RESEARCH ARTICLE

Figures and figure supplements

Neisseria gonorrhoeae LIN codes provide a robust, multi-resolution lineage nomenclature

Anastasia Unitt *et al.*

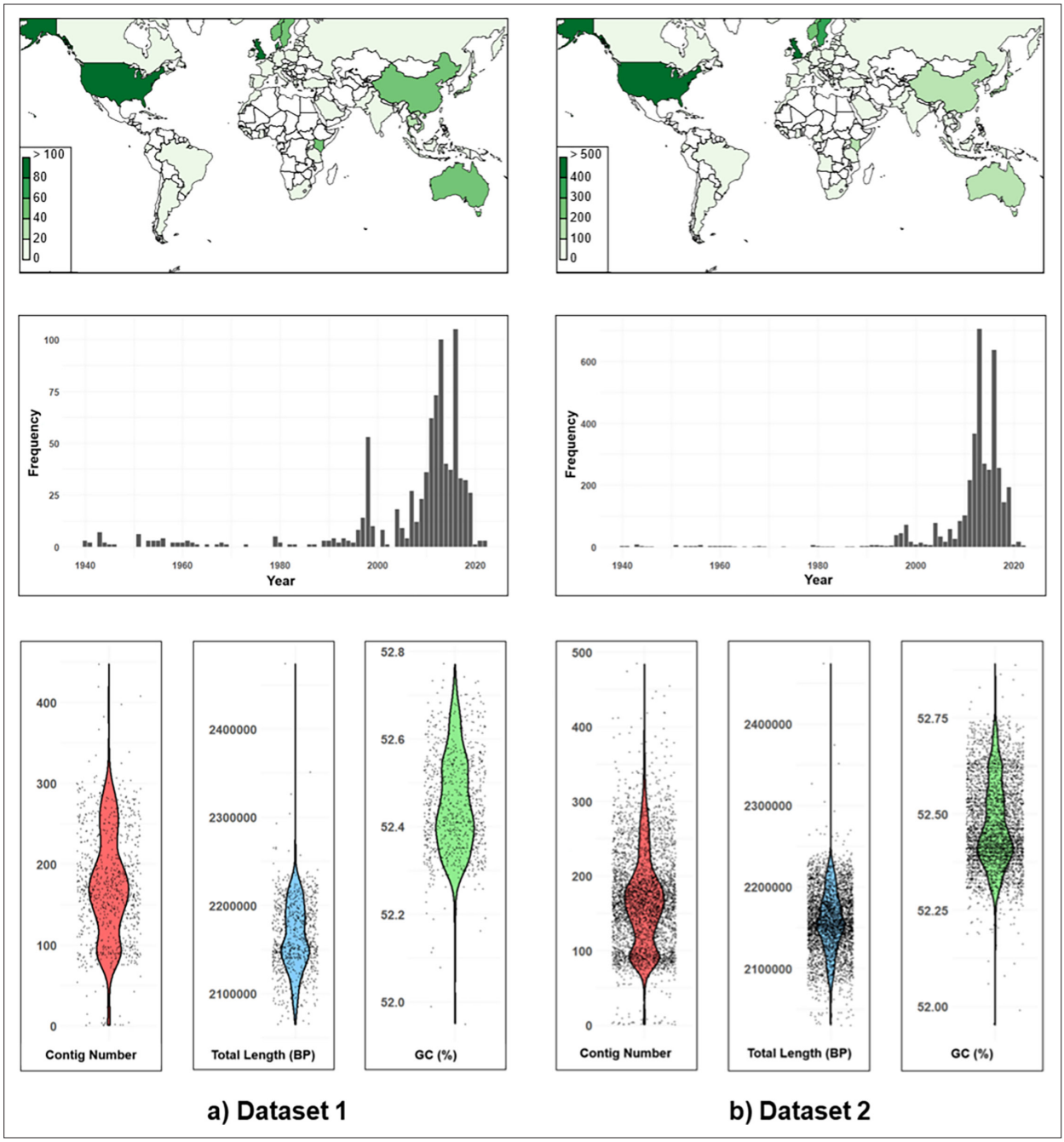
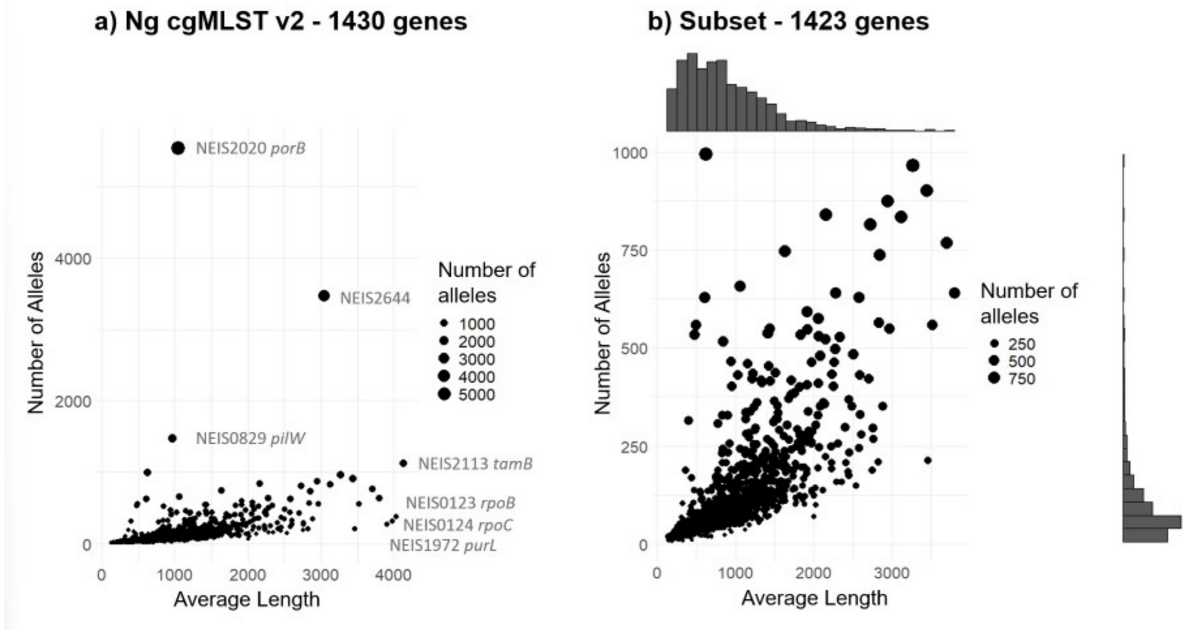


