The Journal of Infectious Diseases

MAJOR ARTICLE

Publicly Available *Neisseria Gonorrhoeae* Genomes Predominantly Represent In Vitro-Derived Nonpiliated Variants

Iryna Boiko,[a] Selma Metaane,[a] and H. Steven Seifert

Department of Microbiology-Immunology, Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA

Background. The *Neisseria gonorrhoeae pilE* gene encodes the PilE protein, the major subunit of the Type IV pilus and a primary colonization and virulence factor. The *pilE* gene undergoes high-frequency diversification mainly through gene conversion from one of many *pilS* copies. These unique molecular processes contribute to gonococcal population diversity, facilitating immune evasion. While the process of pilin variation is understood, the diversity of *pilE* and *pilS* genes from clinical isolates is understudied.

Methods. We analyzed 15 186 *N. gonorrhoeae* genomes, including finished ($n = 65$) and draft ($n = 15\,121$) genomes, in the PubMLST database to characterize *pilE* and *pilS* gene diversity.

Results. The finished genomes had one to nine *pilS* loci at conserved chromosomal locations. Only 52.13% of sequences contained a *pilE* gene, despite all genomes having other Type IV pilus genes. When the *pilE* was present, most defined conserved sequences were preserved. However, most predicted PilE protein sequences contained premature stop codons, which were found in several silent copies.

Conclusions. All *N. gonorrhoeae* strains possess the genes necessary for pilin AV; however, most genomic sequences were derived from nonpiliated variants that emerged during in vitro culture through reversible pilus phase variation and irreversible deletion of the *pilE* gene.

Keywords. *pilE*; *pilS*; antigenic variation; phase variation; *Neisseria gonorrhoeae*.

Neisseria gonorrhoeae (the gonococcus or Gc) is an obligate human pathogen causing gonorrhea, one of the most prevalent sexually transmitted infections [1, 2]. In 2020, over 82 million cases of gonococcal infection were estimated worldwide [3]. Gc colonizes urethral, pharyngeal, and rectal epithelium, causing reproductive complications, disseminated gonococcal infection, and neonatal blindness if untreated [2]. Gonococcal infections are often asymptomatic and serve as a reservoir for transmission [4, 5], underscoring the complex host-pathogen interactions [4].

Gc colonization and pathogenesis require the Type IV pilus [2, 6, 7]. The major pilus subunit, PilE protein, is encoded by the *pilE* gene [8], expressed from a single locus with a conserved promoter [9]. The *pilE* gene's open reading frame (ORF) begins with a conserved seven-amino-acid leader sequence cleaved post-translationally to form mature pilin [10, 11]. Three amino acids in the conserved part of PilE, at positions 38–40, form the *S*-cleavage site [12] responsible for maintaining Gc transformation competence [13]. Two upstream unique *pilE*-associated structures, the Guanine-quadruplex (G4) and transcription of the associated sRNA promoter (*garP*) are conserved and necessary for pilin antigenic variation (AV) [14, 15].

Gc possesses five to six silent loci (*pilS*) located both upstream and downstream of the *pilE* locus, each with one to six silent copies [16–18]. The silent loci lack promoters, ribosome binding sites, leader sequences, and the conserved N-terminal coding sequences. Both *pilE* and *pilS* loci contain conserved sequences, including *cys1*, *cys2,* and the *SmaCla* repeat that is 3′ of *pilE* and each *pilS* locus [16, 19]. Pilin AV occurs through gene conversion reactions where variant sequences transfer unidirectionally from one of the *pilS* copies to the *pilE* gene, resulting in a varied *pilE* sequence [16, 20].

Received 07 July 2025; accepted 24 October 2025; published online 10 December 2025

[a]These authors contributed equally to this work: Iryna Boiko and Selma Metaane.

Conferences: This study was previously presented at the 3rd Annual Pathogen Genomics Symposium on April 23–24, 2024, at Northwestern University, Chicago, IL, USA, and at the 3rd *Neisseria gonorrhoeae* Research Society Virtual Conference held June 3–6, 2024.

Correspondence: H. Steven Seifert, PhD, Department of Microbiology-Immunology, Northwestern University Feinberg School of Medicine, Searle 3-402, 303 East Chicago Avenue, Chicago, IL 60611 (h-seifert@northwestern.edu).

The Journal of Infectious Diseases®

https://doi.org/10.1093/infdis/jiaf557

In addition to AV capacity, the Gc strains harbor one to two *pilC* genes (*pilC1* and *pilC2*), which encode the pilus adhesin PilC [21, 22]. The ORFs of both *pilC* genes contain a poly-G tract within their signal peptide sequences, resulting in frameshift mutations and phase-variable PilC expression [21]. As a result, pilus assembly is affected, and Gc turns to an underpiliated or non-piliated state without genetic changes in the *pilE* locus [23]. Pilin diversification has the highest reported frequency among pathogenic gene conversion systems [24], enhancing Gc diversity and immune evasion [2, 17].

Gc recovered directly from infected patients are almost always piliated, supporting the critical role of pili during infection [25–28]. However, during in vitro propagation, nonpiliated variants rapidly arise, and since nonpiliated variants grow faster than their piliated progenitors, nonpiliated variants can quickly take over a population [25, 26]. Nonpiliated variants can result from several mechanisms, including pilin AV [20, 24], *pilC* phase variation [2, 21], and nonreversible *pilE* deletions [18, 29].

Despite the critical role of pili in Gc pathogenesis, there are no comprehensive analyses of the *pilE* gene and *pilS* loci from clinical isolates. We address these knowledge gaps through systematic analysis of the *pilE* and *pilS* loci from an extensive Gc genome collection, utilizing some novel bioinformatic approaches. Analysis of the *pilE* and *pilS* loci from 15 186 Gc finished and draft genomes in Public Databases for Molecular Typing and Microbial Genome Diversity (PubMLST) [30] revealed conserved *pilS* and *pilE* loci positions but surprisingly demonstrated that most genomes were from DNA isolated from nonpiliated variants. Considering the normal piliated status of fresh clinical isolates, coupled with the increased growth rate of nonpiliated variants, we conclude that these genomic sequences were likely derived from nonpiliated variants that arose during in vitro storage and propagation. We highlight considerations for using genomic data: laboratory-passaged sequences may not reflect clinical phenotypes, potentially underestimating virulence gene functionality in surveillance studies.

METHODS

PubMLST in-silico PCR Tool to Identify *pilE* and *pilS* Loci

We analyzed 15 186 whole-genome Gc sequences from PubMLST, an open-access repository of Gc genomic sequences [30]. Only 0.43% of the genomes were finished using hybrid long and short read sequencing. The other genomic sequences (99.57%) were draft assemblies from short-read sequencing, containing many contigs.

