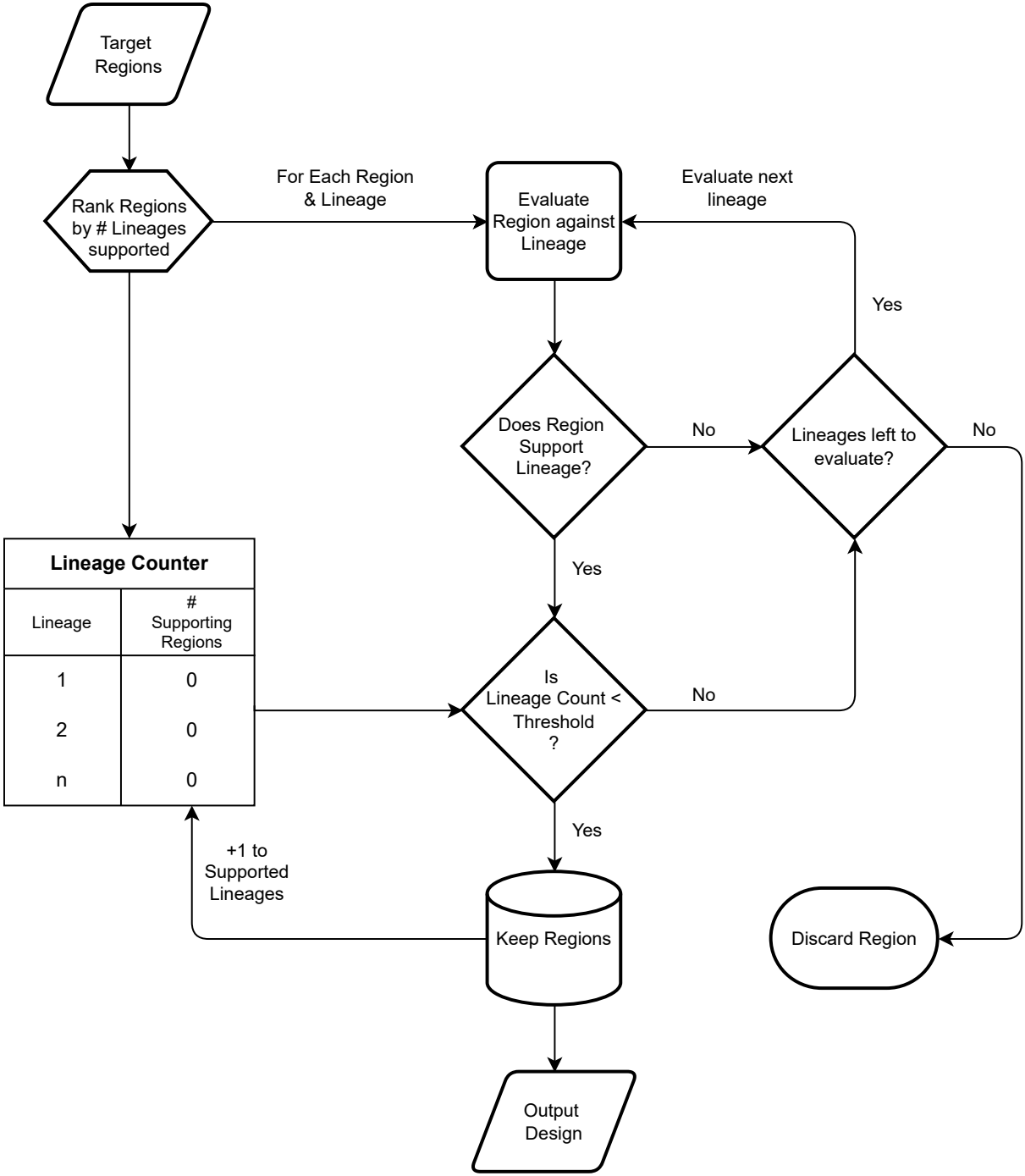


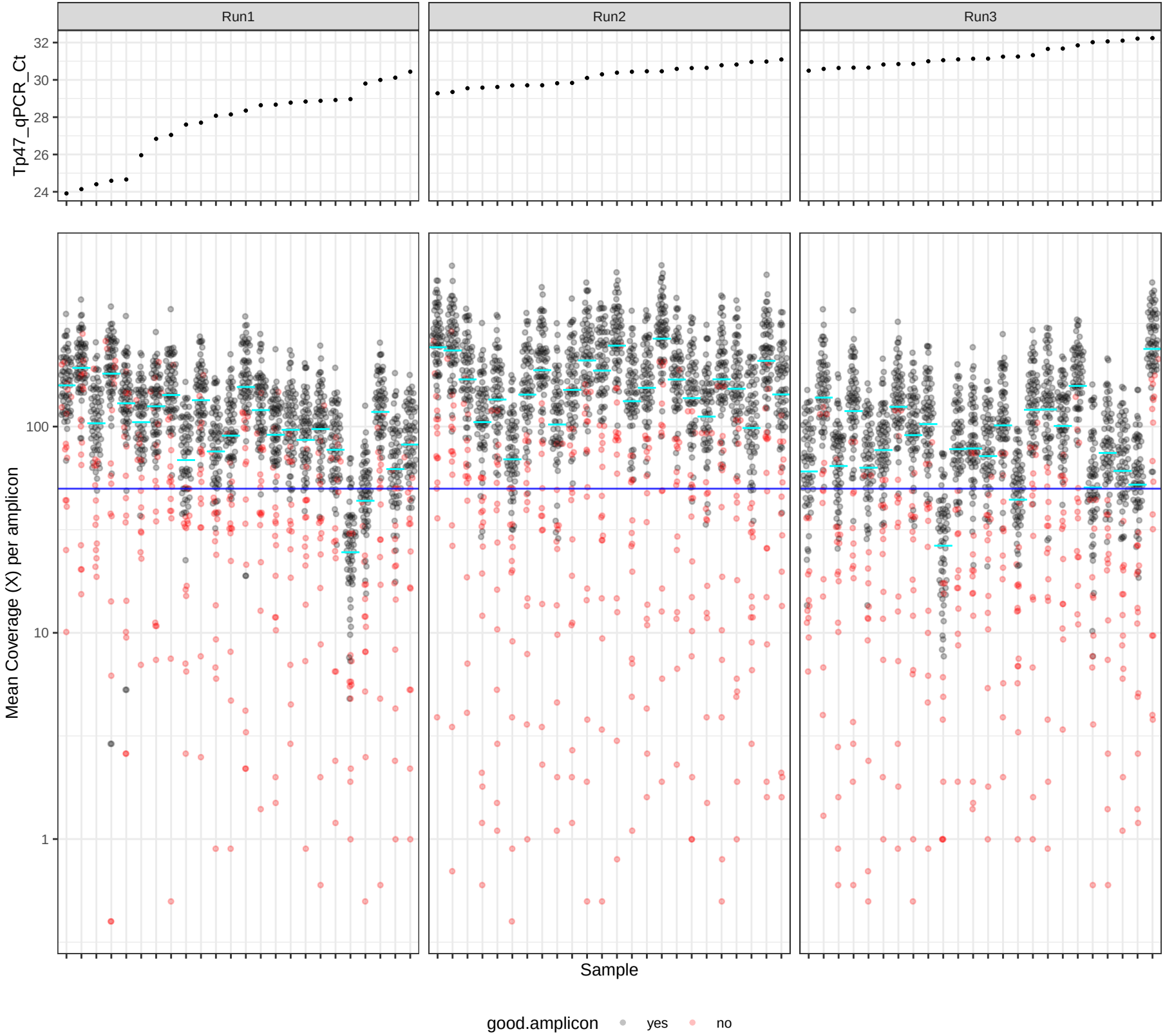
855 discriminating alleles across all 40 defined sublineages

