Neisseria gonorrhoeae LIN codes: a Robust, Multi-Resolution Lineage Nomenclature

Anastasia Unitt[1], Made Krisna[2], Kasia M. Parfitt[2], Keith A. Jolley[2], Martin C.J. Maiden[2], Odile B. Harrison*[1]

1: Nuffield Department of Population Health, University of Oxford.

2: Department of Biology, University of Oxford.

*Corresponding Author: odile.harrison@ndph.ox.ac.uk

Abstract

Investigation of the bacterial pathogen *Neisseria gonorrhoeae* is complicated by extensive horizontal gene transfer: a process which disrupts phylogenetic signals and impedes our understanding of population structure. The ability to identify *N. gonorrhoeae* lineages in a consistent manner is essential for surveillance of this increasingly antimicrobial resistant organism, and to facilitate efficient communication regarding its epidemiology; however, all current typing systems fail to reflect *N. gonorrhoeae* strain taxonomy in a way that is both reliable and stable. Here, a *N. gonorrhoeae* genomic lineage nomenclature based on the barcoding system of Life Identification Number (LIN) codes was developed using a refined 1430 core gene MLST (cgMLST). This hierarchical LIN code nomenclature conveys lineage information at multiple levels of resolution within one code, enabling it to provide immediate context to an isolate's ancestry, and to relate to familiar, previously used typing schemes such as Ng cgMLST v1, 7-locus MLST, or NG-STAR clonal complex (CC). Clustering with LIN codes accurately reflects gonococcal diversity and population structure, providing insight into associations between genotype and phenotype for traits such as antibiotic resistance. These codes are automatically assigned and publicly accessible via the pubmlst.org/organisms/neisseria-spp database.

Introduction

Taxonomic classification is necessary to unravel evolutionary relationships, while also building nomenclatures that facilitate effective communication of where an organism falls on a phylogeny [1]. This is particularly important for lineages within bacterial species, which are often distinct in clinically relevant phenotypes such as antimicrobial resistance (AMR) or virulence [1]. The ability to identify these lineages consistently via a stable nomenclature system is essential for epidemiological surveillance [1].

In *N. gonorrhoeae,* the identification of AMR associated lineages is particularly relevant. Significant surges in the incidence of gonococcal infection have emerged in several regions since the start of the 21st century, including in the UK [2, 3], and outbreaks have occurred associated with antibiotic resistant lineages which have then spread, sometimes globally [4, 5]. Effective surveillance of AMR is key to informing measures to suppress the expansion of these clades. Since the emergence of multi-drug resistance in *N. gonorrhoeae,* and its inclusion in the WHO priority list for AMR, the necessity of surveillance has only increased [3, 6]. Globally, gonorrhoea causes a high burden of disease with an estimated 86.9 million new cases annually, which, if not treated, can cause sequelae such as infertility and pelvic inflammatory disease [6, 7]. Facing the dual challenges of AMR and gonococcal vaccine development [8], accurate characterisation of gonococcal population structure and consistent identification of key lineages is of upmost importance [9, 10].

Numerous approaches have been applied to *N. gonorrhoeae* lineage taxonomy, which can cause confusion and miscommunication. Underlying this disparate nomenclature is the complex population structure of the gonococcus; it is a highly recombinogenic organism, meaning it experiences extensive horizontal gene transfer (HGT) both with itself and with other species [11-14]. Polymorphisms are continually reassorted through this process, leading to low levels of linkage disequilibrium and fundamentally disrupting clonal population structure [11, 15]. As a result, *N. gonorrhoeae* has been described as a “sexual clone”, reflecting the strong impact of HGT and the therefore weak clonal signal present in gonococcal genomes [10, 16]. This low signal means that many molecular typing methods that are effective at capturing lineages in relatives such as *N. meningitidis* are not as reliable in *N. gonorrhoeae* [10]. The result has been a proliferation of differing approaches and nomenclatures seeking to capture *N. gonorrhoeae* taxonomy, particularly since 2010 [17-19].

Most of these methods are Multi Locus Sequence Typing (MLST) based. MLST is a widely used molecular approach for characterising bacterial isolates through the combination of alleles across a set of genes (an ‘allelic profile’) [20]. Each potential combination of alleles across the profile represents a unique sequence type (ST), which can be used, where they

correspond with clonal inheritance, to characterise lineages [1]. Isolates can be further grouped based on these STs, using methods such as goeBURST [21] or single linkage clustering [22].

MLST is a powerful approach that has had widespread applications in the analysis of bacterial lineages [1]; however, conventional MLST approaches such as 7 locus MLST use only a small number of loci, which can make this system more vulnerable to disruptive HGT. In bacteria such as *N. gonorrhoeae*, HGT is extensive enough that it affects housekeeping genes used in 7 locus MLST [10]. This makes 7-locus MLST a sub-optimal tool when applied in *N. gonorrhoeae* taxonomy, as two isolates with the same ST may not actually be closely related, having come by the same ST by HGT rather than clonal inheritance [10]. The same process affects several gonococcal typing systems, including NG-MAST [10, 18] and NG-STAR [10, 19].

Typing systems using large numbers of loci can address this problem. For example, core gene MLST (cgMLST) is an extension of the MLST approach that uses a profile of hundreds of core genes to gain higher resolution while also diluting the taxonomy-jumbling effects of HGT of some of those loci [1, 10, 23, 24]. In *N. gonorrhoeae* cgMLST v1, genes were defined as core if they were present in >95% of genomes, resulting in a core gene list (scheme) of >1600 core genes [10]. If two isolates were identical across this scheme, allowing for up to 50 loci to be unannotated, they would belong to the same core genome sequence type (cgST) [10]. Single-linkage clustering was then applied to define core genome 'groups' based on various threshold levels of allelic differences within and between groups [10]. This clustering was necessary as the large number of loci used in cgMLST results in large numbers of unique cgSTs.

This method provided a high-resolution classification of *N. gonorrhoeae* lineage taxonomy, while being less affected by HGT than 7-locus MLST, NG-STAR or NG-MAST [10]. However, single-linkage clustering is susceptible to group fusion when intermediate isolates are found that bridge the identity thresholds between linkage groups [24, 25]. HGT can exacerbate this issue, homogenising sequences across lineages by spreading polymorphisms throughout the population. As a result, the Ng cgMLST v1 core genome group nomenclature was not stable, and evolved over time as more *N. gonorrhoeae* isolates were sampled, decreasing the resolution provided and potentially leading to confusion as the nomenclature shifts. Therefore, there is still a need for a *N. gonorrhoeae* lineage nomenclature that provides similarly effective discrimination between gonococcal lineages, while also remaining consistent over time.

Here, a definitive nomenclature for *N. gonorrhoeae* is proposed, using LIN codes. This isolate barcoding system has previously been applied to *Klebsiella pneumoniae* [25] and *Streptococcus pneumoniae* [26], where LIN codes were shown to provide a high resolution

nomenclature, enabling accurate insight into the population structure of these organisms. LIN code combines the HGT-diluting effects of a cgMLST scheme with the stability of a numeric barcode [24, 25]. This stability is derived from the fixed nature of these barcodes for each isolate, along with the flexibility of a multi-position numeric code which can increase infinitely as new variants are sequenced [24, 25, 27]. The *N. gonorrhoeae* LIN code developed here was implemented in PubMLST, a freely accessible bacterial genomics database [28]. LIN codes are automatically assigned to all WGS uploaded to PubMLST, making this taxonomy easily applicable to old and new datasets, and encouraging the widespread use of this reliable nomenclature.

Methodology

Representative Isolate Collections

Data were extracted from the PubMLST database (pubmlst.org/organisms/neisseria-spp), a publicly accessible bacterial genomics database [28]. At the time of writing, sequence records from over 28,000 gonococci were available in the database. To facilitate efficient analyses representative sub-datasets were generated.

Dataset 1 consisted of a representative collection of 896 isolates belonging to a range of core genome groups at the 300 allelic mismatch threshold [Figure 1]. This was assembled by including all isolates belonging to core genome groups represented by 6 or fewer isolates, in combination with a random selection of 6 isolates from each of the core genome groups represented by more than 6 isolates (Supplementary table 1). Randomisation was achieved in R, using the sample function [29]. This smaller dataset was used for preliminary analysis, as it required less computational power.

Dataset 2 consisted of 3935 isolates [Figure 1]. This was assembled by including all isolates belonging to core genome groups represented by 20 or fewer isolates (again at allelic mismatch threshold 300), and a random sample of 30% of all core genome groups represented by more than 20 isolates, capped at 350 isolates per core genome group (Supplementary table 2). This larger dataset was used for final analysis once techniques had been explored using dataset 1.

Development of Ng cgMLST v2

Although a 1649 locus Ng cgMLST v1 scheme for *N. gonorrhoeae* has been published [10], an updated, more readily auto-annotated cgMLST scheme was needed to ensure any WGS deposited in the PubMLST database can be assigned a LIN code with little to no curator involvement [25, 26]. This was a consequence of Ng cgMLST v1 including a number of genes

with alternate start codons or length variation, complicating automatic allele annotation by PubMLST.

WGS data from dataset 1 were downloaded from PubMLST and annotated with Prokka using default parameters [30]. The resulting .gff output was further analysed with PIRATE for pangenome analysis including the identification of core genes [31]. PIRATE was run using default parameters with the following options: -a (align all genes), -r (plot summaries using r) and –t 24 (24 threads). A preliminary threshold of 95% presence was chosen as ‘core’ rather than 100% to ensure a large enough gene list was identified. This also allowed for the possibility that some genes may be absent due to factors such as mis-assembly or mutation while still being ‘core’ in nature. Genes with duplication events as detected by PIRATE were excluded.

The new core list was assessed in dataset 2 to identify possible curation issues across a wider range of isolates. Following this analysis, all loci that did not meet a threshold of 98% presence in dataset 2 were excluded. This much-reduced Ng cgMLST v2 therefore prioritised loci that annotate with very little human involvement via manual curation.

The PubMLST Grapetree plugin was used to generate minimum spanning trees based on the Ng cgMLST v2 loci, annotated with core genome groups (based on Ng cgMLST v1), to ensure resolution was high enough with this new reduced scheme to explore established gonococcal lineages [32]. Grapetree compares isolates across a user-defined allelic profile, clustering those that share the most alleles [32].

Analysis of Population Structure and Threshold Selection

Thresholds for LIN code were chosen based on an analysis of natural breaks in population structure, assessed by examining the distribution of pairwise allelic mismatches amongst isolates from dataset 2, combined with testing in a local installation of LIN code [33] and the application of clustering statistics including Rand index and silhouette score.

In order to extract pairwise allelic distances for these analyses, a distance matrix was generated comparing dataset 2 isolates across the loci included in Ng cgMLST v2. This was accomplished using the genome comparator plugin in PubMLST. This tool provides gene-by-gene pairwise analysis of isolates across a list of loci, generating a variety of outputs including a distance matrix, core gene list, and alignments [28]. Here, genome comparator was run with default settings. The distances in this matrix were stacked to form a dataframe with three columns (isolate A, isolate B, distance), allowing the distances to be plotted in a histogram using the ggplot2 package in R [34].

