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: scott.gray.owen@utoronto.ca & 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 defined 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 the surface proteins 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.

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 typically drives a highly inflammatory response with purulent discharge[1]. Without treatment, infection can result in severe morbidities such as pelvic inflammatory disease or infertility[1]. 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[2]. 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[3].

Facilitating its ability to acquire AMR, *Ngo* is naturally competent with many DNA uptake sequences throughout genome, which enables efficient recognition, uptake and incorporation of genetic matter from gonococcal strains and other pathogenic and commensal *Neisseria* species[4,5]. Driving this "promiscuity" are mobile elements, including plasmids containing AMR encoding genes and the Gonococcal Genetic Island (GGI) which encodes a type IV secretion system, together with several proteins of unknown function[6,7]. Further, many gonococcal genes exhibit phase variability in which replication error-produced changes in the number of tandem repeat sequences can alter protein expression[8]. 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*[5]. This population heterogeneity has confounded traditional efforts to define clonal lineages, requiring instead the use of alternative approaches such as multilocus sequence typing (MLST). Complementing such efforts are tools such as PopNet, which take a more agnostic view to selection of loci for determining population structure[9]. Further, PopNet's use of network graphs help reveal shared genetic elements among multiple strains that are otherwise lost in the rigid structure of more traditional bifurcating phylogenetic trees.

Using genomic data from over 30,000 *Ngo* isolates drawn from all publicly available gonococcal genomes deposited on the NCBI Sequence Read Archive, we sought to apply PopNet to

65 systematically define the global population structure of *Ngo*, investigate its pangenome and assess
66 the incidence of AMR mechanisms and the genetic elements helping support their dissemination.
67 Through this analysis we identify and subsequently validate novel phase variable genes and
68 previously unrecognized AMR associated loci.
69
70 **RESULTS**
71 **A representative set of 485 genomes captures temporal and geographic diversity**
72 To begin our analyses of the global population structure of *Ngo*, we first defined a representative
73 set of 485 isolates from an initial set of genome sequence reads from 31,595 gonococcal isolates
74 hosted by the NCBI sequence read archive annotated by year and location of isolation (September
75 26th 2023; **Supplementary Table S1)**. To maximize geographic diversity, we prioritized isolates
76 based on their continent and year of isolation (see Methods; **Figure 1A**). In brief, following
77 alignment of genome sequence reads to the reference genome (*Ngo* strain TUM19854),
78 representative isolates were selected by grouping isolates by their continent and year of isolation.
79 Isolates within each group were then clustered based on average nucleotide identity estimated by
80 the Jaccard Index, resulting in a set of 451 isolates, to which an additional 20 isolates from rare
81 body sites together with a set of 14 isolates representing the WHO reference collection[10,11] were
82 added. The resulting reference collection of 485 isolates represent 36 countries, with 38% of the
83 isolates recovered from Canada (72 isolates), the United States (58), and the United Kingdom (54)
84 (**Figure 1B; Supplementary Table S2**). The reference set captures isolates recovered between
85 1972-2023, with the majority (308, 63.5%) recovered between 2010-2020 (**Figure 1C;**
86 **Supplementary Table S2**). Notably, our prioritization strategy resulting in a substantially higher
87 proportion (148, 30.4%) of isolates from Asia, South America, and Africa compared to the initial
88 set of 31,595 isolates (**Figure 1D**).
89
90 **PopNet defined clades show limited enrichment in specific locales**
91 Having defined our set of representative isolates, we next defined the population structure of these
92 isolates based on shared patterns of single nucleotide polymorphisms (SNPs), relative to the
93 reference genome (*Ngo* strain TUM19854), using PopNet. This revealed 12 clades (designated A-
94 L; **Figure 2A**) ranging in size from 203 isolates (clade A) to 4 isolates (clades K & L). Within
95 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 2Bi)**.

Comparison of PopNet clades with a maximum likelihood tree constructed using ClonalFrameML[12] showed good concordance, with clades broadly remaining intact between the two methods with clear exceptions (**Figure 2Bii**). Comparisons with two other population level clustering methods, cgMLST and PopPunk[13,14], showed that many cgMLST or PopPunk clusters with more than 5 isolates were located within a larger PopNet cluster (**Supplementary Figure S1**). 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[15]. Genomes were subsequently annotated using Prokka[16] to identify genes and orthologs defined using Panaroo[17] (**Supplementary Table S3)**. Across all genomes we identified a total of 2,460 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 (non-core) genes were clustered according to their presence or absence across genomes using the Jaccard index to calculate similarity and MCL clustering with a granularity of 1.10 using Graphia[18] (**Figure 3C)**. Several large, distinct clusters emerged, which we found to represent 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). The GGI was identified in 315 (64.7%) isolates, the cryptic plasmid was identified in 429 (90.1%) isolates, the β-lactamase

127 plasmid was identified in 70 (14.5%) isolates, and the conjugative plasmid was identified in 165
128 (33.9%) isolates, of which 97 (19.9% of total isolates; 58.8% of pConj containing isolates) were a
129 tetM-containing plasmid.
130
131 Notably, clustering of isolates based on their gene complements was largely congruent with
132 PopNet defined clades (**Figure 3B**). This gene-based clustering appears to have been largely
133 driven by the presence or absence of mobile elements, specifically the GGI and the conjugative
134 plasmid, neither of which are present in *Ngo* TUM19854 and were therefore not included in the
135 PopNet analysis (**Figures 3A and 3B**). Further investigation of PopNet defined clades revealed a
136 spectrum of associations with different mobile elements (**Figure 2Cii**). Of note, isolates from
137 Africa were found to be significantly enriched in pConj-TET (34/42 isolates; $p=5.59x10^{-19}$,
138 Fisher's exact) and the β-lactamase plasmid (21/42; $p=1.34x10^{-8}$, Fisher's exact), independent of
139 the year of isolation and likely associated with the continued use of penicillin and tetracycline to
140 treat *Ngo* infections in Africa.
141
142 We used the MCL clustering tool implemented in Graphia, to identify clusters of genes based on
143 their shared presence or absence in our set of representative genomes[18]. While some of these
144 clusters lined up with the mobile elements described above, others did not. We sought to further
145 characterize several of these clusters that were present in <50% of our genomes and included at
146 least 5 genes (**Supplementary Table S4**). Of the 5 clusters meeting these criteria, one (MCL
147 cluster 3) represented the previously identified VirB type IV secretion system while the remaining
148 four represent putatively novel mobile elements (**Figure 3C**). MCL cluster 11 was composed of 5
149 genes which appear to represent a fully intact *Microviridae* genome present in 7 genomes. All the
150 cluster 11 genes are found on a single contig in all isolates where it was present, and do not appear
151 to be integrated into the gonococcal genome. The largest cluster (MCL cluster 5) was composed
152 of 17 genes present in 17 genomes. Interestingly, all 17 genomes were assigned to PopNet clade
153 D, which includes isolates from the United Kingdom, Japan and the United States, indicating either
154 dissemination within the global gonococcal population or multiple independent acquisitions. Many
155 of the genes present in cluster 5 are annotated as putative phage protein, however these genes do
156 not all appear on the same contig and instead are split between 2 contigs flanked by core genes
157 (**Figure 3D**). The last cluster (MCL cluster 13) was the most common (present in 217/485, 44.7%

158 of our representative set of genomes) and was also composed of 5 genes. Unlike the other clusters,
159 MCL cluster 13 does not appear to be phage related, and appears to include a membrane associated
160 ATP-dependent transporter.
161
162 **PopNet defined clades exhibit distinct signatures of AMR mechanisms that are often**
163 **mutually exclusive**
164 With increased concern over the propagation of AMR within *Ngo* populations, we next used the
165 online platform PathogenWatch[19] to predict resistance to six antimicrobials – penicillin (PEN),
166 tetracycline (TET), ciprofloxacin (CIP), azithromycin (AZM), cefixime (CFM) and ceftriaxone
167 (CRO) - for each of the 485 reference isolates (**Supplementary Figure S2**). Incidence of resistance
168 to the three older antimicrobials (PEN, TET & CIP) were higher than rates of resistance to the
169 other three and were significantly enriched in isolates derived from Africa (p=0.029, 0.00011 and
170 0.016 for PEN, TET & CIP respectively, Fisher's exact).
171
172 We found several PopNet clades were significantly associated with resistance to these specific
173 antimicrobials (**Figure 2B).** Beyond the presence/absence of resistance to certain antimicrobials,
174 we also observed clades associated with specific mechanisms of resistance. Clade C was enriched
175 in isolates resistant to TET (23/25 isolates; $p=1.63x10^{-5}$, Fisher's exact) despite the absence of the
176 pConj-TET plasmid. Instead, resistance appears a consequence of a double mutation (*mtrR*
177 promoter -5delA deletion & *rpsJ* V57M). This is contrasted by the high rates of TET resistance in
178 clade B which also carried the pConj-TET plasmid at rates. The association of AMR patterns and
179 mechanisms of resistance with clades suggests that AMR helps shape the global population
180 structure of *Ngo.*
181
182 **Analysis of repeat expansion distributions identifies 57 potentially novel phase variable**
183 **genes**
184 Previous studies have shown that many *Ngo* genes contain repeats that can result in slipped strand
185 mispairing or site-specific recombination altering their expression and providing a reversible
186 process to adapt to rapidly changing environments. Within our collection of 485 reference isolates
187 we searched for candidate phase variable (PV) genes using the repeat motif finding algorithm
188 PERF[20] to detect simple sequence repeats. From this set we identified 85 genes with repeats that

