Network-based population analysis of 31,595 gonococcal genomes reveals phase variability, genetic diversity and mobile element dynamics drive antimicrobial resistance and phenotypic diversity

Duncan Carruthers-Lay[1,2], Jessica Lam [1], Emil Jurga[1,2], Scott D. Gray-Owen[1] and John Parkinson[1,2,3*]

[1] Department of Molecular Genetics, University of Toronto, Toronto ON, Canada

[2] Program in Molecular Medicine, Hospital for Sick Children, Toronto ON, Canada

[3] Department of Biochemistry, University of Toronto, Toronto ON, Canada.

* To whom correspondence should be addressed: jparkin@sickkids.ca

Keywords: *Neisseria gonorrhoeae*, population genomics, antimicrobial resistance, phase variation

ABSTRACT

With a highly plastic genome allowing the accumulation of antimicrobial resistance (AMR) determinants, *Neisseria gonorrhoeae* (*Ngo*) is an urgent public health threat. To better understand the spread of AMR in *Ngo* and its ability to persist in humans, we performed a systematic analysis of its global population structure. From 31,595 publicly available genomes, we curated a maximally diverse representative set of 485 gonococcal genomes, which were found to subdivide into 12 distinct clades on the basis of shared patterns of single nucleotide polymorphisms. Mapping AMR determinants and mobile elements onto these clades revealed significant correlation with gonococcal population structure. Notably, several variants associated with AMR were identified in genes not previously linked with resistance, suggesting roles as new determinants of resistance or compensatory mutations in resistant strains. Analysis of phase variable motifs (a significant mechanism driving regulation of expression in *Ngo*) identified 57 genes not previously identified as phase variable; *in vivo* and *in vitro* passaging validated our sequence-based definition of phase variability and revealed a strong association of phase variation of *opa* and *lgtG* with improved colonization during infection. Together these findings highlight the ability of systematic comparative genomic analyses to shed new light on the drivers of *Ngo* population structure and identify new AMR determinants with implications for the development of novel therapeutics.

INTRODUCTION

The gram-negative bacterial pathogen *Neisseria gonorrhoeae* (*Ngo*) is the causative agent of the sexually transmitted infection gonorrhea. The most common site of gonococcal infection is the urogenital tract, where it is typically asymptomatic in females[1]. Consequently, without treatment, infection can result in severe morbidities such as pelvic inflammatory disease or infertility[1]. Gonococcal infection also increases the risk of coinfections with chlamydia and human immunodeficiency virus[2,3]. With rapidly rising rates of infection and widespread development of antimicrobial resistance (AMR), *Ngo* is now considered a high priority pathogen by the World Health Organization, requiring the development of novel therapeutics and additional treatment options[4]. Since the advent of antimicrobials, *Ngo* has demonstrated the ability to rapidly acquire AMR mechanisms, threatening our ability to treat infections and raising the prospect of untreatable gonorrhea[5].

Supporting its ability to acquire AMR, *Ngo* is naturally competent with many DNA uptake sequences throughout genome which enables easy incorporation of genetic matter between gonococcal strains and from other pathogenic and commensal *Neisseria* species[6,7]. Driving this "promiscuity" are mobile elements, including plasmids which contain AMR encoding genes and the Gonococcal Genetic Island (GGI) which encodes a type IV secretion system, together with several proteins of unknown function[8,9]. Further, many gonococcal genes exhibit phase variability in which changes in the number of tandem repeat sequences can alter gene expression[10]. This genetic promiscuity, phase variability and high rates of recombination support *Ngo's* ability to rapidly adapt to novel environmental perturbations and result in a highly diverse population structure not observed in the closely related *Neisseria meningitidis*[7]. This population heterogeneity has confounded traditional efforts to define clonal lineages, requiring instead the use of alternative approaches such as multilocus sequence typing (MLST). While early attempts at classifying gonococci based on MLST relied on only limited sets of housekeeping genes[11], more recent scheme's adopting larger sets of core loci, such as the core genome MLST scheme, have been more successful in defining relatively stable gonococcal core genome groups[12]. Complementing such efforts are tools such as PopNet which take a more agnostic view to selection of loci for determining population structure[13]. Further, PopNet's use of network graphs help define genetic

relationships across multiple strains that are otherwise lost in the rigid structure of the more traditional bifurcating phylogenetic trees.

With the recent availability of genome data from over 30,000 *Ngo* isolates, we sought to apply PopNet to systematically define the global population structure of *Ngo*, investigate its pangenome and assess the incidence of AMR mechanisms and the genetic elements helping support their dissemination. Through this analysis we identified and subsequently validated novel phase variable genes and AMR associated loci.

RESULTS

A representative set of 485 genomes captures temporal and geographic diversity

To begin our analyses of the global population structure of *Ngo*, we first defined a representative set of 485 isolates from an initial set of genome sequence reads from 31,595 isolates annotated by year and location of isolation hosted by the NCBI sequence read archive (September 26th 2023; **Supplementary Table S1)**. 1,051 isolates were removed after filtering for contaminated, poor quality or mislabelled sequences. Of the remaining isolates, 24,431 (80.0%) were derived from North America or Europe, 3,318 (10.9%) were derived from Oceania, primarily in Australia, while only 2,795 (9.1%) were derived from Asia, South America, or Africa. To maximize geographic diversity, we prioritized isolates based on their continent and year of isolation (see Methods; **Figure 1A**). In brief, following alignment of genome sequence reads to the reference genome (*Ngo* strain TUM19854), representative isolates were selected by first grouping isolates by their country and year of isolation. Isolates within each group were then clustered based on average nucleotide identity estimated by the Jaccard Index, resulting in an initial set of 1,307 isolates. A subsequent round of clustering using the same method with isolates grouped by continent and year of isolation reduced this set to 451 isolates, to which an additional 20 isolates from rare body sites together with a set of 14 isolates representing the WHO reference collection[14,15] were added. The resulting reference collection of 485 isolates represent 36 countries, with 38% of the isolates recovered from Canada (72 isolates), the United States (58), and the United Kingdom (54) (**Figure 1B; Supplementary Table S2**). The reference set captures isolates recovered between 1972-2023, with the majority (308, 63.5%) recovered between 2010-2020. Isolates from Asia appear to be the most evenly temporally spread, while in Africa isolates were exclusively recovered after 2013

(**Figure 1C; Supplementary Table S2**). Notably, our prioritization strategy resulting in a substantially higher proportion (148, 30.4%) of isolates from Asia, South America, and Africa compared to the initial set of 31,595 isolates (**Figure 1D**).

PopNet defined clades show limited enrichment in specific locales

Having defined our set of representative isolates, we next defined the population structure of these isolates based on shared patterns of single nucleotide polymorphisms (SNPs), relative to the reference genome (*Ngo* strain TUM19854), using PopNet. We defined 12 clades (designated A-L; **Figure 2A**) ranging in size from 203 isolates (clade A) to 4 isolates (clades K & L). Within these clades we found an enrichment of isolates from Africa ($p = 1.77x10^{-13}$, Fisher's Exact) and Asia ($p=0.00047$, Fisher's Exact) in clade B (**Figure 2Ci**). Clades K and L were exclusively comprised of isolates from Canada, while 4 of 5 isolates from clade J were isolated from Oceania. No other significant associations between continent of isolation and PopNet clades were observed at a significance level of $p = 0.05$. Together these findings suggest that geography has only a limited impact on *Ngo* diversity, highlighting its ability to rapidly disseminate on a global scale (**Figure 2B)**.

Comparison of PopNet clades with a maximum likelihood tree constructed using ClonalFrameML[16] showed good concordance, with clades largely remaining intact between the two methods (**Figure 2B**). We additionally investigated how the PopNet defined clades compared to those obtained using two other population level clustering methods, cgMLST and PopPunk[12,17]. There was a large discrepancy between the size and number of clades obtained from PopNet (12 clades, median size 10, mean size 37.46) compared to cgMLST (99 clusters, median size 1, mean size 4.87) and PopPunk (161 clusters, median size 1, mean size 3.03). The number and size of clades obtained by PopNet is determined by the inflation and pre-inflation values selected based on the silhouette index within the given set of genomes. Both cgMLST and PopPunk use pre-determined databases in their clustering approaches, which, given the intentionally diverse set of genomes chosen in out representative set, may have contributed to the large number of small clusters observed with these methods. Visual inspection of the clusters (**Supplementary Figure S1**) showed that many cgMLST or PopPunk clusters with more than 5 isolates were located within a larger PopNet cluster. In the context of a smaller, more diverse set of genomes, the agnostic

clustering approach and 2-dimensional network generated by PopNet is better suited to highlight similarities and differences between closely related isolates.

Pangenome analysis reveals PopNet clades reflect patterns of mobile elements

To compare gene complements across the 485 isolates, we generated genome assemblies *de novo* using unicycler v0.5.0[18]. Genomes were subsequently annotated using Prokka[19,20] to identify genes and orthologs defined using Panaroo[19,20] (**Supplementary Table S3)**. Across all genomes we identified a total of 2460 unique genes (i.e. orthogroups), with 1,800 "Core" genes appearing in greater than 95% of isolates, 367 "Shell" genes appearing in between 15% and 95% of isolates and 283 "Cloud" genes appearing in less than 15% of isolates (**Figure 3A**). Accessory genes (non-core genes) were clustered according to their presence or absence across genomes using the Jaccard index to calculate similarity and a Markov Clustering with a granularity of 1.10 (**Figure 3C)**. Several large, distinct clusters emerged, which we theorized represented mobile elements. These mobile elements were identified through sequence similarity searches against a database of reference sequences generated from the GGI, the *Neisseria* cryptic plasmid, the β-lactamase plasmid, and 3 variations of the conjugative plasmid (pConj; the pEP5050 American tetM plasmid, the pEP5233 Dutch plasmid, and the pEP5289 Dutch tetM plasmid). Accessory genes were assigned to mobile elements based on a nucleotide BLAST[21] search against this database with a minimum sequence identity of 85% and an e-value less than $1x10^{-6}$. The GGI was identified in 315 (64.7%) isolates, the cryptic plasmid was identified in 429 (90.1%) isolates, the β-lactamase plasmid was identified in 70 (14.5%) isolates, and the conjugative plasmid was identified in 165 (33.9%) isolates, of which 97 (19.9% of total isolates; 58.8% of pConj containing isolates) were a tetM-containing plasmid.

Interestingly, clustering of isolates based on their gene complements was largely congruent with PopNet defined clades (**Figure 3B**). This gene-based clustering appears to have been largely driven by the presence or absence of mobile elements, specifically the GGI and the conjugative plasmid (**Figures 3A and 3B**). Further investigation of PopNet defined clades revealed a spectrum of associations with different mobile elements (**Figure 2Cii**). For example, clade A was significantly depleted in both the GGI ($p=1.6x10^{-8}$, Fisher's exact) and the pConj (without tetM; p=0.00021, Fisher's exact); clade B was significantly enriched with isolates carrying the β-

lactamase plasmid (p=0.0023, Fisher's exact) as well as the pConj-TET plasmid (p=0.00028, Fisher's exact); clade C was defined by the absence of the pConj-TET plasmid in all 25 isolates (p= 0.0087, Fisher's exact); while all 17 Clade D isolates possessed the GGI (p=0.0021, Fisher's exact) and lacked the cryptic plasmid ($p=3.64x10^{-19}$, Fisher's exact). Of note, isolates from Africa were found to be significantly enriched in pConj-TET (34/42 isolates; $p=5.59x10^{-19}$) and the β-lactamase plasmid (21/42; $p=1.34x10^{-8}$), independent of the year of isolation and likely associated with the continued use of penicillin and tetracycline to treat *Ngo* infections in these countries.