Plots were created based on these data to explore the number of allelic mismatches associated with clusters identified by pre-existing typing systems such as NG-STAR clonal complexes and Ng cgMLST v1 core genome groups. This was done by extracting the pairwise allelic distances associated with each instance of matching NG-STAR CC, rMLST-ST, or matching Ng cgMLST v1 core genome group, from the distance matrix. This was then visualised in a ridgeline plot drawn in R using the packages ggplot2 and ggridges [34, 35].

Based on the observed breaks in population structure, various bin thresholds were tested using a local version of LIN code downloaded from https://gitlab.pasteur.fr/BEBP/LINcoding applied to dataset 2 [33]. The 3935 isolates yielded 3877 unique cgSTs for testing.

Statistics were applied to assess clustering output from the local LIN code tests at various allelic mismatch thresholds, specifically the silhouette score and Rand index. The silhouette score indicates the cohesion of clustering at a particular threshold, varying from -1 to 1 with more positive values indicating higher cluster cohesiveness [36]. The average silhouette score was calculated using MSTClust (v0.21b) (downloaded from https://gitlab.pasteur.fr/GIPhy/MSTclust) for pairwise distance thresholds from 0.01 – 0.7 [37]. Results were plotted using the R package ggplot2 [34].

The Rand index is a statistical measure that indicates the level of similarity between two different partitions of the same dataset, with values approaching 1 indicating higher similarity [38]. The Rand index was applied here to compare local LIN code clustering at various thresholds against Ng cgMLST v1 core genome group clustering at threshold 300 and 400. The index was calculated in R using the package fossil [39].

Finally, LIN codes generated using these thresholds on the representative dataset were visualised on a minimum spanning tree to explore what length of code prefix would be most suited to exploring population structure at various resolutions.

Implementation and Analyses

Once the refined Ng cgMLST v2 scheme was defined in PubMLST and isolates annotated with sequence types, the chosen thresholds were implemented within the database to generate LIN codes. A maximum of 25 missing loci in the Ng cgMLST v2 scheme was tolerated. If this quality threshold was not met, no cgST would be assigned to the isolate, and hence no LIN code. LIN codes were designated in PubMLST ordered based on a minimum spanning tree, using a default batch size of 10,000 isolates [24]. Multilevel single-linkage clustering was used to classify isolates at each threshold within the code. The combination of these clusters with a fixed bar code facilitates LIN code's stability [25, 26]. Subsequently,

isolate records that have a cgST but no LIN code assigned will be scanned once a week, and LIN codes automatically designated.

Once LIN codes were designated within the database, PubMLST was further applied to investigate the distribution of LIN codes and to examine their association with other isolate typing systems such as 7-locus MLST. This was undertaken using the 'Field breakdown' and 'Combinations' analysis tools, alongside exporting datasets for manipulation.

A published dataset of 171 ceftriaxone resistant gonococci from diverse phylogroups [40] was selected for analysis, to validate LIN code's ability to reproduce complex phylogenies. Isolates from this dataset not already present in PubMLST were downloaded from the sequence read archive, and where necessary were assembled using SPAdes/3.15.4-GCC-12.3.0 [41]. These isolates were then uploaded to PubMLST and associated under the appropriate publication record (Fifer et al., 2024). One isolate (H18-368) out of 171 no longer had sequence data available, and so could not be included in this analysis.

Phylogenetic Trees

Minimum spanning trees were drawn using the Grapetree plugin in PubMLST [32]. Neighbour joining trees were constructed using the ITOL plugin within the PubMLST interface.

Nucleotide alignments for other trees were generated via genome comparator in PubMLST [28], using the MUSCLE algorithm, with settings to align all loci, not only variable loci. Maximum likelihood trees were constructed using RaxML [42], and approximate maximum likelihood trees using FastTree [43]. Trees were corrected for recombination using ClonalFrameML [44] and edited using ITOL [45].

Results

Core genome MLST Version 2

The final Ng cgMLST v2 included 1430 loci. Of these, 362 were hypothetical genes of unknown function, 96 encoded transferases, 56 synthases, 45 transporters, 34 50S ribosomal proteins, 29 transcriptional regulators, and 21 30S ribosomal proteins (Supplementary table 3). This is a decrease of 219 loci compared to Ng cgMLST v1, with the largest categories of excluded loci being hypothetical (83 loci) or phage associated (18 loci) (Supplementary table 4). Notable exclusions include *tbpA* and *tbpB,* essential iron acquisition proteins that are hypervariable and so not readily auto-annotated. In total 1418 loci are shared between the two schemes, with 12 new to Ng cgMLST v2 (Supplementary table 4).

Of the 1430 loci included, 99% (1418/1430) had under 1000 alleles, and an average length of less than 3000 base pairs [Figure 2]. Exceptionally varied genes included NEIS2020 *porB* (5544 alleles), NEIS2644 an unnamed phage tail protein (3475 alleles), NEIS0829 *pilW* (1474 alleles) and NEIS2113 *tamB* (1122 alleles) [Figure 2c].

It should be noted that while the new scheme is fundamentally still a core genome MLST scheme, it does not include all core genes, and should not be used as a definitive core gene list. However, the strict inclusion criteria for Ng cgMLST v2 facilitated automated assignment of a cgST to a higher proportion of WGS data than its predecessor, while still providing a similar degree of resolution.

Analysis of Population Structure and Allelic Mismatch Thresholds

The number of allelic mismatches present in a pairwise comparison across the 3935 isolates in dataset 2 was assessed across the 1430 core genes in Ng cgMLST v2. When visualised as a histogram [Figure 3a], an uneven distribution of allelic mismatches was observed, exhibiting a highly varied incidence. The allelic mismatch modes indicate clusters of isolates that share the same proportion of alleles in common across the core genome scheme, reflecting natural breaks in the gonococcal population structure [25].

The most frequently identified number of allelic mismatches between isolates was 840/1430 (59%). This modal mismatch is low compared to that seen in *K. pneumoniae* (within species mode 520/629 mismatches, 83%) [25]. There were no instances of pairwise allelic mismatches exceeding 970/1430 loci (68%), meaning that all isolates shared a minimum of 460 loci with one other isolate in the dataset. This means that compared to *K. pneumoniae, N. gonorrhoeae* isolates shared more of the same alleles across their core genome. This may be due to the homogenising effect of widespread HGT.

When pairwise allelic mismatches were analysed in isolates belonging to the same grouping, such as Ng cgMLST v1 core genome groups or NG-STAR CCs, breaks in population structure associated with these metrics became visible [Figure 3b]. For isolates belonging to the same core genome group, NG-STAR CC, or ribosomal ST, most pairs had under 125/1430 allelic mismatches between them [Figure 3b]. This indicates that shared identity in approximately 1305/1430 (91%) Ng cgMLST v2 loci is required to belong to the same core genome group, NG-STAR CC or ribosomal ST. Interestingly, core genome groups at threshold 400 displayed a second peak at 510/1430 (36%) mismatches. This is a result of group fusion due to intermediate genotypes, which has led to divergent gonococci being included in the same group at this threshold. Isolates from the same country displayed a comparatively high average level of allelic mismatch, with a mode of 855/1430 (60%), indicative of limited geographical association of gonococcal lineages.

Rand index values when comparing LIN code bin thresholds to Ng cgMLST v1 core genome groups (at threshold 300) peaked at 1 for a 300-locus mismatch threshold (21%) [Figure 3c]. Similarly, the silhouette score demonstrated a peak of 0.64 at cut-off 0.225 [Figure 3d].

Based on this evidence, allelic mismatch thresholds were chosen for each bin of the LIN code. Key thresholds at 125 loci (8.74% mismatch) and 300 (20.98% mismatch) were chosen due to their correlation with breaks in population structure [Figure 3a], alongside a Rand index close to 1 demonstrating their relevance to core genome groups, and silhouette score indicating high cluster cohesiveness at these thresholds [Figure 3c & d]. Higher thresholds at 550 loci (38.46% mismatch) and 650 (45.45%) captured superlineage divisions.

To provide the high resolution necessary to identify transmission events or outbreak-related gonococcal variants, thresholds at lower allelic mismatch were added. Together, this resulted in a set of 11 bins, corresponding to mismatch thresholds: 650, 550, 300, 125, 25, 10, 7, 5, 3, 1, and 0 [Figure 4]. Minimum spanning tree analysis indicated that a prefix using the first three thresholds was ideal for exploring lineages [Figure 5] corresponding to Ng cgMLST v1 core genome groups at threshold 300.

Implementation of LIN code

Across the >28,000 publicly available *N. gonorrhoeae* genomes stored in PubMLST, LIN codes were assigned to 25,912 isolates. At the superlineage level (2 threshold prefix: 0_0) there were 118 clusters, at lineage (3 prefix: 0_0_0) 532 clusters, and at sublineage (4 prefix: 0_0_0_0) 1712 clusters. The 3-threshold prefix lineage definitions directly correlated with Ng cgMLST v1 core genome groups at threshold 300 (adjusted [adj.] Rand Index = 1). This was visualised using phylogenetic trees [Figure 6] and minimum spanning trees [Figure 5] which showed good congruence between LIN codes, SNP-based phylogeny, and other clustering methods. The lineages identified by LIN code demonstrated persistence over time, with lineage 0_2_1 isolates spanning 39 years from 1985 to 2024 (Supplementary Figure 1)

N. gonorrhoeae LIN codes can be accessed via an isolate's information page, from the 'allele designations/scheme fields' dropdown box within the pubmlst.org/organisms/neisseria-spp isolate database search form, or through the 'export dataset' link at the bottom of the page following an isolate search. They can also be exported as metadata in Grapetree analyses.

Congruence with Previous Typing Schemes

7-locus MLST-STs showed good association with LIN code lineages (0_0_0) (adj. Rand index = 0.92) [Figure 5], although many STs included multiple lineages. For example, MLST-ST 1901 corresponded with lineages 0_2_0 (1932/2376, 81%) and 0_0_12 (443/2376, 19%), while MLST-ST 7363 was predominantly captured by 0_0_52 (796/1322, 60%) and 0_2_17

(355/1322, 27%) (Supplementary table 5) The fact these LIN codes are from different superlineages (0_0 & 0_2) illustrates how these MLST-STs form polyphyletic groupings, combining isolates that are not closely related under the same ST.

NG-STAR clonal complexes (CCs) showed a stronger association with LIN code sublineages (0_0_0_0) (adj. Rand index 0.96) [Figure 5]. NG-STAR CC 90 belonged predominantly to sublineages 0_2_0_1 (1499/1562, 96%) and to a much lesser extent 0_0_12_2 (25/1562, 2%). NG-STAR CC 63 was represented mainly by 0_2_1_10 (1369/1437, 95%), followed by 1_1_22_9 (21/1437, 1%) and others (Supplementary table 5). Again, the diversity of these LIN codes at the superlineage level, and comparison with their location on the phylogeny [Figure 5 & 6], demonstrates how genetically divergent isolates can belong to the same CC.