189 varied in length, with each repeat length appearing in at least 1% of available sequences for that
190 gene (**Supplementary Table S5**). These genes were compared with a previous set of 34 genes,
191 identified through comparisons of locus tags between the reference strains FA1090 and
192 NCCP11945, with 28 appearing in both sets. This overlap included 11 *opa* genes, 6 genes
193 characterized as *pilC2* or pilin cassettes, and other well-characterized phase variable genes such as
194 *lgtG*, *fetA* and *modA/B*, helping to validate this approach (**Supplementary Table S5**).
195
196 The remaining 57 genes represent putative phase variable genes, with the 10 genes exhibiting the
197 greatest repeat length variation shown in **Table 1**. Among these 10, group_1317 is annotated as a
198 cytosine-specific methyltransferase, consistent with other gonococcal methyltransferases (e.g.
199 *modA* and *modB*) known to be phase variable. Another 2 predicted genes (*argS* and *glyS*) are
200 annotated as arginyl- and glycyl-tRNA synthetases, a system which was previously identified in a
201 computational screen for phase variable genes[21]. We identified 2 genes (group_1584 and *prn*) that
202 were annotated as pertactin or autotransporters, with *prn* appearing to encode a near full length
203 autotransporter while group_1584 represents a truncated form, with only the N-terminal domain
204 encoded[22]. Another autotransporter, *autA* is known to be phase variable, however the identified
205 repeat for *autA* is AAGC, while we identified poly-A tracts in group_1584 and *prn*[23]. While all the
206 genes so far described are chromosomally encoded "Core" genes, we also identified a gene (*kfrB*)
207 carried on the conjugative plasmid that possesses a poly-G repeat tract immediately after the start
208 codon; the protein this encodes plays a role in plasmid partitioning[24].
209
210 **Passaging gonococci *in vivo* reveals phase variable gene selection**
211 We aimed to evaluate the potential impact that phase variable genes would have on enabling
212 gonococci to adapt to and survival in a more natural environment. To accomplish this, we utilized
213 a lower genital tract colonization model with female FVB mice expressing human CEACAM3,
214 CEACAM5, and CEACAM6[25]. A schematic of the mouse passaging work is shown in Figure 4A.
215
216 For initial infection (Passage 1; P1), 12 mice were each inoculated with 10^7 bacteria of N2009
217 (laboratory-passaged MS11 derivative; denoted parental strain), and vaginal washes were
218 performed every two days to recover viable bacteria. On Day 2, only 2 of the 12 mice remained
219 colonized by the bacteria; these were denoted as M1 and M2 (Figure 4A). Mice infected from the

220 M1 lineage demonstrated progressively higher rates of colonization, 75% and 88% in Passage 2
221 (P2) and Passage 3 (P3), respectively (**Supplementary Figure S3**). Moreover, we also observed
222 a significant increase in bacterial burden recovered from the P3 isolates compared to the initial
223 inoculum (Wilcoxon ranked-sum, p= 0.0047) (**Figure 4B**). These findings suggest that gonococcal
224 isolates passaged *in vivo* are better adapted for survival during the initial stages of infection
225 **(Figure 4B**).
226
227 To differentiate mutations that confer a fitness advantage *in vivo* from those selected by simple
228 subculture, we performed complimentary *in vitro* passaging using the parental strain. From this
229 complementary screen, we identified 9 genes with signatures of phase variability including the
230 known phase variable genes *pilC2*, pilin, *opa* (4), *fetA,*, *lgtG* and *modA* (**Figure 4C**). We observed
231 that the predicted number of repeats in group_1283, which encodes an Opa protein, dropped from
232 12 to 11 in P2, P3_580 and P3_698. No shift was observed in any of the *in vitro* strains, suggesting
233 that this shift might contribute to the improved colonization of *in vivo* passage 3 isolates. Due to
234 the inherent population heterogeneity in *Ngo*, especially when it comes to phase variable genes,
235 we used capillary gel electrophoresis to determine the bacterial population proportion containing
236 11 or 12 repeats based on peak density. We found that the parental strain and the 3 tested *in vitro*
237 passage strains showed no variation in repeat length, remaining at 12 repeats the entire time, while
238 the passage 2 and 3 strains were more heterogeneous, with approximately 32.5% of the recovered
239 passage 2 isolates and 100% of P3_676 and P3_698 at 11 repeats (**Figure 4D**). We were unable to
240 obtain sufficient genetic material from recovered P3_680 for analysis. These findings indicate that
241 the majority of the *in vivo*-passaged population has been selected to express this Opa variant (with
242 11 repeats), consistent with its key role in adhesion to host CEACAMs, though a subset of the
243 population remains Opa off in the tissues[25].
244
245 The selection within *lgtG* (group_1111) was also notable, as the parental and all the *in vitro* strains
246 contained a 10bp-long poly-C repeat, while the P2 and all the P3 isolates shifted to an 11bp long
247 repeat. This shift enables complete transcription of *lgtG*, resulting in the attachment of the β-chain
248 to the gonococcal lipooligosaccharide and production of the 2C7 epitope (**Figure 4E**)[26]. Due to
249 the challenges of differentiating single nucleotide length repeats with capillary electrophoresis, we
250 did not perform this analysis for *lgtG*, however we did confirm the presence of the 2C7 epitope

251 exclusively in the *in vivo* passaged strains through an immunoblot assay. Together these findings
252 highlight the ability of our method to identify phase variable genes in an *in vivo* setting, and also
253 identified phase variation in 2 genes, *opa* and *lgtG*, which may play a substantial role in improving
254 *Ngo*'s ability to colonize and infect a humanized mouse model.

255 **Analysis of sequence variants reveals that novel sites of genes involved in AMR are under**
256 **positive selection**

257 In the previous sections our genomic surveys identified several phase variable genes that may help
258 *Ngo* adapt to environmental perturbations. Another signature of genes supporting adaptation is
259 whether they are under positive selection. To identify such genes we used the BUSTED algorithm
260 as implemented in HyPhy[27,28]. Of the 1,853 core genes, 1,724 were sufficiently variable to test for
261 selection. We identified 355 genes with significant evidence for positive selection (likelihood ratio
262 test statistic > 0, p-value < 0.05). As expected, given their direct contact with the environment,
263 positively selected genes were significantly enriched in outer membrane proteins (Benjamini-
264 Hochberg (BH) corrected p-value 0.00615) (**Figure 5A**). In addition, we also identified several
265 AMR-associated genes as being under positive selection, including *porB*, *penA*, *parC*, *mtrD*, and
266 *gyrA*. To identify specific sites in AMR genes under positive selection, we applied the FUBAR
267 algorithm as implemented in HyPhy[29]. Sites with a posterior probability of positive selection > 0.9
268 were considered as sites with strong evidence for selection.
269

270 Focusing on the gyrase subunit GyrA, our analyses identified residues S91 and D95, previously
271 characterized as conferring resistance to ciprofloxacin[30], as being under positive selection
272 (**Supplementary Table S6**). Ciprofloxacin resistance is also associated with mutations in residues
273 D86, S87, S88, and E91 in the type IV topoisomerase subunit, ParC. In addition to sites D86, S87
274 and E91, which interact directly with ciprofloxacin[31,32], we also identified residues T386, G481,
275 and S598 as being under positive selection. Residues T386 and G481 reside in the α-helices that
276 connect the catalytic and dimerization domains, while residue S598 is located in the C-terminal
277 domain (**Figure 5B**)[33]. Independent of these, we identified 8 residues (310, 642, 742, 922, 926,
278 1052, 1065, 1067) with signatures indicating strong positive selection in *mtrD*, likely reflecting
279 the acquisition of mosaic alleles from other *Neisseria* sp. associated with azithromycin
280 resistance[34]. For the known penicillin resistance gene, *penA*, we identified 6 residues with strong
281 signatures of positive selection (**Figure 5C**). We identified residues A550, P553, K556, and V567

282 within the α11 helix, which also includes the P551 variant previously linked to penicillin resistance
283 through inhibition of acylation by β-lactamases[35]. We also identified a residue (A502) in the region
284 between the β3 and β4 loops that has been linked to ceftriaxone resistance[36]. The final residue
285 under positive selection was Y201, which is located between the α3n helix and β6n loop in the N-
286 terminal domain of PenA, which to date has not been linked to AMR[35]. Cumulatively, these finding
287 reveal signatures of strong positive selection in individual residues of AMR-associated genes,
288 which may represent new or complementary sites contributing to resistance. This reflects prior
289 work that has used signatures of positive selection to identify genes contributing to antimicrobial
290 resistance in *Escherichia coli*[37].
291
292 **Genome wide association study identifies previously unreported sequence variants**
293 **associated with AMR**
294 Beyond signatures of positive selection, genome wide association studies offer a powerful
295 approach to identify SNPs associated with specific traits such as AMR[38]. Here we used the
296 software tool, pyseer, with the linear mixed model mode to identify SNPs significantly associated
297 with predicted resistance to antimicrobials (**Supplementary Table S7)**[39]. We identified several
298 well-known variants together with several less characterized mutations significantly associated
299 with AMR. For ciprofloxacin resistance, the former includes the well studied S91F and D95A/N
300 mutations in GyrA, as well as an S87R mutation in ParC (**Figure 5Di).** However, we also identified
301 two mutations (G9S & I10V) in LrgB, a cytoplasmic membrane-associated murein hydrolase
302 transporter, as being associated with ciprofloxacin resistance. We found multiple known and
303 uncharacterized variants with *penA* associated with resistance to penicillin, ceftriaxone and
304 cefixime (**Supplementary Figure 4A, 4Ci & 4Cii).** For tetracycline resistance, our analysis
305 identified a variant in the known resistance gene, RpsJ (M57V), as well as multiple mutations in
306 *rsmB* which encodes a methylase that interacts with the target of tetracycline, the 16S portion of
307 the small ribosomal subunit (**Supplementary Figure 4B**). Finally, we identified multiple
308 significantly associated variants associated with azithromycin resistance in genes associated with
309 the MTR multi-drug efflux pump (*mtrC, mtrD, & mtrE*) (**Figure 5Dii**). Interestingly, we also
310 identified multiple variants in *mqo*, a malate-quinone oxidoreductase involved in the TCA cycle
311 to be significantly associated with azithromycin resistance. Azithromycin is not known to target
312 *mqo* or the TCA cycle, and pairwise structural alignments from AlphaFold predicted structures