We designed PCR primers using SnapGene software (version 8.0.1) (Supplementary Table 1). Primers were validated using Gc reference genomes FA1090 (CP115654.1) [31] and MS11 (CP115904.1). We utilized the PubMLST in silico PCR tool (Supplementary Table 2) to detect *pilE* loci by upstream *garP*-G4, the complete *pilE* sequence, the *cys1* and *cys2* sequences within the *pilE* gene, and downstream *SmaCla* repeat (Figure 1). Combining upstream (RS1-*garP*) and downstream (*pilE* start codon—*SmaCla*) in silico PCR reactions provided 10.26% complete *pilE* sequences (1558 out of 15 186 genomes) for analysis (Figure 1).

We used SnapGene with 80% identity to reference sequences (Supplementary Table 1) to identify *pilS* loci in 65 finished genomes. Draft genomes were excluded from the *pilS* loci analysis due to their high fragmentation. *pilS* loci were defined by the presence of the *cys1*, *cys2*, and a 3′ *SmaCla* repeat without the adjacent Class I Leader sequence and *garP*-G4. The closest *pilS* locus to the *pilE* gene lacks a *SmaCla* repeat and was identified by the presence of *cys1* and *cys2* only. We designated this locus *pilS8*, extending the published *pilS1-7* nomenclature (Supplementary Table 3) [16, 18, 32]. We measured the distance between the *dnaA* gene (the conventional chromosomal starting point) and each identified *SmaCla* repeat in finished Gc genomes using a custom R script (*SmaCla* to *dnaA* relative distance algorithm). The *SmaCla* repeat positions were normalized to the total genome length.

To identify *pilC*1 and *pilC*2 genes, we used SnapGene with 80% identity threshold to reference sequences NEIS0371 and NEIS0033, respectively [30]. We analyzed five upstream and downstream loci of the *pilE* and *pilC* genes in finished Gc genomes using the SnapGene software and annotated them with the PubMLST "NEIS" designation [30].

Sequence Analysis

The *pilE* gene and *pilS* loci sequences were translated using the Geneious Prime software (version 2025.0.3). Nucleotide and protein multiple alignments were performed using the MAFFT fast progressive FFT-NS-2 2 method with default parameters [33]. The alignments were visualized using Geneious Prime and Jalview (version 2.11) [34]. The *pilE* gene nucleotide variation index was defined as the cumulative percentage of nucleotides in all tested sequences differing from the 1-81-S2 FA1090 reference [27]. Potential stop codon donors from *pilS* loci were identified through multiple alignments of the PilE sequences against translated FA1090 *pilS* copies (Supplementary Table 3). The stop codon rate was calculated as the percentage of stop codon events at a specific amino acid position relative to the total sequences examined.

PilE Protein Analysis

We used the existing MS11 PilE crystal structure (PDB:2HI2) [35] to predict the FA1090 PilE structure using ColabFold 1.5.5 [36], which employs MMseqs2 [37] and AlphaFold2

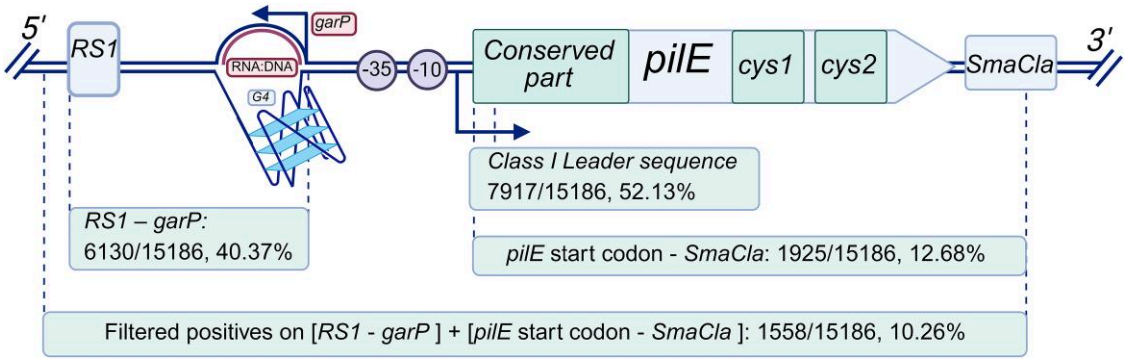

Figure 1. The *pilE* locus The figure shows the *pilE* gene's main structural elements and related upstream and downstream sequences. RS1—repetitive sequence 1; G4 and *garP*—G4-*garP*-*pilE* associated structures; formed R-loop (RNA:DNA hybrid); −35 and −10 indicate the *pilE* gene promoter; *cys1*, *cys2*, and *SmaCla*—repetitive sequences; DS—downstream. The primary PCR reactions targeting the Class I Leader sequence and the *pilE* gene, from the start codon to the downstream SmaCla repeat, yielded 52.13% (7917/15 186) and 12.68% (1925/15 186) positive reactions, respectively. Secondary in silico PCRs verified upstream RS1 and *garP*-G4 sequences. Combining PCR targets for upstream (RS1 and *garP*) and downstream (*SmaCla*) sequences provided 10.26% (1558/15 186) complete *pilE* gene sequences for analysis. This reduced yield was primarily due to the fragmented nature of draft genome assemblies from short-read sequencing data. Created with BioRender.

[38]. Nucleotide variation at the *pilE* locus was translated to amino acid and mapped onto the predicted PilE protein structure using a custom biopython script (PilE structure variation mapping algorithm). Visualization of PilE structure with mapped variation index was performed using the PyMOL Molecular Graphics System, Version 3.1.4.1, Schrödinger, LLC.

Identification of *pilS* Donors

We performed a comparative nucleotide analysis between the *pilE* and *pilS* loci from the finished Gc genomes using a custom biopython script (*pilS* copies identification and extraction tool). Extracted *pilS* sequences were estimated based on the longest silent locus (2661 bp, *pilS2* FA1090, Supplementary Table 3), which included the *SmaCla* repeat sequence and 3000 bps adjacent sequence. The *pilE* gene and *pilS* loci were annotated as above. We extracted the *pilE* gene fragments and then manually trimmed the sequence from the predicted start codon. We aligned the extracted *pilE* gene sequence against each *pilS* loci from the same genome using MAFFT algorithm with default parameters [33]. Sequence similarity and identical region length were used to identify potential donor-recipient relationships between the *pilE* gene and *pilS* copies.

Genome Rearrangement Analysis

We performed multiple alignments of all finished genomes against FA1090 and MS11 reference genomes using the Mauve progressive algorithm with default parameters [39]. The resulting alignments were manually inspected to identify potential structural genomic variations.

Statistical Analysis

Statistical analyses were performed in R version 4.4.2. The Shapiro-Wilk test assessed data normality, the Wilcoxon rank-sum test—pairwise comparisons, and Cohen's d—size effect. *SmaCla*'s genomic position distributions were evaluated via Kolmogorov-Smirnov and strand bias via Pearson's chi-square tests. The significance level was set as $P < .05$.

