THE LANCET Infectious Diseases

Supplementary webappendix

Supplement to: Town K, Field N, Harris SR. Phylogenomic analysis of *Neisseria gonorrhoeae* transmission to assess sexual mixing and HIV transmission risk in England: a cross-sectional, observational, whole-genome sequencing study. *Lancet Infect Dis* 2020; published online Jan 21. https://doi.org/10.1016/S1473-3099(19)30610-3.

Phylogenomic analysis of Neisseria gonorrhoeae transmission to assess sexual mixing and HIV transmission risk in England: a cross-sectional, observational, whole-genome sequencing study: Appendix

Methods

Whole genome sequencing

At the Wellcome Sanger Institute, WGS was conducted using the Illumina HiSeq X Ten system,[1,2] and put through the routine Sanger WGS data management pipeline.[3] The following measures were used to assess the quality of the WGS data for each isolate included in the phylogenetic analyses: a quality score >30 for the nucleotides called during the sequencing process, the majority of raw reads to be identified as *N. gonorrhoeae* when cross-referenced to a public database of pathogen genomes (Kraken),[4] the assembly length similar to the *N. gonorrhoeae* reference genome (FA1090)[5] – 2,153,922 nucleotides, the assembly guanine and cytosine content similar to the *N. gonorrhoeae* reference genome (FA1090) – 53%, a high percentage (>90%) of the reference genome covered by reads.

After passing quality control, the raw reads were aligned to the reference genome (FA1090) in order to create a consensus whole genome sequence for each isolate. We used the Burrows-Wheeler Aligner Maximal Exact Match (BWA-MEM) algorithm[6] with the option to flag duplicate shorter reads that match as secondary for removal (option –M). The Sequence Alignment/Map (SAM) file output was converted into a Binary Alignment/Map (BAM) file using SAMTools[7] in order to reduce the size of the file for faster computer processing. The Genome Analysis Toolkit (GATK)[8] was used to realign indels, which helps the process of identifying SNPs. SAMTools mpileup was used to identify the variant nucleotides identified in each read and the haploid option of Binary Call Format (BCF) tools from SAMTools filtered this information to select the variant nucleotides based on the following conditions: the minimum base call quality was ≥50 (quality of the base was previously determined using the Phred score system in SAMTools); the minimum mapping quality score by BWA-MEM was 20; at least eight reads have the same variant and at least three of these are from each strand direction (forward and back); that the specific variant called is the same in 80% of the reads used. The consensus sequence for each isolate was compiled into one multiple fasta file and used for the analyses.

Phylogenetic tree construction

Gubbins (Version 2.4.0)[9] was used with the default settings to remove regions of high SNP density that were potentially introduced by recombination (five iterations and a minimum number of three base substitutions to identify a recombination event) with the tree building option that uses Randomized Axelerated Maximum Likelihood (RAxML) (Version 8.2.8).[10] The detected recombination events were removed from the alignment together with the *opa* and *pil* genes, phages[11] and the Gonoccocal Genetic Island (GGI),[12] and SNP sites were obtained with snp-sites[13] and used to create a final phylogenetic tree with RAxML.

Data availability

Sequence data available on the European Nucleotide Archive using study accession ID: ERP022090 (https://www.ebi.ac.uk/ena/data/search?query=ERP022090)

Metadata for sequences available in the supplementary data table.

Results

SNP threshold sensitivity analysis

To assess if the clusters change when different single nucleotide polymorphism (SNP) thresholds were used, we conducted a sensitivity analysis using three SNP thresholds (≤3 (Table S2), ≤10 (Table S3), ≤14 (Table S4)) in addition to the ≤5 SNP threshold presented in the main manuscript (Table 1). As the SNP threshold increased, a higher percentage of isolates clustered, as expected. The distribution of cluster types by sexual risk (Figure S1) and HIV status (Figure S2) was similar for all SNP thresholds. The only notable difference was the slight increase in clusters containing isolates from women, heterosexual men and MSM as the SNP threshold increased.