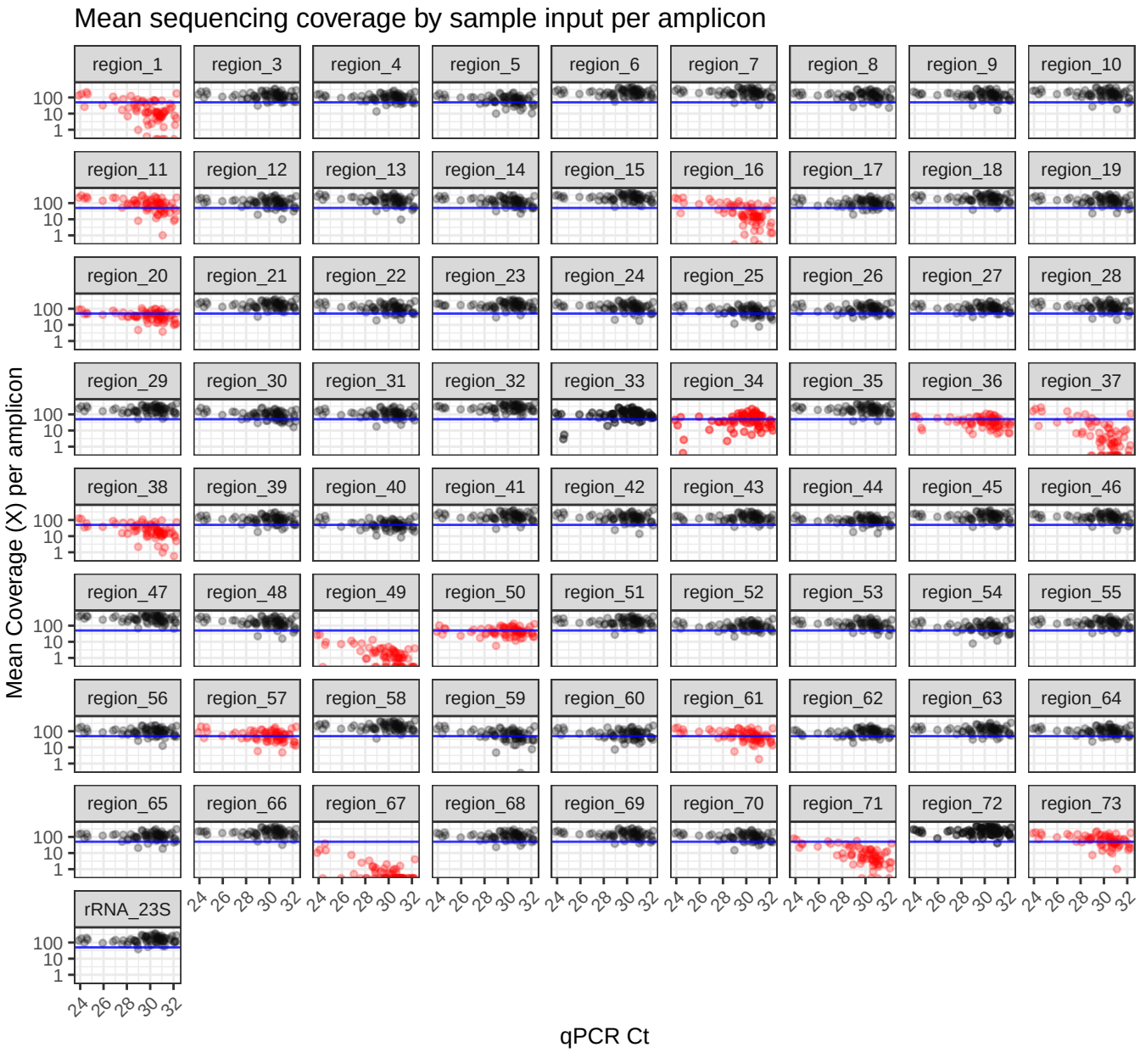

All Sites
Selected Regions
Final PCR
Discriminatory Site Count
Genomic Sublineage
Nichols
SS14
sublin−1
sublin−10
sublin−11
sublin−12
sublin−13
sublin−14
sublin−15
sublin−16
sublin−17
sublin−2
sublin−3
sublin−4
sublin−5
sublin−6
sublin−7
sublin−8
sublin−9
sublin−TPE20
sublin−TPE21
sublin−TPE22
sublin−TPE23
sublin−TPE24
sublin−TPE25
sublin−TPE26
sublin−TPE27
sublin−TPE28
sublin−TPE29
sublin−TPE30
sublin−TPE31
sublin−TPE32
sublin−TPE33
sublin−TPE34
sublin−TPE35
sublin−TPE36
sublin−TPE37
sublin−TPE38
sublin−TPE39
sublin−TPE40
sublin−TPE41
sublin−TPE42