Data Visualization

Data visualization was performed using BioRender (https://www.biorender.com) and R (version 4.4.2) with RStudio (RStudio Team, 2024).

Ethical Considerations

This study analyzed publicly available gonococcal genomic data from the PubMLST database, adhering to its data use policy [30]. No human subjects, personal patient information, or biological specimens were directly involved; thus, institutional review board approval and informed consent were not required.

RESULTS

Analysis of *pilE* and *pilS* Loci in Gc Genomic Sequences

We unexpectedly found that the *pilE* gene was deleted in 41.54% (27/65) of finished genomes. These deletions consistently removed the entire *pilE* gene with upstream structures (G4-*garP* and RS1), while maintaining the downstream *SmaCla* repeat (Figure 2*A*). We found a similar deletion frequency (47.89%, 7242/15 121) in draft genomes. In total, the *pilE* gene was present in only 52.13% (7917/15 186) of all examined Gc genomes (Figure 1).

We identified eight common *pilE* deletion variants in the finished genomes (Supplementary Figure 1). The predominant variant (63%, 17/27 genomes) deleted *pilE* with the upstream *cys2*, while leaving an upstream *pilS8* locus. A second variant (15%, 4/27 genomes) retained both the upstream *pilS8* locus and a *cys2* repeat (Supplementary Figure 1*A*). Six rare variants

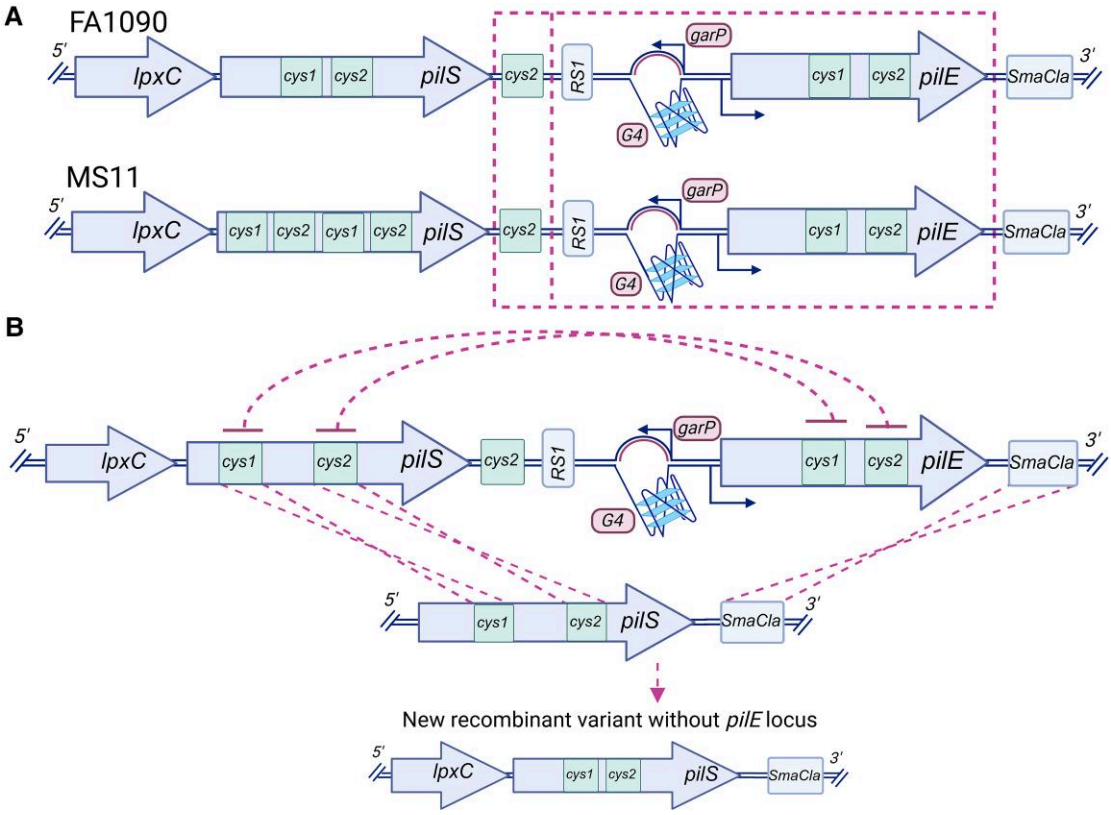

Figure 2. Structural analysis of *pilE* deletions in Gc finished genomes. *A*,FA1090 and MS11 *pilE* loci organization. The complete *pilE* gene and associated elements (G4-*garP*, RS1) targeted for deletion are highlighted within the magenta dashed box. *lpxC*—UDP-3-O-[3-hydroxymyristoyl] N- acetylglucosamine deacetylase (essential gene); RS1—repetitive sequence 1; G4 and *garP*—G4-*garP-pilE* associated structures; *pilS*—silent gene; *pilE*—*pilE* gene; *cys1*, *cys2*, and *SmaCla*—repetitive sequences. *B*, Maps of *pilE* deletions. Two pathways lead to irreversible piliation loss: recombination between repeated upstream and downstream sequences (magenta dashed crossed lines indicate gene conversion events), or transformation between *cys1 or cys2* elements between *pilS* and *pilE* loci (upper circled dashed line). Both deletion processes result in genetic variants lacking the *pilE* locus. Created with BioRender.

(22%, 6/27 genomes) display unique recombination patterns with variable conservation of the upstream *pilS* and *cys2* sequences (Supplementary Figure 1*B*). Notably, all variants maintained the *pilE* downstream *SmaCla* repeat. Genomic architecture surrounding deletion sites suggests that recombination events involved the *cys1* and *cys2* of upstream *pilS* and the downstream *SmaCla* (Figure 2*B*). These events might proceed through endogenous DNA by gene conversion or transformation involving exogenous DNA, a mechanism previously described for L-pilin variants [40] (Supplementary Figure 1*B*).

Loci surrounding the *pilE* gene exhibited distinct patterns. The three upstream loci closest to *pilE* were conserved: *pilS8* (86.15%, 56/65), NEIS0001 (84.62%, 55/65), and *pilS2* (81.54%, 53/65) at positions 1–3 (Supplementary Table 4). Further upstream of *pilE*, diversity increased with *pilS1* (44.62%, 29/65) and NEIS0235 (27.69%, 18/65) being the most common loci at the fourth and fifth positions. In contrast, downstream loci were more highly conserved, with NEIS2021 being the closest downstream locus in 92.31% (60/65) of genomes. The four other closest downstream genes (NEIS0020, NEIS0019, NEIS0018, and NEIS0017) are identical in 90.77% (59/65) of genomes. This genomic organization remains consistent regardless of *pilE*'s presence.