Several accessory genes clusters remained that did not match our set of known mobile elements. Using the MCL clustering tool implemented in Graphia, we identified several clusters of genes based on their presence or absence in our set of representative genomes[22]. While some of these clusters lined up with the mobile elements described above, others did not. We sought to further characterize several of these clusters that were present in <50% of our genomes and included at least 5 genes (**Supplementary Table S4**). We identified 5 clusters meeting these criteria, with one of them (MCL cluster 3) representing the previously identified VirB type IV secretion system (**Figure 3C)**. MCL cluster 11 was composed of 5 genes which appear to represent a fully intact *Microviridae* genome present in 7 genomes. All the cluster 11 genes are found on a single contig, and do not appear to be integrated into the gonococcal genome. MCL cluster 12 was also composed of 5 genes present in 31 genomes, 2 of these genes were annotated as putative phage proteins however unlike cluster 11 we were unable to identify a fully intact phage genome. The largest cluster we identified was cluster 5, composed of 17 genes present in 17 genomes. Interestingly, these 17 genomes all fall within PopNet clade D, and include isolates from 3 BioProjects from the United Kingdom, Japan and the United States. Many of the genes present in cluster 5 are annotated as putative phage protein, however, unlike the other elements described so far, these genes do not all appear on the same contig and instead are split between 2 contigs flanked by core genes (**Figure 3D**). The cluster present in the largest number of genomes was MCL cluster 13, also composed of 5 genes present in roughly half (217/485, 44.7%) of our set of representative genomes. Unlike the previously described clusters, cluster 13 does not appear to be phage related, with the only 2 annotated genes listed as colicin V and hemolysin D.

PopNet defined clades exhibit distinct signatures of AMR mechanisms that are often mutually exclusive

With increased concern over the propagation of AMR within *Ngo* populations, we next predicted resistance to six antimicrobials – penicillin (PEN), tetracycline (TET), ciprofloxacin (CIP), azithromycin (AZM), cefixime (CFM) and ceftriaxone (CRO) - for each of the 485 reference isolates using the online platform, PathogenWatch[23] (**Supplementary Figure S2**). The degree of resistance (resistant, intermediate, susceptible) was classified based on the EUCAST clinical breakpoints v9.0[24]. Incidence of resistance to the three older antimicrobials (PEN, TET & CIP) were higher than rates of resistance to the other three and were significantly enriched in isolates derived from Africa (p=0.029, 0.00011 and 0.016 for PEN, TET & CIP respectively, Fisher's exact).

We found several PopNet clades were also significantly associated with resistance to specific antimicrobials (**Figure 2B**). For example, clade A was enriched in isolates resistant to AZM (61/202 isolates; $p=1.15x10^{-17}$, Fisher's exact) and isolates with intermediate resistance to TET (108/202 isolates; $p=8.44x10^{-6}$, Fisher's exact); clade B isolates were significantly more resistant to CFM (44/176 isolates; $p=2.31x10^{-5}$, Fisher's exact) and PEN (71/176 isolates; $p=4.97x10^{-5}$, Fisher's exact); while all 13 isolates in clade F, exhibited intermediate resistance to PEN (p=0.036, Fisher's exact), resistance to CIP (p=0.00093, Fisher's exact) and resistance to TET (p=0.00027, Fisher's exact). Interestingly, clade C was enriched in isolates resistant to TET (23/25 isolates; $p=1.63x10^{-5}$, Fisher's exact) despite the absence of the pConj-TET plasmid. Instead, resistance appears a consequence of a double mutation (*mtrR* promoter -5delA deletion & *rpsJ* V57M). Clade C was also enriched in isolates with resistance to CIP (20/25 isolates; p=0.021, Fisher's exact). The association of AMR patterns with clades suggests that AMR helps shape the global population structure of *Ngo*.

When evaluating susceptibility to older antimicrobials (PEN, TET & CIP), isolates demonstrating resistant to one were significantly more likely to be resistant to another. When comparing isolates resistant to the newer antimicrobials (AZM, CFM & CRO) no significant association was observed between AZM resistance and resistance to either CFM or CRO. Isolates resistant to cefixime were significantly more likely to be resistant to ceftriaxone ($p=8.64x10^{-6}$, Fisher's exact).

Analysis of repeat expansion distributions identifies 57 potentially novel phase variable genes

Previous studies have shown that many *Ngo* genes contain repeats that can result in slipped strand mispairing or site-specific recombination altering their expression and providing a reversible process to adapt to rapidly changing environments. Within our collection of 485 reference isolates we searched for candidate phase variable (PV) genes using PERF[25] to detect simple sequence repeats. In total, we identified 1554 genes with at least one simple sequence repeat. From this set we identified 85 genes with repeats that varied in length, with each repeat length appearing in at least 1% of available sequences for that gene (**Supplementary Table S5**). These genes were compared to those identified by Zelewska et. al in their 2016 screen based on locus tags from the reference strains FA1090 and NCCP11945, with 28 appearing in both sets. This overlap included 11 *opa* genes, 6 genes characterized as *pilC2* or Pilin cassette, and other well characterized phase variable genes such as *lgtG*, *fetA* and *modA/B* (**Supplementary Table S5**).

The remaining 57 genes represent putative phase variable genes, with the 10 genes with the greatest repeat length variation shown in **Table 1**. Among these 10, group_1317 is annotated as cytosine-specific methyltransferase, consistent with other gonococcal methyltransferases (e.g. *modA* and *modB*) known to be phase variable. Another 2 predicted genes (*argS* and *glyS*) are annotated as arginyl- and glycyl-tRNA synthetases, a system which was previously identified in a computational screen for phase variable genes[26]. We identified 2 genes (group_1584 and *prn*) which were annotated as pertactin or autotransporters, with *prn* appearing to encode a mostly intact autotransporter while group_1584 appears to only encode the N-terminal domain. Another autotransporter, *autA* is known to be phase variable, however the identified repeat for *autA* is AAGC, while we identified poly-A tracts in group_1584 and *prn*[27]. While all the genes so far described are chromosomally encoded "Core" genes, we also identified a gene (*kfrB*) carried on the conjugative plasmid. This gene possesses a poly-G repeat tract immediately after the start codon, and plays a role in plasmid partitioning[28].

Passaging gonococci *in vivo* improves colonization in a mouse model

Passaging of gonococci *in vitro* has previously been used to identify phase variable genes by observing repeat length changes in a single strain over time[10,29]. Consequently, we wanted to evaluate the potential impact that phase variable genes would have on enabling gonococci to adapt to and survival in a more natural environment. To accomplish this, we utilized a lower genital tract colonization model with female FVB mice expressing human CEACAM3, -5, and -6[30]. Prior to infection, mice in the diestrus stage of the estrous cycle began a hormone and antibiotic regimen to suppress commensal microflora A schematic of the mouse passaging work is shown in **Figure 4A**.

For initial infection (Passage 1; P1), 12 mice were each inoculated with 10^7 bacteria of N2009 (laboratory-passaged MS11 derivative; denoted parental strain), and vaginal washes were performed every two days to recover viable bacteria. On Day 2, only 2 of the 12 mice remained colonized by the bacteria; these were denoted as M1 and M2 (**Figure 4A**). Mice infected from the M1 lineage demonstrated progressively higher rates of colonization, 75% and 88% in Passage 2 (P2) and Passage 3 (P3), respectively (**Supplementary Figure S3**). Moreover, we also observed an increase in bacterial burden recovered from the P3 isolates compared to previous rounds of infection (**Figure 4B**). These findings suggest that gonococcal isolates passaged *in vivo* are better adapted for survival during the initial stages of infection **(Figure 4B**).

Phase variation of *opa* and *lgtG* genes contributes to enhanced colonization of passaged isolates

To delineate mutations solely conferring a fitness advantage *in vivo*, we performed complimentary *in vitro* passaging using the parental strain to control for random variations introduced during the bacterial replication process. Bacteria was sub-cultured in nutrient-rich media in 5 separate tubes every 10-12 hours for 5 consecutive days to generate 5 independent in vitro passaged lineages.

The parental isolate, *in vivo* P2 isolate (M1.05), 7 isolates from the *in vivo* P3 (P3_580, P3_521, P3_581, P3_676, P3_680, P3_691 & P3_698), and 5 *in vitro* passaged isolates (T1D5-T5D5) were sequenced and analyzed for the presence of PV genes as described previously. From this complementary screen, we identified 9 genes with signatures of phase variability including the known phase variable genes *pilC2*, pilin, *opa* (4), *fetA*,, *lgtG* and *modA* (**Table 2**).

We examined the repeat length for each gene to identify any shifts corresponding to the improved survival of the P3 isolates. We observed that the predicted number of repeats in group_1283 dropped from 11 to 10 in P2, P3_580 and P3_698. No shift was observed in any of the *in vitro* strains, leading us to hypothesize that this shift may be contributing to the improved colonization of *in vivo* passage 3 isolates. Due to the inherent population heterogeneity in *Ngo*, especially when it comes to phase variable genes, we used capillary gel electrophoresis to determine the bacterial population proportion containing 11 or 12 repeats based on peak density. We found that the parental strain and the 3 tested *in vitro* passage strains showed no variation in repeat length, remaining at 12 repeats the entire time, while the passage 2 and 3 strains were more heterogeneous (**Figure 4C**). Approximately 32.5% of the recovered passage 2 isolates had 11 repeats, with the remaining 67.5% staying at 12. Passage 3 isolate P3_691 also did not vary, while P3_521 (9.1% 11, 90.9% 12), and P3_581 (37.3% 11, 62.7% 12) remained majority at 12 repeats. However, passage 3 isolates P3_676 and P3_698 were found to be 100% at 11 repeats, and the majority of P3_580 (54.2% 11, 45.8% 12) was found to be at 11 repeats. We were unable to recover sufficient genetic material from P3_680 for analysis. These findings indicate that the isolate populations containing a majority of the 11 repeat *opa* gene group_1283 are turning on expression of an Opa protein which may contribute to improved colonization in a mouse model, consistent with its key role in adhesion to host CEACAMs[30].

We also identified *lgtG* (group_1111) as a promising candidate, as the parental and all the *in vitro* strains contained a 10bp long poly-C repeat, while the P2 and all the P3 isolates shifted to an 11bp long repeat. This shift enables complete transcription of *lgtG*, resulting in the attachment of the β-chain to the gonococcal lipooligosaccharide and production of the 2C7 epitope (**Figure 4D**)[31]. We confirmed the presence of the 2C7 epitope exclusively in the *in vivo* passaged strains through an immunoblot assay. Due to the challenges of differentiating single nucleotide length repeats with capillary electrophoresis, we did not perform this analysis for *lgtG*. We show here the ability of our method to identify phase variable genes in an *in vivo* setting. We also identified phase variation in 2 genes, *opa* and *lgtG* which may play a substantial role in improving *Ngo*'s ability to colonize and infect a humanized mouse model.