NG-MAST-STs showed some correlation with LIN code groups (0_0_0_0_0) (adj. Rand index 0.50). For example, NG-MAST-ST 2992 belonged mostly to group 0_2_1_0_73 (409/653, 63%) and to others such as 0_2_1_1_4 (72/653, 11%). NG-MAST-ST 1407 predominantly belonged to group 0_2_0_1_0 (498/787, 63%) and to 0_2_0_1_27 (77/787, 10%).

Resistance associated lineage A, as described using hierarchical Bayesian analyses (BAPs) [5], was previously shown to associate well with Ng cgMLST v1 core genome groups [10]. As LIN code lineages correspond to core genome groups at threshold 300 [Figure 6], the same is true of this nomenclature. Lineage 0_2_0, 0_2_1, and 0_2_17 correspond to core genome group 3, 16 and 8, which in turn correlate with BAPS 8, 7, and 9 respectively, all within the resistance associated lineage A [5, 10].

Mosaic *penA* type 34 alleles, an important antimicrobial resistance determinant, were strongly associated with LIN code lineage, with 89.43% of these alleles found in isolates belonging to lineage 0_2_0 (Supplementary Table 6).

Example Analysis

The gonococcal LIN code was applied in a reproduction of an analysis of 170 ceftriaxone resistant gonococci initially performed by Fifer et al., 2024 [40] (Supplementary Table 7). A maximum likelihood tree of these isolates was reconstructed and labelled using LIN codes [Figure 7], and demonstrated that LIN code lineages recaptured the diverse clades identified by phylogenetic methods, while also classifying several isolates that were not assigned a phylogroup in the original publication. Furthermore, 15 pairs or triplets of isolates sharing a full-length LIN code were identified within this dataset. Comparison with a random selection of isolates from across the PubMLST database [Figure 6] suggested an overrepresentation of 0_14_0 isolates in the ceftriaxone-resistant dataset. This dataset can be accessed by filtering by publication “Fifer et al., 2024” on the PubMLST *Neisseria* isolate search page.

Discussion

Accurate identification of bacterial lineages is necessary to examine links between bacterial population structure and characteristics such as AMR, and to enable surveillance of emergent or outbreak-associated variants [1]. However, in *N. gonorrhoeae,* widespread HGT reassorts DNA between isolates, disrupting clonal inheritance, and consequently distorting phylogenetic trees [10, 46]. This makes it difficult to accurately characterise gonococcal lineages, particularly when using conventional typing systems based on small numbers of genes, such as 7-locus MLST [10, 15]. Previously, the most reliable approach applied core genome MLST (cgMLST) to define core genome groups, using over 1000 genes to dilute the effects of HGT and facilitate high resolution analyses [1, 10, 47]. Unfortunately, this method suffers from cluster instability [25, 46]. LIN code provides an alternative means to leverage the benefits of cgMLST within a stable classification system, which can accommodate new genotypes without the risk of disrupting established clustering [24, 25]. Here, we developed a *N. gonorrhoeae* LIN code: an effective nomenclature for categorising, exploring, and discussing gonococcal population structure.

Evidence for the existence of distinct, persistent gonococcal lineages pre-dates whole genome sequencing, in the characterisation of auxotypes: related isolates that exhibit similar patterns of growth in media containing different combinations of key nutrients [48]. This may be due to metabolic competition between gonococcal strains, causing the population structure to segregate into distinct metabolic types [49]. Previous research has theorised that selection can preserve such lineages even in the face of extensive HGT, as has been the case in immune selection for differing antigenic types in other bacterial species [49-51]. Later sequence-based approaches including 7-locus MLST, NG-STAR and cgMLST confirmed that the *N. gonorrhoeae* population contains identifiable groups of related isolates, some of which are associated with clinically relevant phenotypes such as AMR [10, 19, 52, 53], with outbreak events [4], or with specific at-risk groups such as men who have sex with men (MSM) [5, 54]. However, some of these typing systems are prone to providing misleading classifications, where isolates that are not closely related appear to be, due to the effects of HGT within the small numbers of loci used to classify them. Consequently, while equivalent systems have proved highly effective at categorising the closely related *N. meningitidis* [55], the extensive HGT observed across the gonococcal genome necessitates a whole genome approach to reliably identify related groups of *N. gonorrhoeae* isolates [10]. LIN code achieves this by applying cgMLST, using hundreds of loci belonging to a wide range of functional categories, including genes involved in metabolic pathways, genetic processing, and AMR. This allows the LIN code to capture gonococcal population structure with increased consistency, and

higher resolution, than many previous approaches, providing fresh insight into the biology that underlies this species' genetic composition.

The LIN nomenclature conveys lineage information in the form of hierarchical clustering at sequential thresholds of tolerated allelic mismatch within a cgMLST scheme (i.e. the number of alleles differing out of the total list of loci in the scheme) [25, 26]. Here, 11 thresholds were used, resulting in an 11-bin barcode. The leftmost numbers of the barcode represent clustering at the highest thresholds of allelic mismatch tolerated, here corresponding with LIN super-lineage (up to 550 mismatches), LIN lineage (300 mismatches), and LIN sublineage (125 mismatches). Progressing across the barcode to the right, each number indicates clustering at an increasing level of allelic similarity, ultimately resulting in the delineation of highly related isolates that differ in a very small number of core loci [Figure 4] [24-26]. Subsets of thresholds, for example a prefix of the first three or four, may be used to define clustering down to a particular level of taxonomic similarity [25, 26]. If two isolates are highly related and therefore identical across their core gene profile, they will share the same LIN code.

The use of multiple thresholds enables the LIN code to relate back to several familiar nomenclatures, such as core genome groups and 7-locus MLST, which are both broadly equivalent to LIN lineages, and NG-STAR CCs, approximately equivalent to LIN sublineages. The difference lies in the accuracy and stability of the LIN code versus its predecessors. Visualisation via phylogenetic trees demonstrated that the LIN codes related strongly to phylogenetic structure, capturing distinct clades without overlap. This is in contrast to previous nomenclatures such as 7-locus MLST, which often form polyphyletic groups due to the misleading effects of HGT [10].

Furthermore, the multi-resolution, hierarchical nature of LIN codes also allows this nomenclature to encode more information about the relationship between isolates than conventional typing systems [24]. For example, if a pair of isolates belong to MLST 1 and 2, it can only be deduced that they differ in between 1 and 7 housekeeping genes. Meanwhile, if the same pair of isolates belong to LIN lineages 0_2_1 and 0_2_2, this communicates that these isolates differ by no more than 550/1430 core gene alleles, but by more than 300/1430, quickly defining the degree of their relatedness [Figure 4]. This is useful when comparing isolates, for example those associated with AMR. Previous approaches used to track AMR strains include both 7-locus MLST and NG-STAR, which types isolates based on their alleles across seven AMR genes (*penA*, *mtrR*, *porB*, *ponA*, *gyrA*, *parC*, and 23S rRNA) [56]. This typing system has been successfully used in many analyses of AMR associated gonococci, and NG-STAR CCs have been recommended as a tool for characterising gonococcal lineages [19, 52]. However, due to HGT specific NG-STAR types or clonal complexes do not always

represent closely related isolates. LIN sublineages were able to discern isolates with a similar level of resolution to NG-STAR CCs, but higher reliability. This will facilitate more accurate tracking of AMR associated isolates, such as those belonging to LIN sublineage 0_2_0_1, when compared to the equivalent NG-STAR CC 90, which also includes unrelated isolates belonging to LIN superlineages 0_0 and 1_1. While NG-STAR effectively characterises an isolate's AMR genotype [56], it cannot consistently distinguish related isolates. LIN code fulfils this role, while also providing increased information about the degree of relatedness of the isolates in question. Accurately defining this lineage structure will be important to elucidate how AMR emerges in distinct gonococcal lineages.

As an allele-based method, LIN codes enable analysis of a large number of isolates quickly in a readily reproducible manner [24, 25, 27]. This is an improvement on phylogenetic tools used to identify gonococcal lineages which rely on nucleotide sequence alignments, requiring significant time and expertise to compute [57]. LIN codes can replicate similar clustering, identifying the same lineages while being fully automated [24]. For example, hierarchical Bayesian analyses (BAPs) have previously been used to delineate gonococcal population structure into two key lineages: lineage A, associated with high-risk sexual networks and AMR, and lineage B, associated with lower risk networks and susceptibility [5]. In our LIN code analysis, we observed that lineage A (BAPS 8, 7 and 9) corresponds to LIN lineages 0_2_0, 0_2_1, and 0_2_17, reaffirming that cgMLST is able to provide the same resolution as SNP-based techniques [10]. While BAPs analysis can be complex and time intensive to run, LIN codes can be used to rapidly identify lineage A gonococci simply by uploading WGS data to PubMLST. Within 24 hours of upload, any sequence data in which 1405/1430 core loci can be annotated will be assigned a Ng cgMLST v2 cgST; if this is a previously identified cgST a LIN code will be assigned immediately, if a new cgST the new LIN code will be assigned within 7 days. Furthermore, these isolates can then be analysed in the context of the wider PubMLST gonococcal genome collection, which currently includes >28,000 public records and integrates accessible plugins such as GrapeTree [32] and Genome Comparator [28] for additional investigation.

To further illustrate the efficacy of LIN code we reproduced an analysis of 170 global ceftriaxone resistant isolates [40]. The original article's methodology involved a WGS alignment and generation of a maximum likelihood tree in order to characterise eight major phylogroups [40]. The gonococcal LIN code instantly reproduced these clusters, while providing additional detail about each clade in the form of superlineage/sublineage divisions and simultaneously contextualising the isolates amongst the wider PubMLST database. For example, lineage 0_14_0 appeared to be over-represented in this ceftriaxone-resistant dataset when compared against 1000 randomly selected *N. gonorrhoeae* isolates from

PubMLST. Also, LIN code was able to classify several divergent isolates that were not assigned to any phylogroup in the original publication, and detected 15 instances of matching full-length LIN codes, meaning these isolates were identical across their core genome and could therefore represent isolates associated with transmission events [10]. The PubMLST interface facilitates in-depth analysis of these instances of shared LIN codes, allowing users to explore related isolates at various thresholds of allelic dissimilarity by viewing a table of similar isolates (as defined by LIN code) on an isolate's information page [Figure 8]. This enables quick investigation of any related isolates, including their location, year, allele at a particular locus, or classification by other typing schemes such as NG-STAR.