Figure 1. Characteristics of isolates within representative development datasets 1 and 2. This includes geographical distribution (top panels), frequency of isolates sampled over time (middle panels), and genome quality statistics (lower panels) including (i) contig number, (ii) total genome length and (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. (**a**) All 1430 genes included in Ng cgMLST v2 are plotted. (**b**) 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. (**c**) Table summarising the average length in base pairs, number of alleles and gene name/function of the seven outlier loci.

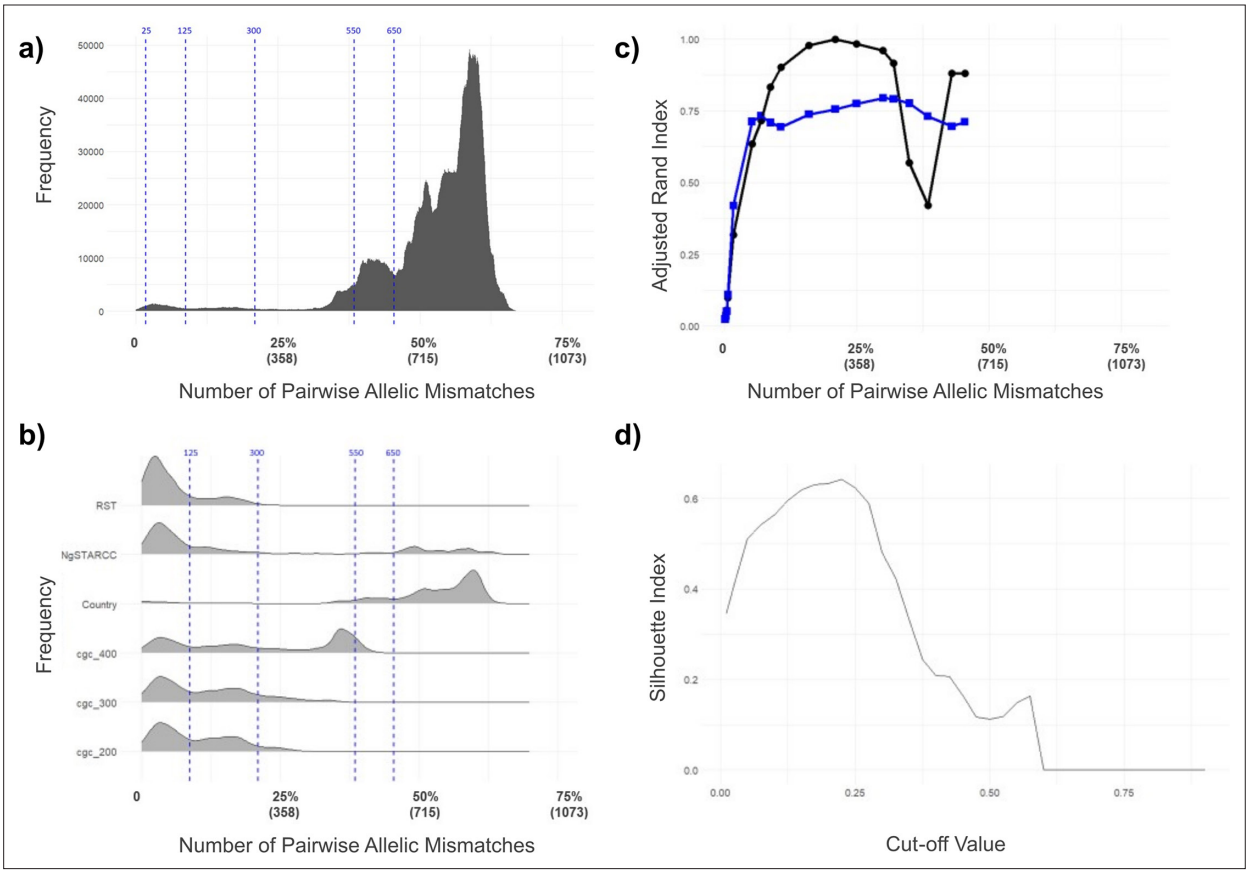


Figure 3. Plots used in the selection of allelic mismatch thresholds for the LIN code. (**a**) Histogram showing the frequency of pairwise allelic mismatches within dataset 2. A subset of the allelic mismatch thresholds applied in the gonococcal LIN code is shown (blue dashed lines) at 25 mismatches (1.75%), 125 (8.74%), 300 (20.98%), 550 (38.46%), and 650 (45.45%). (**b**) 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). (**c**) 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. (**d**) 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.