Conservation of *pilS* and *pilC* Loci

All 65 finished genomes contained 1–9 *SmaCla* sequences (mean ± SD: 5.35 ± 1.24) (Table 1). Intact *pilE* genomes showed fewer *SmaCla* repeats than *pilE*-negative genomes (5.08 ± 1.32 vs 5.74 ± 1.02), suggesting *pilE* deletion does not remove downstream *SmaCla* (Table 1). No differences were detected in relative *pilS* positions or strand distribution (Table 1). Two distinct genomic clades were observed: one (46.15%, 30/65) with *pilE* and surrounding *pilS* loci at relative position 0.80 from the *dnaA* gene, and another (52.31%, 34/65) at position 0.95 (Figure 3). These differential positions resulted from a large genomic inversion upstream of the *pilS3* locus (Supplementary Figure 2) previously described [16, 41]. The *pilC1* gene was present in all 65 finished Gc genomes, regardless of *pilE* status, and the prevalence of flanking loci was similar for *pilE*-positive and *pilE*-negative genomes (Supplementary

Table 1. SmaCla Sequence Analysis in Finished Genomes, *n* = 65

Category	Statistical Parameter	Genomes With *pilE* Gene, *n* = 38	Genomes Without *pilE* Gene, *n* = 27	*P* Value
***SmaCla* copy number**	Range (min—max)	1–9	4–8	
	Mean ± SD	5.08 ± 1.32	5.74 ± 1.02	*P* = .025[a]
	Median (IQR)	5 (5–6)	6 (5–6.5)	
Relative genomic position	Mean ± SD	0.73 ± 0.21	0.71 ± 0.24	*P* = .20[b]
	Median (IQR)	0.80 (0.58–0.94)	0.80 (0.59–0.82)	
Strand distribution	Forward, n (%)	77 (39.9)	51 (32.9)	*P* = .22[c]
	Reverse, n (%)	116 (60.1)	104 (67.1)	

Abbreviations: IQR; interquartile range between first and third quartile; SD; standard deviation.

[a]Wilcoxon rank-sum test. Effect size measured by Cohen's d = −0.55 indicated a moderate difference in *SmaCla* counts between groups.

[b]Kolmogorov-Smirnov test.

[c]Pearson's chi-square test ($\chi^2 = 1.52$, $df = 1$).

Table 5). Both *pilC* genes (*pilC1* and *pilC2*) were present together in only two of the 65 finished genomes, one of which lacked *pilE*.

The Regulatory Pilin AV Elements—*garP* and G4

We analyzed the complete *pilE* gene sequences, and their upstream and downstream associated elements in 1558 out of 15 186 genomes (10.26%) (Figure 1). The *garP* and G4 were conserved in all 1558 genomes with complete *pilE* gene sequence (Supplementary Figure 3). Most genomes contained FA1090, MS11 (11.10%, 173/1558), or a C-to-T substitution within the *garP* sequence (Supplementary Table 6) [42]. We also found a rare fourth allele, with a single nucleotide substitution (A-to-C) in the *garP* −10 element.

Similarly, the wild-type G4 motif [G3:T:G3:TT:G3:T] was preserved in a majority of the genomes, with rare variants potentially affecting G4 formation found in only 4.36% of isolates (Supplementary Table 7) [15]. Whether these sequences alter AV or represent sequencing errors is unknown.

Variability of the *pilE* ORF

The *pilE* gene exhibited high sequence variation (Supplementary Figure 4). There were 1439 unique alleles identified among 1558 genomes (Supplementary Table 8). Only eight *pilE* alleles were repeated 4–25 times; the remaining 1431 alleles appeared 1–3 times. Nucleotides encoding the first 38 amino acids, including the leader sequence, were conserved. Only Class I Leader sequence was found among all 1558 Gc genomes, in contrast to the appearance of Class II Leader sequences in *Neisseria meningitidis* (Nm). The first 150 nucleotides, encoding the leader sequence and conserved N-terminus of the mature PilE protein, showed minimal variation (0.90 ± 7.4%) (Figure 4). We detected three synonymous nucleotide mutations (54T (86.52%), 87C (19.83%), and 90C (19.83%)) within the conserved region that do not alter PilE amino acid sequences or structure (Figure 5) [27]. The semi-variable region (from the N-terminal conserved coding sequences to the *cys1* repeat) displayed intermediate variability (12.71 ± 24.26%). The hypervariable loop (HV_L; between *cys1* and *cys2*) and tail (HV_T; from the end of the *cys2* to stop codon) regions exhibited the highest diversity (45.78 ± 29.37% and 41.25 ± 17.04%, respectively) (Figure 4). The *cys1* and *cys2* sequences essential for AV maintained low variability (2.17 ± 9.88% and 9.09 ± 25.01%, respectively).

Identification of *pilS* Donors

We identified potential *pilS* donor sequences for *pilE* variants [20, 24, 43] in 35/38 (92.10%) finished genomes with *pilE* genes and mapped their genomic positions (Supplementary Table 9). In the remaining genomes (7.90%, 3/38), scattered single-nucleotide indels in the *pilE* sequences prevented the identification of specific *pilS* donors.

The Prevalence of Nonfunctional PilE Proteins

PilE ORF length, including the seven-amino acid signal peptide, averaged 170.11 ± 3.60 amino acids (range: 165–183). Most Gc genomes retained the defined *S*-cleavage site first described by Haas *et al* [12].: A38 (100%), Q39 (98.97%), and Q40 (98.78%). While 36.26% had normal C-terminal stop codons, 0.13% lacked a stop codons, and likely were L-pilin variants [40], but most *pilE* sequences (63.61%) contained internal stop codons (Supplementary Figure 5). Of these truncated proteins, 51.22% had multiple stop codons (2.49 ± 1.35), distributed from the semi-variable region through the C-terminus of the PilE protein.

The earliest truncation occurred at amino acid G85 (7.12%) (Figure 5*B*), while the most frequent PilE stop codon positions were at T134 (53.40%), V103 (51.48%), or M98 (29.08%). All premature stop codons occurred beyond essential N-terminal elements, including the Class I Leader sequence and the S-cleavage site [12] required for transformation competence [13]. Comparisons with FA1090 reference *pilS* sequences [24] identified *pilS1* copy 4, *pilS2* copy 4, and *pilS6* copy 2 as the primary donors of premature stop codons (Table 2).