313 using TM-align show *mqo* shares minimal similarity with *rplV* (TM-score 0.11, 7% identity) or
314 *rplD* (TM-score 0.17, 6% identity), known ribosomal targets of azithromycin, indicating if a direct
315 interaction exists it must be unique[40–43] (**Figure 5E)**. Together these findings reveal the potential
316 of genetic analyses to reveal novel AMR determinants within and beyond known AMR-associated
317 genes.
318
319 **DISCUSSION**
320 *Neisseria gonorrhoeae* are not generally considered to be a clonal species based upon classical
321 phylogenetic analyses[5]. Here, we have used PopNet to analyse the population structure based upon
322 localized patterns of SNPs to consider whether temporal, geographic, or clinical metadata related
323 to diversity. We also sought to determine whether AMR or other genetic determinants have a
324 previously unrecognized distribution. Our set of 485 representative genomes successfully captured
325 the genetic diversity of the larger set of 31,595 isolates, increasing the relative contribution of
326 isolates collected from outside North America and Europe. Ensuring representation across the
327 globe is critical if we are to understand the impact of global public health programs regarding
328 surveillance and treatment options, factors that are acutely elevated for a pathogen that is restricted
329 to a single host[2]. Applying the population genomics tool PopNet, we define 12 clades, which, with
330 the exceptions of clades B, K and L, contain isolates that are globally dispersed, suggesting other,
331 non-geographical, factors drive gonococcal population structure. While different, we did observe
332 a high degree of overlap between clades defined by other tools (cgMLST & PopPunk)[13,14],
333 supporting the utility of PopNet to define genetically distinct sub-populations.
334
335 Given the lack of overlap between the geographical location of isolation and clade membership,
336 we investigated how global population structure may be influenced by other factors. Most
337 identified clades were not significantly associated with other available metadata, including sex of
338 patient, site of isolation, year of isolation and country or continent of isolation. We then
339 investigated other factors not captured in isolate metadata: AMR and the presence of mobile
340 genetic elements. In keeping with previous findings, we observed PopNet-defined clades correlate
341 with a strong connection between the presence of specific mobile elements and AMR
342 determinants[44,45]. The association with mobile elements is particularly striking as PopNet
343 clustering only accounts for genetic variation in core chromosomal genes, revealing an association

344 between lineages and specific mobile elements. From these findings, it appears that AMR and
345 mobile elements are the primary drivers of the gonococcal population structure as opposed to
346 traditional metadata like geographic location or site of isolation. Clade B appears to be an
347 exception, as it is significantly associated with isolates recovered from Africa and Asia. Hoewver,
348 this association can be explained by the elevated presence of the β-lactamase plasmid and the TET-
349 containing conjugative plasmid in group B. This is in line with previous studies that have suggested
350 that isolates from lower and middle income countries are more likely to carry plasmids encoding
351 resistance to penicillin or tetracycline[46].
352
353 Phase variability is a key mechanism used by *Ngo* to evade the immune system by independently
354 switching on or off (or up-/down-regulating) protein expression of a variety of different genes[22].
355 While extensive work has been done to characterize these genes in reference strains, and/or by
356 identifying homologs in the closely related species, *N. meningitidis*, recent evidence has suggested
357 that the phase variable repertoire of *Ngo* could be much larger than previously thought[21,22]. In our
358 analysis, we identified 57 genes containing tandem repeats of varying lengths which represent
359 putative phase variable genes. To complement our population-based approaches, we performed an
360 *in vitro* and *in vivo* passaging experiment in an attempt to drive the emergence of phase variability
361 in the reference gonococcal strain MS11. Passaging of gonococci *in vitro* has previously been used
362 to identify phase variable genes[10,27], however this strategy likely selects for expression of factors
363 required for tissue colonization to phase "off". Our *in vivo* passaging in the lower genital tract of
364 CEACAM-humanized mice aimed to identify phase variable genes that help support *Ngo*'s ability
365 to infect and propagate within this mucosal niche. During the passaging experiment, we observed
366 improved colonization of gonococcal isolates with each successive passage, indicating adaptation
367 to the mouse host. Through this tandem approach, we observed several previously characterized
368 phase variable genes (*pilC*, *opa, modA, lgtG* & *fetA*) to phase vary during passaging, of which two
369 (group_1283, *opa* & group_1111, *lgtG*) varied in repeat number only in the *in vivo* passaged
370 isolates. We confirmed the change in repeat number within the Opa protein encoded by
371 group_1283 of *in vivo* passaged 2 and 3 isolates; since this turns on the expression of this adhesin,
372 it likely contributes to the observed improvement in colonization[47]. We also observed a shift in the
373 number of repeats in all *in vivo* passage 2 and 3 isolates, resulting in transcription of *lgtG*, with no
374 shift observed in the *in vitro* passaged strains. Expression of LgtG leads to the production of the

375 2C7 epitope, which is present in a high proportion of clinical isolates, indicative of a role in
376 gonococcal pathogenesis and colonization[48]. The selection for LgtG expression exclusively in all
377 *in vivo*-passaged isolates supports this conclusion, and demonstrates the utility of passaging and
378 sequencing for identifying genomic variation linked to improved gonococcal pathogenicity.
379
380 Given the genomic plasticity of *Ngo* and selection constraints imposed by the host immune system,
381 we hypothesized many *Ngo* genes may be subject to positive, diversifying selective pressure. We
382 identified 355 of 1,853 core genes to be under positive selection. This number dramatically
383 expands on the previous estimate of only 10 *Ngo* genes under positive selection, likely reflecting
384 differences in methodology as well as the larger number of genetically diverse isolates used here[49].
385 We found no functional enrichment for the positively selected genes, though we did observe a
386 significant enrichment in predicted outer membrane proteins. Beyond identification of novel genes
387 responding to host pressure, we also identified several novel sites within known AMR genes
388 subject to positive selection. These novel sites may represent complimentary variants, emerging
389 resistance-conferring variants or adaptive responses to the primary mutations and require further
390 exploration.
391
392 We also considered whether variants were associated with the carriage of mobile elements or AMR
393 through a genome-wide association analysis. We successfully identified known variants in
394 multiple AMR-associated genes, including *gyrA* and *parC* for ciprofloxacin, *penA* for penicillin,
395 ceftriaxone, and cefixime, *rpsJ* for tetracycline, and multiple *mtr* genes for azithromycin,
396 validating our approach. Enticingly, we additionally identified several highly significant variants
397 associated with AMR in genes with no known links to AMR. These include a variant in *lrgB*
398 associated with ciprofloxacin resistance; this protein has previously been indirectly linked to
399 ciprofloxacin resistance in *Ngo* and may play a role in the oxidative stress response[50]. Notably, we
400 also found multiple significant variants in *mqo*, a malate-quinone oxidoreductase associated with
401 azithromycin resistance. Despite appearing to have minimal interaction with known azithromycin
402 targets, studies testing the efficacy of azithromycin against the protozoan parasite *Plasmodium*
403 *falciparum* found reduced levels of TCA-associated metabolites, including malate, upon
404 administration of the drug[51]. It is possible that azithromycin exerts a similar effect in gonococci,
405 impacting its growth and that mutations in *mqo* may serve to alleviate this effect. Additional work

406 has been conducted in *P. falciparum* to validate *mqo* as a target for drug development efforts, and
407 our finding suggest similar efforts in *Ngo* may be pursuing[52].
408
409 In this study we curated a representative set of 485 genetically and globally diverse gonococcal
410 genomes from the 52,874 deposited gonococcal sequences on the NCBI SRA. Using PopNet, we
411 assessed the population structure of our representative set, and uncovered significant relationships
412 between population structure, geography, mobile elements, and AMR. This analysis suggests
413 global gonococcal genetic diversity is driven by the pressure applied by antimicrobials and links
414 the carriage of mobile elements with chromosomal variants. We used our representative set of
415 genomes to identify putative novel phase variable genes and were able to identify phase variation
416 of *opa* and *lgtG* potentially contributing to improved survival in the context of infection. Using a
417 combined approach of GWAS and tests for selective pressure, we identified new sites and variants
418 potentially associated with AMR in *Ngo*, demonstrating the utility of these large-scale analyses for
419 studying AMR in pathogen populations.
420

421 **MATERIALS & METHODS**

422 **Selection of sequences**

423 An initial set of 52,874 accessions of sequencing runs were obtained from the NCBI Sequence
424 Read Archive on September 26th, 2023 based on a keyword search for "*Neisseria gonorrhoeae*".
425 This set was reduced to 31,595 by keeping only those accessions containing paired-end Illumina
426 sequences with associated metadata for the country and year of isolation. Additional metadata
427 recorded when available included the site of isolation, associated BioProject, and sex of the patient.
428 All sequence data was obtained using the SRA Toolkit v3.0.0 using the default options for the
429 prefetch and fasterq-dump commands. Further metadata, such as minimum inhibitory
430 concentration values were obtained from pubMLST or publications associated with the relevant
431 BioProjects[53]. Kraken2 v2.1.3 was run to identify erroneously annotated or contaminated
432 sequences[54]. Any sequence datasets with less than 75% of its reads annotated by Kraken as being
433 derived from *Ngo* were discarded. Any sequences with between 75 and 85% gonococcal reads
434 were kept if over 95% of the reads were assigned to the genus *Neisseria.*
435