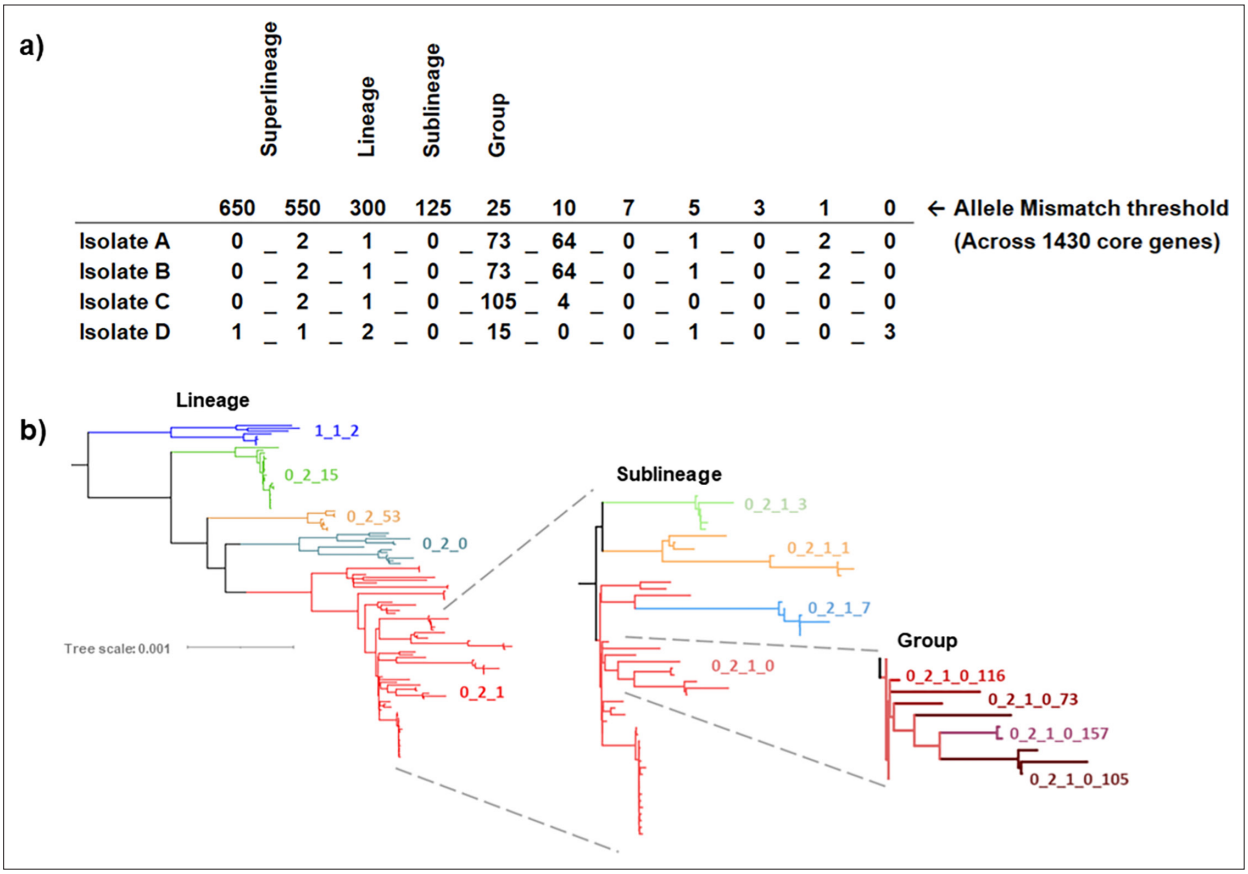


Figure 4. Illustration of the gonococcal LIN code nomenclature. (**a**) Each successive allelic mismatch threshold dictates clustering at a specific position within the code. This clustering is hierarchical, such that isolates sharing a larger proportion of the code (from the left across) are of higher 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". (**b**) 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 et al., 2024***; ***R Development Core Team, 2018***).