Analysis of sequence variants reveal novel sites of genes involved in AMR as being under positive selection

In the previous sections our genomic surveys identified several phase variable genes that may help *Ngo* adapt to environmental perturbations. Another signature of genes supporting adaptation is whether they are under positive selection. To identify such genes we used the BUSTED algorithm as implemented in HyPhy[32,33]. Of the 1,853 core genes, 1,724 were sufficiently variable to test for selection. We identified 355 genes with significant evidence for positive selection (likelihood ratio test statistic > 0, p-value < 0.05). As expected, given their direct contact with the environment, positively selected genes were significantly enriched in outer membrane proteins (Benjamini-Hochberg (BH) corrected p-value 0.00615) (**Figure 5A**). In addition, we also identified several AMR-associated genes as being under positive selection, including *porB*, *penA*, *parC*, *mtrD*, and *gyrA*. To identify specific sites in AMR genes under positive selection we applied the FUBAR algorithm as implemented in HyPhy[34]. Sites with a posterior probability of positive selection > 0.9 were considered as sites with strong evidence for selection. Focusing on the gyrase, *gyrA*, our analyses identified residues S91 and D95, previously characterized as conferring resistance to ciprofloxacin[35], as being under positive selection (**Supplementary Table S6**). Ciprofloxacin resistance is also associated with mutations in residues D86, S87, S88, and E91 in the type IV topoisomerase, *parC*. In addition to sites D86, S87 and E91 which interact directly with ciprofloxacin[36,37], we also identified residues T386, G481, and S598 as being under positive selection. Residues T386 and G481 reside in the α-helices which connect the catalytic and dimerization domains, while residue S598 is located in the C-terminal domain (**Figure 5B**)[38]. We identified 8 residues (310, 642, 742, 922, 926, 1052, 1065, 1067) with signatures indicating strong positive selection in *mtrD*, likely reflecting the acquisition of mosaic alleles from other *Neisseria* sp. associated with azithromycin resistance[39]. For the known penicillin resistance gene, *penA*, we identified 6 residues with strong signatures of positive selection (**Figure 5C**). We identified residues 550, 553, 556, and 567 within the α11 helix which includes the known P551 residue previously linked to penicillin resistance through inhibition of acylation by β-lactamases[40]. We also identified a residue (A502) in the region between the β3-β4 loops which has been linked to ceftriaxone resistance[41]. The final residue under positive selection was Y201, which is located between the α3n helix and β6n loop in the N-terminal domain of penA, which to date has not been linked to AMR[40]. Previous work has used signatures of positive selection to identify genes

contributing to antimicrobial resistance in *Escherichia coli*, here we identified signatures of strong positive selection in individual residues of AMR-associated genes to identify new or complementary sites that may contribute to resistance[42].

Genome wide association study

Beyond signatures of positive selection, genome wide association studies offer a powerful approach to identify SNPs associated with specific traits such as AMR[43]. Here we used the software tool, pyseer, with the linear mixed model mode to identify SNPs significantly associated with predicted resistance to antimicrobials (**Supplementary Table S7)**[44].

We identified several well-known SNPs together with several less characterized mutations significantly associated with AMR. For ciprofloxacin resistance, the former include the well studied S91F and D95A/N mutations in *gyrA*, as well as an S87R mutation in *parC* (**Figure 5Di)** However, we also identified two mutations in (G9S & I10V) in *lrgB*, a cytoplasmic membrane associated murein hydrolase transporter, as being associated with ciprofloxacin resistance. Of the four variants with the greatest predicted effect size associated with penicillin resistance, two were identified in *penA* (1719C>T & 1713A>C), while the other two were found in *murE* (216C>T, 71G>A), which lies 25bp upstream of *penA* and plays a role in cell wall biogenesis (**Supplementary Figure 4A**). A highly significant missense variant (Arg89Glu) was also identified in *ychF*, a ribosome binding ATPase. For tetracycline resistance, our analysis identified a variant in the known resistance gene, *rpsJ* (Met57Val), as well as multiple mutations in *rsmB*, a methylase which interacts with the target of tetracycline, the 16S portion of the small ribosomal subunit (**Supplementary Figure 4B**). For both ceftriaxone and cefixime resistance, we identified multiple variants in *penA*, consistent with known resistance mechanisms (**Supplementary Figure 4Ci & 4Cii**). Finally, we identified multiple significant variants associated with azithromycin resistance in genes associated with the MTR multi-drug efflux pump (*mtrC*, *mtrD*, and *mtrE*) (**Figure 5Dii**). Interestingly, we also identified multiple significant variants in *mqo*, a malate-quinone oxidoreductase involved in the TCA cycle. Azithromycin is not known to target *mqo* or the TCA cycle, and pairwise structural alignments from AlphaFold predicted structures using TM-align show *mqo* shares minimal similarity with *rplV* (TM-score 0.11, 7% identity) or *rplD* (TM-score 0.17, 6% identity), known ribosomal targets of azithromycin[45–48] (**Figure 5E)**.

DISCUSSION

With the emergence of strains resistant to all known antibiotics, *Ngo* is a threat to global health. To better understand how AMR mechanisms arise and disseminate on a global scale, we conducted a systematic analysis of *Ngo* population genomics. Our set of 485 representative genomes successfully captured the genetic diversity of the larger set of 31,595 isolates, increasing the relative contribution of isolates collected from outside North America and Europe. Ensuring representation across the globe is critical if we are to understand the impact of geopolitical considerations regarding surveillance and treatment options, factors that are acutely elevated for a pathogen that is restricted to a single host[5]. Applying the population genomics tool - PopNet, we define 12 clades, which, with the exceptions of clades B, K and L, contain isolates that are geographically dispersed, suggesting another mechanism my be influencing gonococcal population structure. We observed a high degree of overlap between clades defined by other tools based on analysis of more limited sets of loci (e.g. cgMLST), supporting the utility of PopNet to define genetically distinct sub-populations.

Given the lack of overlap between the place of isolation and clade membership, we investigated how global population structure may be influenced by other factors, specifically mobile elements and AMR mechanisms. In keeping with previous findings, we observed PopNet defined clades correlate with the presence of specific mobile elements and AMR determinants[49,50]. By contrast, most identified clades were not significantly associated with other available metadata, including sex of patient, site of isolation, year of isolation and country or continent of isolation. A noteworthy exception was clade B, which was significantly associated with isolates collected from Africa and Asia. This association can be explained by the elevated presence of the β-lactamase plasmid and the TET-containing conjugative plasmid in group B. This is in line with previous studies that have suggested that isolates from lower income countries are more likely to carry plasmids encoding resistance to penicillin or tetracycline[51]. Further, penicillin resistance in African isolates and nearly all of the observed tetracycline resistance (94.2%) could be attributed to the presence of the β-lactamase and TET-containing conjugative plasmids, respectively, In keeping with previous findings, we observed PopNet defined clades correlate with a strong connection between the presence of specific mobile elements and AMR determinants with PopNet clades[49,50]. By contrast,

most identified clades were not significantly associated with other available metadata, including sex of patient, site of isolation, year of isolation and country or continent of isolation. A noteworthy exception was that of clade B, which was significantly associated with isolates collected from collection in Africa and or Asia. This association can be explained by the elevated presence of the β-lactamase plasmid and the TET-containing conjugative plasmid in group B, both of which isolates from Africa were significantly more likely to carry. This is in line with previous studies that have suggested that disparity has been previously noted, with isolates from lower income countries are significantly more likely to carry plasmids encoding resistance to penicillin or tetracycline[51]. Going further, all the penicillin resistance in African isolates and nearly all of the observed tetracycline resistance (94.2%) could be attributed to the presence of the β-lactamase and TET-containing conjugative plasmids, respectively. For both penicillin and tetracycline, resistance in non-African isolates was driven by a range of chromosomal mutations, particularly in *penA*, *mtrR* and *rpsJ*. It is likely that this difference is driven by the continued usage of penicillin and tetracycline to treat *Ngo* infections in lower income countries in Africa, and may indicate some additional fitness cost to plasmid carriage in gonococci in the absence of this pressure[52].

Phase variability is a key mechanism used by *Ngo* to evade the immune system through antigenic variation and increasing genetic variance either by directly affecting gene transcription and translation or by impacting downstream processes like methylation. Several well characterized phase variable genes have been identified in *Ngo*, including those encoding for outer membrane proteins such as *opa* and *pilC*, metabolic genes such as *lgtA*, and methyltransferases like *modA*[29]. While extensive work has been done to characterize these genes in reference strains, as well as by identifying homologs in the closely related species, *N. meningitidis*, recent evidence has suggested that the phase variable repertoire of *Ngo* could be much larger than we previously thought[26,29]. We identified 57 genes containing tandem repeats of varying lengths which represent putative phase variable genes. To complement these analyses we performed a passaging experiment in an attempt to drive the emergence of phase variability. *In vitro* passaging experiments have previously been successful in the identification of novel phase variable genes[29]. Here we used a tandem *in vitro* & *in vivo* passaging strategy to identify phase variable genes that help support *Ngo*'s ability to infect and propagate. During the passaging experiment, we observed improved colonization of gonococcal isolates with each successive passage, indicating some level of adaptation to the mouse

host. Through this tandem approach, we identified several previously characterized phase variable genes (*pilC*, *opa*, *modA*, *lgtG* & *fetA*) in our passaged isolates, of which two (group_1283, *opa* & group_1111, *lgtG*) varied in repeat number only in *in vivo* passaged isolates. Through capillary gel electrophoresis, we were able to confirm that through phase variation the opa protein encoded by group_1283 varied in passage 2 and 3 isolates, likely contributing to the observed improved colonization. We also observed a shift from 10 to 11 bp repeats in all *in vivo* passage 2 and 3 isolates, resulting in transcription of *lgtG*, with no shift observed in the *in vitro* passaged strains. Expression of lgtG leads to the production of the 2C7 epitope, which is present in a high proportion of clinical isolates, indicative of a role in gonococcal pathogenesis and colonization[53]. Our results support this conclusion, and demonstrate the utility of passaging and sequencing for identifying genomic variation linked to improved gonococcal pathogenicity.

Given the genomic plasticity of *Ngo* and selection constraints imposed by the host immune system, we hypothesized many *Ngo* genes may be subject to positive, diversifying selective pressure. We identified 355 of 1,853 core genes to be under positive selection. This number dramatically expands on the previous estimate of only 10 *Ngo* genes under positive selection, likely reflecting differences in methodology as well as the larger number of genetically diverse isolates used here[54]. A similar approach focusing on 2 isolates of *Burkholderia glumae* revealed a comparable number of genes under positive selection (259/3,777 tested genes)[55], although we note the proportion of gonococcal genes under positive selection is substantially higher. We found no functional enrichment for positively selected genes, though we did observe a significant enrichment in predicted outer membrane proteins, as would be expected given immune pressure exercised by the host. Beyond identification of novel genes responding to host pressure, we also identified several novel sites in known AMR genes associated with positive selection. These novel sites likely represent complimentary variants or emerging resistance-conferring variants and require further exploration.