436 **Variant discovery for raw sequence data**

437 The reduced set of 31,595 isolates were aligned to the chromosomal genome of *Ngo* strain
438 TUM19854, the current recommended reference genome for gonococci from the NCBI (obtained
439 on October 13th, 2023). Alignment was performed using the Genome Analysis Toolkit (GATK)
440 v4.2.5.0[55]. Quality control and trimming was performed using fastp v0.23.4 with minimum
441 qualified phred quality of 20, a 40 percent limit on unqualified bases, a mean quality requirement
442 of 20, a cut window size of 5, both front and tail cutting, and with correction enabled[56]. Trimmed
443 reads were then passed to GATK for alignment using bwa v0.7.17 with default settings. The GATK
444 Haplotype caller was then run with default settings to identify variants. Dependencies for GATK
445 included samtools v1.17 and picard v2.26.3. Variants were then filtered using bcftools v1.11 and
446 htslib v1.11 with the following parameters: allelic depth 8, allelic depth forward/reverse 3,
447 minimum base quality 50, minimum mapping quality score 20 and percent of reads containing
448 variant 80[57].
449
450 **Generation of a representative set of sequences**
451 Isolates were grouped by their country of isolation and year of isolation and the Jaccard statistic
452 was calculated for each pair of sequences using the bedtools v2.31.0 jaccard function with default
453 settings[58]. An initial clustering step was performed using the k-medoids fastPAM algorithm
454 implemented in the R packages cluster v2.1.5 and factoextra v1.0.7[59–61]. For each group, a value
455 for k was determined using the silhouette method, with the maximum value of k capped at 15. The
456 identified medoids of the clusters were selected as representative isolates. This resulted in an initial
457 set of 1,307 isolates. Isolates from this set were then grouped by their continent and year of
458 isolation, and k-medoids clustering was performed as before without a cap on the value of k.
459 Strains identified as disseminated or ocular as well as the WHO global reference collection were
460 added to the final set, bringing the total number of isolates to 485[10,11].
461
462 **PopNet analysis and phylogenetic tree construction**
463 PopNet was used to cluster the reference set of genomes using the variants identified from the
464 previous analysis[9]. A variant table was generated using the GATK VariantsToTable function from
465 a merged vcf file of identified variants. The parameters used for the PopNet analysis were
466 section_length= 5000, S1_iVal = 8, S1_piVal = 19, S2_iVal= 4.5, and S2_piVal= 2.5. These
467 parameters were determined using the silhouette index readout provided by the tool. Initial

468 phylogenetic tree reconstruction was performed with IQTree v2.2.2.7 from a sequence alignment
469 created from the previously identified variants[62]. To begin, Model Finder Plus was run which
470 identified TPM2u+F+I+R7 as the best substitution model[63]. IQTree was then run using this model
471 with 1000 ultrafast bootstrap replicates[64]. All other parameters were left as default. This initial tree
472 was used as a basis for ClonalFrameML v 1.12, which was run with the standard model with 100
473 pseudo-bootstrap replicates[12]. Assembled genomes were uploaded to pubMLST and automatically
474 assigned to a cgMLST Ng_300 clade.
475
476 **Pangenome analysis**
477 Trimmed reads were assembled *de novo* using unicycler v0.5.0 with default settings[15,65]. Gene
478 prediction and initial annotation was performed using Prokka v1.14.5 with recommended settings
479 and an added protein database containing annotated proteins from *Ngo* strains MS11, FA19,
480 FA1090 obtained from the NCBI on March 7th, 2022[16]. The pangenome was calculated using
481 panaroo v1.3.4 on the strict setting with otherwise default options[17]. Representative gene
482 sequences from panaroo were supplied to eggNOG-mapper v2.19 with no restrictions on orthology
483 identification and permitting electronic GO term annotation[66]. AMR genes were predicted from
484 sequences uploaded to PathogenWatch on January 22nd, 2024[19]. The degree of resistance (resistant,
485 intermediate, susceptible) was classified based on the EUCAST clinical breakpoints v9.0[67].
486 Mobile elements were predicted by BLAST+ v2.0.13 against reference sequences AY803022.1,
487 NC_001377, NC_002098, GU479464.1, GU479465, and GU479466 which were obtained from
488 the NCBI on December 16th, 2023[68]. Accessory genes were assigned to mobile elements based on
489 a nucleotide BLAST[68] search against this database with a minimum sequence identity of 85% and
490 an e-value less than $1x10^{-6}$. The Graphia v5.0 MCL clustering tool was used to cluster accessory
491 genes based on shared presence or absence in our representative collection with a granularity of
492 1.5[18]. Structures were predicted based on consensus protein sequences using AlphaFold4, and
493 structural homology searches were performed with FoldSeek release 10[69,70]. Predicted structures
494 were visualized with UCSF ChimeraX v1.9[71].
495
496 **Prediction of phase variable genes**
497 Simple sequence repeats were identified from the *de novo* assemblies and the GFF files generated
498 by Prokka using PERF v0.4.6[20]. Genes with repeats that varied in the number of tandem repeats in

499 more than 2 isolates where each repeat length was present in at least 5% of genomes were
500 considered as potentially phase variable. Mutations classified as intergenic were manually verified
501 to ensure they impacted the associated gene.
502
503 **Lower genital tract colonization for in vivo passaging**
504 Bacterial growth conditions: *Ngo* MS11, parental strain, was grown on GC agar (Becton,
505 Dickinson and Company, Sparks, USA) plates supplemented with 1% Isovitalex (ISO, Becton,
506 Dickinson and Company, Sparks, USA) at 37˚C with 5% CO_2 for 16-20 hours. For *Ngo* isolates
507 used for in vivo passaging, agar plates included a vancomycin, colistin, nystatin, and
508 trimethoprim antibiotic mix (VCNT, Becton, Dickinson and Company, Sparks, USA) for *Ngo*
509 selection.
510
511 Mouse strains: 6-8 week old female FVB mice containing a human chromosome-derived BAC
512 expressing CEACAM3, -5 and -6, were used for all infection experiments[25]. Animals were housed
513 in cages with sterile rodent chow, sterile water, enrichment material, and 12-hour light-dark cycles.
514
515 Gonococcal *in vivo* passaging: Mice were staged by cytological vaginal smears observed under
516 light microscopy. Those in the diestrus stage of the estrous cycle were started on an antibiotic and
517 hormone regimen to control impact of commensal microflora on infection. Antibiotics consisted
518 of 0.6 mg vancomycin hydrochloride (Bioshop, Burlington, ON, Canada) and 2.4 mg streptomycin
519 sulfate (Sigma-Aldrich, Oakville, Canada) resuspended in 200 µL of PBS administered by
520 intraperitoneal injection daily from 2 days pre-infection (-2) to 5 days post infection (+5), where
521 an dose was provided on Day -1. Trimethoprim (Sigma-Aldrich, Oakville, Canada) was dissolved
522 in the drinking water at 40 mg/mL for ad libitum consumption. Water soluble β-estradiol
523 (SigmaAldrich, Oakville, Canada) was constituted in PBS (0.5 mg/200 µL), filter-sterilized and
524 administered via subcutaneous injection on Days -2, 0 and +2.
525
526 On Day 0, bacteria were resuspended in Dulbecco's phosphate-buffered saline (DPBS) with
527 calcium and magnesium (PBS++ , Life Technologies, Burlington, Canada)and mice were
528 inoculated with 10^7 bacteria in 2.5 µL Vaginal lavages were performed every two days by washing
529 the genital tract with 10 µL of PBS++ and plated for bacterial recovery and enumeration. After

530 ~20 hrs bacterial isolates were frozen in 50% brain heart infusion (BHI, Becton, Dickinson and
531 Company, Sparks, USA) and 25% glycerol (EMD Millipore, Darmstadt, Germany) and stored at
532 -80˚C. Clearance of infection was characterized by 2 consecutive vaginal lavages that tested
533 negative for *Ngo*.
534
535 Gonococcal *in vitro* passaging: Five culture tubes containing 5 mL of BHI supplemented with 1%
536 ISO was inoculated with the parental strain with a starting optical density at 550 nm (OD550) of
537 0.05. Bacterial cultures were placed in a 37˚C shaker at 180 rpm. Every 12 hrs, bacterial stocks
538 were generated and sub-cultured for 5 consecutive days.
539
540 *Ngo* whole genome sequencing: *Ngo* isolates were resuspended in 200 µL of PBS++ and genomic
541 DNA was extracted by a QIAGEN DNeasy Blood and Tissue Kit (QIAGEN, Hilden, Germany)
542 following manufacturer instructions. DNA quality and quantity was checked by NanoDrop ND-
543 1000 Spectrophotometer (NanoDrop Technologies, Wilmington, USA) and Quant-iT™
544 PicoGreen® dsDNA Assay Kit (Life Technologies, Oregon, USA) was used to quantify amount
545 of dsDNA.Genomic DNA was prepared for whole genome sequencing by a Nextera XT library
546 preparation kit (Illumina, San Diego, USA). Libraries were run on a 1% agarose gel to ensure
547 libraries were of the correct length prior to poolingand submission to the Donnelly Sequencing
548 Centre, University of Toronto for sequencing using the Illumina MiSeq platform.
549
550 **Capillary electrophoresis for confirmation of *opa* repeat lengths**
551 Capillary electrophoresis was performed as described previously for *Neisseria meningitidis*[72].
552 Briefly, strains were cultured overnight on GC agar as described above, and 3 distinct colonies
553 per plate were selected and added to 20µl of Lucigen QuickExtract buffer (Mandel, Canada).
554 Suspended colonies were then heated at 66°C for 6 minutes, then 98°C for 2 minutes to complete
555 the extraction. Extracted DNA was amplified with PCR using the primers in **Table 2** listed
556 below, with the forward primer conjugated to a 6-FAM dye. Samples were then diluted to
557 approximately 1-2ng/µl before being sent for capillary electrophoresis. The results were
558 analyzed with the ThermoFisher MicroSatellite Analysis tool.
559
560 **Identification of genes under positive selection**