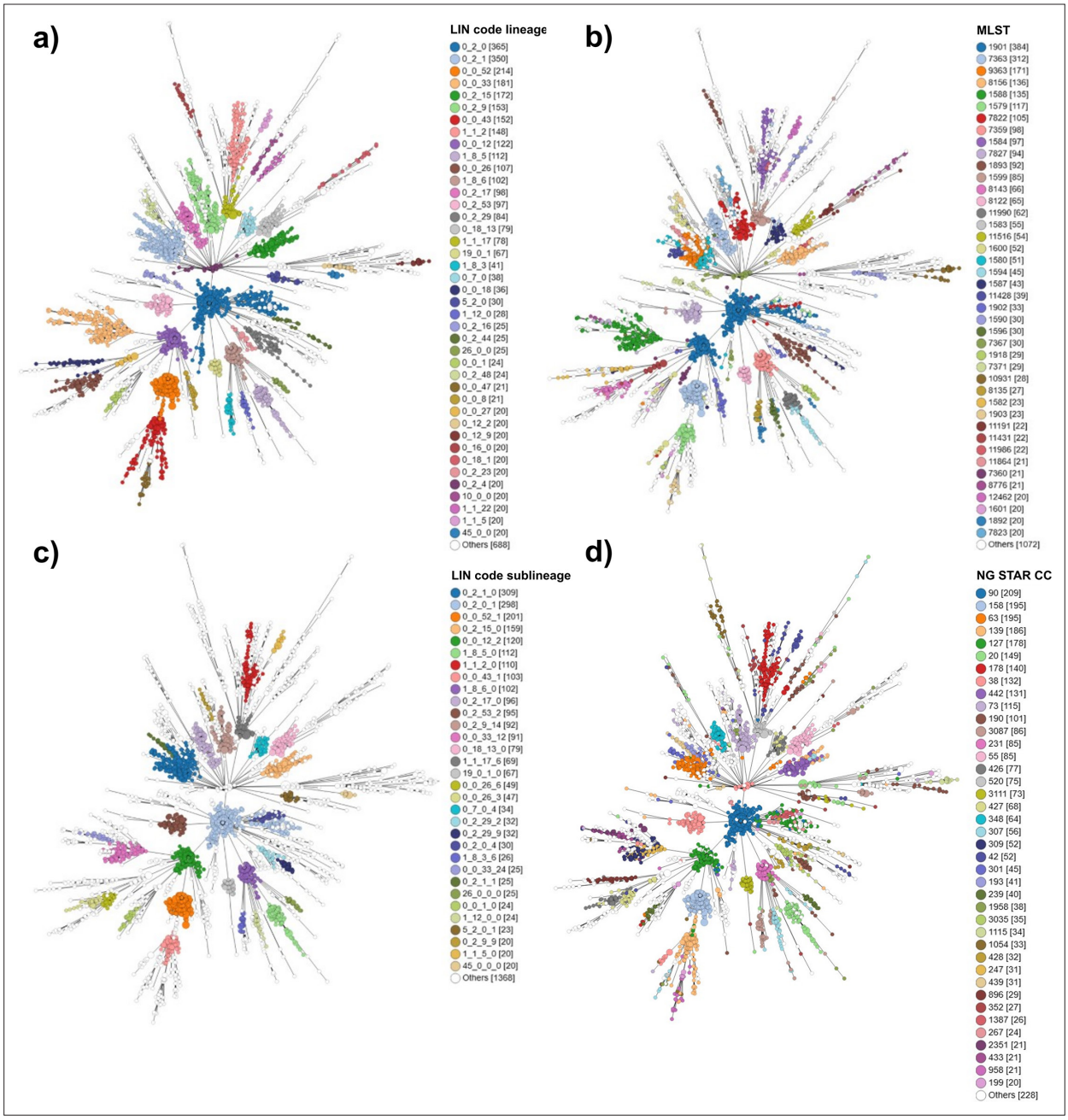


Figure 5. Minimum spanning tree showing clustering of 3935 isolates from dataset 2 based on Ng cgMLST v2. LIN code lineages (**a**) and 7-locus MLST (**b**) demonstrate similar levels of resolution for characterising clustering. LIN codes sublineages (**c**) provide higher resolution, similar to that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