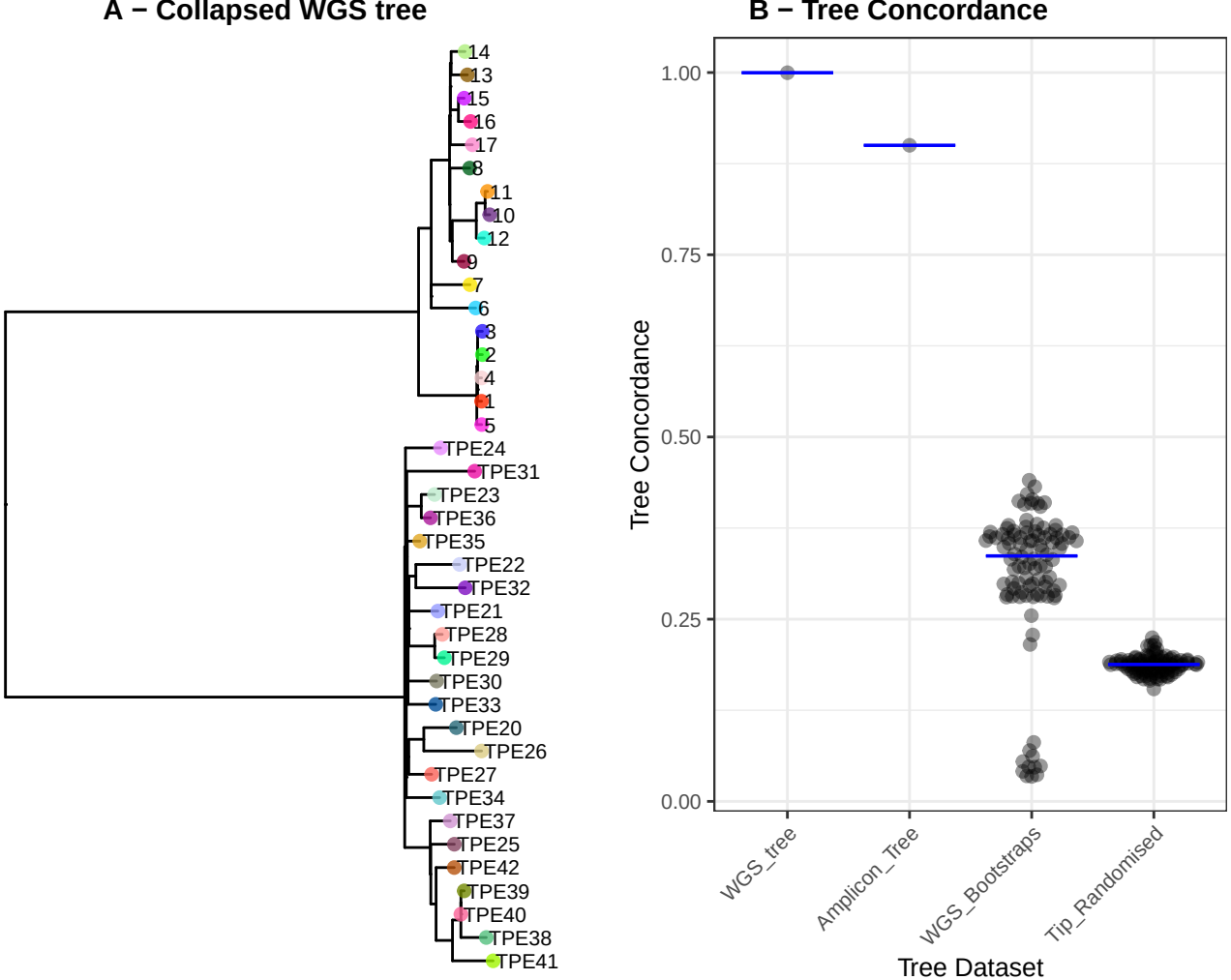

Population Specific Sites
sublin−TPE42
sublin−TPE41
sublin−TPE40
sublin−TPE39
sublin−TPE38
sublin−TPE37
sublin−TPE36
sublin−TPE35
sublin−TPE34
sublin−TPE33
sublin−TPE32
sublin−TPE31
sublin−TPE30
sublin−TPE29
sublin−TPE28
sublin−TPE27
sublin−TPE26
sublin−TPE25
sublin−TPE24
sublin−TPE23
sublin−TPE22
sublin−TPE21
sublin−TPE20
sublin−17
sublin−16
sublin−15
sublin−14
sublin−13
sublin−12
sublin−11
sublin−10
sublin−9
sublin−8
sublin−7
sublin−6
sublin−5
sublin−4
sublin−3
sublin−2
sublin−1
All Sites
1e+00
1e+02
1e+04
1e+06
Distance between adjacent discriminatory SNPs (bp)

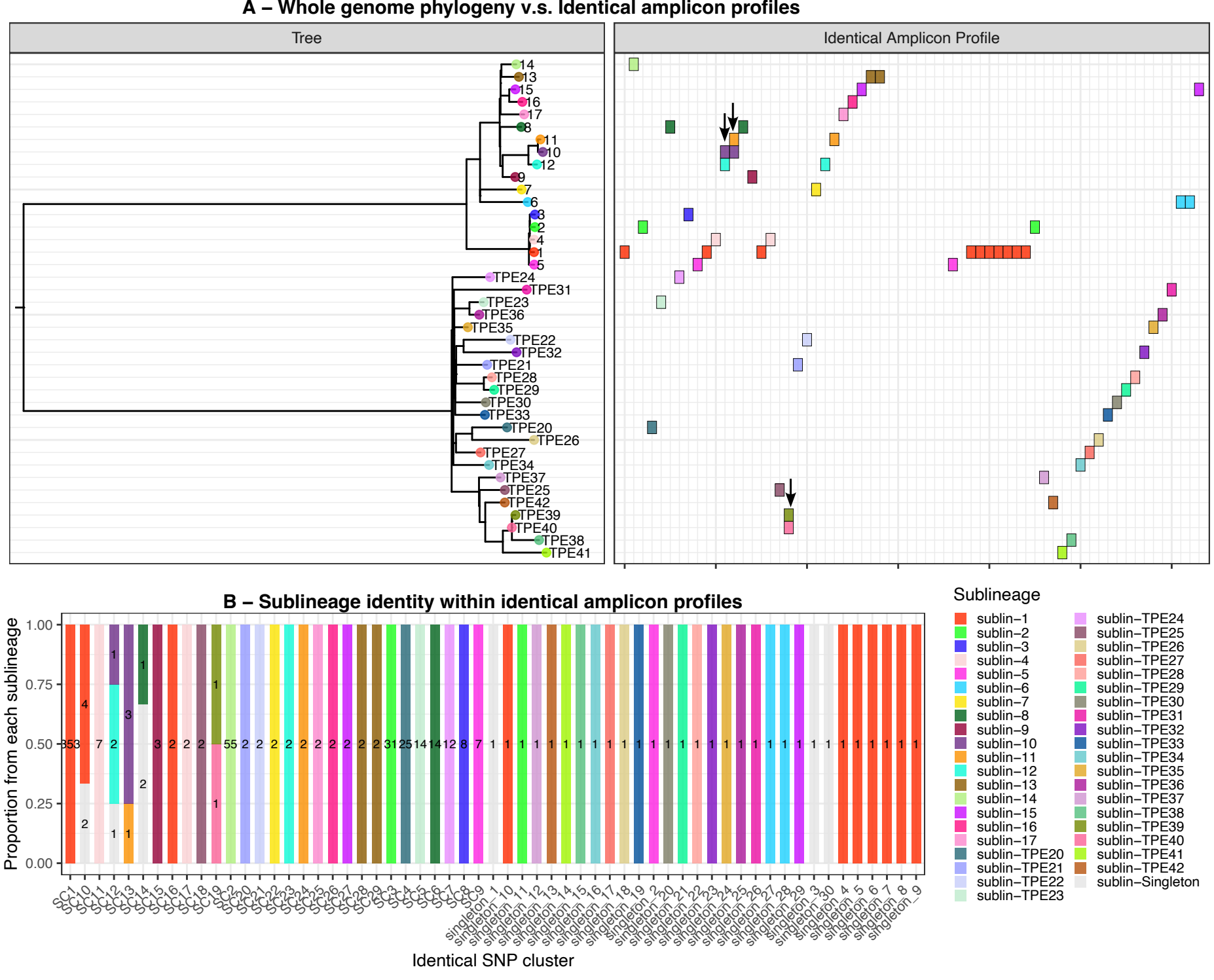

A
Cluster count
300
200
100
0
B
SNPs/Amplicon
60
50
40
30
20
10
0
C
Length of region spanning SNPs (bp)
10000
1000
100
10
1
0
100
200
300
400
500
600
700
800
900
1000
Cluster linkage size (bp)