The genetic mix-and-matching performed by *N. gonorrhoeae* can make characterisation of its population structure difficult [10]. However, the LIN code nomenclature proposed here provides clarity, consistency, and stability in its description of *N. gonorrhoeae* lineages. The multi-resolution clustering intrinsic to LIN code facilitates a common language around lineage nomenclature at different epidemiological levels, from high divisions such as superlineage down to unique clones [24]. In conclusion, *N. gonorrhoeae* LIN codes represent a portable, publicly available taxonomic nomenclature that has the potential to enhance surveillance of *N. gonorrhoeae* in order to benefit public health.

Author statements

Author contributions:

A.U created the updated Ng cgMLST v2 scheme, performed the allelic mismatch and statistical analyses, selected the LIN code thresholds, wrote the manuscript, and created all figures. O.B.H. and M.C.J.M conceptualised and supervised this work, and edited the manuscript. M.A.K contributed python and linux scripts and provided guidance on pangenome analysis. K.A.J implemented the cgMLST scheme and LIN code within the PubMLST database. K.M.P provided assistance with recombination correction.

Funding information:

A.U was funded by the BBSRC Interdisciplinary Biosciences DTP (Grant no. BB/M011224/1), and subsequently an ECR Fellowship through the Nuffield Department of Population Health (NDPH), University of Oxford. M.C.J.M, K.A.J and PubMLST are funded by the Wellcome trust (Grant no. 218205/Z/19/Z). O.B.H was funded by the Wellcome trust (Grant no. 214374/Z/18/Z), and subsequently though a Senior Research Fellowship at NDPH, University of Oxford. M.A.K's studentship was funded by the Ministry of Education, Indonesia, in collaboration with the Medical Science Division, University of Oxford. K.M.P was funded by the Wellcome trust (Grant no. 214374/Z/18/Z).

Conflicts of interest:

The authors declare that there are no conflicts of interest.

Data availability:

Isolate IDs and metadata of relevant datasets are included in the Supplementary documents. Ng cgMLST v2 and the gonococcal LIN code is publicly accessible via PubMLST (https://pubmlst.org/organisms/neisseria-spp).

Acknowledgements:

The authors would like to thank Nazreen Hadjirin and Iman Yassine for their valuable advice. Computational aspects of this work were enabled by the Oxford University Biomedical Research Computing (BMRC) facility.

Figures

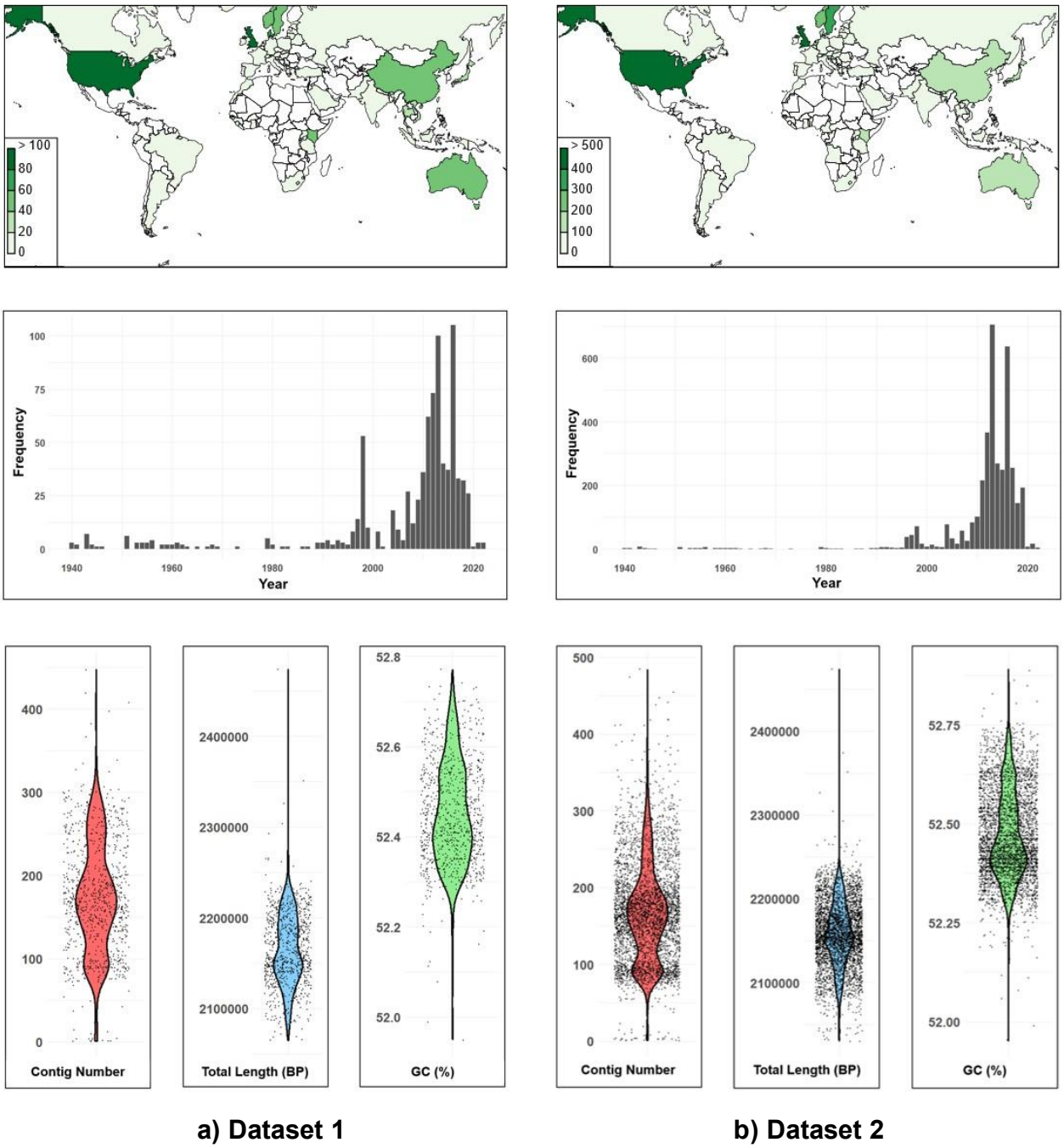
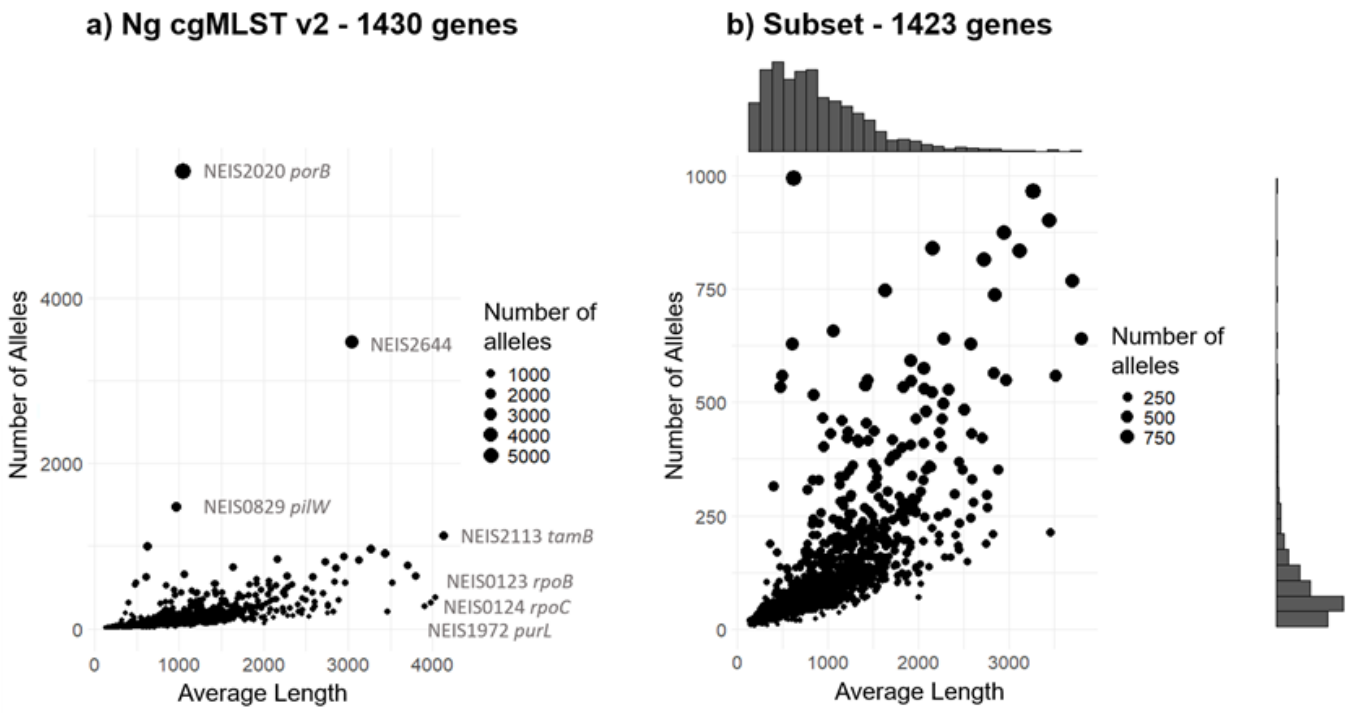


Figure 1) Characteristics of isolates within representative development datasets 1 and 2. *This includes geographical distribution (top panels), frequency of isolates sampled over time (middle panels), and genome quality statistics (lower panels) including i) contig number, ii) total genome length and iii) % GC content.*



c) Table of outliers

Locus number	Average length (BP)	Number of alleles	Gene name/function
NEIS2113	4137	1122	*tamB* Translocation protein
NEIS0123	4025	381	*rpoB* RNA polymerase subunit beta
NEIS0124	3980	314	*rpoC* RNA polymerase subunit beta'
NEIS1972	3909	273	*purL* Purine metabolism
NEIS2644	3047	3475	Phage Tail protein
NEIS2020	1047	5544	*porB* Porin
NEIS0829	968	1474	*pilW* Pilus assembly

Figure 2) Allele number vs allele length for the 1430 genes in Ng cgMLST v2*. Figure* ***2a)*** *All 1430 genes included in Ng cgMLST v2 are plotted.* ***2b****) The four genes with the highest average length, and the four with the highest number of alleles, were excluded as outliers. The figure was compiled excluding these genes in order to allow a closer examination of the distribution of allele length vs number. One gene, NEIS2113, appeared in both lists.* ***2c)*** *Table summarising the average length in base pairs, number of alleles, and gene name/function of the seven outlier loci.*