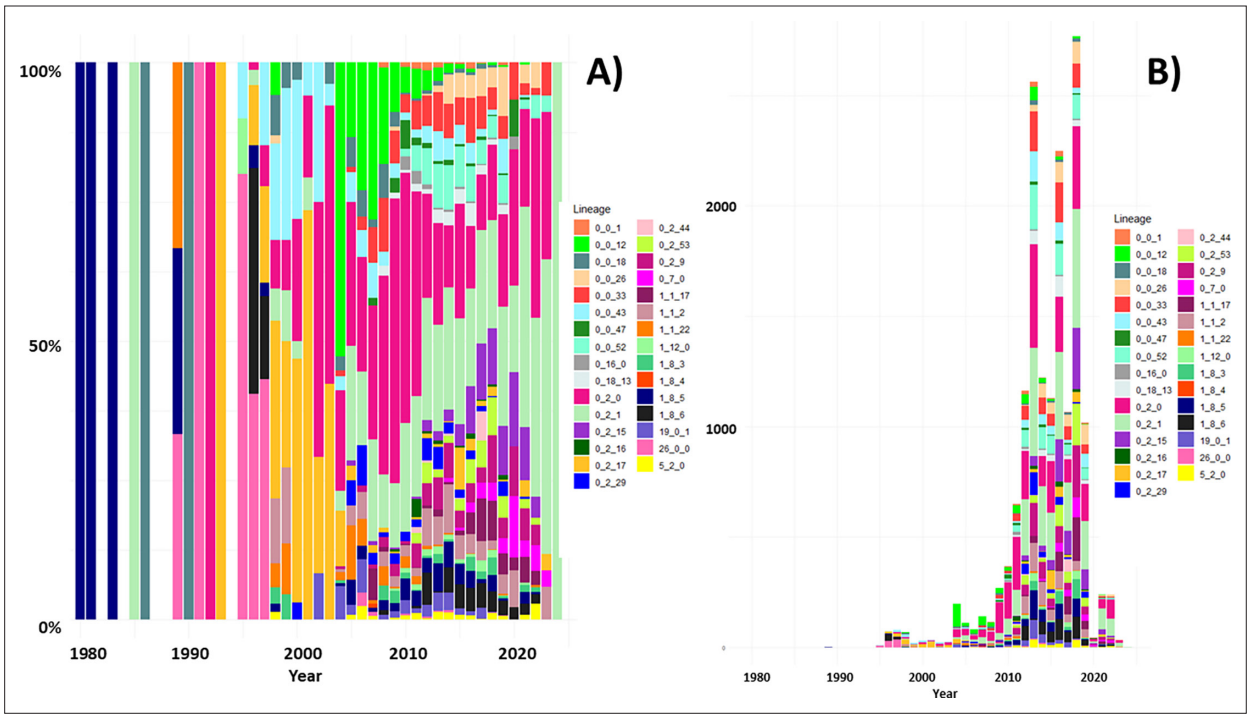


Figure 5—figure supplement 1. LIN lineage by year. (**A**) Stacked percentage barplot showing the 30 most common lineages (out of 19035 gonococcal isolates in the PubMLST database) by year. (**B**) Stacked barplot showing the 30 most common lineages (out of 19035 gonococcal isolates in the PubMLST database) by year.