Target Regions
Rank Regions by # Lineages supported
For Each Region & Lineage
Evaluate Region against Lineage
Evaluate next lineage
Yes
Does Region Support Lineage?
No
Lineages left to evaluate?
No
Yes
Lineage Counter
Lineage
Supporting Regions
1
0
2
0
n
0
Is Lineage Count < Threshold ?
No
Yes
+1 to Supported Lineages
Keep Regions
Discard Region
Output Design

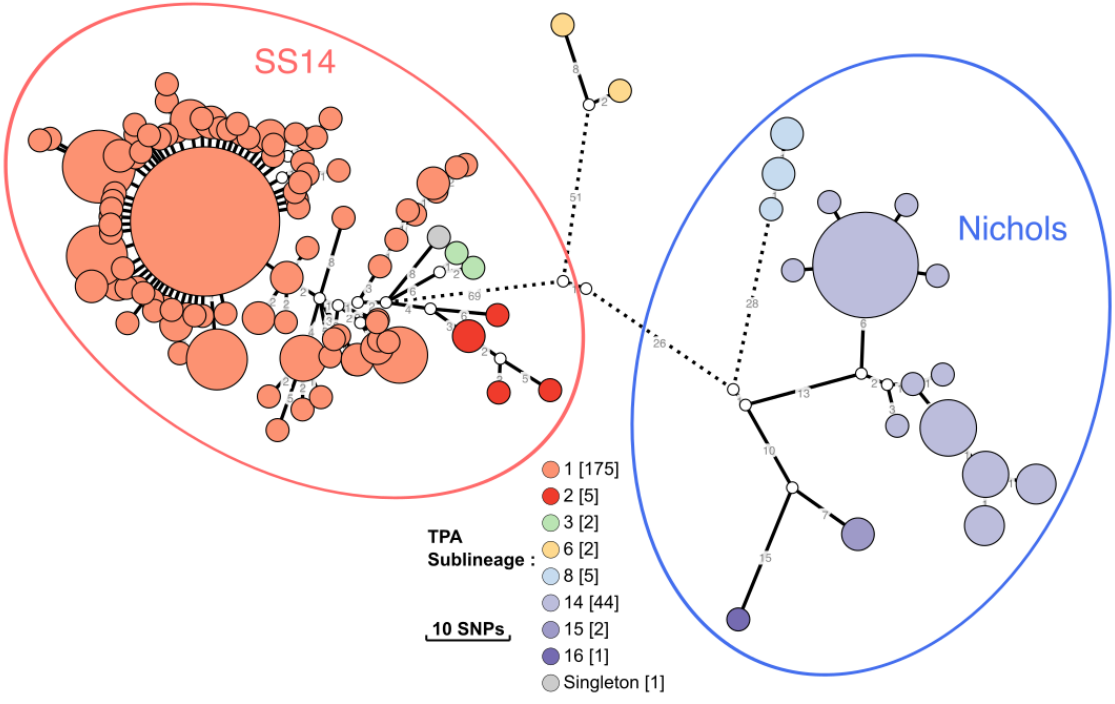

Run1
Run2
Run3
Tp47_qPCR_Ct
32
30
28
26
24
Mean Coverage (X) per amplicon
100
10
1
Sample
good.amplicon
yes
no


Mean sequencing coverage by sample input per amplicon
region_1
region_3
region_4
region_5
region_6
region_7
region_8
region_9
region_10
region_11
region_12
region_13
region_14
region_15
region_16
region_17
region_18
region_19
region_20
region_21
region_22
region_23
region_24
region_25
region_26
region_27
region_28
region_29
region_30
region_31
region_32
region_33
region_34
region_35
region_36
region_37
region_38
region_39
region_40
region_41
region_42
region_43
region_44
region_45
region_46
region_47
region_48
region_49
region_50
region_51
region_52
region_53
region_54
region_55
region_56
region_57
region_58
region_59
region_60
region_61
region_62
region_63
region_64
region_65
region_66
region_67
region_68
region_69
region_70
region_71
region_72
region_73
rRNA_23S
100
10
1
24
26
28
30
32
Mean Coverage (X) per amplicon
qPCR Ct
good.amplicon
yes
no