Table S1 Epidemiological characteristics of the study sample compared to all gonorrhoea diagnoses in England during the study period (2013-2016)

	Study sample		**England***		**Two sample proportions z-test p-value****
	n	**%**	**n**	**%**	
Total	1277	100	**146,369**	**100**	
Year					
2013	326	25.5	31,213	21.3	**<0.001**
2014	333	26.1	37,178	25.4	0.580
2015	367	28.7	41,396	28.3	0.718
2016	251	19.7	36,582	25.0	**<0.001**
Geographical location					
London	572	44.8	72,809	49.7	**<0.001**
Outside London	705	55.2	73,560	50.2	**<0.001**
Gender & sexual risk					
MSM	766	60.0	72,660	49.6	**<0.001**
Heterosexual men	304	23.8	34,330	23.5	0.768
Women	206	16.1	36,178	24.7	**<0.001**
Missing	1	<0.1	3,201	2.2	N/A
Age group (years)					
≤24	384	30.1	55,029	37.6	**<0.001**
25-34	503	39.4	54,143	37.0	0.077
≥35	390	30.5	37,197	25.4	**<0.001**
Ethnicity					
White	824	64.5	104,028	71.1	**<0.001**
Black Caribbean	132	10.3	8,280	5.7	**<0.001**
Black African	47	3.7	5,858	4.0	0.559
Black Other	10	0.8	3,238	2.2	**<0.001**
Asian	74	5.8	5,750	3.9	**0.026**
Other	32	2.5	4,747	3.2	0.138
Mixed	105	8.2	8,614	5.9	**<0.001**
Missing	53	4.2	5,815	4.0	0.747
Country of birth					
UK	782	61.2	96,189	65.7	**<0.001**
Not UK	407	31.9	38,334	26.2	**<0.001**
Missing	88	6.8	11,846	8.1	0.117
Diagnoses with a new STI (excluding HIV) in the past year					
No/Unknown	1,015	79.5	117,493	80.3	0.481
Yes	262	20.5	28,876	19.7	0.481
HIV status					
Negative/Unknown	1,051	82.3	130,198	89.0	**<0.001**
Positive	226	17.7	16,171	11.0	**<0.001**

Table S2 Number of clusters and number of people within each cluster type stratified by sexual risk of the person providing the isolate (SNP threshold of ≤3)

Cluster description	Number of Clusters (col %)	Number of isolates in cluster by patient sexual risk (N=523)		
		Women N (col %)	**Het. men N (col %)**	**MSM N (col %)**
Total	**191 (100)**	**72 (100)**	**103 (100)**	**348 (100)**
Only women	8 (4.2)	17 (23.6)	-	-
Only het. men	9 (4.2)	-	19 (18.5)	-
Only MSM	100 (52.4)	-	-	281 (80.8)
Only women & het. men	43 (22.5)	50 (69.4)	55 (53.4)	-
Only women & MSM	3 (1.6)	3 (4.2)	-	3 (0.9)
Only het. men & MSM	26 (13.6)	-	27 (26.2)	60 (17.2)
Women, het. men & MSM	2 (1.)	2 (2.8)	2 (1.9)	4 (1.2)

Table S3 Number of clusters and number of people within each cluster type stratified by sexual risk of the person providing the isolate (SNP threshold of ≤10)

Cluster description	Number of Clusters (col %)	Number of isolates in cluster by patient sexual risk (N=786)		
		Women N (col %)	**Het. men N (col %)**	**MSM N (col %)**
Total	**210 (100)**	**119 (100)**	**164 (100)**	**503 (100)**
Only women	4 (1.9)	12 (10.1)	-	-
Only het. men	9 (4.3)	-	19 (11.6)	-
Only MSM	96 (45.7)	-	-	315 (63.2)
Only women & het. men	57 (27.0)	86 (72.3)	82 (50.0)	-
Only women & MSM	3 (1.4)	3 (2.5)	-	19 (3.8)
Only het. men & MSM	34 (16.2)	-	44 (26.8)	132 (26.2)
Women, het. men & MSM	7 (3.3)	18 (15.1)	19 (11.6)	34 (6.8)

Table S4 Number of clusters and number of people within each cluster type stratified by sexual risk of the person providing the isolate (SNP threshold of ≤14)