561 Aligned gene sequences from panaroo for core genes with 5 or more unique sequences were used
562 as input for selection analyses after removing identical sequences and sequences containing
563 premature stop codons using Hyphy rmv[27]. First, a tree for each gene was built using IQTree
564 v2.2.2.7 with 1000 ultrafast bootstraps and otherwise default settings[62]. The BUSTED algorithm
565 was then run as implemented in HyPhy with default settings for each gene using the generated tree
566 and panaroo alignment[28]. Genes identified as being under positive selection were then analyzed
567 with FUBAR as implemented in HyPhy with default settings using the same inputs to identify
568 specific sites under positive selection[29]. To assess the location of predicted residues under positive
569 selection, structural models were generated with AlphaFold3 using the AlphaFold server based on
570 the consensus gene sequences from panaroo[17,69]. Predicted structures were visualized with UCSF
571 ChimeraX v1.9[71].
572
573 **Genome wide association study**
574 The genome wide association study was performed using pyseer with the linear mixed model
575 setting using the VCF file generated as described above[39]. The distance matrix was calculated
576 using the ClonalFrameML generated tree described above. All other pyseer settings were kept at
577 default. The count_patterns script was used to establish a significance threshold for each tested
578 phenotype. The impact of each variant was predicted with SnpEff using the *Ngo* TUM19854
579 reference genome obtained as previously described[73]. Structural alignments and comparisons
580 were performed using TM-align with default settings[40].
581
582 **ACKNOWLEDGEMENTS**
583 We are grateful to Odile Harrison, University of Oxford, for helpful advice and technical
584 support, and the Division of Comparative Medicine & Donnelly Sequencing Centre at the
585 University of Toronto for technical support. This study was supported by operating grants from
586 the Canadian Institutes for Health Research (PJT-159677) and National Institutes of Health
587 (1R01AI146941-01A1). S.D.G. is supported by the Canada Research Chairs Program. D.C. is a
588 recipient of the EPIC Doctoral Award and the DSI Doctoral Fellowship.
589
590 **ETHICS STATEMENT**

591 All animal procedures conducted in this study were approved by the Local Animal Care
592 Committee (LACC) at the University of Toronto (Protocol #20011775), which is in compliance
593 with ethical and legal requirements under Ontario's Animals for Research Act and the federal
594 Canadian Council on Animal Care.
595

596 **DATA & CODE AVAILABILITY**

597 The data sets generated in this study are provided in the Supplementary Information and Source
598 Data files. Codes for analysis and data are available on Github
599 [https://github.com/ParkinsonLab/Neisseria_work.git].
600

601 **AUTHOR CONTRIBUTIONS**

602 J.P. and S.D.G conceived the idea of the research and supervised it. D.C and E.J. performed the
603 computational and data analysis. D.C. and J.M. conducted the mouse passaging and phase
604 variation experiments and analyzed the data. D.C. wrote the manuscript with contributions from
605 J.P., S.G.D. & J.L.
606

607 **REFERENCES**

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

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

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

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

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

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

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

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

626 9. Zhang, J., Khan, A., Kennard, A., Grigg, M. E. & Parkinson, J. PopNet: A Markov
627 Clustering Approach to Study Population Genetic Structure. *Mol. Biol. Evol.* **34**, 1799–1811
628 (2017).

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

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

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

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

640 14. Harrison, O. B. *et al.* Neisseria gonorrhoeae Population Genomics: Use of the
641 Gonococcal Core Genome to Improve Surveillance of Antimicrobial Resistance. *J. Infect. Dis.*
642 **222**, 1816–1825 (2020).

643 15. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial
644 genome assemblies from short and long sequencing reads. *PLOS Comput. Biol.* **13**, e1005595
645 (2017).

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

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

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

652 19. Sánchez-Busó, L. *et al.* A community-driven resource for genomic epidemiology and
653 antimicrobial resistance prediction of Neisseria gonorrhoeae at Pathogenwatch. *Genome Med.*
654 **13**, 61 (2021).

655 20. Avvaru, A. K., Sowpati, D. T. & Mishra, R. K. PERF: an exhaustive algorithm for ultra-
656 fast and efficient identification of microsatellites from large DNA sequences. *Bioinformatics* **34**,
657 943–948 (2018).

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

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

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

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

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

670 26. Erwin, A. L., Haynes, P. A., Rice, P. A. & Gotschlich, E. C. Conservation of the
671 lipooligosaccharide synthesis locus lgt among strains of Neisseria gonorrhoeae: requirement for
672 lgtE in synthesis of the 2C7 epitope and of the beta chain of strain 15253. *J. Exp. Med.* **184**,
673 1233–1241 (1996).

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

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

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

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

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

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

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

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

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

697 36. Fenton, B. A. *et al.* Mutations in PBP2 from ceftriaxone-resistant Neisseria gonorrhoeae
698 alter the dynamics of the β3–β4 loop to favor a low-affinity drug-binding state. *J. Biol. Chem.*
699 **297**, 101188 (2021).

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

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

706 39. Lees, J. A., Galardini, M., Bentley, S. D., Weiser, J. N. & Corander, J. pyseer: a
707 comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics* **34**, 4310–
708 4312 (2018).

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

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

714 42. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**,
715 583–589 (2021).

716 43. Ng, L.-K., Martin, I., Liu, G. & Bryden, L. Mutation in 23S rRNA Associated with
717 Macrolide Resistance in Neisseria gonorrhoeae. *Antimicrob. Agents Chemother.* **46**, 3020–3025
718 (2002).

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

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

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

728 47. Kupsch, E. M., Knepper, B., Kuroki, T., Heuer, I. & F.meyer, T. Variable opacity (Opa)
729 outer membrane proteins account for the cell tropisms displayed by Neisseria gonorrhoeae for
730 human leukocytes and epithelial cells. *EMBO J.* **12**, 641–641 (1993).

731 48. Gulati, S., McQuillen, D. P., Mandrell, R. E., Jani, D. B. & Rice, P. A. Immunogenicity
732 of Neisseria gonorrhoeae Lipooligosaccharide Epitope 2C7, Widely Expressed In Vivo with No
733 Immunochemical Similarity to Human Glycosphingolipids. *J. Infect. Dis.* **174**, 1223–1237
734 (1996).

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

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

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

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

744 53. Jolley, K. A., Bray, J. E. & Maiden, M. C. J. Open-access bacterial population genomics:
745 BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res.* **3**, 124
746 (2018).

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

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

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

753 57. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping
754 and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–
755 2993 (2011).

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

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

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

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

764 62. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic
765 Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

766 63. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S.
767 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–
768 589 (2017).

769 64. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2:
770 Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).

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

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

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

778 68. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421
779 (2009).

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

782 70. Fast and accurate protein structure search with Foldseek | Nature Biotechnology.
783 https://www.nature.com/articles/s41587-023-01773-0.