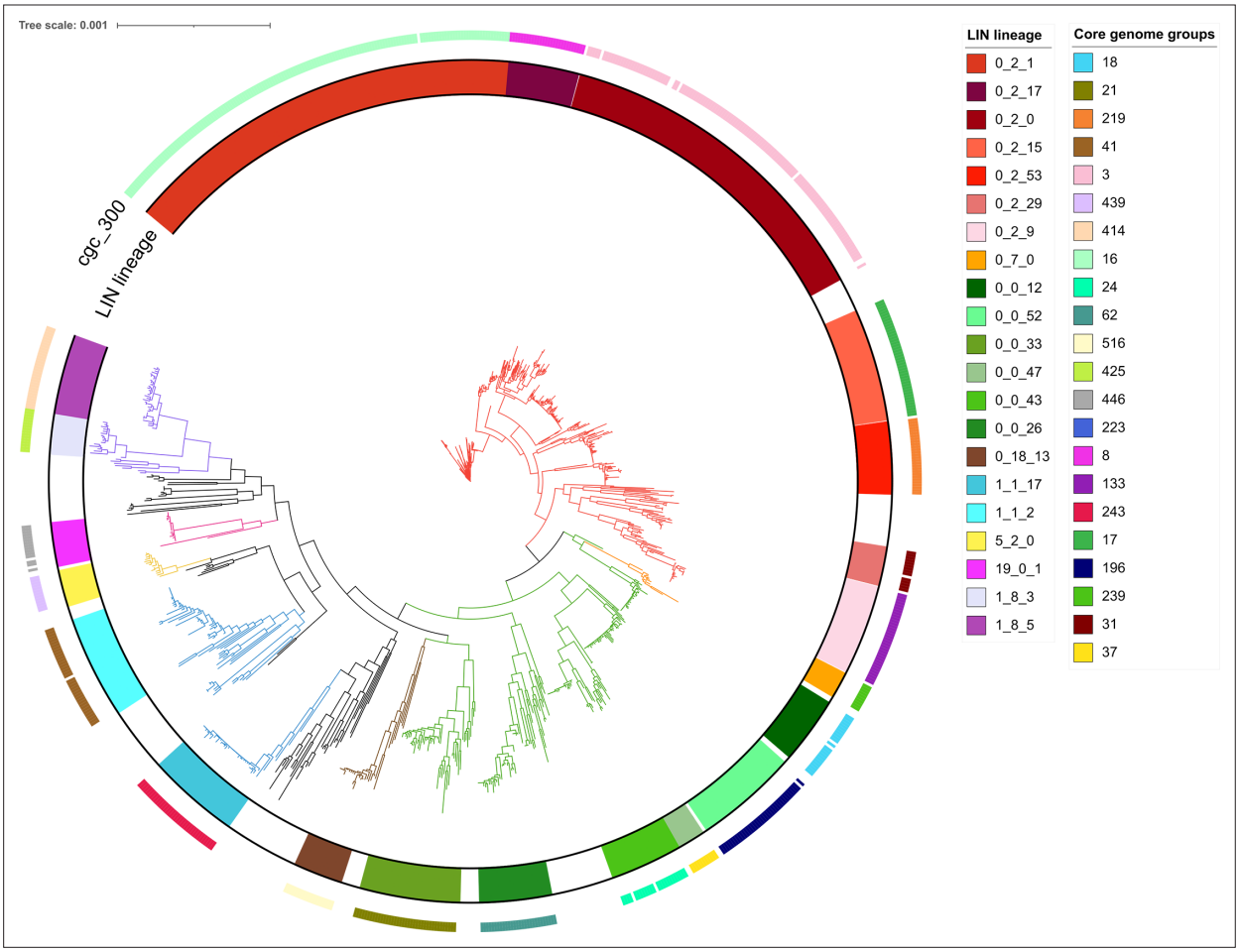


Figure 6. Unrooted FastTree of 1000 randomly chosen *N. gonorrhoeae* isolates with LIN codes assigned. Constructed using 1430 loci from Ng cgMLST v2. Branches are coloured by the 8 most frequent superlineages (0_2=red, 0_7=orange, 0_0=green, 0_18=brown, 1_1=blue, 5_2=yellow, 19_01=pink, and 1_8=purple.) The 21 highest frequency LIN lineages are labelled, 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. All isolates used in this tree had a LIN code assigned.