DISCUSSION

Analysis of the *pilE* and *pilS* loci in 15 186 publicly available Gc genomes revealed that approximately half lacked a *pilE* gene,

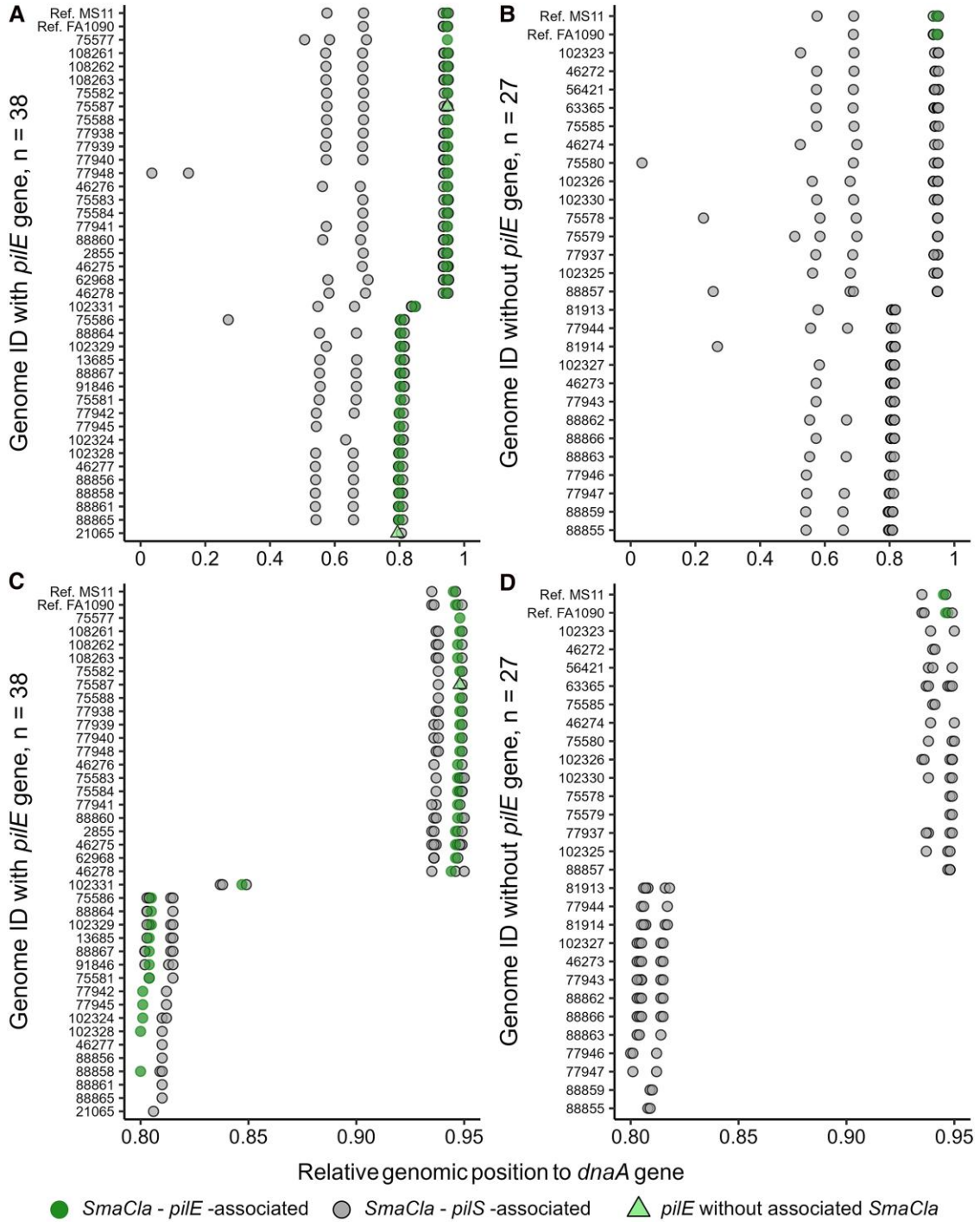

Figure 3. Relative *SmaCla* repeat positions. *SmaCla* relative genomic position in genomes that maintain the *pilE* gene (*A*) and in genomes without the *pilE* gene (*B*). Panels *C* and *D* show zoomed relative genome segments (0.80–0.95) enriched by the *SmaCla* repeats in genomes with the *pilE* gene and those with *pilE* gene deletion, respectively. *SmaCla* repeats associated with *pilE* genes are distinquished from those associated with *pilS* loci. Triangles mark two instances of *pilE* genes lacking downstream *SmaCla* repeats.

and most genomes with *pilE* harbored premature stop codons, indicating most of these genomic sequences are from nonpiliated Gc. This prevalence of *pilE* deletion and nonfunctional pilin variants is surprising, given piliation's critical roles in attachment, DNA transformation, and immune evasion [1, 2], and that piliated gonococci are almost always isolated from infected patients [44–46]. Although nonpiliated Gc can invade cells [47], grow on fallopian tube tissue [28], and appear

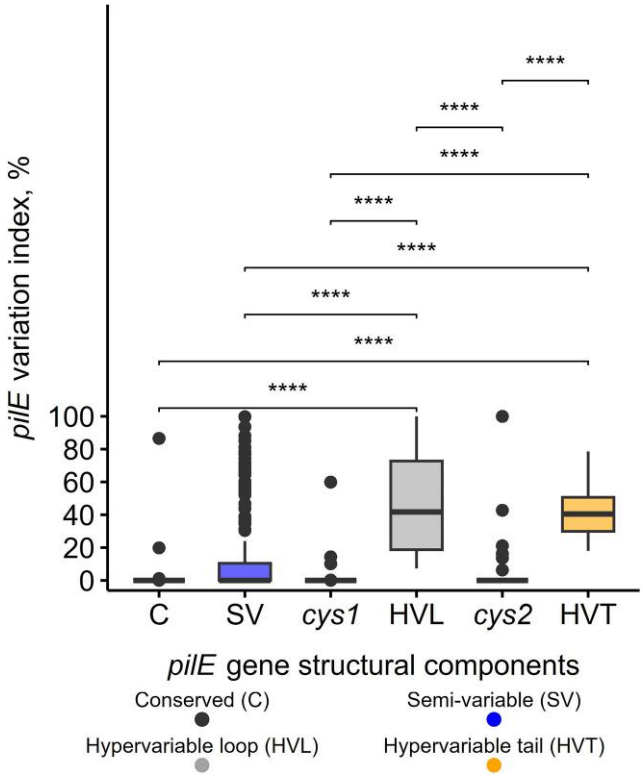

Figure 4. Nucleotide variation of the *pilE* gene in *N. gonorrhoeae* genomes, $n = 1558$. Pairwise comparisons of the nucleotide variation index of *pilE* gene structural components. **** $P < .0001$, Wilcoxon rank-sum test.

later in infection [25, 48], they represent <10% of Gc populations when present [46], which contrasts with their predominance in the genomic database. Our findings confirm that, alongside pilin AV [20, 24, 27] and phase variation [21, 23, 48], *pilE* deletion [29] represents a third mechanism of pilin variation in Gc [40] that is more prevalent in the genomic database than previously recognized.

The *pilC* locus is conserved across all 65 finished Gc genomes, contrasting with frequent *pilE* deletions. However, the G-tract in *pilC* ORF subjects its expression to reversible phase variation [2, 21, 48], consequently affecting piliation levels. The role of pilus phase variation in infection remains unclear. The fact that pilus phase variation is mediated by pilin AV and *pilC* phase variation strongly suggests that nonpiliated Gc plays unknown roles in infection. We have postulated that the sensitivity of nonpiliated variants to neutrophil killing explains the paucity of nonpiliated Gc in purulent exudates containing numerous neutrophils [49].