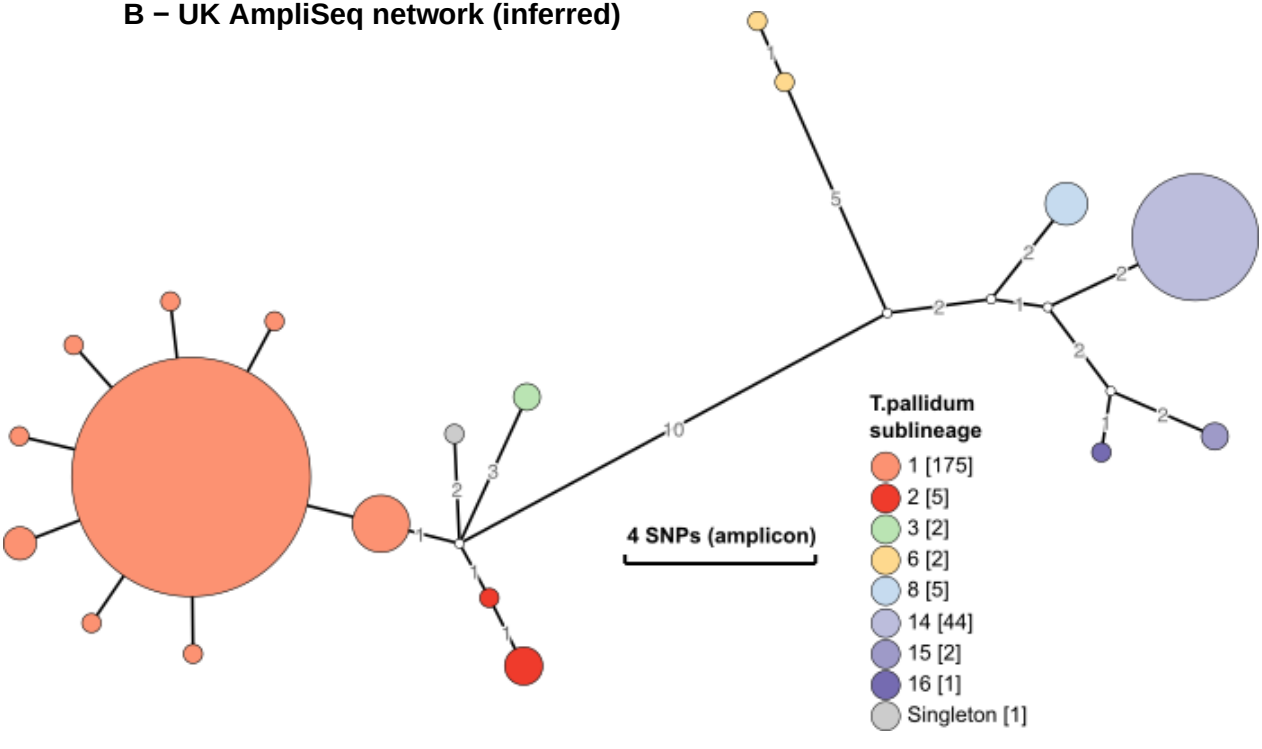

A – Collapsed WGS tree
14
13
15
16
17
8
11
10
12
9
7
6
3
2
4
1
5
TPE24
TPE31
TPE23
TPE36
TPE35
TPE22
TPE32
TPE21
TPE28
TPE29
TPE30
TPE33
TPE20
TPE26
TPE27
TPE34
TPE37
TPE25
TPE42
TPE39
TPE40
TPE38
TPE41
B – Tree Concordance
Tree Concordance
1.00
0.75
0.50
0.25
0.00
WGS_tree
Amplicon_Tree
WGS_Bootstraps
Tip_Randomised
Tree Dataset

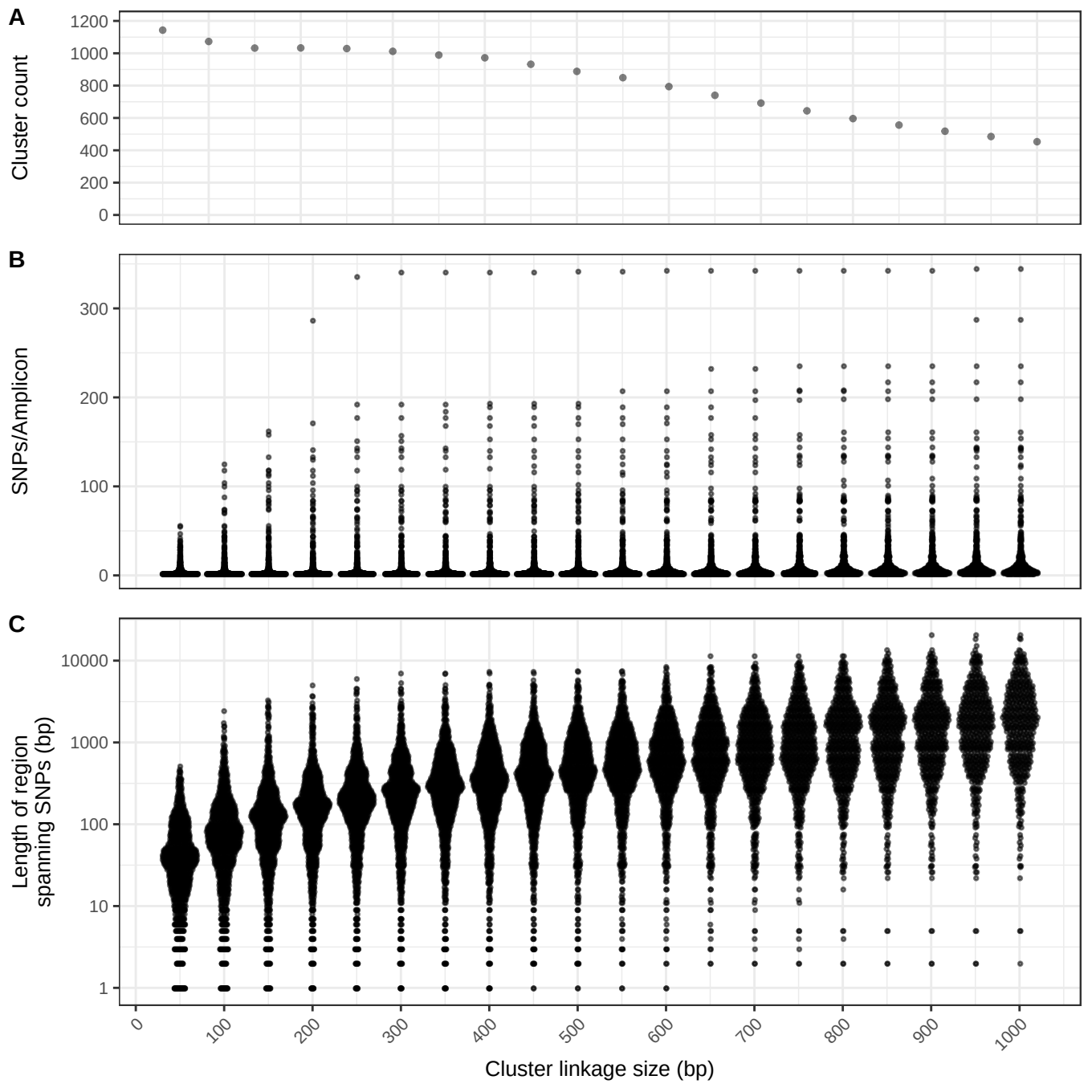

A – Whole genome phylogeny v.s. Identical amplicon profiles
Tree
Identical Amplicon Profile
B – Sublineage identity within identical amplicon profiles
Proportion from each sublineage
Identical SNP cluster
Sublineage
sublin−1
sublin−2
sublin−3
sublin−4
sublin−5
sublin−6
sublin−7
sublin−8
sublin−9
sublin−10
sublin−11
sublin−12
sublin−13
sublin−14
sublin−15
sublin−16
sublin−17
sublin−TPE20
sublin−TPE21
sublin−TPE22
sublin−TPE23
sublin−TPE24
sublin−TPE25
sublin−TPE26
sublin−TPE27
sublin−TPE28
sublin−TPE29
sublin−TPE30
sublin−TPE31
sublin−TPE32
sublin−TPE33
sublin−TPE34
sublin−TPE35
sublin−TPE36
sublin−TPE37
sublin−TPE38
sublin−TPE39
sublin−TPE40
sublin−TPE41
sublin−TPE42
sublin−Singleton