Beyond using selection pressure to identify genes and variants associated with host adaptation and environmental perturbations, we also sought to identify variants associated with the carriage of mobile elements and AMR through a genome-wide association analysis. We successfully identified known variants in multiple AMR-associated genes, including *gyrA* and *parC* for

ciprofloxacin, *penA* for penicillin, ceftriaxone, and cefixime, *rpsJ* for tetracycline, and multiple *mtr* genes for azithromycin. We additionally identified several highly significant variants associated with AMR in genes with no known links to AMR. These include a variant in *lrgB* associated with ciprofloxacin, which has previously been indirectly linked to ciprofloxacin resistance in *Ngo* and may play a role in the oxidative stress response[56]. We also found multiple non-synonymous and missense mutations in *rsmB* associated with tetracyline resistance. This finding is of particular interest, as *rsmB* is a methyltransferase which interacts directly with the 16S rRNA portion of the small ribosomal subunit, which is targeted by tetracycline[57]. Little is known about *rsmB* in *Ngo*, and while it is unlikely that it represents a novel resistance mechanism, it may complement additional tetracycline resistance-conferring mutations. Excitingly, we also found multiple significant variants in *mqo*, a malate-quinone oxidoreductase, associated with azithromycin resistance. Despite appearing to have minimal interaction with known azithromycin targets, studies testing the efficacy of azithromycin against the protozoan parasite *Plasmodium falciparum* found reduced levels of TCA associated metabolites, including malate, upon administration of the drug[58]. It is possible a similar effect is produced in gonococci and the mutations in *mqo* may play a role in gonococcal survival during azithromycin treatment. Additional work has been conducted in *P. falciparum* to validate *mqo* as a target for drug development efforts, and our finding suggest similar efforts in *Ngo* may be worth pursuing[59].

We have curated a representative set of 485 genetically and globally diverse gonococcal genomes from the 52,874 deposited gonococcal sequences on the NCBI SRA. Using PopNet, we assessed the population structure of our representative set, and uncovered significant relationships between population structure, geography, mobile elements, and AMR. This analysis suggests global gonococcal genetic diversity is driven by the pressure applied by antimicrobials and links the carriage of mobile elements with chromosomal variants. We used our representative set of genomes to identify putative novel phase variable genes and were able to identify phase variation of *opa* potentially contributing to improved survival in the context of infection. Using a combined approach of GWAS and tests for selective pressure, we identified new sites and variants potentially associated with AMR in *Ngo*, demonstrating the utility of these large-scale analyses for studying AMR in pathogen populations.

MATERIALS & METHODS

Selection of sequences

An initial set of 52,874 accessions of sequencing runs were obtained from the NCBI Sequence Read Archive on September 26th, 2023 based on a keyword search for "*Neisseria gonorrhoeae*". This set was reduced to 31,595 by keeping only those accessions containing paired-end Illumina sequences with associated metadata for the country and year of isolation. Additional metadata recorded when available included the site of isolation, associated BioProject, and sex of the patient. All sequence data was obtained using the SRA Toolkit v3.0.0 using the default options for the prefetch and fasterq-dump commands. Further metadata, such as minimum inhibitory concentration values were obtained from pubMLST or publications associated with the relevant BioProjects[60]. Kraken2 v2.1.3 was run to identify erroneously annotated or contaminated sequences[61]. Any sequence datasets with less than 75% of its reads annotated by Kraken as being derived from *Ngo* were discarded. Any sequences with between 75 and 85% gonococcal reads were kept if over 95% of the reads were assigned to the genus *Neisseria.*

Variant discovery for raw sequence data

The reduced set of 31,595 isolates were aligned to the chromosomal genome of *Ngo* strain TUM19854, the current recommended reference genome for gonococci from the NCBI (obtained on October 13th, 2023). Alignment was performed using the Genome Analysis Toolkit (GATK) v4.2.5.0[62]. Quality control and trimming was performed using fastp v0.23.4 with minimum qualified phred quality of 20, a 40 percent limit on unqualified bases, a mean quality requirement of 20, a cut window size of 5, both front and tail cutting, and with correction enabled[63]. Trimmed reads were then passed to GATK for alignment using bwa v0.7.17 with default settings. The GATK Haplotype caller was then run with default settings to identify variants. Dependencies for GATK included samtools v1.17 and picard v2.26.3. Variants were then filtered using bcftools v1.11 and htslib v1.11 with the following parameters: allelic depth 8, allelic depth forward/reverse 3, minimum base quality 50, minimum mapping quality score 20 and percent of reads containing variant 80[64].

Generation of a representative set of sequences

Isolates were grouped by their country of isolation and year of isolation and the Jaccard statistic was calculated for each pair of sequences using the bedtools v2.31.0 jaccard function with default settings[65]. An initial clustering step was performed using the k-medoids fastPAM algorithm implemented in the R packages cluster v2.1.5 and factoextra v1.0.7[66–68]. For each group, a value for k was determined using the silhouette method, with the maximum value of k capped at 15. The identified medoids of the clusters were selected as representative isolates. This resulted in an initial set of 1,307 isolates. Isolates from this set were then grouped by their continent and year of isolation, and k-medoids clustering was performed as before without a cap on the value of k. Strains identified as disseminated or ocular as well as the WHO global reference collection were added to the final set, bringing the total number of isolates to 485[14,15].

PopNet analysis and phylogenetic tree construction

PopNet was used to cluster the reference set of genomes using the variants identified from the previous analysis[13]. A variant table was generated using the GATK VariantsToTable function from a merged vcf file of identified variants. The parameters used for the PopNet analysis were section_length= 5000, S1_iVal = 8, S1_piVal = 19, S2_iVal= 4.5, and S2_piVal= 2.5. These parameters were determined using the silhouette index readout provided by the tool. Initial phylogenetic tree reconstruction was performed with IQTree v2.2.2.7 from a sequence alignment created from the previously identified variants[69]. To begin, Model Finder Plus was run which identified TPM2u+F+I+R7 as the best substitution model[70]. IQTree was then run using this model with 1000 ultrafast bootstrap replicates[71]. All other parameters were left as default. This initial tree was used as a basis for ClonalFrameML v 1.12, which was run with the standard model with 100 pseudo-bootstrap replicates[16]. Assembled genomes were uploaded to pubMLST and automatically assigned to a cgMLST Ng_300 clade.

Pangenome analysis

Trimmed reads were assembled *de novo* using unicycler v0.5.0 with default settings[18,72]. Gene prediction and initial annotation was performed using Prokka v1.14.5 with recommended settings and an added protein database containing annotated proteins from *Ngo* strains MS11, FA19, FA1090 obtained from the NCBI on March 7th, 2022[20]. The pangenome was calculated using panaroo v1.3.4 on the strict setting with otherwise default options[19]. Representative gene

sequences from panaroo were supplied to eggNOG-mapper v2.19 with no restrictions on orthology identification and permitting electronic GO term annotation[73]. AMR genes were predicted from sequences uploaded to PathogenWatch on January 22nd, 2024[23]. Mobile elements were predicted by BLAST+ v2.0.13 against reference sequences AY803022.1, NC_001377, NC_002098, GU479464.1, GU479465, and GU479466 which were obtained from the NCBI on December 16th, 2023[21]. The Graphia MCL clustering tool was used to cluster accessory genes based on shared presence or absence in our representative collection with a granularity of 1.5[22].

Prediction of phase variable genes

Simple sequence repeats were identified from the *de novo* assemblies and the GFF files generated by Prokka using PERF v0.4.6[25]. Genes with repeats that varied in the number of tandem repeats in more than 2 isolates where each repeat length was present in at least 5% of genomes were considered as potentially phase variable. Mutations classified as intergenic were manually verified to ensure they impacted the associated gene.

Lower genital tract colonization for in vivo passaging

Bacterial growth conditions: *Ngo* MS11, parental strain, was grown on GC agar (Becton, Dickinson and Company, Sparks, USA) plates supplemented with 1% Isovitalex (ISO, Becton, Dickinson and Company, Sparks, USA) at 37˚C with 5% CO_2 for 16-20 hours. For *Ngo* isolates used for in vivo passaging, agar plates included a vancomycin, colistin, nystatin, and trimethoprim antibiotic mix (VCNT, Becton, Dickinson and Company, Sparks, USA) for *Ngo* selection.

Mouse strains: 6-8 week old female FVB mice containing a human chromosome-derived BAC expressing CEACAM3, -5 and -6, were used for all infection experiments[30]. Animals were housed in cages with sterile rodent chow, sterile water, enrichment material, and 12-hour light-dark cycles.

Gonococcal *in vivo* passaging: Mice were staged by cytological vaginal smears observed under light microscopy. Those in the diestrus stage of the estrous cycle were started on an antibiotic and hormone regimen to control impact of commensal microflora on infection. Antibiotics consisted of 0.6 mg vancomycin hydrochloride (Bioshop, Burlington, ON, Canada) and 2.4 mg streptomycin

sulfate (Sigma-Aldrich, Oakville, Canada) resuspended in 200 µL of PBS administered by intraperitoneal injection daily from 2 days pre-infection (-2) to 5 days post infection (+5), where an dose was provided on Day -1. Trimethoprim (Sigma-Aldrich, Oakville, Canada) was dissolved in the drinking water at 40 mg/mL for ad libitum consumption. Water soluble β-estradiol (SigmaAldrich, Oakville, Canada) was constituted in PBS (0.5 mg/200 µL), filter-sterilized and administered via subcutaneous injection on Days -2, 0 and +2.

On Day 0, bacteria were resuspended in Dulbecco's phosphate-buffered saline (DPBS) with calcium and magnesium (PBS++ , Life Technologies, Burlington, Canada)and mice were inoculated with 10^7 bacteria in 2.5 µL Vaginal lavages were performed every two days by washing the genital tract with 10 µL of PBS++ and plated for bacterial recovery and enumeration. After ~20 hrs bacterial isolates were frozen in 50% brain heart infusion (BHI, Becton, Dickinson and Company, Sparks, USA) and 25% glycerol (EMD Millipore, Darmstadt, Germany) and stored at -80˚C. Clearance of infection was characterized by 2 consecutive vaginal lavages that tested negative for *Ngo*.

Gonococcal *in vitro* passaging: Five culture tubes containing 5 mL of BHI supplemented with 1% ISO was inoculated with the parental strain with a starting optical density at 550 nm (OD550) of 0.05. Bacterial cultures were placed in a 37˚C shaker at 180 rpm. Every 12 hrs, bacterial stocks were generated and sub-cultured for 5 consecutive days.

Ngo whole genome sequencing: *Ngo* isolates were resuspended in 200 µL of PBS++ and genomic DNA was extracted by a QIAGEN DNeasy Blood and Tissue Kit (QIAGEN, Hilden, Germany) following manufacturer instructions. DNA quality and quantity was checked by NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies, Wilmington, USA) and Quant-iT™ PicoGreen® dsDNA Assay Kit (Life Technologies, Oregon, USA) was used to quantify amount of dsDNA.Genomic DNA was prepared for whole genome sequencing by a Nextera XT library preparation kit (Illumina, San Diego, USA). Libraries were run on a 1% agarose gel to ensure libraries were of the correct length prior to poolingand submission to the Donnelly Sequencing Centre, University of Toronto for sequencing using the Illumina MiSeq platform.

Capillary electrophoresis for confirmation of *opa* repeat lengths

Capillary electrophoresis was performed as described previously for *Neisseria meningitidis*[74]. Briefly, strains were cultured overnight on GC agar as described above, and 3 distinct colonies per plate were selected and added to 20µl of Quick Extract buffer. Suspended colonies were then heated at 66°C for 6 minutes, then 98°C for 2 minutes to complete the extraction. Extracted DNA was amplified with PCR using the primers in **Table 3** listed below, with the forward primer conjugated to a 6-FAM dye. Samples were then diluted to approximately 1-2ng/µl before being sent for capillary electrophoresis. The results were analyzed with the ThermoFisher MicroSatellite Analysis tool.