The genomic sequences of bacterial clinical isolates are presumed to represent the organisms present in the patient or subject at the time of isolation. However, clinical laboratories frequently utilize a single colony, which may fail to capture the full spectrum of subtypes present in an individual and may overlook critical epidemiological indicators. Nonpiliated colonies are larger with significantly more colony-forming units per colony than piliated colonies grown on solid media (Supplementary Figure 6). We postulate that the predominance of nonpiliated variants in the genomic database likely results from variants arising during laboratory storage and cultivation, rather than nonpiliated Gc isolated directly from patients [23, 26]. Moreover, the propagation of any bacterial isolate on standard media may select for genetic mutants or variants that grow better under laboratory conditions than in the host environment. These fitness limits imposed by in-vitro cultivation would lead to variation in metabolism and possibly many other phenotypic changes.

Despite variations in *pilE* presence and functionality, key elements essential for pilin diversity and functionality were conserved. The G4 and *garP* sequences, necessary for AV, are conserved in over 95% of genomes [14, 15, 42]. We identified Class I Leader sequence, required for pilin processing [11], in all *pilE*-retaining genomes, but did not detect any Class II leader sequences. When Gc relies primarily on AV, Nm employs a dual strategy to avoid immune response: using AV in strains with Class I Leader sequences and pilin glycosylation in Class II strains [50]. We cannot rule out the possibility that Class II leader sequences were present in the isolates deleted for *pilE*, but we consider this possibility unlikely. Although most PilE proteins were truncated, the conserved *S*-cleavage site [12] suggests that Gc isolates with intact *pilE* loci can maintain transformation competency [13]. The *cys1* and *cys2* cysteine motifs necessary for AV [16, 43] and stabilizing pilin [35] showed minimal variation. These findings show how Gc preserves essential pilin functions while allowing AV for immune evasion.

Most Gc genomes harbored unique *pilE* gene alleles, showing high sequence diversity. The highest sequence diversity occurred in the previously identified highly variable HV_L and HV_T regions, which are exposed on the pilus surface and subject to immune selection [8, 35]. This result is the first global analysis of *pilE* variability and suggests there may be limits to the variability of the PilE protein and the pilus.

The conservation of *pilS* loci location, irrespective of *pilE* presence, indicates strong evolutionary selection for maintaining silent repositories to preserve AV capacity in Gc [16, 18, 19]. Our results are in line with previous studies, showing the unidirectional recombination reaction from *pilS* to *pilE*, with only *pilE* sequence variation [16, 17, 20]. Higher *SmaCla* counts in *pilE*-negative genomes suggest *pilE* deletion events retain downstream *SmaCla*, which serve as homologous sequences facilitating AV (Figure 2*B*) [16, 17, 19]. Upstream *pilS* genes provide the homologous substrate (*cys1* and *cys2*) for gene conversion events. Two distinct clades show *pilE*/*pilS* loci at 0.80 or 0.95 relative positions from *dnaA*, resulting from chromosomal inversions near *pilS3* [41], demonstrating that Gc balances genome rearrangements while preserving essential immune evasion systems.

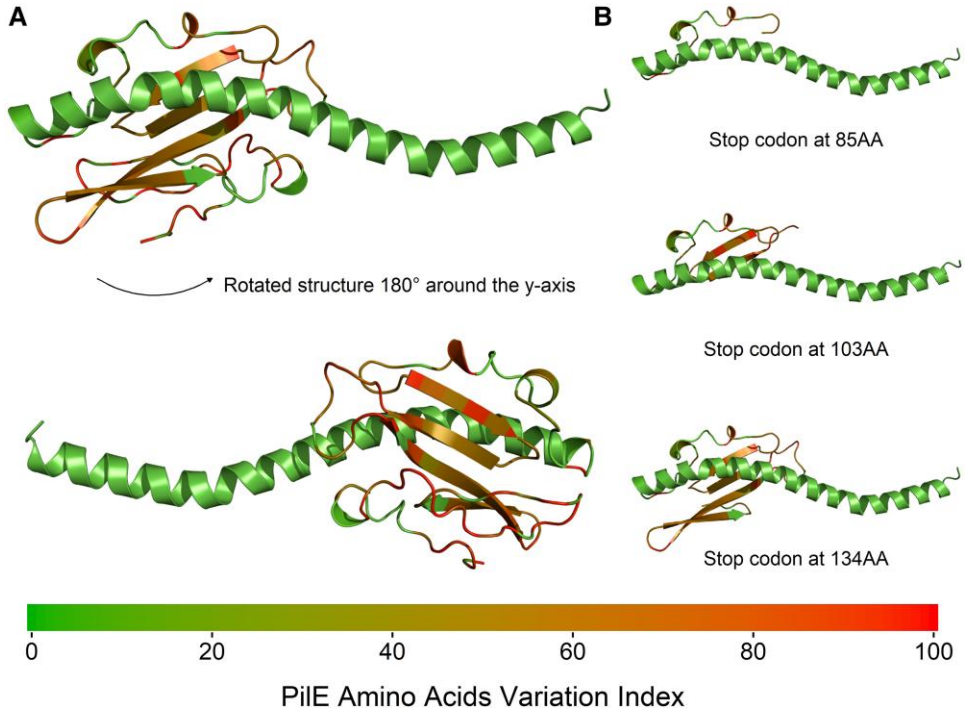

Figure 5. PilE structure of *N. gonorrhoeae* FA1090 showing sequence variability and major truncation sites. *A*, Predicted mature PilE protein structure in *N. gonorrhoeae* FA1090. The N-terminal α-helices indicate conservation; the C-terminal globular head highlights variability. The heat map below shows conservation-to-variation gradient. *B*, Predicted truncated PilE structures showing the earliest truncation (G85 amino acid (AA) position) and most frequent premature stop codons (V103 AA and T134 AA positions).