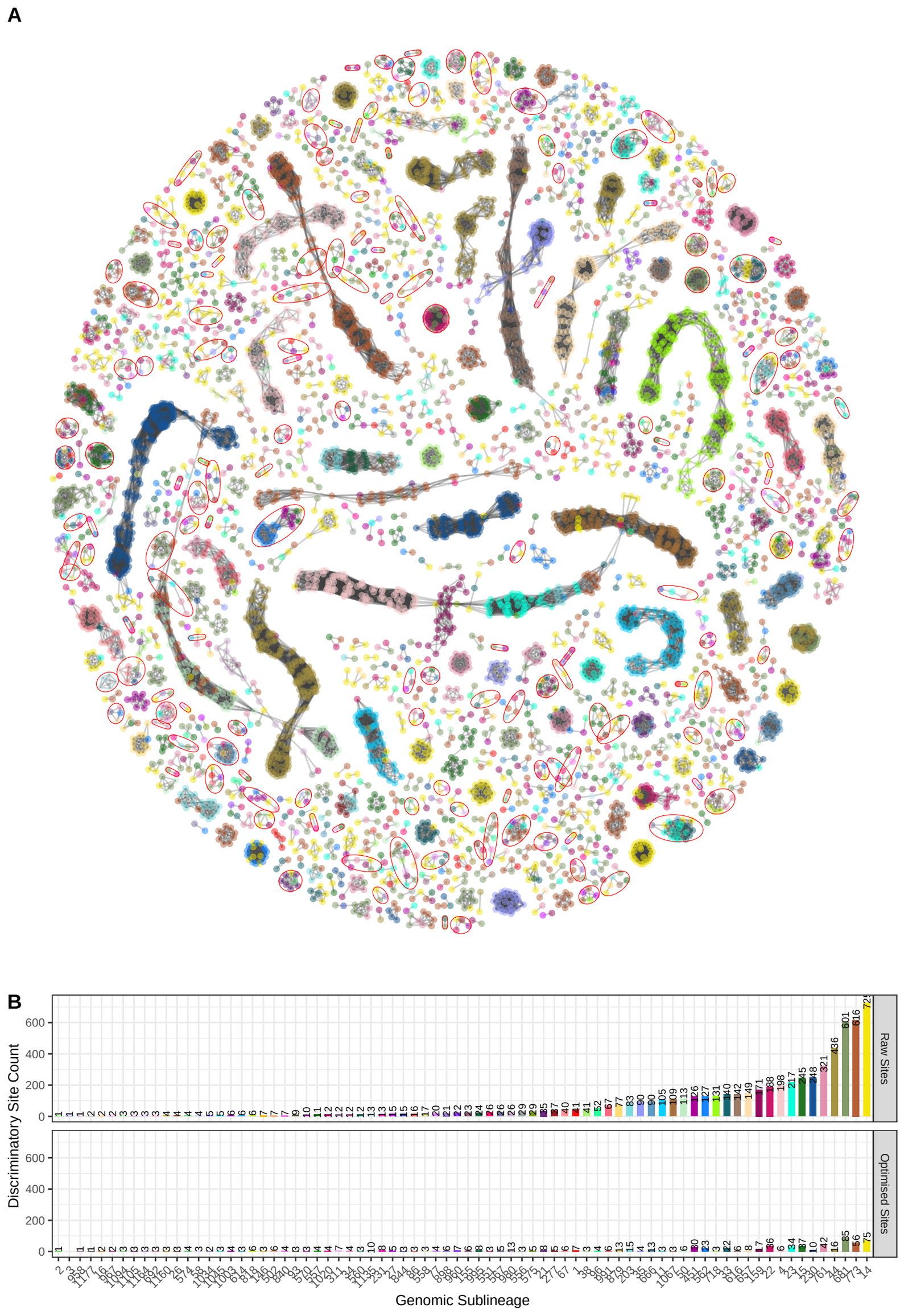
A – UK WGS network

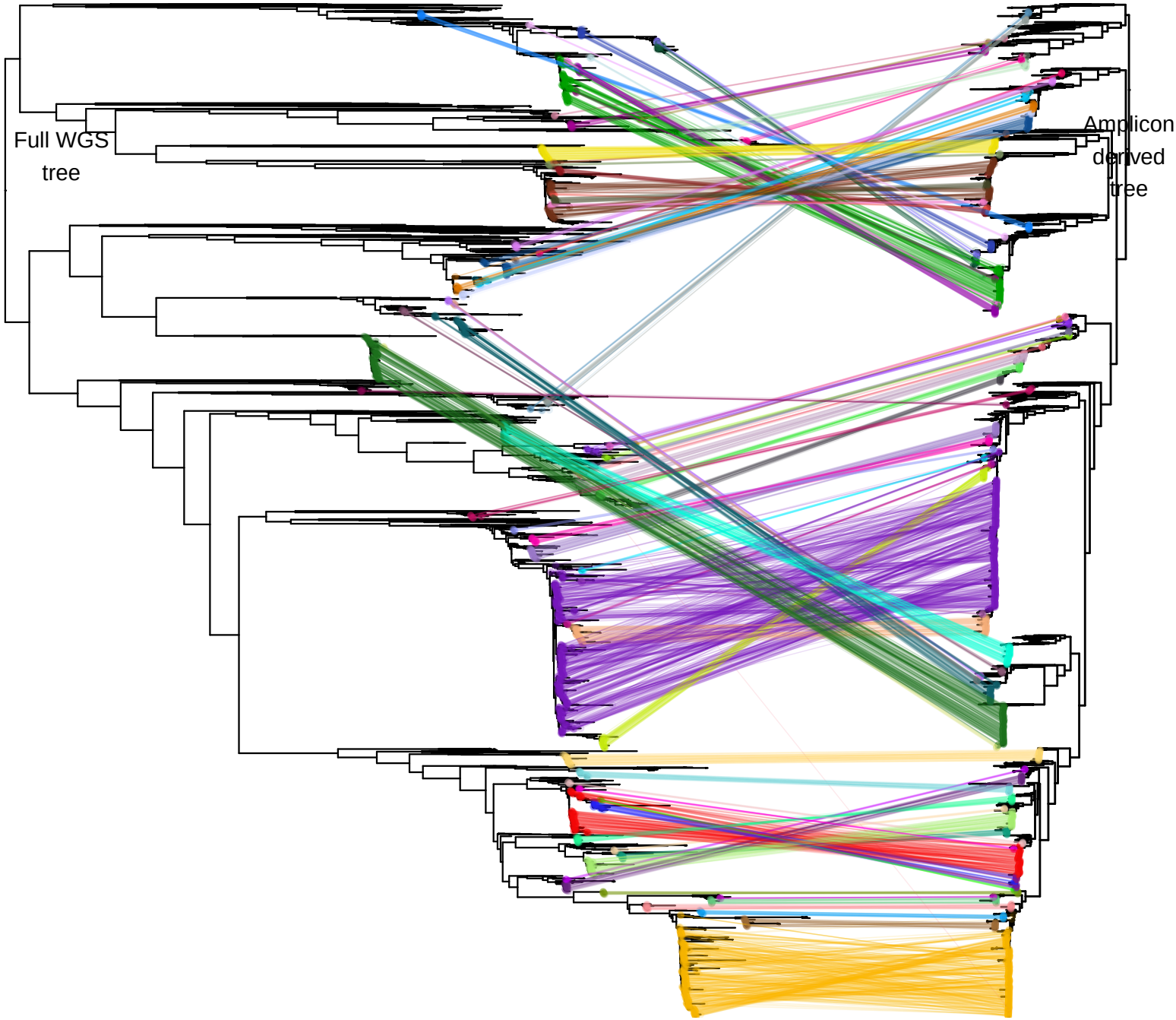


B – UK AmpliSeq network (inferred)