Cluster description	Number of Clusters (col %)	Number of isolates in cluster by patient sexual risk (N=853)		
		Women N (col %)	**Het. men N (col %)**	**MSM N (col %)**
Total	**201 (100)**	129 (100)	183 (100)	541 (100)
Only women	4 (2.0)	10 (7.7)	-	-
Only het. men	8 (4.0)	-	18 (9.8)	-
Only MSM	89 (44.3)	-	-	302 (55.8)
Only women & het. men	57 (28.3)	93 (72.1)	95 (51.9)	-
Only women & MSM	3 (1.5)	3 (2.3)	-	2 (2.3)
Only het. men & MSM	29 (14.4)	-	36 (19.7)	148 (27.4)
Women, het. men & MSM	11 (5.5)	23 (17.8)	34 (18.6)	70 (12.9)

Figure S1 Comparison of SNP cut-off for cluster definition by sexual risk cluster type

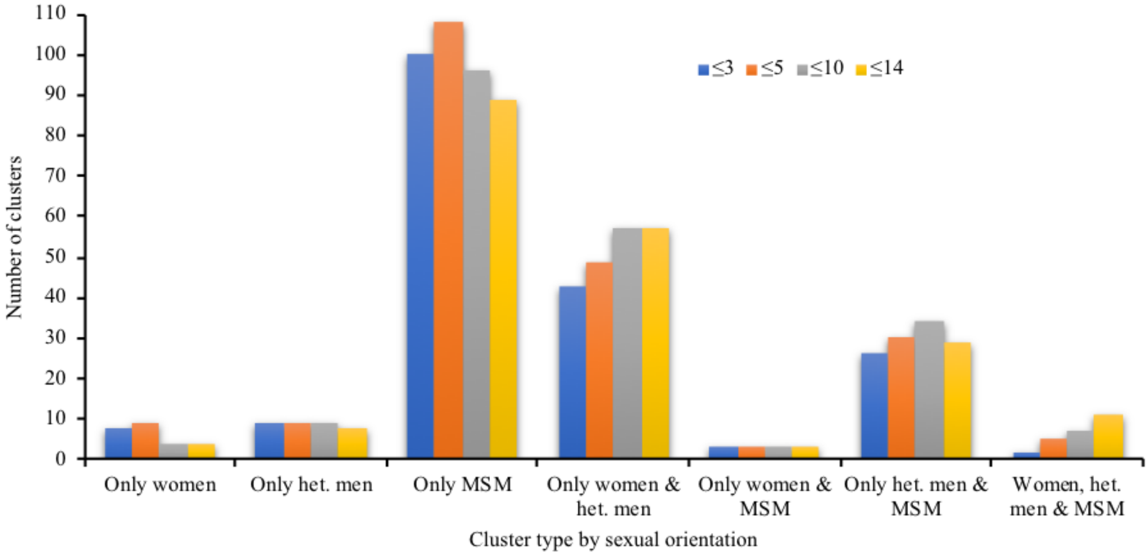

Figure S2 Comparison of SNP cut-off for cluster definition by HIV status cluster type

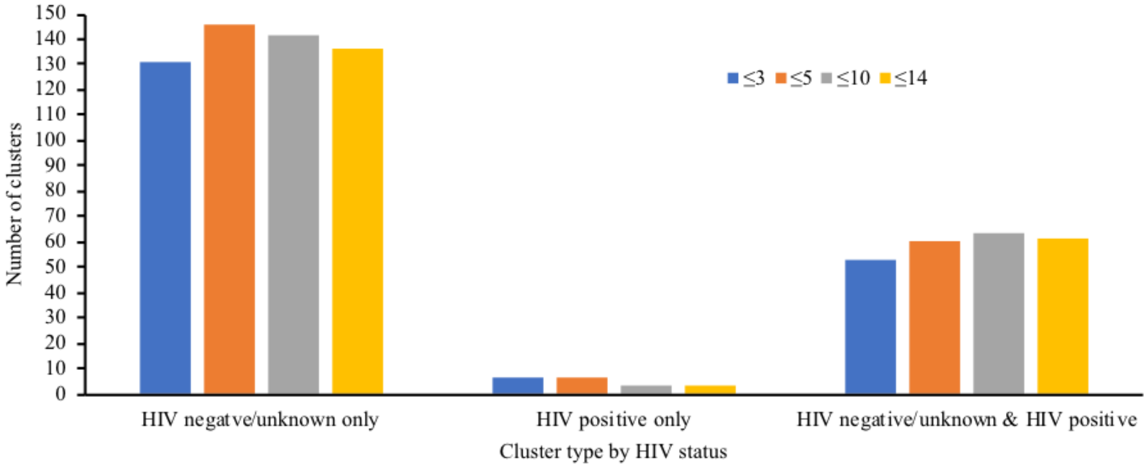

Table S5 Description of isolates in the two largest clusters identified in the study sample