Table 2. Distinct pilS Copies May Contribute to the PilE Internal Stop Codon(s) Acquisition in Gc Genomes, $n = 1558$

PilE Amino Acids Position and Its Substitution to Stop Codon (*)	Location Within PilE	Stop Codon, n (%)	*pilS* Copies With Homologous Sequences and Stop Codons
Glutamine, (Q), 85→*	Semi-variable	111 (7.12%)	*pilS1* copy 3
Glutamic acid (E), 88→*	Semi-variable	15 (0.96%)	*pilS1* copy 3
Alanine (A), 90→*	Semi-variable	2 (0.13%)	*pilS1* copy 3
Methionine (M), 98→*	Semi-variable	453 (29.08%)	*pilS6* copy 2
Valine (V), 103→*	Semi-variable	802 (51.48%)	*pilS1* copy 4, *pilS2* copy 4, *pilS6* copy2
Lysine (K), 105→*	Semi-variable	14 (0.90%)	*pilS1* copy 3
Valine (V), 131→*	*cys1*	71 (4.56%)	*pilS1* copy 3
Threonine (T), 134→*	Hypervariable	832 (53.40%)	*pilS1* copy 4, *pilS2* copy 4, *pilS6* copy 2
Threonine (T), 156→*	*cys2*	69 (4.43%)	*pilS1* copy 3

We didn't investigate the presence of phase variation in the *opa* gene family or the ∼80 other phase variable genes in the Gc genome [22]. If any of these phase variants provide an in vitro growth advantage, they may also be misrepresented in these genomic sequences. Although in silico PCR detected the *pilE* gene in draft genomes, the fragmentation of the contigs could underestimate its presence. However, the similar *pilE* deletion frequency in draft and finished genomes strongly suggests that the *pilE* deletion frequency in draft genomes is correct. Gc genome assembly is complicated by repeated sequences, but hybrid or long-read sequencing enables more accurate assembly. We lacked access to sample collection protocols, which may have contributed to the emergence of nonpiliated variants. Storage or transport of cultures, varying incubation times, and the number of passages all could lead to a takeover of nonpiliated variants.

This study presents the first systematic analysis of the *pilE* and *pilS* loci across large-scale gonococcal genome collections, utilizing an in silico PCR approach and bioinformatics tools to identify and characterize all the pilus loci. While it is impossible to prove that most of these clinical isolates originated as piliated, the predominance of irreversible, non-piliated *pilE* deletions strongly suggests that these were in vitro-derived variants. This study also provides important insights into the stability and variability of pilin loci, raising essential questions about how we interpret all genomic data. We suggest that clinical laboratories

and genomic surveillance efforts should focus on protocols that minimize laboratory passage time and consider Gc colony morphology to utilize piliated Gc in genomic sequencing.

Supplementary Data

Supplementary materials are available at *The Journal of Infectious Diseases* online (http://jid.oxfordjournals.org/). Supplementary materials consist of data provided by the author that are published to benefit the reader. The posted materials are not copyedited. The contents of all supplementary data are the sole responsibility of the authors. Questions or messages regarding errors should be addressed to the author.

Notes

Author Contributions. IB and SM—conceptualization, formal analysis, investigation, methodology, visualization, and original draft writing. HSS—conceptualization, funding acquisition, supervision, methodology, data interpretation, project administration, original draft writing. All authors approved the final version of the manuscript and agreed to be accountable for all aspects of this study.

Acknowledgments. This study utilized the PubMLST website (https://pubmlst.org), developed by Keith Jolley and located at the University of Oxford [30]. The Wellcome Trust funded the development of that website. We are grateful to Dr. Robert Nicholas for providing the Gc clinical isolate 35/02. We greatly appreciate Dr. Odile B. Harrison for her skillful, extensive training in using the PubMLST database. We are grateful to Dr. Odile B. Harrison, Dr. Rachel M. Exley, and Dr. Ellen L. Aho for their discussions and helpful advice during manuscript preparation. We thank Dr. Kathleen Nicholson for her valuable and constructive feedback during the manuscript revision. We are also grateful to members of the Seifert lab for helpful discussions throughout this study.

Financial support. The National Institutes of Health (NIH) supported this work by grants R37AI033493 and R01AI146073 to HSS. The funder had no role in study design, data collection, analysis or interpretation, manuscript writing, or the decision to submit the article for publication. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. It is subject to the NIH Public Access Policy. Through acceptance of this federal funding, NIH has been given the right to make this manuscript publicly available in PubMed Central upon the Official Date of Publication, as defined by NIH.

Ethics approval and consent to participate. Not applicable.

Consent for publication. All authors consented to the publication of this study.

Data availability. All custom scripts are publicly available from https://github.com/ibboiko/Neisseria_gonorrhoeae_genomes_pilE_pilS. The associated raw data supporting the conclusions of this study are openly available from the following repository: Boiko, I., Metaane, S., & Seifert, H. S. (2025). Publicly available *Neisseria gonorrhoeae* genomes predominantly represent in vitro-derived nonpiliated variants (Version 2) [Data set]. Prism. Galter Health Sciences Library. Northwestern University. https://doi.org/10.18131/z9e54-v0237.

Potential conflicts of interest. The authors declare no conflict of interest.

References

1. Unemo M, Seifert HS, Hook EW III, Hawkes S, Ndowa F, Dillon J-AR. Gonorrhoea. Nat Rev Dis Primer **2019**; 5:79.
2. Quillin SJ, Seifert HS. Neisseria gonorrhoeae host adaptation and pathogenesis. Nat Rev Microbiol **2018**; 16:226–40.
3. World Health Organization. Gonorrhoea (Neisseria gonorrhoeae infection) [Internet]. 2025. Available at: https://www.who.int/news-room/fact-sheets/detail/gonorrhoea-(neisseria-gonorrhoeae-infection). Accessed 4 October 2025.
4. Lovett A, Duncan JA. Human immune responses and the natural history of Neisseria gonorrhoeae infection. Front Immunol **2019**; 9:3187.
5. Walker E, van Niekerk S, Hanning K, Kelton W, Hicks J. Mechanisms of host manipulation by Neisseria gonorrhoeae. Front Microbiol **2023**; 14:1119834.
6. Tønjum T, Koomey M. The pilus colonization factor of pathogenic neisserial species: organelle biogenesis and structure/function relationships–a review. Gene **1997**; 192:155–63.
7. Punsalang AP Jr, Sawyer WD. Role of pili in the virulence of Neisseria gonorrhoeae. Infect Immun **1973**; 8:255–63.
8. Craig L, Forest KT, Maier B. Type IV pili: dynamics, biophysics and functional consequences. Nat Rev Microbiol **2019**; 17:429–40.
9. Fyfe JA, Davies JK. An AT-rich tract containing an integration host factor-binding domain and two UP-like elements enhances transcription from the pilEp1 promoter of Neisseria gonorrhoeae. J Bacteriol **1998**; 180:2152–9.
10. Dupuy B, Pugsley AP. Type IV prepilin peptidase gene of Neisseria gonorrhoeae MS11: presence of a related gene in other piliated and nonpiliated Neisseria strains. J Bacteriol **1994**; 176:1323–31.
11. Wörmann ME, Horien CL, Bennett JS, et al. Sequence, distribution and chromosomal context of class I and class II pilin genes of Neisseria meningitidis identified in whole genome sequences. BMC Genomics **2014**; 15:253.
12. Haas R, Schwarz H, Meyer TF. Release of soluble pilin antigen coupled with gene conversion in Neisseria gonorrhoeae. Proc Natl Acad Sci U S A **1987**; 84:9079–83.
13. Obergfell KP, Seifert HS. The pilin N-terminal domain maintains Neisseria gonorrhoeae transformation competence during pilus phase variation. PLoS Genet **2016**; 12: e1006069.