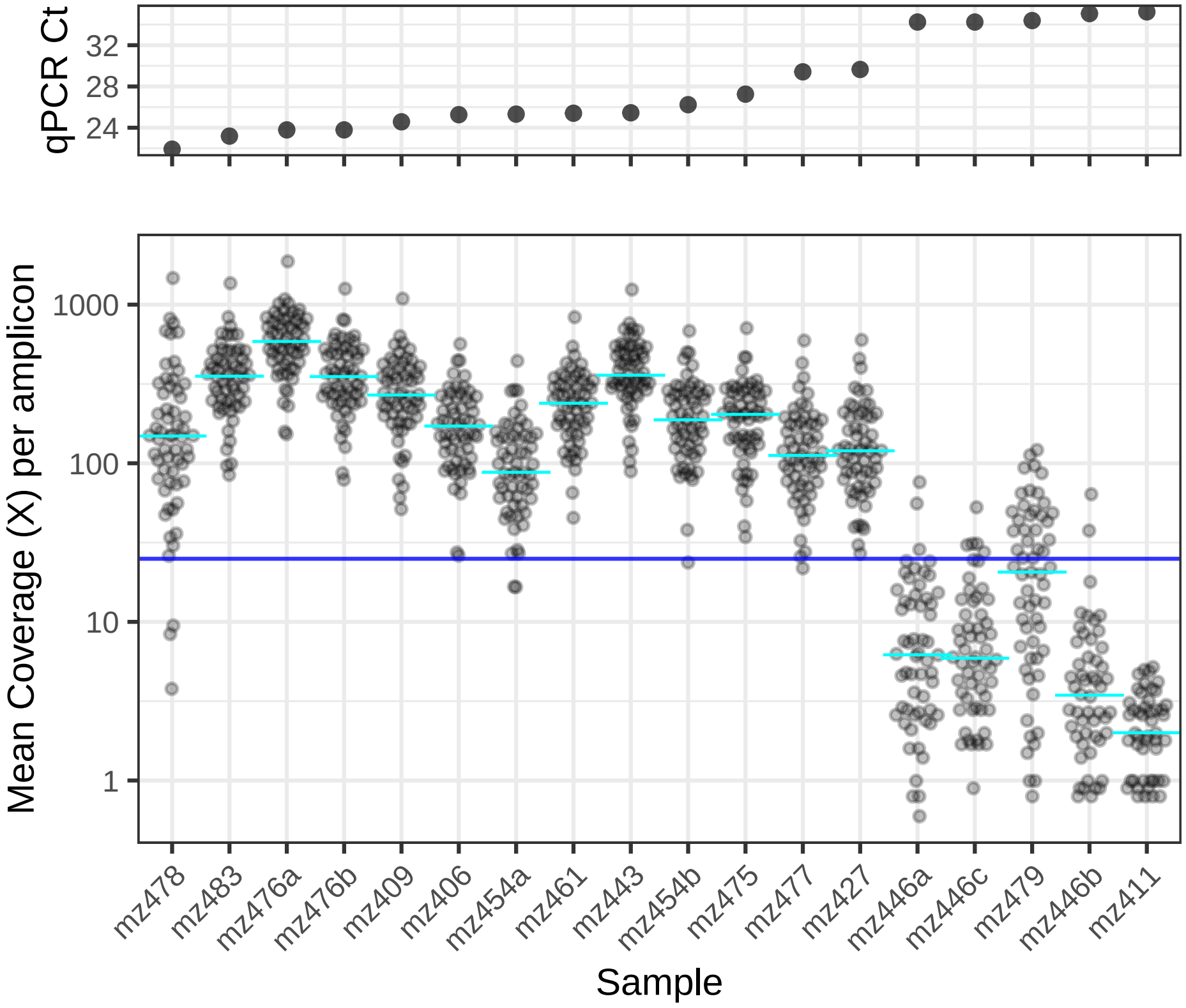

A
Cluster count
1200
1000
800
600
400
200
0
B
SNPs/Amplicon
300
200
100
0
C
Length of region spanning SNPs (bp)
10000
1000
100
10
1
0
100
200
300
400
500
600
700
800
900
1000
Cluster linkage size (bp)

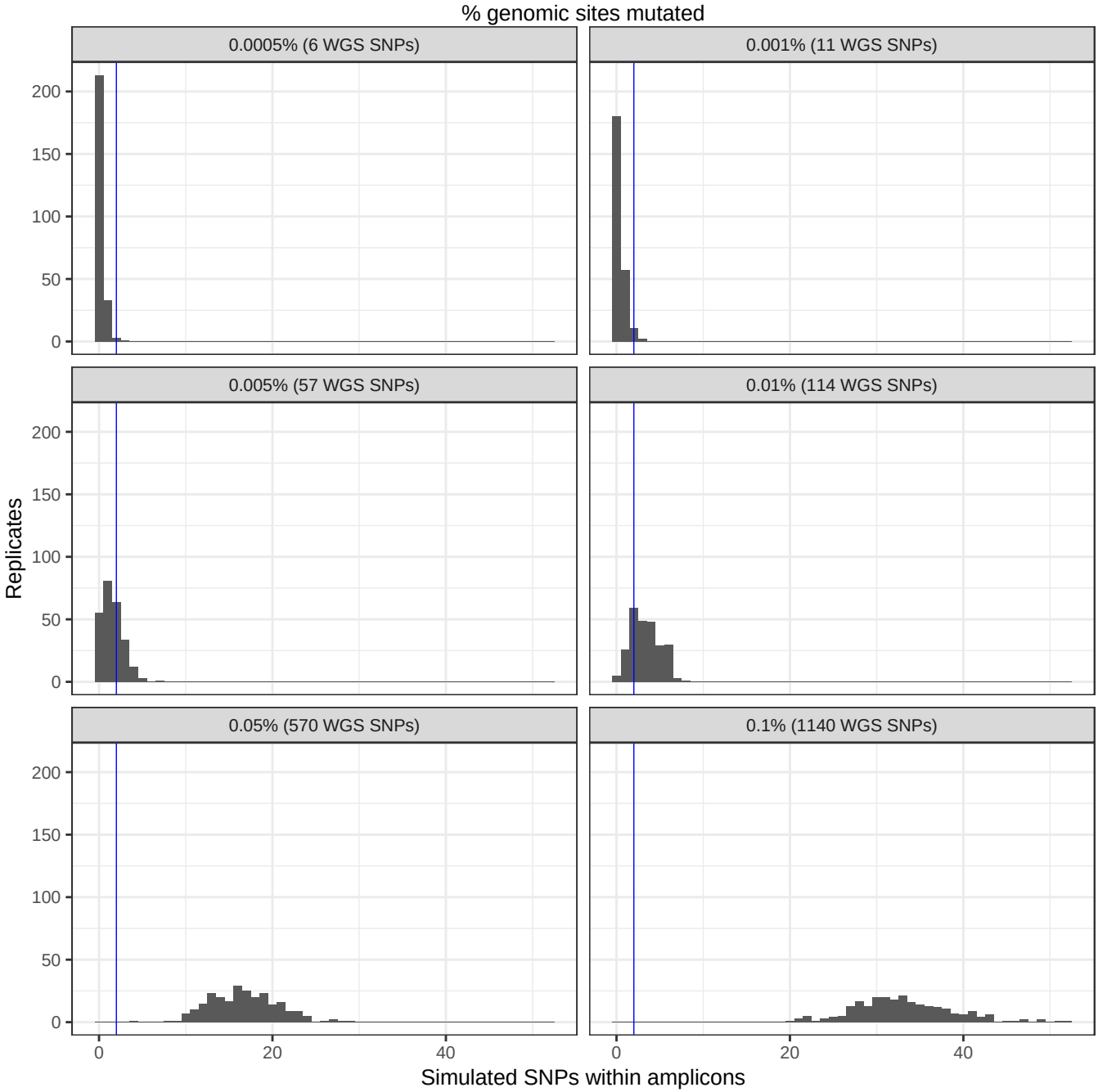

A
B
Discriminatory Site Count
Genomic Sublineage
Raw Sites
Optimised Sites
0
200
400
600