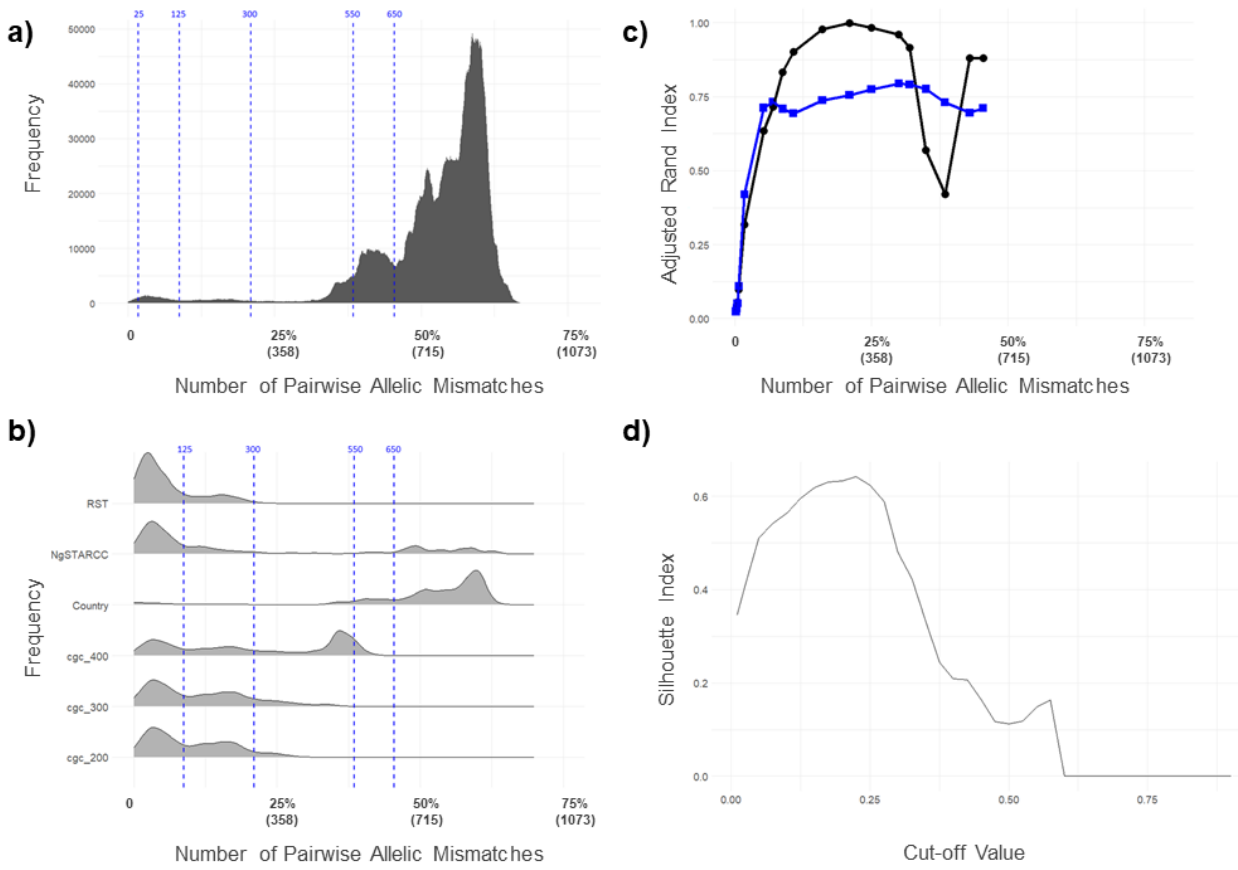


Figure 3) Plots used in the selection of allelic mismatch thresholds for the LIN code. *Figure **3a)** Histogram showing the frequency of pairwise allelic mismatches within dataset 2. A subset of the allelic mismatch thresholds applied in the gonococcal LIN code are shown (blue dashed lines) at 25 mismatches (1.75%), 125 (8.74%), 300 (20.98%), 550 (38.46%), and 650 (45.45%). **3b)** Ridgeline plots depicting the frequency of allelic mismatches amongst pairs of isolates that belong to the same category of different metrics from dataset 2. From top to bottom: Ribosomal MLST (RST), NG-STAR Clonal Complex (NgSTARCC), Country, Ng cgMLST v1 core genome group at threshold 400 (cgc_400), threshold 300 (cgc_300), and threshold 200 (cgc_200)). **3c)** Plot of adjusted Rand index comparing LIN code clustering at various allelic mismatch thresholds to Ng cgMLST v1 core genome groups at threshold 300 (black dots) and NG-STAR CC (blue squares). Clustering was compared using dataset 2. **3d)** Plot of silhouette index (score) at various cutoff values, based on MSTclust analysis of 1430 core loci across 3935 representative N. gonorrhoeae isolates (dataset 2). Silhouette score peaked at 0.64 at a cutoff value of 0.225.*

a)

	Superlineage		Lineage		Sublineage		Group											← Allele Mismatch threshold
	650	550	300	125	25	10	7	5	3	1	0							
Isolate A	0	2	1	0	73	64	0	1	0	2	0							(Across 1430 core genes)
Isolate B	0	2	1	0	73	64	0	1	0	2	0							
Isolate C	0	2	1	0	105	4	0	0	0	0	0							
Isolate D	1	1	2	0	15	0	0	1	0	0	3							

b)

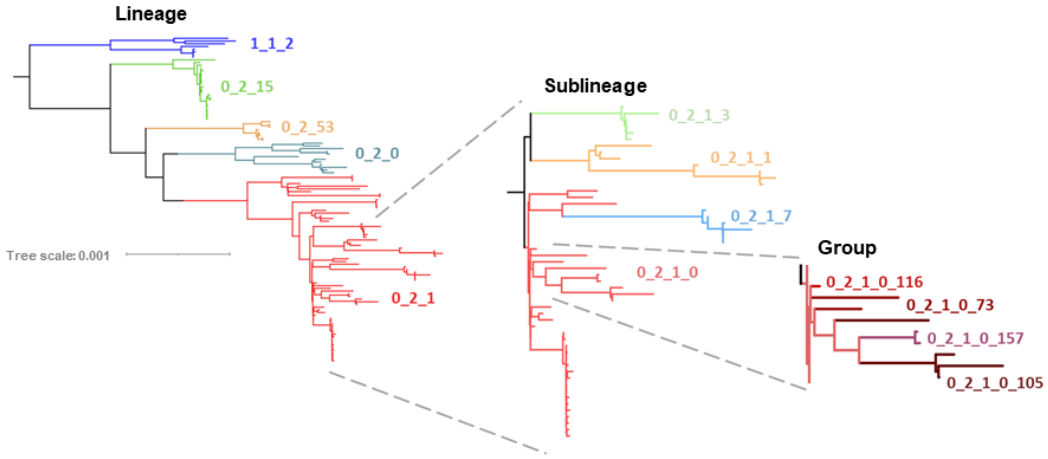


***Figure 4) Illustration of the gonococcal LIN code nomenclature**. **4a)** Each successive allelic mismatch threshold dictates clustering at a specific position within the code. This clustering is hierarchical, such that isolates sharing a larger proportion the code (from the left across) are of higher genetic similarity. For example, Isolate A and B share a complete LIN code, meaning they have 0 allelic mismatches in their Ng cgMLST v2 loci. Isolate B and C share the first three digits of their LIN code; they belong to the same clusters at these thresholds, and therefore differ in less than 300 alleles out of the 1430 core genes in Ng cgMLST v2 i.e., they belong to the same LIN code "lineage". **4b)** Rooted Maximum likelihood tree demonstrating how LIN codes reflect phylogenetic relationships. The first tree shows a subset of LIN code lineages within superlineage 0_2, with lineage 1_1_2 as the outgroup. Moving to the right, the figure focuses in on lineage 0_2_1, showing the higher resolution provided by LIN sublineages, and then groups. (Figure inspired by Figure 3 in Van Rensburg, Berger et al., 2024 [26])*

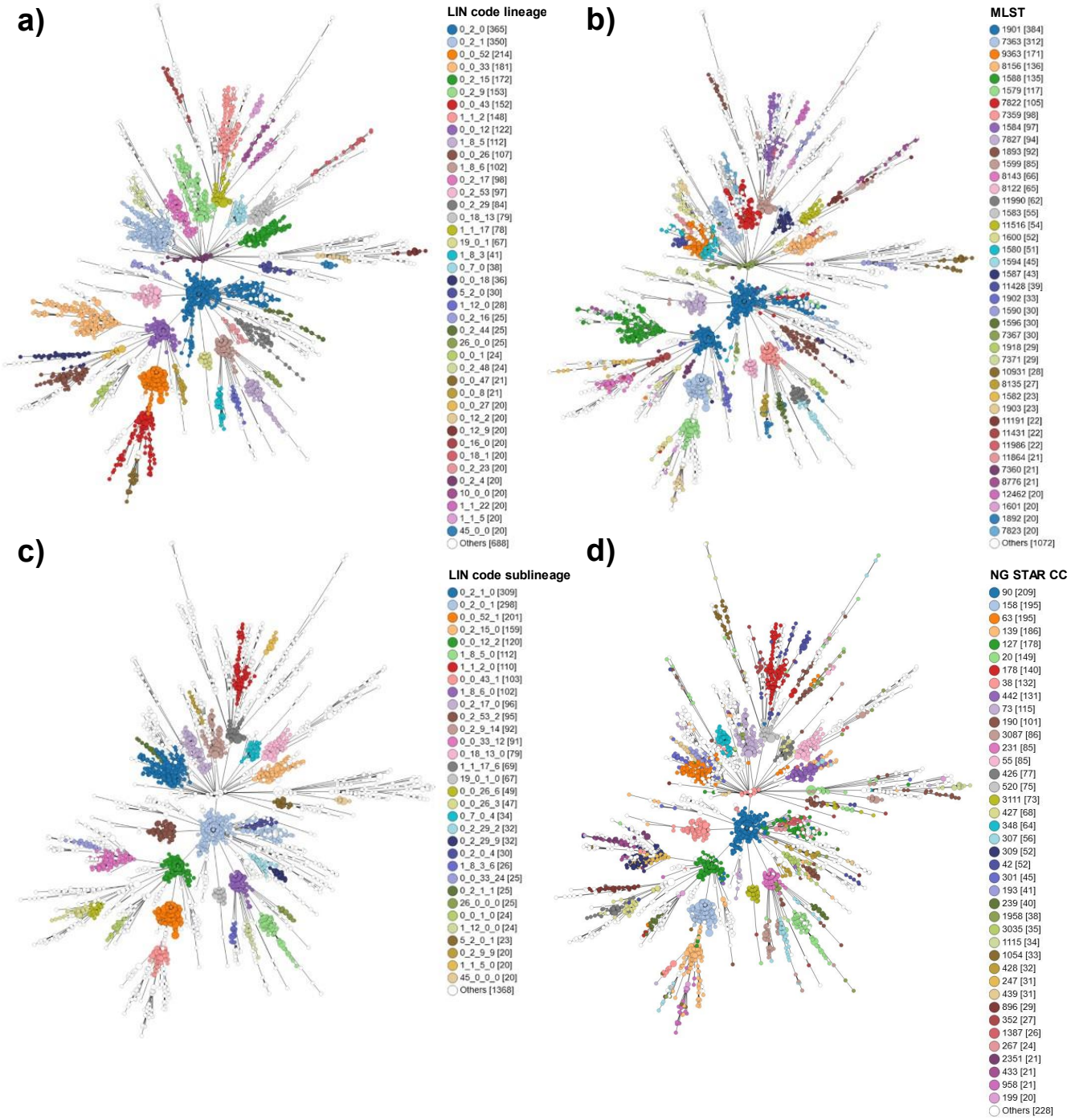


Figure 5) Minimum spanning tree showing clustering of 3935 isolates from dataset 2 based on Ng cgMLST v2*. LIN code lineages* ***(5a)*** *and 7 locus MLST* ***(5b****) demonstrate similar levels of resolution for characterising clustering. LIN codes sublineages* ***(5c)*** *provide higher resolution, similar to that provided by NG-STAR clonal complexes* ***(5d).*** *Only categories including 20 or more isolates are coloured.*