Identification of genes under positive selection

Aligned gene sequences from panaroo for core genes with 5 or more unique sequences were used as input for selection analyses after removing identical sequences and sequences containing premature stop codons using Hyphy rmv[32]. First, a tree for each gene was built using IQTree v2.2.2.7 with 1000 ultrafast bootstraps and otherwise default settings[69]. The BUSTED algorithm was then run as implemented in HyPhy with default settings for each gene using the generated tree and panaroo alignment[33]. Genes identified as being under positive selection were then analyzed with FUBAR as implemented in HyPhy with default settings using the same inputs to identify specific sites under positive selection[34]. To assess the location of predicted residues under positive selection, structural models were generated with AlphaFold3 using the AlphaFold server based on the consensus gene sequences from panaroo[19,75]. Predicted structures were visualized with UCSF ChimeraX v1.9[76].

Genome wide association study

The genome wide association study was performed using pyseer with the linear mixed model setting using the VCF file generated as described above[44]. The distance matrix was calculated using the ClonalFrameML generated tree described above. All other pyseer settings were kept at default. The count_patterns script was used to establish a significance threshold for each tested phenotype. The impact of each variant was predicted with SnpEff using the *Ngo* TUM19854 reference genome obtained as previously described[77]. Structural alignments and comparisons were performed using TM-align with default settings[45].

ACKNOWLEDGEMENTS

We are grateful to Odile Harrison, University of Oxford, for helpful advice and technical support, and the Division of Comparative Medicine & Donnelly Sequencing Centre at the University of Toronto for technical support. This study was supported by operating grants from the Canadian Institutes for Health Research (PJT-159677) and National Institutes of Health (1R01AI146941-01A1). S.D.G. is supported by the Canada Research Chairs Program. D.C. is a recipient of the EPIC Doctoral Award and the DSI Doctoral Fellowship.

ETHICS STATEMENT

All animal procedures conducted in this study were approved by the Local Animal Care Committee (LACC) at the University of Toronto (Protocol #20011775), which is in compliance with ethical and legal requirements under Ontario's Animals for Research Act and the federal Canadian Council on Animal Care.

DATA & CODE AVAILABILITY

The data sets generated in this study are provided in the Supplementary Information and Source Data files. Codes for analysis and data are available on Github [https://github.com/ParkinsonLab/Neisseria_work.git].

AUTHOR CONTRIBUTIONS

J.P. and S.D.G conceived the idea of the research and supervised it. D.C and E.J. performed the computational and data analysis. D.C. and J.M. conducted the mouse passaging and phase variation experiments and analyzed the data. D.C. wrote the manuscript with contributions from J.P., S.G.D. & J.L.

REFERENCES

1. Mayor, M. T., Roett, M. A. & Uduhiri, K. A. Diagnosis and Management of Gonococcal Infections. *Am. Fam. Physician* **86**, 931–938 (2012).

2. Bernstein, K. T., Marcus, J. L., Nieri, G., Philip, S. S. & Klausner, J. D. Rectal Gonorrhea and Chlamydia Reinfection Is Associated With Increased Risk of HIV Seroconversion. *JAIDS J. Acquir. Immune Defic. Syndr.* **53**, 537–543 (2010).

3. Zhang, J. *et al. Neisseria gonorrhoeae* Enhances Infection of Dendritic Cells by HIV Type 1. *J. Immunol.* **174**, 7995–8002 (2005).

4. *WHO Bacterial Priority Pathogens List, 2024: Bacterial Pathogens of Public Health Importance to Guide Research, Development and Strategies to Prevent and Control Antimicrobial Resistance*. https://iris.who.int/bitstream/handle/10665/376776/9789240093461-eng.pdf?sequence=1 (2024).

5. Quillin, S. J. & Seifert, H. S. *Neisseria gonorrhoeae* host adaptation and pathogenesis. *Nat. Rev. Microbiol.* **16**, 226–240 (2018).

6. Ohnishi, M. *et al.* Spread of a chromosomal cefixime-resistant penA gene among different *Neisseria gonorrhoeae* lineages. *Antimicrob. Agents Chemother.* **54**, 1060–1067 (2010).

7. Harrison, O. B. *et al. Neisseria* genomics: current status and future perspectives. *Pathog. Dis.* **75**, 60–60 (2017).

8. Cehovin, A., Lewis, S. B. & William, S. Mobile genetic elements in *Neisseria gonorrhoeae*: movement for change. **75**, 71–71 (2017).

9. Ramsey, M. E., Woodhams, K. L. & Dillard, J. P. The gonococcal genetic island and type IV secretion in the pathogenic *Neisseria*. *Front. Microbiol.* **2**, 61–61 (2011).

10. Siena, E. *et al.* In-silico prediction and deep-DNA sequencing validation indicate phase variation in 115 *Neisseria meningitidis* genes. *BMC Genomics* **17**, 843 (2016).

11. Jolley, K. A. Multi-locus sequence typing. *Methods Mol. Med.* **67**, 173–186 (2001).

12. Harrison, O. B. *et al. Neisseria gonorrhoeae* Population Genomics: Use of the Gonococcal Core Genome to Improve Surveillance of Antimicrobial Resistance. *J. Infect. Dis.* **222**, 1816–1825 (2020).

13. Zhang, J., Khan, A., Kennard, A., Grigg, M. E. & Parkinson, J. PopNet: A Markov Clustering Approach to Study Population Genetic Structure. *Mol. Biol. Evol.* **34**, 1799–1811 (2017).

14. Unemo, M. *et al.* The novel 2016 WHO *Neisseria gonorrhoeae* reference strains for global quality assurance of laboratory investigations: phenotypic, genetic and reference genome characterization. *J. Antimicrob. Chemother.* **71**, 3096–3096 (2016).

15. Phenotypic and genetic characterization of the 2008 WHO *Neisseria gonorrhoeae* reference strain panel intended for global quality assurance and quality control of gonococcal antimicrobial resistance surveillance for public health purposes | Journal of Antimicrobial Chemotherapy | Oxford Academic. https://academic.oup.com/jac/article/63/6/1142/747767.

16. Didelot, X. & Wilson, D. J. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLOS Comput. Biol.* **11**, e1004041–e1004041 (2015).

17. Lees, J. A. *et al.* Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res.* **29**, 304–316 (2019).

18. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Comput. Biol.* **13**, e1005595 (2017).

19. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* **21**, 180 (2020).

20. Prokka: rapid prokaryotic genome annotation | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/30/14/2068/2390517.

21. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

22. Freeman, T. C. *et al.* Graphia: A platform for the graph-based visualisation and analysis of high dimensional data. *PLoS Comput. Biol.* **18**, e1010310 (2022).

23. Sánchez-Busó, L. *et al.* A community-driven resource for genomic epidemiology and antimicrobial resistance prediction of *Neisseria gonorrhoeae* at Pathogenwatch. *Genome Med.* **13**, 61 (2021).

24. European Committee on Antimicrobial Susceptibility Testing (EUCAST). Breakpoint tables for interpretation of MICs and zone diameters. Version 9.0.

25. Avvaru, A. K., Sowpati, D. T. & Mishra, R. K. PERF: an exhaustive algorithm for ultra-fast and efficient identification of microsatellites from large DNA sequences. *Bioinformatics* **34**, 943–948 (2018).

26. Wanford, J. J., Green, L. R., Aidley, J. & Bayliss, C. D. Phasome analysis of pathogenic and commensal *Neisseria* species expands the known repertoire of phase variable genes, and highlights common adaptive strategies. *PLoS ONE* **13**, e0196675 (2018).

27. The meningococcal autotransporter AutA is implicated in autoaggregation and biofilm formation - PubMed. https://pubmed.ncbi.nlm.nih.gov/25059714/.

28. Pachulec, E. & van der Does, C. Conjugative Plasmids of *Neisseria gonorrhoeae*. *PLoS ONE* **5**, e9962 (2010).

29. Zelewska, M. A. *et al.* Phase variable DNA repeats in *Neisseria gonorrhoeae* influence transcription, translation, and protein sequence variation. *Microb. Genomics* **2**, e000078 (2016).

30. Islam, E. A. *et al.* Specific binding to differentially expressed human carcinoembryonic antigenrelated cell adhesion molecules determines the outcome of *Neisseria gonorrhoeae* infections along the female reproductive tract. *Infect. Immun.* **86**, (2018).

31. Erwin, A. L., Haynes, P. A., Rice, P. A. & Gotschlich, E. C. Conservation of the lipooligosaccharide synthesis locus lgt among strains of *Neisseria gonorrhoeae*: requirement for lgtE in synthesis of the 2C7 epitope and of the beta chain of strain 15253. *J. Exp. Med.* **184**, 1233–1241 (1996).

32. Pond, S. L. K., Frost, S. D. W. & Muse, S. V. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**, 676–679 (2005).

33. Gene-Wide Identification of Episodic Selection | Molecular Biology and Evolution | Oxford Academic. https://academic.oup.com/mbe/article/32/5/1365/1134918.

34. Murrell, B. *et al.* FUBAR: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection. *Mol. Biol. Evol.* **30**, 1196–1205 (2013).

35. Belland, R. J., Morrison, S. G., Ison, C. & Huang, W. M. *Neisseria gonorrhoeae* acquires mutations in analogous regions of gyrA and parC in fluoroquinolone-resistant isolates. *Mol. Microbiol.* **14**, 371–380 (1994).

36. Collins, J. A., Oviatt, A. A., Chan, P. F. & Osheroff, N. Target-Mediated Fluoroquinolone Resistance in *Neisseria gonorrhoeae*: Actions of Ciprofloxacin against Gyrase and Topoisomerase IV. *ACS Infect. Dis.* **10**, 1351–1360 (2024).

37. The InterPro database, an integrated documentation resource for protein families, domains and functional sites - PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC29841/.

38. Corbett, K. D., Schoeffler, A. J., Thomsen, N. D. & Berger, J. M. The Structural Basis for Substrate Specificity in DNA Topoisomerase IV. *J. Mol. Biol.* **351**, 545–561 (2005).

39. Wadsworth, C. B., Arnold, B. J., Sater, M. R. A. & Grad, Y. H. Azithromycin resistance through interspecific acquisition of an epistasis-dependent efflux pump component and transcriptional regulator in *Neisseria gonorrhoeae*. *mBio* **9**, (2018).

40. Powell, A. J., Tomberg, J., Deacon, A. M., Nicholas, R. A. & Davies, C. Crystal Structures of Penicillin-binding Protein 2 from Penicillin-susceptible and -resistant Strains of *Neisseria gonorrhoeae* Reveal an Unexpectedly Subtle Mechanism for Antibiotic Resistance. *J. Biol. Chem.* **284**, 1202–1212 (2009).

41. Fenton, B. A. *et al.* Mutations in PBP2 from ceftriaxone-resistant *Neisseria gonorrhoeae* alter the dynamics of the β3–β4 loop to favor a low-affinity drug-binding state. *J. Biol. Chem.* **297**, 101188 (2021).

42. Malekian, N., Agrawal, A. A., Berendonk, T. U., Al-Fatlawi, A. & Schroeder, M. A genome-wide scan of wastewater *E. coli* for genes under positive selection: focusing on mechanisms of antibiotic resistance. *Sci. Rep.* **12**, 8037 (2022).