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

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

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

791

792 **TABLES**

793 **Table 1: Predicted new phase variable genes.**

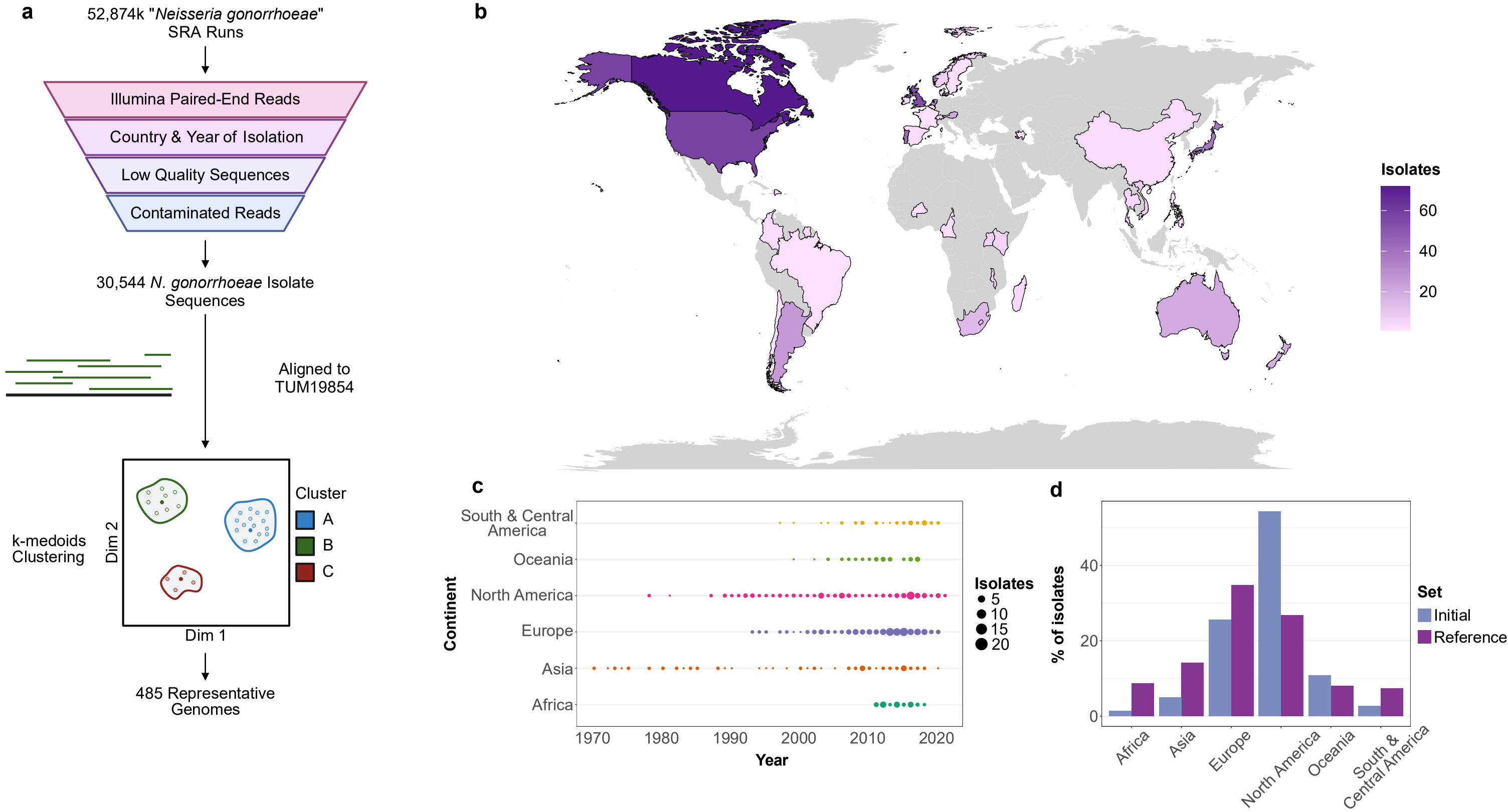
Gene	Repeat	No. of repeats	Annotation	NGK locus	NGO locus
group_1584	A	33	Pertactin	NGK_2016	NGO_1158
group_280	AACGC	24	Hypothetical protein	-	NGO_0366
argS	AAGC	18	Arginyl-tRNA synthetase	NGK_0832	NGO_0963
group_1317	AACGC	17	C-5 cytosine-specific DNA methylase	NGK_0522	NGO_0365
group_1449	AT	11	Spermidine 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	-

794

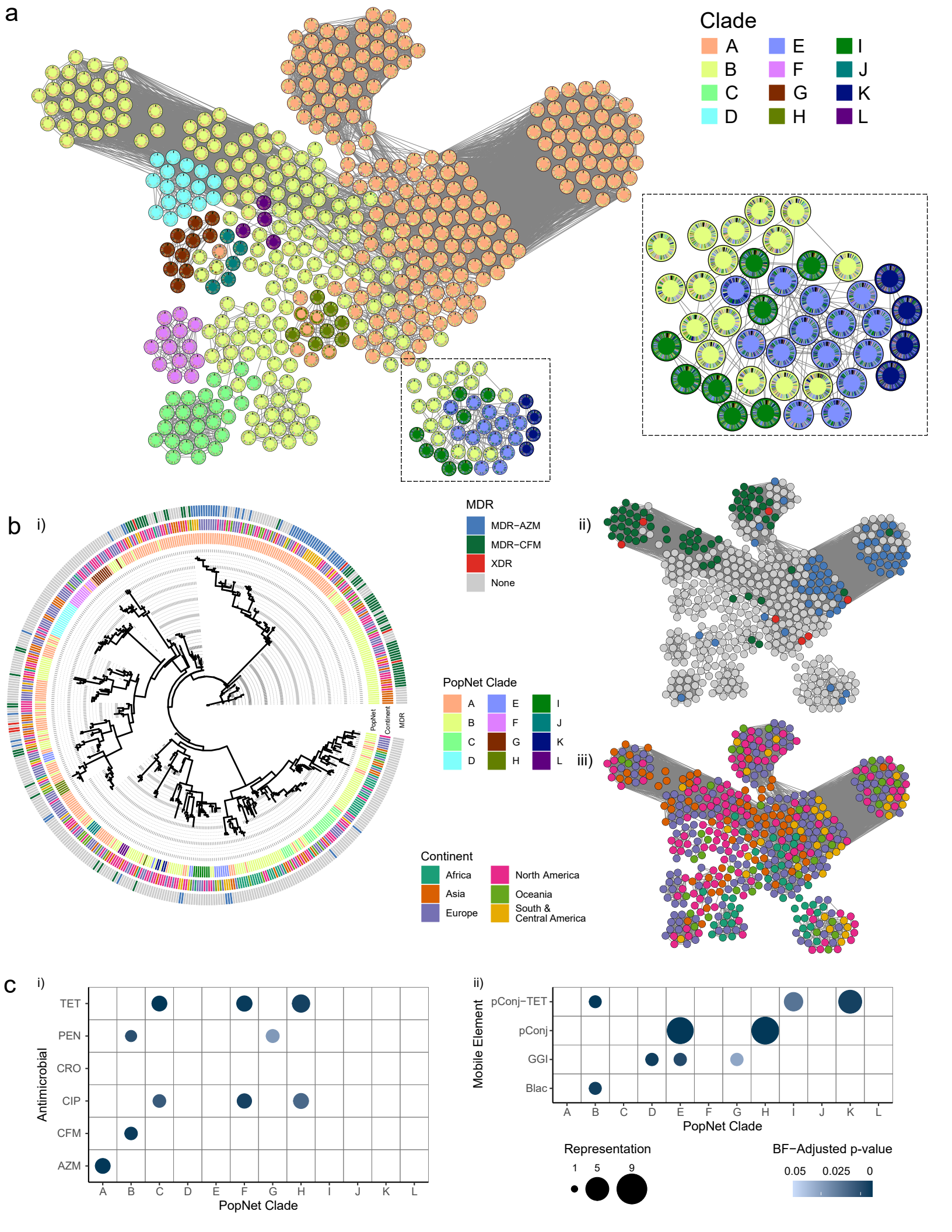
798 **Table 2: Primers used for capillary electrophoresis.**