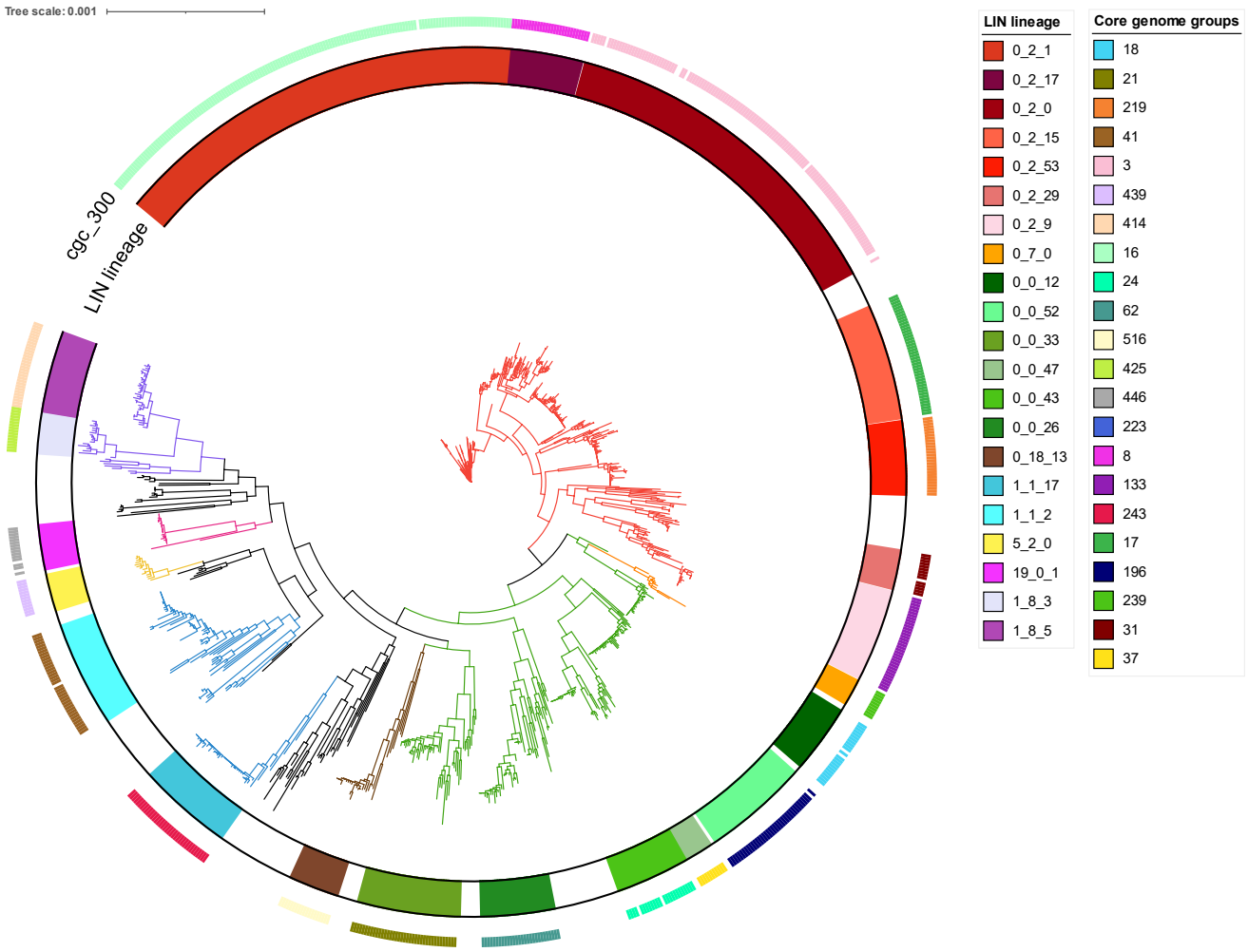


Figure 6) FastTree of 1000 randomly selected N. gonorrhoeae isolates. *Constructed using 1430 loci from Ng cgMLST v2. Branches are coloured by superlineage (0_2 = red, 0_7 = orange, 0_0 = green, 0_18 = brown, 1_1 = blue, 5_2 = yellow, 19_01 = pink, and 1_8 = purple.) The 21 highest frequency LIN lineages are represented by the inner bar, in colour ranges corresponding to their superlineage colour. LIN codes form monophyletic groupings, indicating that there is a good degree of congruence between the allelic profile clustering method used in cgMLST LIN code and nucleotide sequence alignment-based phylogeny. Core genome groups at threshold 300 (cgc_300) are represented by the outer coloured bar, and show good concordance with LIN lineages, although fewer isolates were able to be annotated with core genome groups than LIN codes due to use of the larger and more poorly auto-annotated cgMLST v1.*

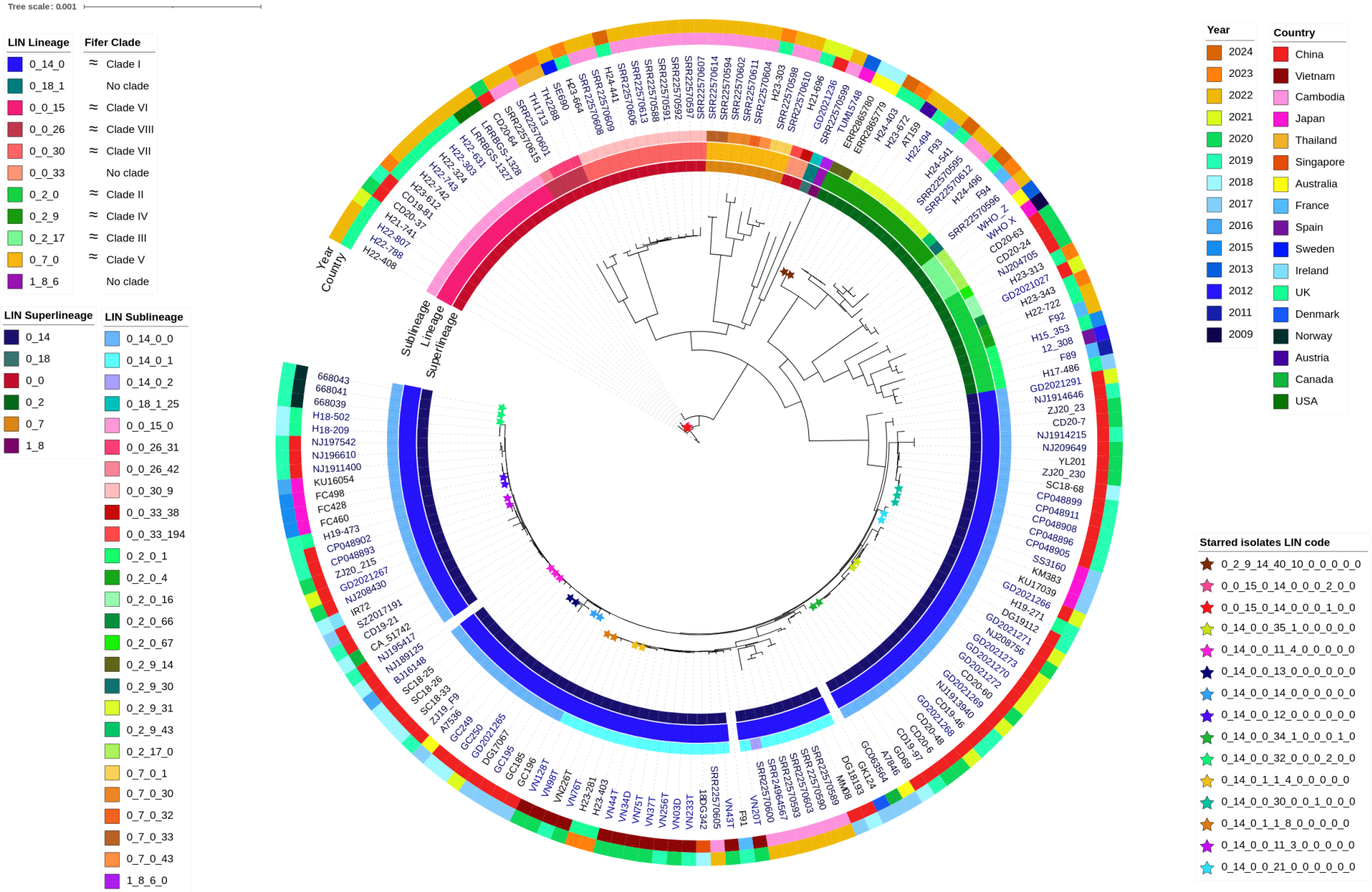


Figure 7) Maximum Likelihood tree of 170 Ceftriaxone-resistant isolates previously analysed in Fifer et al., (2024). *Constructed using 1430 loci from Ng cgMLST v2. LIN code lineages were able reproduce the clades identified in Fifer et al., while being readily accessible and providing additional detail about each clade in the form of superlineage & sublineage divisions. Groups of isolates that share the same full length LIN code are highlighted as coloured stars.*