43. A genome-wide association analysis reveals a potential role for recombination in the evolution of antimicrobial resistance in *Burkholderia multivorans* | PLOS Pathogens. https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1007453.

44. Lees, J. A., Galardini, M., Bentley, S. D., Weiser, J. N. & Corander, J. pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics* **34**, 4310–4312 (2018).

45. Bittrich, S., Segura, J., Duarte, J. M., Burley, S. K. & Rose, Y. RCSB protein Data Bank: exploring protein 3D similarities via comprehensive structural alignments. *Bioinformatics* **40**, btae370 (2024).

46. Varadi, M. *et al.* AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* **52**, D368–D375 (2024).

47. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

48. Ng, L.-K., Martin, I., Liu, G. & Bryden, L. Mutation in 23S rRNA Associated with Macrolide Resistance in *Neisseria gonorrhoeae*. *Antimicrob. Agents Chemother.* **46**, 3020–3025 (2002).

49. de Korne-Elenbaas, J., Bruisten, S. M., van Dam, A. P., Maiden, M. C. J. & Harrison, O. B. The *Neisseria gonorrhoeae* Accessory Genome and Its Association with the Core Genome and Antimicrobial Resistance. *Microbiol. Spectr.* **10**, e02654-21.

50. Pinto, M. *et al. Neisseria gonorrhoeae* clustering to reveal major European whole-genome-sequencing-based genogroups in association with antimicrobial resistance. *Microb. Genomics* **7**, 000481 (2020).

51. Cehovin, A., Jolley, K. A., Maiden, M. C. J., Harrison, O. B. & Tang, C. M. Association of *Neisseria gonorrhoeae* Plasmids With Distinct Lineages and The Economic Status of Their Country of Origin. *J. Infect. Dis.* **222**, 1826–1836 (2020).

52. San Millan, A. & MacLean, R. C. Fitness Costs of Plasmids: a Limit to Plasmid Transmission. *Microbiol. Spectr.* **5**, (2017).

53. Gulati, S., McQuillen, D. P., Mandrell, R. E., Jani, D. B. & Rice, P. A. Immunogenicity of *Neisseria gonorrhoeae* Lipooligosaccharide Epitope 2C7, Widely Expressed In Vivo with No Immunochemical Similarity to Human Glycosphingolipids. *J. Infect. Dis.* **174**, 1223–1237 (1996).

54. Yu, D. *et al.* A Genome-Wide Identification of Genes Undergoing Recombination and Positive Selection in *Neisseria*. *BioMed Res. Int.* **2014**, 815672 (2014).

55. Wang, S. *et al.* Next Generation Sequencing and Comparative Genomic Analysis Reveal Extreme Plasticity of Two *Burkholderia glumae* Strains HN1 and HN2. *Pathogens* **11**, 1265 (2022).

56. Schubert, B., Maddamsetti, R., Nyman, J., Farhat, M. R. & Marks, D. S. Genome-wide discovery of epistatic loci affecting antibiotic resistance in *Neisseria gonorrhoeae* using evolutionary couplings. *Nat. Microbiol.* **4**, 328–338 (2019).

57. Chopra, I. & Roberts, M. Tetracycline Antibiotics: Mode of Action, Applications, Molecular Biology, and Epidemiology of Bacterial Resistance. *Microbiol. Mol. Biol. Rev.* **65**, 232 (2001).

58. Burns, A. L. *et al.* Retargeting azithromycin analogues to have dual-modality antimalarial activity. *BMC Biol.* **18**, 133 (2020).

59. Wang, X. *et al.* Identification of Plasmodium falciparum Mitochondrial Malate: Quinone Oxidoreductase Inhibitors from the Pathogen Box. *Genes* **10**, 471 (2019).

60. Jolley, K. A., Bray, J. E. & Maiden, M. C. J. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res.* **3**, 124 (2018).

61. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).

62. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1297 (2010).

63. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

64. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

65. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

66. Schubert, E. & Rousseeuw, P. J. Fast and eager k-medoids clustering: O (k) runtime improvement of the PAM, CLARA, and CLARANS algorithms. *Inf. Syst.* **101**, 101804 (2021).

67. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. & Hornik, K. cluster: Cluster Analysis Basics and Extensions. (2023).

68. Kassambara, A. & Mundt, F. factoextra: Extract and Visualize the Results of Multivariate Data Analyses. (2020).

69. 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).

70. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

71. 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).

72. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

73. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale | Molecular Biology and Evolution | Oxford Academic. https://academic.oup.com/mbe/article/38/12/5825/6379734.

74. Jen, F. E.-C., Seib, K. L. & Tan, A. Screening DNA Repeat Tracts of Phase Variable Genes by Fragment Analysis. in *Neisseria meningitidis: Methods and Protocols* (eds. Seib, K. L. & Peak, I. R.) 93–104 (Springer, New York, NY, 2019). doi:10.1007/978-1-4939-9202-7_6.

75. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

76. Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

77. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)* **6**, 80–92 (2012).

TABLES

Table 1: Predicted new phase variable genes.

Gene	Repeat	No. of repeats	Annotation	NGK locus	NGO locus
group_1584	A	33	Pertactin	NGK_2016	NGO_1158
group_280	AACGC	24	Hypothetical protein	-	NGO_0366
argS	AAGC	18	Arginyl-tRNA synthetase	NGK_0832	NGO_0963
group_1317	AACGC	17	C-5 cytosine-specific DNA methylase	NGK_0522	NGO_0365
group_1449	AT	11	Spermidine synthase	NGK_2156	NGO_1752
group_1206	C	11	ISXO2-like transposase domain	NGK_1147	NGO_1157
kfrB	C	10	-	-	-
glyS	C	9	Glycyl-tRNA synthetase beta	NGK_2629	NGO_2154
prn	A	8	Autotransporter	NGK_1206	NGO_0694
yigL	C	7	Hydrolase	NGK_1967	-

Table 2: Predicted phase variable genes from mouse passaging experiment.

Gene	Repeat	No. of repeats	Annotation	NGK locus	NGO locus
group_1059	C	3	Pilus assembly protein	NGK_0074	NGO_0055
group_1287	C	2	Pilus assembly protein	NGK_0074	NGO_0055
group_399	CCTCT	2	Opacity protein	NGK_0693	NGO_07725
group_579	CCTCT	2	Opacity protein	NGK_0693	NGO_1513
group_784	C	3	Enterobactin receptor *fetA*	NGK_2557	NGO_2093
group_1283	CCTCT	3	Opacity-associated protein	NGK_0749	NGO_08230
group_973	AGCC	2	Adenine specific DNA methylase	NGK_1273	NGO_0641
group_1471	CCTCT	2	Opacity-associated protein	NGK_2410	NGO_1513
group_1111	C	2	Glycosyl transferase family 1	NGK_2534	NGO_2072

Table 3: Primers used for capillary electrophoresis.

Primer	Modification	Sequence
Opa-For	5’ 6-FAM	CAACATCAGTGAAAATCGTTCTTTT
Opa-Rev	None	TCGTGGGTAATGTGTTCGTAG

FIGURES

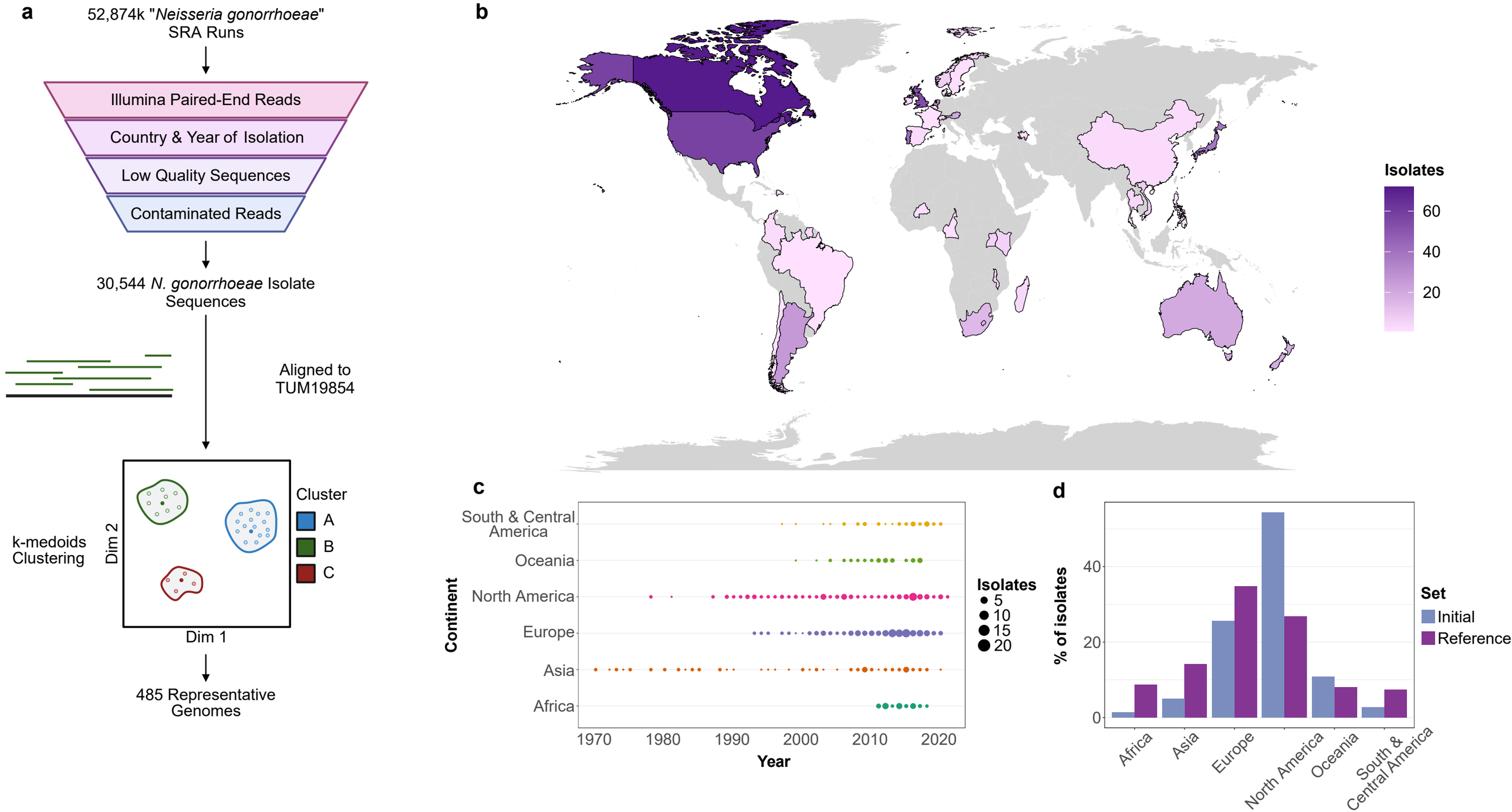
a
52,874k "Neisseria gonorrhoeae" SRA Runs
Illumina Paired-End Reads
Country & Year of Isolation
Low Quality Sequences
Contaminated Reads
30,544 N. gonorrhoeae Isolate Sequences
Aligned to TUM19854
k-medoids Clustering
Dim 2
Dim 1
Cluster
A
B
C
485 Representative Genomes
b
Isolates
60
40
20
c
Continent
South & Central America
Oceania
North America
Europe
Asia
Africa
1970
1980
1990
2000
2010
2020
Year
Isolates
5
10
15
20
d
% of isolates
0
20
40
Africa
Asia
Europe
North America
Oceania
South & Central America
Set
Initial
Reference

Figure 1: Overview of representative set of 485 *N. gonorrhoeae* genomes.

a Schematic of selection of representative genomes. **b** Geographic locations of isolates in the representative set. **c** Geographic and temporal distribution of isolates in the representative set. **d** Comparison of the geographic distribution of isolates between the initial set of 52,784 SRA runs and the representative set of 485.