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

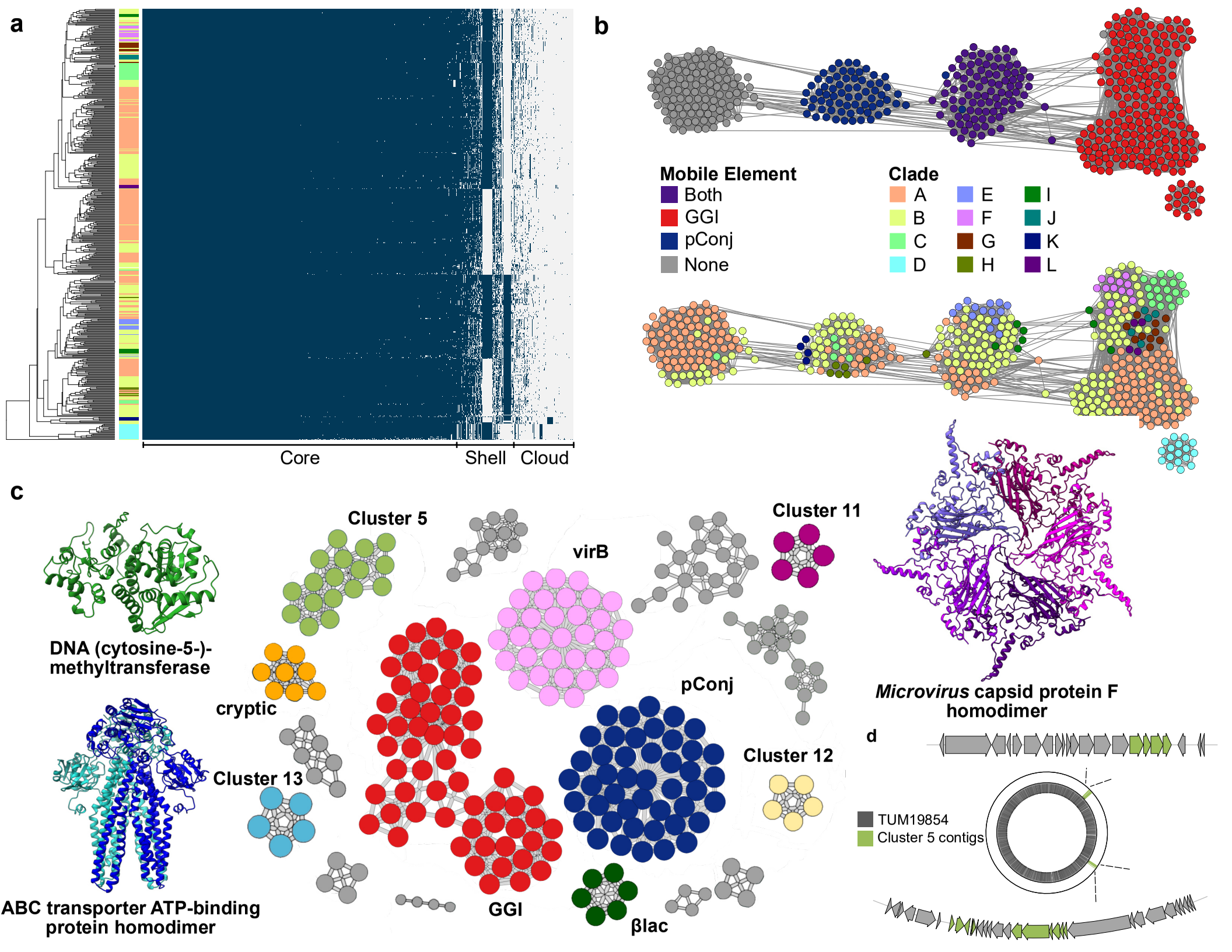
799


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

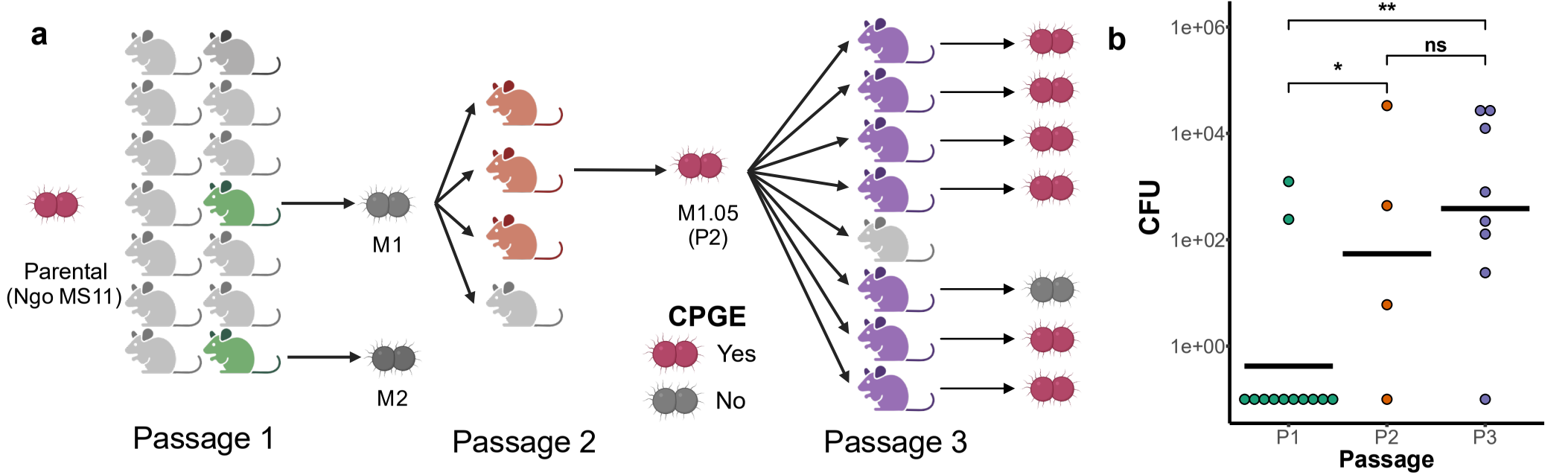
800 **Figure 1: Overview of representative set of 485 *N. gonorrhoeae* genomes.**
801 **a** Schematic of selection of representative genomes. **b** Geographic locations of isolates in the
802 representative set. **c** Geographic and temporal distribution of isolates in the representative set. **d**
803 Comparison of the geographic distribution of isolates between the initial set of 52,784 SRA runs
804 and the representative set of 485.


a
Clade
A
B
C
D
E
F
G
H
I
J
K
L
b
i)
PopNet
Continent
MDR
MDR
MDR−AZM
MDR−CFM
XDR
None
PopNet Clade
A
B
C
D
E
F
G
H
I
J
K
L
Continent
Africa
Asia
Europe
North America
Oceania
South & Central America
ii)
iii)
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

805 **Figure 2: Population structure of *N. gonorrhoeae* is influenced by AMR and the presence of**
806 **mobile elements.**
807 **a** PopNet network view of 485 representative *Ngo* genomes. Nodes represent genomes, coloured
808 by their assigned clades. Annuli show the chromosome painting generated for each genome, with
809 regions sharing a colour showing overlap between genomes. A selection of genomes from clades
810 B, E, I & K are shown to illustrate the chromosome painting visualization. **b** ClonalFrameML
811 generate phylogenetic tree of representative *Ngo* genomes. Outer rings show the assigned
812 PopNet clade, MDR (multi-drug-resistant) status and continent of isolation for each genome. **i**:
813 MDR status overlaid on the PopNet network view. **ii**: Continent of isolation overlaid on the
814 PopNet network view. AZM: azithromycin, CFM: cefixime, XDR: extreme drug resistant. **c**
815 Association between antimicrobial resistance (left, i) or mobile elements (right, ii) and PopNet
816 clades. Only significant associations with a p-value < 0.05 as determined by Fisher's exact test
817 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
DNA (cytosine-5-)-methyltransferase
ABC transporter ATP-binding protein homodimer
Cluster 5
virB
Cluster 11
cryptic
pConj
Cluster 13
Cluster 12
GGI
βlac
Microvirus capsid protein F homodimer
d
TUM19854
Cluster 5 contigs

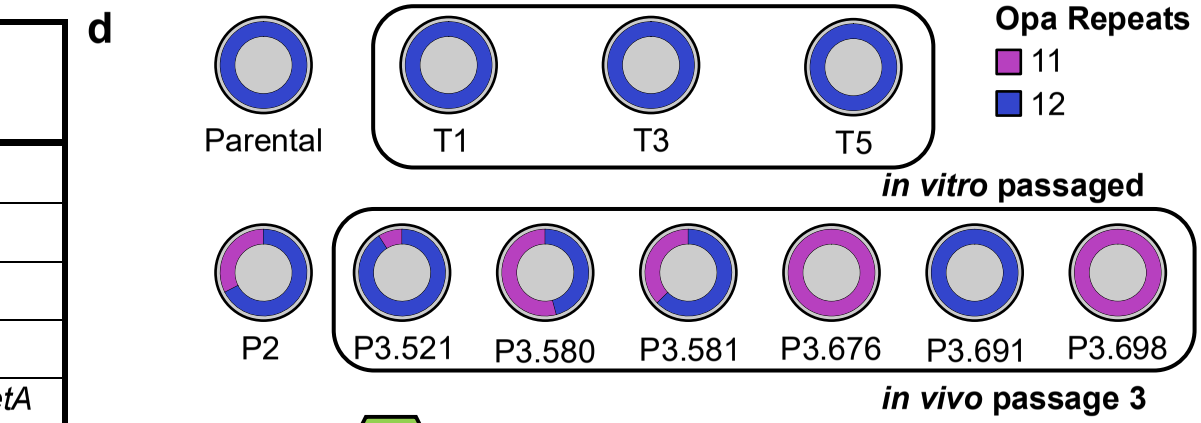
818 **Figure 3: The pangenome of *N. gonorrhoeae* is dominated by large mobile elements, which**
819 **are associated with PopNet clades determined by chromosomal variants.**
820 **a** Pangenome of *Ngo* is composed of 2460 genes, with 1800 Core (present in > 95% of
821 genomes), 367 Shell (present in between 15% and 95% of genomes) & 283 Cloud (present in
822 <15% of genomes) genes. Dendrogram and ordering of genomes on the y-isolates calculated
823 using complete linkage from a Euclidian distance matrix generated based on shared gene content.
824 **b** Top: Network view of isolates clustered by shared gene content coloured by the presence or
825 absence of 2 large mobile elements: the GGI (gonococcal genetic island) and the pConj
826 (conjugative plasmid). Bottom: Genomes coloured by PopNet clade overlaid on the network
827 view of isolates clustered by shared gene content. **c** Clustering of genes based on their co-
828 occurrence in genomes. Clusters defined by MCL clustering as implemented in Graphia which
829 are present in <50% of genomes and composed of 5 or more genes are coloured. Predicted
830 structures from AlphaFold3. **d** Organization of genes from MCL cluster 11 associated with clade
831 D.