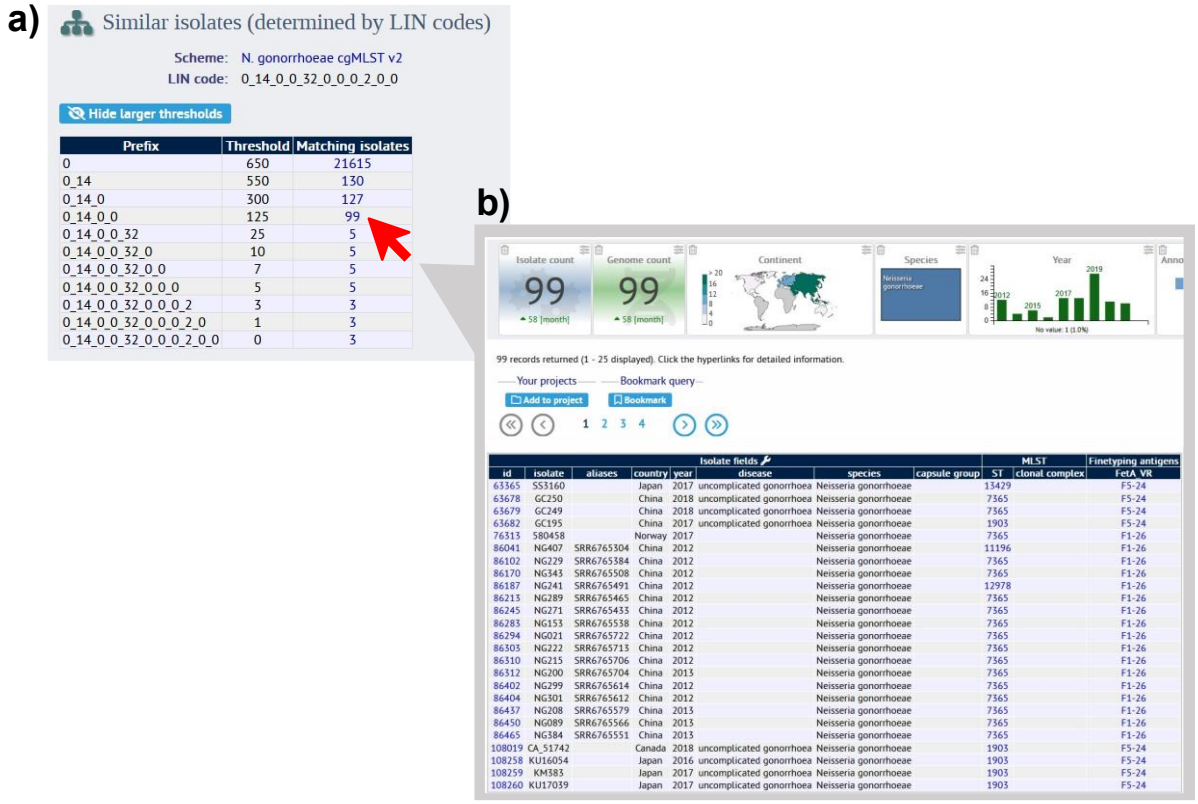


Figure 8) Using PubMLST to explore related isolates by LIN code. *Within an isolate's information page, here using isolate SC18-25 (PubMLST id: 165303) (https://pubmlst.org/bigsdb?page=info&db=pubmlst_neisseria_isolates&id=165303), it is possible to view a breakdown table of similar isolates by LIN code. This isolate shares a complete LIN code with 2 other isolates, meaning they are identical in their core genome.* ***(8a).*** *Clicking on the "matching isolates" number at a certain LIN code threshold then takes the user to the dataset of matching isolates for further analysis* ***(8b).*** *This feature can be applied in the investigation of transmission chains, outbreak events and the dissemination of AMR through clonal expansion.*

Bibliography

1. Maiden, M.C.J., M.J.J. Van Rensburg, J.E. Bray, S.G. Earle, S.A. Ford, K.A. Jolley, and N.D. McCarthy, *MLST revisited: the gene-by-gene approach to bacterial genomics.* Nature Reviews Microbiology, 2013. **11**(10): p. 728-736 DOI: 10.1038/nrmicro3093.

2. *Sexually transmitted infections and screening for Chlamydia in England: 2022 report*. 2022 [cited 2024 10/03/2024]; Available from: https://www.gov.uk/government/statistics/sexually-transmitted-infections-stis-annual-data-tables/sexually-transmitted-infections-and-screening-for-chlamydia-in-england-2022-report#conclusions.

3. WHO. *Report on global sexually transmitted infection surveillance 2018*. 2018; Available from: https://www.who.int/reproductivehealth/publications/stis-surveillance-2018/en/

4. Smolarchuk, C., A. Wensley, S. Padfield, H. Fifer, A. Lee, and G. Hughes, *Persistence of an outbreak of gonorrhoea with high-level resistance to azithromycin in England, November 2014–May 2018.* Eurosurveillance, 2018. **23**(23) DOI: 10.2807/1560-7917.es.2018.23.23.1800287.

5. Sánchez-Busó, L., D. Golparian, J. Corander, Y.H. Grad, M. Ohnishi, R. Flemming, J. Parkhill, S.D. Bentley, M. Unemo, and S.R. Harris, *The impact of antimicrobials on gonococcal evolution.* Nature Microbiology, 2019. **4**(11): p. 1941-1950 DOI: 10.1038/s41564-019-0501-y.

6. Unemo, M. and R.A. Nicholas, *Emergence of multidrug-resistant, extensively drug-resistant and untreatable gonorrhea.* Future Microbiology, 2012. **7**(12): p. 1401-1422 DOI: 10.2217/fmb.12.117.

7. Rowley, J., S. Vander Hoorn, E. Korenromp, N. Low, M. Unemo, L.J. Abu-Raddad, R.M. Chico, A. Smolak, L. Newman, S. Gottlieb, S.S. Thwin, N. Broutet, and M.M. Taylor, *Chlamydia, gonorrhoea, trichomoniasis and syphilis: global prevalence and incidence estimates, 2016.* Bulletin of the World Health Organization, 2019. **97**(8): p. 548-562P DOI: 10.2471/blt.18.228486.

8. Lyu, Y., A. Choong, E.P.F. Chow, K.L. Seib, H.S. Marshall, M. Unemo, A. de Voux, B. Wang, A.E. Miranda, S.L. Gottlieb, M.B. Mello, T. Wi, R. Baggaley, C. Marshall, L.J. Abu-Raddad, W.E. Abara, X.-S. Chen, and J.J. Ong, *Vaccine value profile for Neisseria gonorrhoeae.* Vaccine, 2023 DOI: https://doi.org/10.1016/j.vaccine.2023.01.053.

9. Gottlieb, S.L., A.E. Jerse, S. Delany-Moretlwe, C. Deal, and B.K. Giersing, *Advancing vaccine development for gonorrhoea and the Global STI Vaccine Roadmap.* Sexual Health, 2019. **16**(5): p. 426 DOI: 10.1071/sh19060.

10. Harrison, O.B., A. Cehovin, J. Skett, K.A. Jolley, P. Massari, C.A. Genco, C.M. Tang, and M.C.J. Maiden, *Neisseria gonorrhoeae Population Genomics: Use of the Gonococcal Core Genome to Improve Surveillance of Antimicrobial Resistance.* The Journal of Infectious Diseases, 2020 DOI: 10.1093/infdis/jiaa002.

11. O'Rourke, M. and E. Stevens, *Genetic structure of Neisseria gonorrhoeae populations: a non-clonal pathogen.* J Gen Microbiol, 1993. **139**(11): p. 2603-11 DOI: 10.1099/00221287-139-11-2603.

12. Unitt, A., M. Maiden, and O. Harrison, *Characterizing the diversity and commensal origins of penA mosaicism in the genus Neisseria.* Microb Genom, 2024. **10**(2) DOI: 10.1099/mgen.0.001209.

13. Manoharan-Basil, S.S., Gonzalez, Natalia. , Laumen, Jolein. , Kenyon, Chris., *Horizontal Gene Transfer of Fluoroquinolone Resistance-Conferring Genes From Commensal Neisseria to Neisseria gonorrhoeae: A Global Phylogenetic Analysis of 20,047 Isolates.* Frontiers in Microbiology, 2022 DOI: 10.3389/fmicb.2022.793612.

14. Bennett, J.S., H.B. Bratcher, C. Brehony, O.B. Harrison, and M.C.J. Maiden, *The Genus Neisseria*, in *The Prokaryotes: Alphaproteobacteria and Betaproteobacteria*, E. Rosenberg, et al., Editors. 2014, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 881-900.

15. Hanage, W.P., *Not So Simple After All: Bacteria, Their Population Genetics, and Recombination.* Cold Spring Harbor Perspectives in Biology, 2016. **8**(7): p. a018069 DOI: 10.1101/cshperspect.a018069.

16. Harrison, O., *Recent advances in understanding and combatting Neisseria gonorrhoeae: a genomic perspective.* 2021 DOI: 10.12703/r/10-65.

17. Harrison, O.B., M. Clemence, J.P. Dillard, C.M. Tang, D. Trees, Y.H. Grad, and M.C.J. Maiden, *Genomic analyses of Neisseria gonorrhoeae reveal an association of the gonococcal genetic island with antimicrobial resistance.* Journal of Infection, 2016. **73**(6): p. 578-587 DOI: 10.1016/j.jinf.2016.08.010.

18. Kwong, J.C., A. Gonçalves Da Silva, K. Dyet, D.A. Williamson, T.P. Stinear, B.P. Howden, and T. Seemann, *NGMASTER: in silico multi-antigen sequence typing for Neisseria gonorrhoeae.* Microbial Genomics, 2016. **2**(8) DOI: 10.1099/mgen.0.000076.

19. Golparian, D., L. Sánchez-Busó, M. Cole, and M. Unemo, *Neisseria gonorrhoeae Sequence Typing for Antimicrobial Resistance (NG-STAR) clonal complexes are consistent with genomic phylogeny and provide simple nomenclature, rapid visualization and antimicrobial resistance (AMR) lineage predictions.* Journal of Antimicrobial Chemotherapy, 2021. **76**(4): p. 940-944 DOI: 10.1093/jac/dkaa552.

20. Maiden, M.C.J., J.A. Bygraves, E. Feil, G. Morelli, J.E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D.A. Caugant, I.M. Feavers, M. Achtman, and B.G. Spratt, *Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms.* Proceedings of the National Academy of Sciences, 1998. **95**(6): p. 3140-3145 DOI: 10.1073/pnas.95.6.3140.

21. Francisco, A.P., M. Bugalho, M. Ramirez, and J.A. Carriço, *Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach.* BMC Bioinformatics, 2009. **10**(1): p. 152 DOI: 10.1186/1471-2105-10-152.

22. Everitt, B.S., S. Landau, M. Leese, and D. Stahl, *An Introduction to Classification and Clustering*, in *Cluster Analysis* 2011. p. 1-13 DOI: https://doi.org/10.1002/9780470977811.ch1.

23. Maiden, M.C.J. and O.B. Harrison, *Population and Functional Genomics of Neisseria Revealed with Gene-by-Gene Approaches.* Journal of Clinical Microbiology, 2016. **54**(8): p. 1949-1955 DOI: 10.1128/jcm.00301-16.

24. Palma, F., M. Hennart, K.A. Jolley, C. Crestani, K.L. Wyres, S. Bridel, C.A. Yeats, B. Brancotte, B. Raffestin, S. David, M.M.C. Lam, R. Izdebski, V. Passet, C. Rodrigues, M. Rethoret-Pasty, M.C.J. Maiden, D.M. Aanensen, K.E. Holt, A. Criscuolo, and S. Brisse, *Bacterial strain nomenclature in the genomic era: Life Identification Numbers using a gene-by-gene approach*. 2024, Cold Spring Harbor Laboratory DOI: 10.1101/2024.03.11.584534.