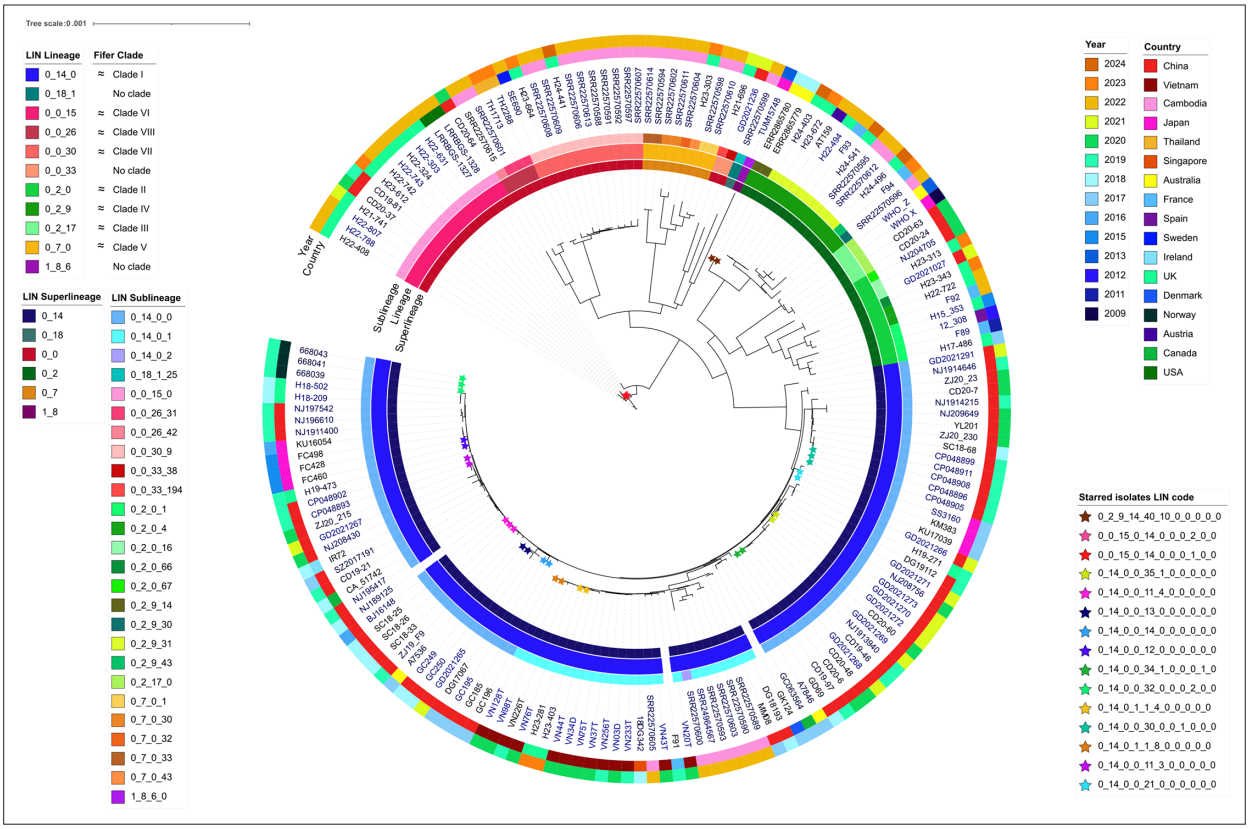


Figure 7. Maximum Likelihood tree of 170 Ceftriaxone-resistant isolates previously analysed in ***Fifer et al., 2015***. Constructed using 1430 loci from Ng cgMLST v2. LIN code lineages were able to reproduce the clades identified in Fifer et al., while being readily accessible and providing additional detail about each clade in the form of superlineage and sublineage divisions. Groups of isolates that share the same full length LIN code are highlighted as coloured stars.