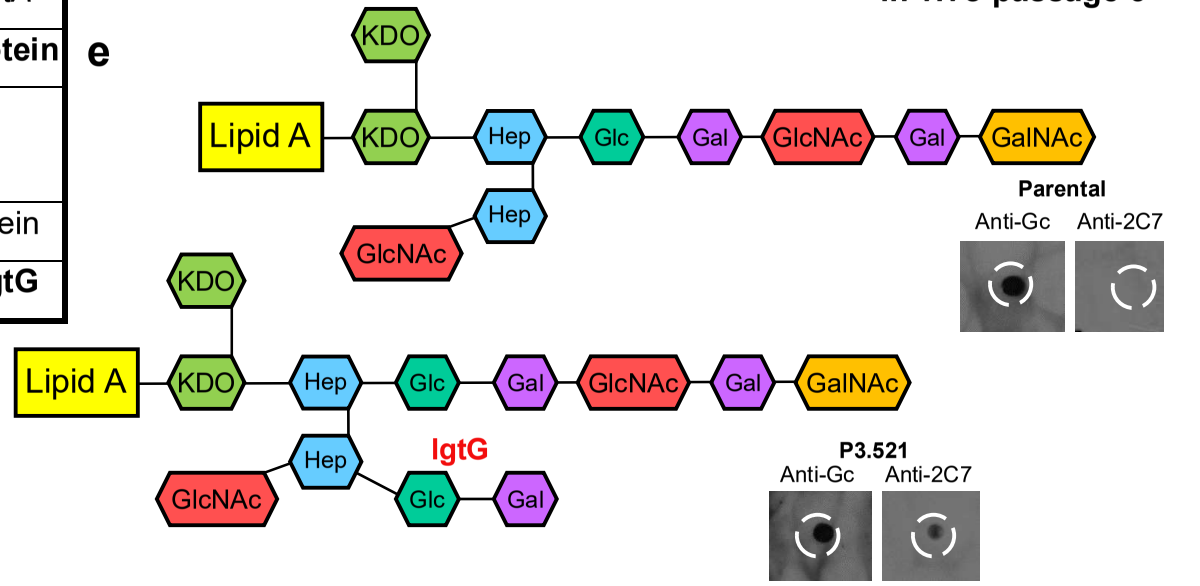


c

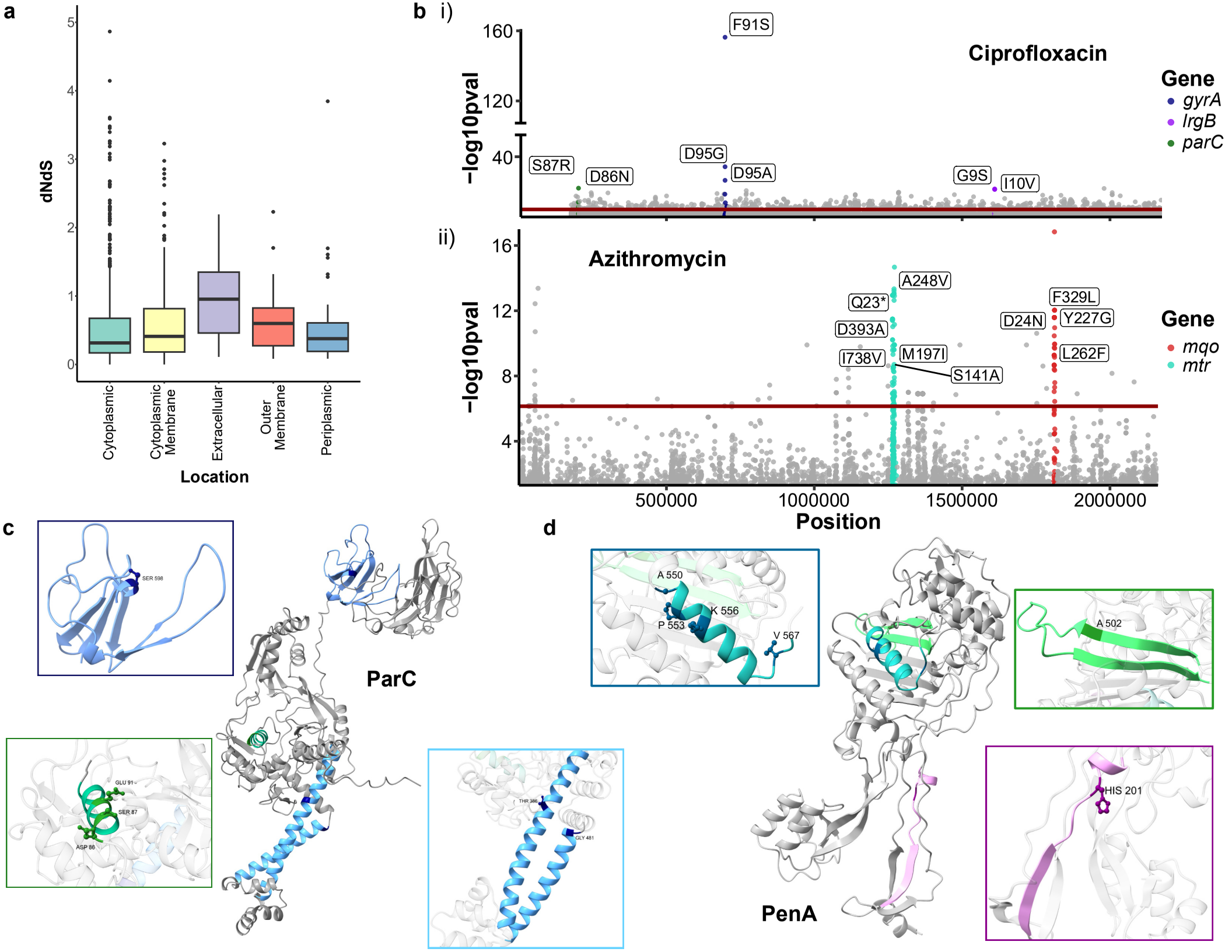
Gene	Repeat	#of repeats	Annotation
group_1059	C	3	Pilus assembly protein
group_1287	C	2	Pilus assembly protein
group_399	CCTCT	2	Opacity protein
group_579	CCTCT	2	Opacity protein
group_784	C	3	Enterobactin receptor *fetA*
group_1283	**CCTCT**	**3**	**Opacity-associated protein**
group_973	AGCC	2	Adenine specific DNA methylase
group_1471	CCTCT	2	Opacity-associated protein
group_1111	**C**	**2**	**Glycosyl transferase lgtG**

*Bolded genes varied between *in vitro* & *in vivo* passages





832 **Figure 4: Passaging of *N. gonorrhoeae* improves colonization in a mouse model of infection**
833 **which may be associated with phase variation of *opa*.**
834 **a** Schematic overview of *in vivo* and *in vitro* passaging experiment. Coloured mice represent
835 mice successfully colonized by *Ngo* 2 days post infection. Coloured gonococci indicate isolates
836 sequenced by capillary electrophoresis to confirm the repeat lengths of *opa*. **b** Recovered CFU
837 (colony-forming units) of *Ngo* recovered 2 days post infection. Black bars represent the mean
838 CFUs recovered for each passage. **c** Predicted phase variable genes from mouse passaging
839 experiment. **d** Ratio of *opa* repeats in each isolate sequenced by capillary electrophoresis. **e**
840 Schematic of *Ngo* lipooligosaccharide with (bottom) and without (top) lgtG expression, resulting
841 in the production of the 2C7 epitope. Isolates passaged *in vivo* express lgtG, while no expression
842 is observed in isolates passaged *in vitro.* Immunoblot assay for the 2C7 isotope confirms the
843 expression of lgtG in a passage 3 isolate, which is absent from the parental isolate. KDO: 3-
844 Deoxy-D-manno-oct-2-ulosonic. Hep: Heptose. Glc: Glucose. Gal: Galactose. GlcNAc: N-
845 acetylglucosamine. GalNAc: N-Acetylgalactosamine.


a
dNdS
Cytoplasmic
Cytoplasmic Membrane
Extracellular
Outer Membrane
Periplasmic
Location
b i)
Ciprofloxacin
−log10pval
F91S
S87R
D86N
D95G
D95A
G9S
I10V
Gene
gyrA
lrgB
parC
ii)
Azithromycin
A248V
Q23*
D393A
I738V
M197I
S141A
D24N
F329L
Y227G
L262F
Gene
mqo
mtr
Position
c
SER 598
ParC
GLU 91
SER 87
ASP 86
THR 386
GLY 481
d
A 550
K 556
P 553
V 567
A 502
PenA
HIS 201

846 **Figure 5: GWAS and tests for positive selection identify sites linked to AMR in *N.***
847 ***gonorrhoeae*.**
848 **a** Boxplot of dN/dS ratio for positive selection of *Ngo* genes by predicted cellular localization. **b**
849 Manhattan plots of variants with significant associations with ciprofloxacin (top) and
850 azithromycin (bottom) resistance. Only variants with an LRT p-value <0.05 are shown. Red lines
851 indicate significance thresholds determined by Bonferroni correction with the number of unique
852 variant patterns as the number of multiple tests. **c** Predicted residues with strong signatures of
853 positive selection in *parC*. Residues highlighted in green are under positive selection and known
854 to contribute to ciprofloxacin resistance. Residues in blue are under positive selection with no
855 known links to AMR. Structural prediction from AlphaFold3 visualized with USCF ChimeraX. **d**
856 Predicted residues with strong signatures of positive selection in *penA*. Highlighted residues are
857 under positive selection with no known links to AMR. Structural prediction from Alphafold3
858 visualized with USCF ChimeraX.

859 **SUPPLEMENTARY INFORMATION**

860 **Supplementary Figure S1: Overlap of clustering from PopNet, cgMLST & PopPunk.** Tree
861 (left) from ClonalFrameMLST overlaid with PopNet clusters, 11 largest clusters from cgMLST
862 and PopPunk. Network view (right) generated by PopNet overlaid with 11 largest clusters from
863 cgMLST and PopPunk.

864

865 **Supplementary Figure S2: Predicted AMR from PathogenWatch.** Predictions based on
866 assembled genomes using PathogenWatch. Degree of resistance based on EUCAST clinical
867 breakpoints v9.0.

868

869 **Supplementary Figure S3: Colonization of female transgenic FVB mice by *in vivo* passaged**
870 ***Ngo*.** Mice were checked every 2 days after 1-day post-infection for evidence of colonization.
871 Mice contained a human chromosome-derived BAC expressing CEACAM3, -5 and -6.

872

873 **Supplementary Figure S4: Variants associated with AMR identified through GWAS.**
874 Manhattan plots of variants with significant associations with penicillin (A), tetracycline (B),
875 ceftriaxone (Ci), and cefixime (Cii) resistance. Only variants with an LRT p-value <0.05 are
876 shown. Red lines indicate significance thresholds determined by Bonferroni correction with the
877 number of unique variant patterns as the number of multiple tests.

878

879 **Supplementary Table S1: SRA metadata from initial set of 31,595 gonococcal genomes.**

880

881 **Supplementary Table S2: Isolate metadata for representative set of 485 genomes.**

882

883 **Supplementary Table S3: Metadata for all genes identified within the gonococcal**
884 **pangenome.**

885

886 **Supplementary Table S4: Clusters of accessory genes determined through Graphia MCL.**

887

888 **Supplementary Table S5: Phase variable genes identified through genetic screening of**
889 **representative genomes.**

890

891 **Supplementary Table S6: Evidence for signatures of positive selection at the gene and**
892 **residue level.**

893

894 **Supplementary Table S7: Variants significantly associated with AMR identified through**
895 **GWAS.**