14. Cahoon LA, Seifert HS. Transcription of a cis-acting, noncoding, small RNA is required for pilin antigenic variation in Neisseria gonorrhoeae. PLoS Pathog **2013**; 9:e1003074.
15. Prister LL, Yin S, Cahoon LA, Seifert HS. Altering the Neisseria gonorrhoeae pilE guanine quadruplex loop bases affects pilin antigenic variation. Biochemistry **2020**; 59: 1104–12.
16. Hamrick TS, Dempsey JAF, Cohen MS, Cannon JG. Antigenic variation of gonococcal pilin expression in vivo: analysis of the strain FA1090 pilin repertoire and identification of the pilS gene copies recombining with pilE during experimental human infection. Microbiology (Reading) **2001**; 147:839–49.
17. Hill SA, Davies JK. Pilin gene variation in Neisseria gonorrhoeae: reassessing the old paradigms. FEMS Microbiol Rev **2009**; 33:521–30.
18. Haas R, Meyer TF. The repertoire of silent pilus genes in Neisseria gonorrhoeae: evidence for gene conversion. Cell **1986**; 44:107–15.
19. Wainwright LA, Pritchard KH, Seifert HS. A conserved DNA sequence is required for efficient gonococcal pilin antigenic variation. Mol Microbiol **1994**; 13:75–87.
20. Cahoon LA, Seifert HS. Focusing homologous recombination: pilin antigenic variation in the pathogenic Neisseria. Mol Microbiol **2011**; 81:1136–43.
21. Jonsson AB, Pfeifer J, Normark S. Neisseria gonorrhoeae PilC expression provides a selective mechanism for structural diversity of pili. Proc Natl Acad Sci U S A **1992**; 89:3204–8.
22. Snyder LAS, Butcher SA, Saunders NJ. Comparative whole-genome analyses reveal over 100 putative phase-variable genes in the pathogenic Neisseria spp. Microbiology **2001**; 147:2321–32.
23. Long CD, Madraswala RN, Seifert HS. Comparisons between colony phase variation of Neisseria gonorrhoeae FA1090 and pilus, pilin, and S-pilin expression. Infect Immun **1998**; 66:1918–27.
24. Criss AK, Kline KA, Seifert HS. The frequency and rate of pilin antigenic variation in Neisseria gonorrhoeae. Mol Microbiol **2005**; 58:510–9.
25. James JF, Swanson J. Studies on gonococcus infection. XIII. Occurrence of color/opacity colonial variants in clinical cultures. Infect Immun **1978**; 19:332–40.
26. Swanson J, Kraus SJ, Gotschlich EC. Studies on gonococcus infection. I. Pili and zones of adhesion: their relation to gonococcal growth patterns. J Exp Med **1971**; 134:886–906.
27. Seifert HS, Wright CJ, Jerse AE, Cohen MS, Cannon JG. Multiple gonococcal pilin antigenic variants are produced during experimental human infections. J Clin Invest **1994**; 93:2744–9.
28. Lenz JD, Dillard JP. Pathogenesis of Neisseria gonorrhoeae and the host defense in ascending infections of human fallopian tube. Front Immunol **2018**; 9:2710.
29. Segal E, Billyard E, So M, Storzbach S, Meyer TF. Role of chromosomal rearrangement in N. gonorrhoeae pilus phase variation. Cell **1985**; 40:293–300.
30. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res **2018**; 3:124.
31. Hu LI, Yin S, Ozer EA, et al. Discovery of a new Neisseria gonorrhoeae type IV pilus assembly factor, TfpC. mBio **2020**; 11:e02528-20.
32. Rotman E, Webber DM, Seifert HS. Analyzing Neisseria gonorrhoeae pilin antigenic variation using 454 sequencing technology. J Bacteriol **2016**; 198:2470–82.
33. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform **2019**; 20:1160–6.
34. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics **2009**; 25:1189–91.
35. Craig L, Volkmann N, Arvai AS, et al. Type IV pilus structure by cryo-electron microscopy and crystallography: implications for pilus assembly and functions. Mol Cell **2006**; 23:651–62.
36. Kim G, Lee S, Levy Karin E, et al. Easy and accurate protein structure prediction using ColabFold. Nat Protoc **2025**; 20: 620–42.
37. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol **2017**; 35:1026–8.
38. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature **2021**; 596: 583–9.
39. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One **2010**; 5:e11147.
40. Manning PA, Kaufmann A, Roll U, Pohlner J, Meyer TF, Haas R. L-pilin variants of Neisseria gonorrhoeae MS11. Mol Microbiol **1991**; 5:917–26.
41. Gibbs CP, Meyer TF. Genome plasticity in Neisseria gonorrhoeae. FEMS Microbiol Lett **1996**; 145:173–9.
42. Prister LL, Ozer EA, Cahoon LA, Seifert HS. Transcriptional initiation of a small RNA, not R-loop stability, dictates the frequency of pilin antigenic variation in Neisseria gonorrhoeae. Mol Microbiol **2019**; 112: 1219–34.
43. Howell-Adams B, Seifert HS. Molecular models accounting for the gene conversion reactions mediating gonococcal pilin antigenic variation. Mol Microbiol **2000**; 37: 1146–58.
44. James-Holmquest AN, Swanson J, Buchanan TM, Wende RD, Williams RP. Differential attachment by piliated and nonpiliated Neisseria gonorrhoeae to human sperm. Infect Immun **1974**; 9:897–902.

45. Fichorova RN, Desai PJ, Gibson FC III, Genco CA. Distinct proinflammatory host responses to Neisseria gonorrhoeae infection in immortalized human cervical and vaginal epithelial cells. Infect Immun **2001**; 69:5840–8.
46. Kovalchik MT, Kraus SJ. Neisseria gonorrhoeae: colonial morphology of rectal isolates. Appl Microbiol **1972**; 23: 986–9.
47. Shaw JH, Falkow S. Model for invasion of human tissue culture cells by Neisseria gonorrhoeae. Infect Immun **1988**; 56:1625–32.
48. Ilver D, Källström H, Normark S, Jonsson A-B. Transcellular passage of Neisseria gonorrhoeae involves pilus phase variation. Infect Immun **1998**; 66:469–73.
49. Stohl EA, Dale EM, Criss AK, Seifert HS. Neisseria gonorrhoeae metalloprotease NGO1686 is required for full piliation, and piliation is required for resistance to H2O2- and neutrophil-mediated killing. mBio **2013**; 4:e00399-13.
50. Gault J, Ferber M, Machata S, et al. Neisseria meningitidis type IV pili composed of sequence invariable pilins are masked by multisite glycosylation. PLoS Pathog **2015**; 11:e1005162.