25. Hennart, M., J. Guglielmini, M.C.J. Maiden, K.A. Jolley, A. Criscuolo, and S. Brisse, *A dual barcoding approach to bacterial strain nomenclature: Genomic taxonomy of Klebsiella pneumoniae strains.* Mol Biol Evol, 2022: p. 2;39(7) DOI: 10.1093/molbev/msac135.

26. Rensburg, M., D. Berger, I. Yassine, D. Shaw, A. Fohrmann, J. Bray, K. Jolley, M. Maiden, and A. Brueggemann, *Development of the Pneumococcal Genome Library, a core genome multilocus sequence typing scheme, and a taxonomic life identification number barcoding system to investigate and define pneumococcal population structure.* Microbial genomics, 2024. **10** DOI: 10.1099/mgen.0.001280.

27. Vinatzer, B.A., L. Tian, and L.S. Heath, *A proposal for a portal to make earth's microbial diversity easily accessible and searchable.* Antonie van Leeuwenhoek, 2017. **110**(10): p. 1271-1279 DOI: 10.1007/s10482-017-0849-z.

28. Jolley, K.A., J.E. Bray, and M.C.J. Maiden, *Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications.* Wellcome Open Research, 2018. **3**: p. 124 DOI: 10.12688/wellcomeopenres.14826.1.

29. R Core Team. *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna,*. 2018; 4.0:[Available from: https://www.R-project.org/.

30. Seemann, T., *Prokka: rapid prokaryotic genome annotation.* Bioinformatics, 2014. **30**(14): p. 2068-2069 DOI: 10.1093/bioinformatics/btu153.

31. Bayliss, S.C., H.A. Thorpe, N.M. Coyle, S.K. Sheppard, and E.J. Feil, *PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria.* Gigascience, 2019. **8**(10) DOI: 10.1093/gigascience/giz119.

32. Zhou, Z., N.-F. Alikhan, M.J. Sergeant, N. Luhmann, C. Vaz, A.P. Francisco, J.A. Carriço, and M. Achtman, *GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens.* Genome Research, 2018. **28**(9): p. 1395-1404 DOI: 10.1101/gr.232397.117.

33. Hennart, M. *LINcoding*. 2019 [cited 2024 10/03/2024]; Available from: https://gitlab.pasteur.fr/BEBP/LINcoding.

34. Wickham, H., *ggplot2: Elegant Graphics for Data Analysis*. 2016, Springer International Publishing DOI: https://doi.org/10.1007/978-0-387-98141-3.

35. Wilke, C. *ggridges: Ridgeline Plots in 'ggplot2'*. 2024; R package version 0.5.6:[Available from: https://wilkelab.org/ggridges/.

36. Rousseeuw, P.J., *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.* Journal of Computational and Applied Mathematics, 1987. **20**: p. 53-65 DOI: https://doi.org/10.1016/0377-0427(87)90125-7.

37. Criscuolo, A. *MSTclust*. 2021 [cited 2024 10/03/2024]; Available from: https://gitlab.pasteur.fr/GIPhy/MSTclust.

38. Rand, W.M., *Objective Criteria for the Evaluation of Clustering Methods.* Journal of the American Statistical Association, 1971. **66**(336): p. 846-850 DOI: 10.1080/01621459.1971.10482356.

39. Vavrek, M., *fossil: palaeoecological and palaeogeographical analysis tools*. 2011: Palaeontologia Electronica DOI: 10.32614/CRAN.package.fossil.

40. Fifer, H., M. Doumith, L. Rubinstein, L. Mitchell, M. Wallis, S. Singh, G.J. Singh, M. Rayment, J. Evans-Jones, A. Blume, O. Dosekun, K. Poon, A. Nori, M. Day, R. Pitt, S. Sun, P. Narayanan, E. Callan, A. Vickers, J. Minshull, K. Bennet, J.E.C. Johnson, J. Saunders, S. Alexander, H. Mohammed, N. Woodford, K. Sinka, and

M. Cole, *Ceftriaxone-resistant Neisseria gonorrhoeae detected in England, 2015 to 2024; an observational study.* medRxiv, 2024: p. 2024.08.12.24311674 DOI: 10.1101/2024.08.12.24311674.

41. Prjibelski, A., D. Antipov, D. Meleshko, A. Lapidus, and A. Korobeynikov, *Using SPAdes De Novo Assembler.* Current Protocols in Bioinformatics, 2020. **70**(1): p. e102 DOI: https://doi.org/10.1002/cpbi.102.

42. Stamatakis, A., *RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.* Bioinformatics, 2014. **30**(9): p. 1312-1313 DOI: 10.1093/bioinformatics/btu033.

43. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments.* PLoS ONE, 2010. **5**(3): p. e9490 DOI: 10.1371/journal.pone.0009490.

44. Didelot, X. and D.J. Wilson, *ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes.* PLOS Computational Biology, 2015. **11**(2): p. e1004041 DOI: 10.1371/journal.pcbi.1004041.

45. Letunic, I. and P. Bork, *Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation.* Nucleic Acids Research, 2021. **49**(W1): p. W293-W296 DOI: 10.1093/nar/gkab301.

46. Turner, K.M.E. and E.J. Feil, *The secret life of the multilocus sequence type.* International Journal of Antimicrobial Agents, 2007. **29**(2): p. 129-135 DOI: https://doi.org/10.1016/j.ijantimicag.2006.11.002.

47. de Korne-Elenbaas, J., S.M. Bruistena, A.P. van Dama, M.C.J. Maiden, and O.B. Harrison, *The Neisseria gonorrhoeae Accessory Genome and Its Association with the Core Genome and Antimicrobial Resistance.* Microbiology Spectrum, 2022 DOI: 10.1128/spectrum.02654-21.

48. Catlin, B.W., *Chapter XIII Characteristics and Auxotyping of Neisseria gonorrhoeae*, in *Methods in Microbiology*, T. Bergan and J.R. Norris, Editors. 1978, Academic Press. p. 345-380.

49. Gupta, S., *Darwin review: the evolution of virulence in human pathogens.* Proceedings of the Royal Society, 2024. **291** DOI: 10.1098/rspb.2023.2043.

50. Watkins, E.R., M.M.C. J., and S. and Gupta, *Metabolic Competition as a Driver of Bacterial Population Structure.* Future Microbiology, 2016. **11**(10): p. 1339-1357 DOI: 10.2217/fmb-2016-0079.

51. Buckee, C.O., M. Recker, E.R. Watkins, and S. Gupta, *Role of stochastic processes in maintaining discrete strain structure in antigenically diverse pathogen populations.* Proc Natl Acad Sci U S A, 2011. **108**(37): p. 15504-9 DOI: 10.1073/pnas.1102445108.

52. Golparian, D., M.J. Cole, L. Sánchez-Busó, M. Day, S. Jacobsson, T. Uthayakumaran, R. Abad, B. Bercot, D.A. Caugant, D. Heuer, K. Jansen, S. Pleininger, P. Stefanelli, D.M. Aanensen, B. Bluemel, and M. Unemo, *Antimicrobial-resistant Neisseria gonorrhoeae in Europe in 2020 compared with in 2013 and 2018: a retrospective genomic surveillance study.* Lancet Microbe, 2024. **5**(5): p. e478-e488 DOI: 10.1016/s2666-5247(23)00370-1.

53. Sánchez-Busó, L., M.J. Cole, G. Spiteri, M. Day, S. Jacobsson, D. Golparian, N. Sajedi, C.A. Yeats, K. Abudahab, A. Underwood, B. Bluemel, D.M. Aanensen, M. Unemo, S. Pleininger, A. Indra, I. De Baetselier, W. Vanden Berghe, B. Hunjak, T.N. Blažić, P. Maikanti-Charalambous, D. Pieridou, H. Zákoucká, H. Žemličková, S. Hoffmann, S. Cowan, L.J. Schwartz, R. Peetso, J. Epstein, J. Viktorova, N. Ndeikoundam, B. Bercot, C. Bébéar, F. Lot, S. Buder, K. Jansen, V. Miriagou, G.

Rigakos, V. Raftopoulos, E. Balla, M. Dudás, L.R. Ásmundsdóttir, G. Sigmundsdóttir, G.S. Hauksdóttir, T. Gudnason, A. Colgan, B. Crowley, S. Saab, P. Stefanelli, A. Carannante, P. Parodi, G. Pakarna, R. Nikiforova, A. Bormane, E. Dimina, M. Perrin, T. Abdelrahman, J. Mossong, J.-C. Schmit, F. Mühlschlegel, C. Barbara, F. Mifsud, A. Van Dam, B. Van Benthem, M. Visser, I. Linde, H. Kløvstad, D. Caugant, B. Młynarczyk-Bonikowska, J. Azevedo, M.-J. Borrego, M.L.R. Nascimento, P. Pavlik, I. Klavs, A. Murnik, S. Jeverica, T. Kustec, J. Vázquez Moreno, A. Diaz, R. Abad, I. Velicko, M. Unemo, H. Fifer, J. Shepherd, and L. Patterson, *Europe-wide expansion and eradication of multidrug-resistant Neisseria gonorrhoeae lineages: a genomic surveillance study.* The Lancet Microbe, 2022. **3**(6): p. e452-e463 DOI: 10.1016/s2666-5247(22)00044-1.

54. Williamson, D.A., E.P.F. Chow, C.L. Gorrie, T. Seemann, D.J. Ingle, N. Higgins, M. Easton, G. Taiaroa, Y.H. Grad, J.C. Kwong, C.K. Fairley, M.Y. Chen, and B.P. Howden, *Bridging of Neisseria gonorrhoeae lineages across sexual networks in the HIV pre-exposure prophylaxis era.* Nat Commun, 2019. **10**(1): p. 3988 DOI: 10.1038/s41467-019-12053-4.

55. Maiden, M.C., *Population genomics: diversity and virulence in the Neisseria.* Curr Opin Microbiol, 2008. **11**(5): p. 467-71 DOI: 10.1016/j.mib.2008.09.002.

56. Demczuk, W., S. Sidhu, M. Unemo, D.M. Whiley, V.G. Allen, J.R. Dillon, M. Cole, C. Seah, E. Trembizki, D.L. Trees, E.N. Kersh, A.J. Abrams, H.J.C. De Vries, A.P. Van Dam, I. Medina, A. Bharat, M.R. Mulvey, G. Van Domselaar, and I. Martin, *Neisseria gonorrhoeae Sequence Typing for Antimicrobial Resistance, a Novel Antimicrobial Resistance Multilocus Typing Scheme for Tracking Global Dissemination of N. gonorrhoeae Strains.* Journal of Clinical Microbiology, 2017. **55**(5): p. 1454-1468 DOI: 10.1128/jcm.00100-17.

57. Tonkin-Hill, G., J.A. Lees, S.D. Bentley, S.D.W. Frost, and J. Corander, *Fast hierarchical Bayesian analysis of population structure.* Nucleic Acids Research, 2019. **47**(11): p. 5539-5549 DOI: 10.1093/nar/gkz361.