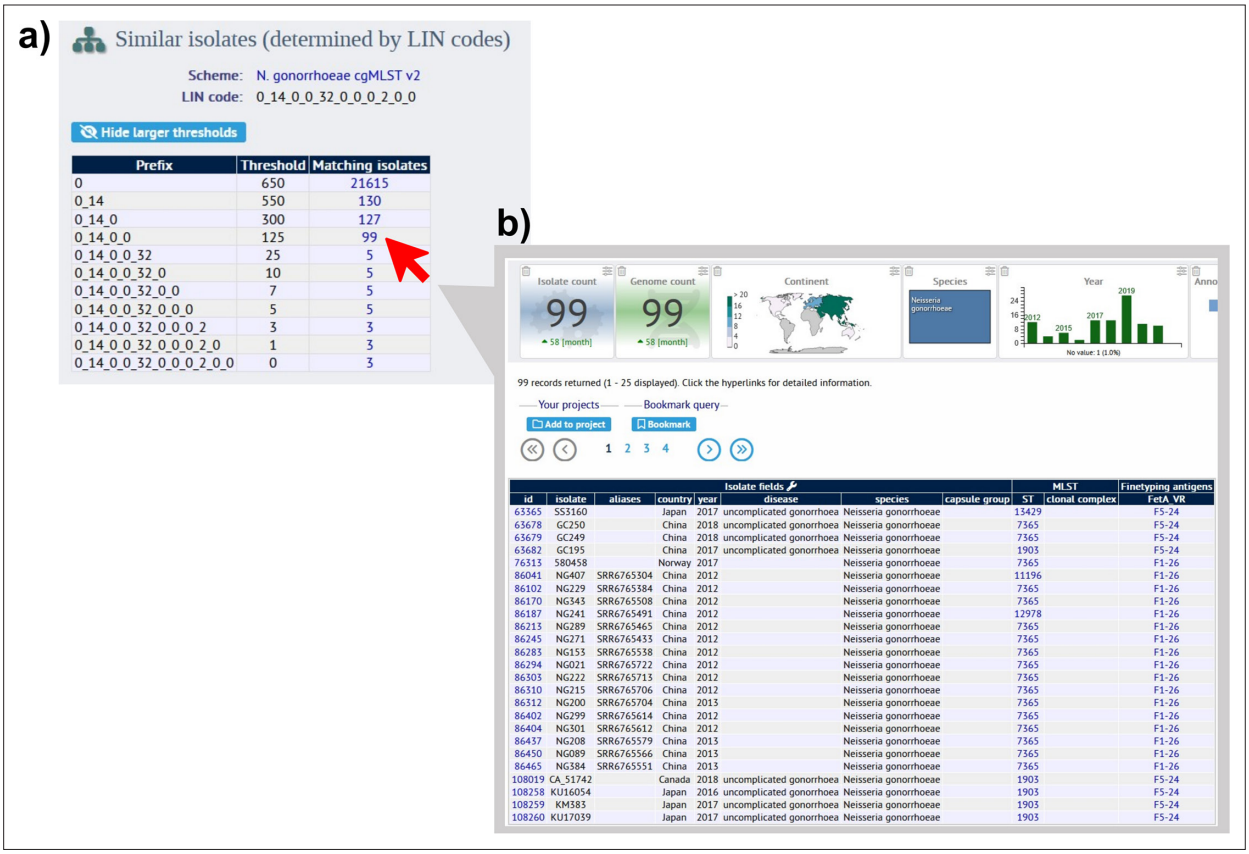


Figure 8. Using PubMLST to explore related isolates by LIN code. Within an isolate's information page, here using isolate SC18-25 (PubMLST id: 165303; https://pubmlst.org/bigsdb?page=info&db=pubmlst_neisseria_isolates&id=165303), it is possible to view a breakdown table of similar isolates by LIN code.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