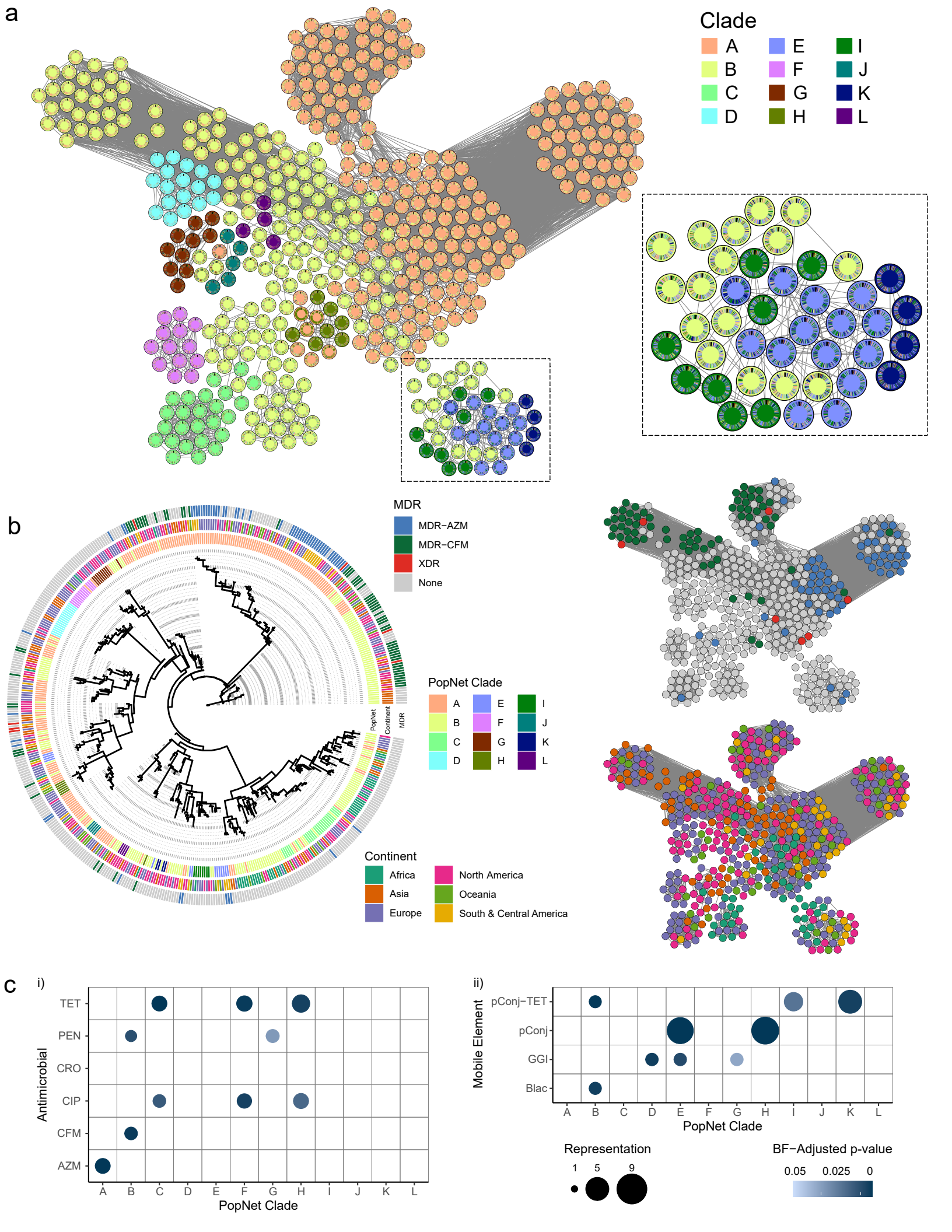
a
Clade
A
B
C
D
E
F
G
H
I
J
K
L
b
MDR
MDR−AZM
MDR−CFM
XDR
None
PopNet
Continent
MDR
PopNet Clade
A
B
C
D
E
F
G
H
I
J
K
L
Continent
Africa
Asia
Europe
North America
Oceania
South & Central America
c
i)
Antimicrobial
TET
PEN
CRO
CIP
CFM
AZM
A B C D E F G H I J K L
PopNet Clade
ii)
Mobile Element
pConj−TET
pConj
GGI
Blac
A B C D E F G H I J K L
PopNet Clade
Representation
1 5 9
BF−Adjusted p-value
0.05 0.025 0

Figure 2: Population structure of *N. gonorrhoeae* is influenced by AMR and the presence of mobile elements.

a PopNet network view of 485 representative *Ngo* genomes. Nodes represent genomes, coloured by their assigned clades. Annuli show the chromosome painting generated for each genome, with regions sharing a colour showing overlap between genomes. A selection of genomes from clades B, E, I & K are shown to illustrate the chromosome painting visualization. **b** ClonalFrameML generate phylogenetic tree of representative *Ngo* genomes. Outer rings show the assigned PopNet clade, MDR (multi-drug-resistant) status and continent of isolation for each genome. Top: MDR status overlaid on the PopNet network view. Bottom: Continent of isolation overlaid on the PopNet network view. AZM: azithromycin, CFM: cefixime, XDR: extreme drug resistant. **c** Association between antimicrobial resistance (left, i) or mobile elements (right, ii) and PopNet clades. Only significant associations with a p-value < 0.05 as determined by Fisher's exact test with Bonferonni multiple testing correction are shown.

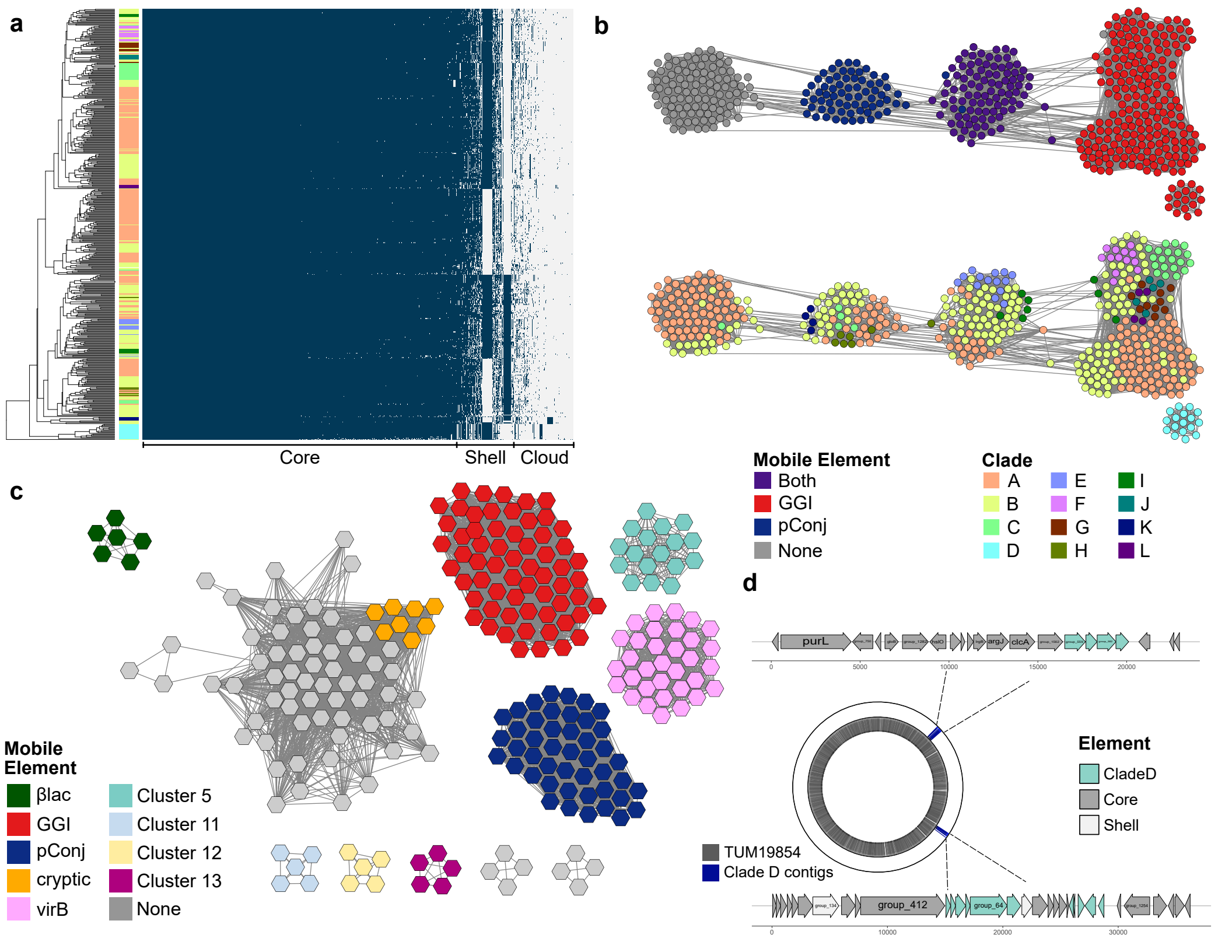
a
Core
Shell
Cloud
b
Mobile Element
Both
GGI
pConj
None
Clade
A
B
C
D
E
F
G
H
I
J
K
L
c
Mobile Element
βlac
GGI
pConj
cryptic
virB
Cluster 5
Cluster 11
Cluster 12
Cluster 13
None
d
purL
argJ
clcA
group_412
group_64
0
5000
10000
15000
20000
30000
TUM19854
Clade D contigs
Element
CladeD
Core
Shell

Figure 3: The pangenome of *N. gonorrhoeae* is dominated by large mobile elements, which are associated with PopNet clades determined by chromosomal variants.

a Pangenome of *Ngo* is composed of 2460 genes, with 1800 Core (present in > 95% of genomes), 367 Shell (present in between 15% and 95% of genomes) & 283 Cloud (present in <15% of genomes) genes. Dendrogram and ordering of genomes on the y-isolates calculated using complete linkage from a Euclidian distance matrix generated based on shared gene content. **b** Top: Network view of isolates clustered by shared gene content coloured by the presence or absence of 2 large mobile elements: the GGI (gonococcal genetic island) and the pConj (conjugative plasmid). Bottom: Genomes coloured by PopNet clade overlaid on the network view of isolates clustered by shared gene content. **c** Clustering of genes based on their co-occurrence in genomes. Clusters defined by MCL clustering as implemented in Graphia which are present in <50% of genomes and composed of 5 or more genes are coloured. **d** Organization of genes from MCL cluster 11 associated with clade D.