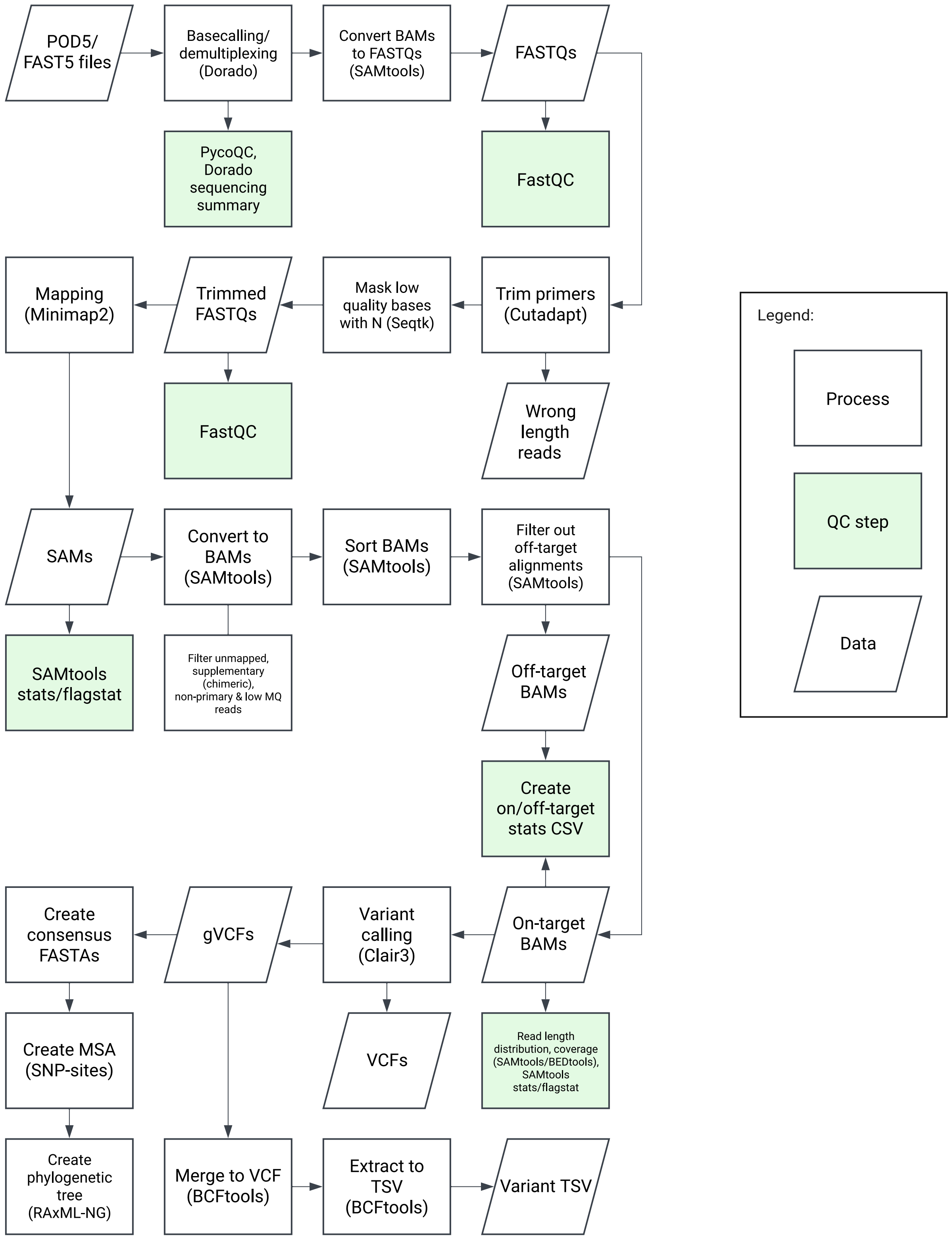


Sublineage

1 15 33 67 159 451 562 657 818 967 1045 1164
2 16 34 77 180 502 567 666 840 980 1067 1177
3 21 35 86 203 551 570 681 844 991 1094
4 22 38 93 230 556 574 692 879 995 1105
7 23 44 96 231 557 575 718 898 1003 1126
11 26 50 146 277 558 614 761 958 1020 1135
14 27 58 155 311 560 616 773 960 1034 1160


qPCR Ct
32
28
24
Mean Coverage (X) per amplicon
1000
100
10
1
mz478
mz483
mz476a
mz476b
mz409
mz406
mz454a
mz461
mz443
mz454b
mz475
mz477
mz427
mz446a
mz446c
mz479
mz446b
mz411
Sample


% genomic sites mutated
0.0005% (6 WGS SNPs)
0.001% (11 WGS SNPs)
0.005% (57 WGS SNPs)
0.01% (114 WGS SNPs)
0.05% (570 WGS SNPs)
0.1% (1140 WGS SNPs)
Replicates
0
50
100
150
200
0
20
40
Simulated SNPs within amplicons


POD5/ FAST5 files
Basecalling/ demultiplexing (Dorado)
Convert BAMs to FASTQs (SAMtools)
FASTQs
PycoQC, Dorado sequencing summary
FastQC
Mapping (Minimap2)
Trimmed FASTQs
Mask low quality bases with N (Seqtk)
Trim primers (Cutadapt)
FastQC
Wrong length reads
SAMs
Convert to BAMs (SAMtools)
Sort BAMs (SAMtools)
Filter out off-target alignments (SAMtools)
SAMtools stats/flagstat
Filter unmapped, supplementary (chimeric), non-primary & low MQ reads
Off-target BAMs
Create on/off-target stats CSV
Create consensus FASTAs
gVCFs
Variant calling (Clair3)
On-target BAMs
Create MSA (SNP-sites)
VCFs
Read length distribution, coverage (SAMtools/BEDtools), SAMtools stats/flagstat
Create phylogenetic tree (RAxML-NG)
Merge to VCF (BCFtools)
Extract to TSV (BCFtools)
Variant TSV
Legend:
Process
QC step
Data