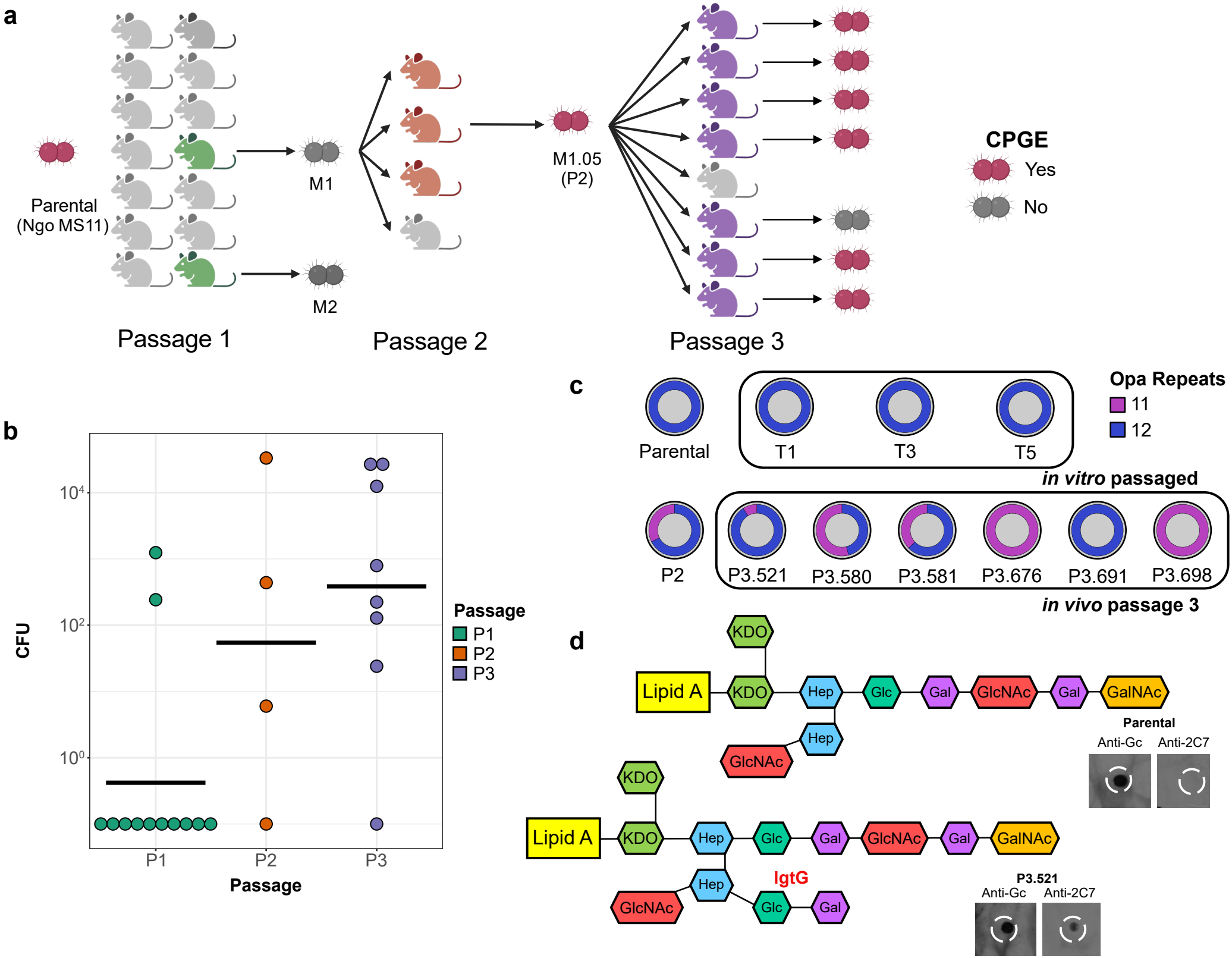
a
Parental
(Ngo MS11)
M1
M2
M1.05
(P2)
Passage 1
Passage 2
Passage 3
CPGE
Yes
No
b
CFU
10^4
10^2
10^0
P1
P2
P3
Passage
Passage
P1
P2
P3
c
Parental
T1
T3
T5
Opa Repeats
11
12
in vitro passaged
P2
P3.521
P3.580
P3.581
P3.676
P3.691
P3.698
in vivo passage 3
d
KDO
Lipid A
KDO
Hep
Glc
Gal
GlcNAc
Gal
GalNAc
Hep
GlcNAc
Parental
Anti-Gc
Anti-2C7
KDO
Lipid A
KDO
Hep
Glc
Gal
GlcNAc
Gal
GalNAc
Hep
IgtG
GlcNAc
Glc
Gal
P3.521
Anti-Gc
Anti-2C7

Figure 4: Passaging of *N. gonorrhoeae* improves colonization in a mouse model of infection which may be associated with phase variation of *opa*.

a Schematic overview of *in vivo* and *in vitro* passaging experiment. Coloured mice represent mice successfully colonized by *Ngo* 2 days post infection. Coloured gonococci indicate isolates sequenced by capillary electrophoresis to confirm the repeat lengths of *opa*. **b** Recovered CFU (colony-forming units) of *Ngo* recovered 2 days post infection. Black bars represent the mean CFUs recovered for each passage. **c** Ratio of *opa* repeats in each isolate sequenced by capillary electrophoresis. **d** Schematic of *Ngo* lipooligosaccharide with (bottom) and without (top) lgtG expression, resulting in the production of the 2C7 epitope. Isolates passaged *in vivo* express lgtG, while no expression is observed in isolates passaged *in vitro.* Immunoblot assay for the 2C7 isotope confirms the expression of lgtG in a passage 3 isolate, which is absent from the parental isolate. KDO: 3-Deoxy-D-manno-oct-2-ulosonic. Hep: Heptose. Glc: Glucose. Gal: Galactose. GlcNAc: N-acetylglucosamine. GalNAc: N-Acetylgalactosamine.

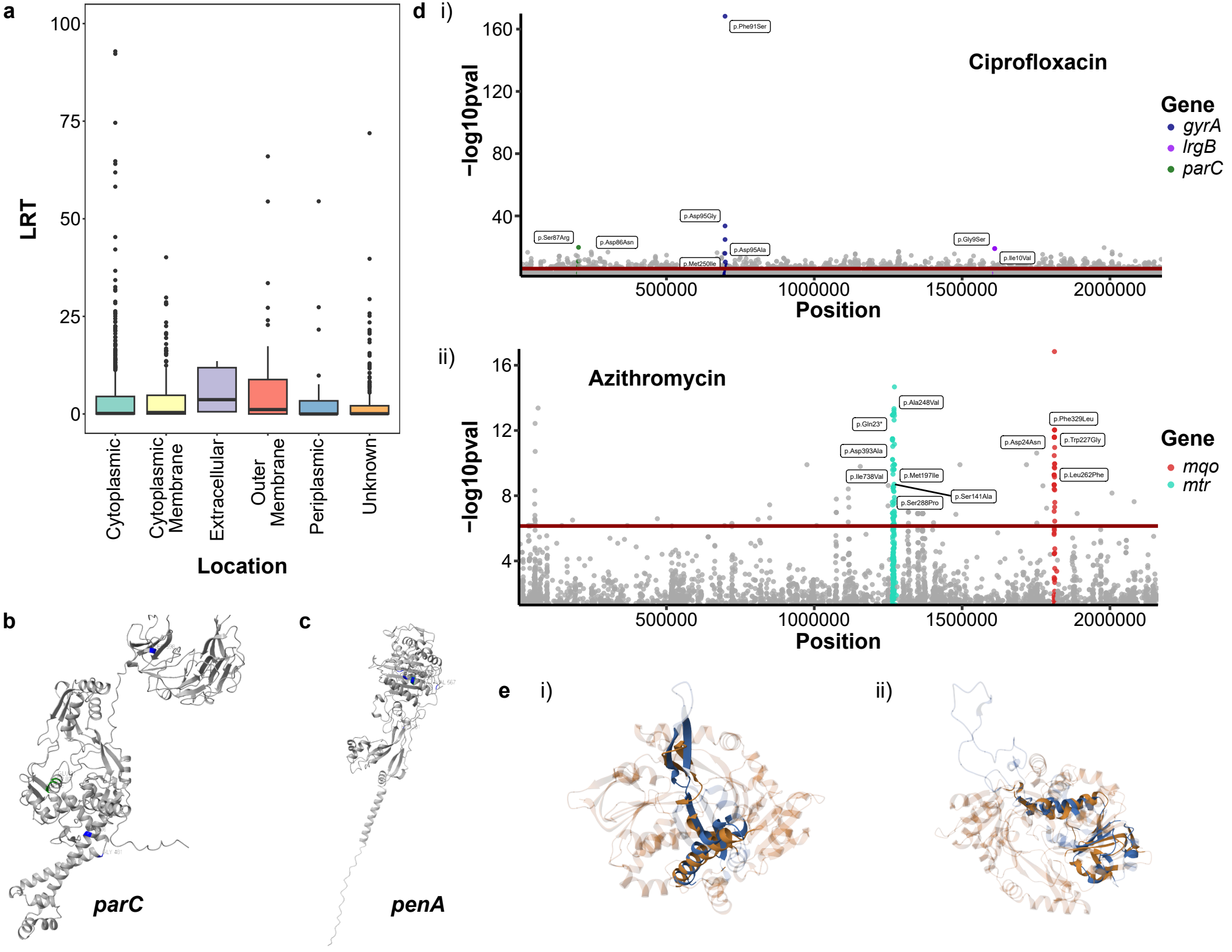
a
LRT
100
75
50
25
0
Cytoplasmic
Cytoplasmic Membrane
Extracellular
Outer Membrane
Periplasmic
Unknown
Location
d
i)
Ciprofloxacin
−log10pval
160
120
80
40
p.Phe91Ser
p.Asp95Gly
p.Ser87Arg
p.Asp86Asn
p.Asp95Ala
p.Met250Ile
p.Gly9Ser
p.Ile10Val
500000
1000000
1500000
2000000
Position
Gene
gyrA
lrgB
parC
ii)
Azithromycin
−log10pval
16
12
8
4
p.Ala248Val
p.Gln23*
p.Asp393Ala
p.Ile738Val
p.Met197Ile
p.Ser288Pro
p.Ser141Ala
p.Phe329Leu
p.Asp24Asn
p.Trp227Gly
p.Leu262Phe
500000
1000000
1500000
2000000
Position
Gene
mqo
mtr
b
parC
c
penA
e
i)
ii)

Figure 5: GWAS and tests for positive selection identify sites linked to AMR in *N. gonorrhoeae*.

a Boxplot of LRT (likelihood ratio test statistic) of positive selection for *Ngo* genes for predicted cellular localization. **b** Predicted residues with strong signatures of positive selection in *parC*. Residues highlighted in green are under positive selection and known to contribute to ciprofloxacin resistance. Residues in blue are under positive selection with no known links to AMR. Structural prediction from AlphaFold3 visualized with USCF ChimeraX. **c** Predicted residues with strong signatures of positive selection in *penA*. Residues in blue are under positive selection with no known links to AMR. Structural prediction from Alphafold3 visualized with USCF ChimeraX. **d** Manhattan plots of variants with significant associations with ciprofloxacin (top) and azithromycin (bottom) resistance. Only variants with an LRT p-value <0.05 are shown. Red lines indicate significance thresholds determined by Bonferroni correction with the number of unique variant patterns as the number of multiple tests. **e** Pairwise alignment of AlphaFold predicted structures for Mqo (yellow) with RplV (left, blue) and RplD (right, blue) using RSCPD TM-align. Darker residues indicate regions of structural overlap.