	Cluster N=21		Cluster N=11	
	N	**%**	**N**	**%**
Total	**21**	**100**	**11**	**100**
Year				
2013	0	0.0	4	36.4
2014	14	66.7	0	0.0
2015	7	33.3	4	36.4
2016	0	0.0	3	27.3
Clinic				
Outside London	20	95.2	7	63.6
London	1	4.8	4	36.4
Sexual risk				
Heterosexual men	3	14.3	3	27.3
MSM	18	85.7	8	72.7
Age (years)				
≤24	2	9.5	2	18.2
25-34	9	42.	5	45.4
≥35	10	47.6	4	36.2
Ethnicity				
White	16	76.2	8	72.7
Black Caribbean	3	14.3	1	9.1
Black African	1	4.8	0	0.0
Black Other	0	0.0	0	0.0
Asian	0	0.0	1	9.1
Other	0	0.0	0	0.0
Mixed	1	4.8	1	9.1
Country of birth				
UK	18	85.7	7	63.6
Not UK	3	14.3	4	36.4
Number of sexual partners in the UK in the three months prior to diagnosis				
0	6	28.6	4	36.4
1	9	42.9	4	36.4
≥2	4	19.0	0	0.0
Unknown	2	9.5	3	27.3
Symptoms				
No	5	23.8	3	27.3
Yes	12	57.1	7	63.6
Unknown	4	19.0	1	9.1
Diagnosed with an STI (excluding HIV) in the year prior to gonorrhoea diagnosis				
No/Unknown	16	76.2	8	72.7
Yes	5	23.8	3	27.3
HIV status				
Negative/Unknown	13	61.9	9	81.8
Positive	8	38.1	2	18.2
Travel-associated sexual partnership in the three months prior to diagnosis				
No	18	85.7	7	63.6
Yes	1	4.8	1	9.0
Unknown	2	9.5	3	27.3

MSM = men who reported sex with men

Table S6 Comparison of epidemiological characteristics of isolates from heterosexual men that clustered with isolates from women only or isolates from MSM only

	Heterosexual men clustered only with isolates from women		Heterosexual men clustered only with isolates from MSM		P Value
	N	%	N	%	
Total	**63**	**100.0**	**36**	**100.0**	
Year					
2013	17	27.0	6	16.7	0.578
2014	16	25.4	8	22.2	
2015	17	27.0	13	36.1	
2016	13	20.6	9	25.0	
Clinic					
Outside London	58	92.1	30	83.3	0.319
London	5	7.9	6	16.7	
Age (years)					
≤24	29	46.0	8	22.2	0.06
25-34	23	36.5	18	50.0	
≥35	11	17.5	10	27.8	
Ethnicity					
White	28	46.7	17	51.5	0.252*
Black Caribbean	17	28.3	3	9.1	
Black African	2	3.3	3	9.1	
Black Other	1	1.7	0	0	
Asian	5	8.3	5	15.2	
Other	2	3.3	1	3.0	
Mixed	5	8.3	4	12.1	
Country of birth					
UK	46	73.0	24	66.7	0.516
Not UK	14	22.2	10	27.8	
Symptoms					
No	10	16.1	6	17.6	0.849
Yes	52	83.9	28	82.4	
Diagnosed with an STI (excluding HIV) in the year prior to gonorrhoea diagnosis					
No/Unknown	57	90.5	30	83.3	0.345*
Yes	6	9.5	6	16.7	
HIV status					
Negative/Unknown	63	100.0	36	100.0	N/A
Positive	0	0.0	0	0.0	
Number of sexual partners in the UK in the three months prior to diagnosis					
0	4	7.0	1	3.2	0.618*
1	24	42.1	11	35.5	
≥2	29	50.9	19	61.3	
Travel-associated sexual partnership in the three months prior to diagnosis					
No	50	87.7	30	96.8	0.251*
Yes	7	12.3	1	3.2	

* Fisher’s Exact test used instead of Chi^2 test, MSM = men who reported sex with men

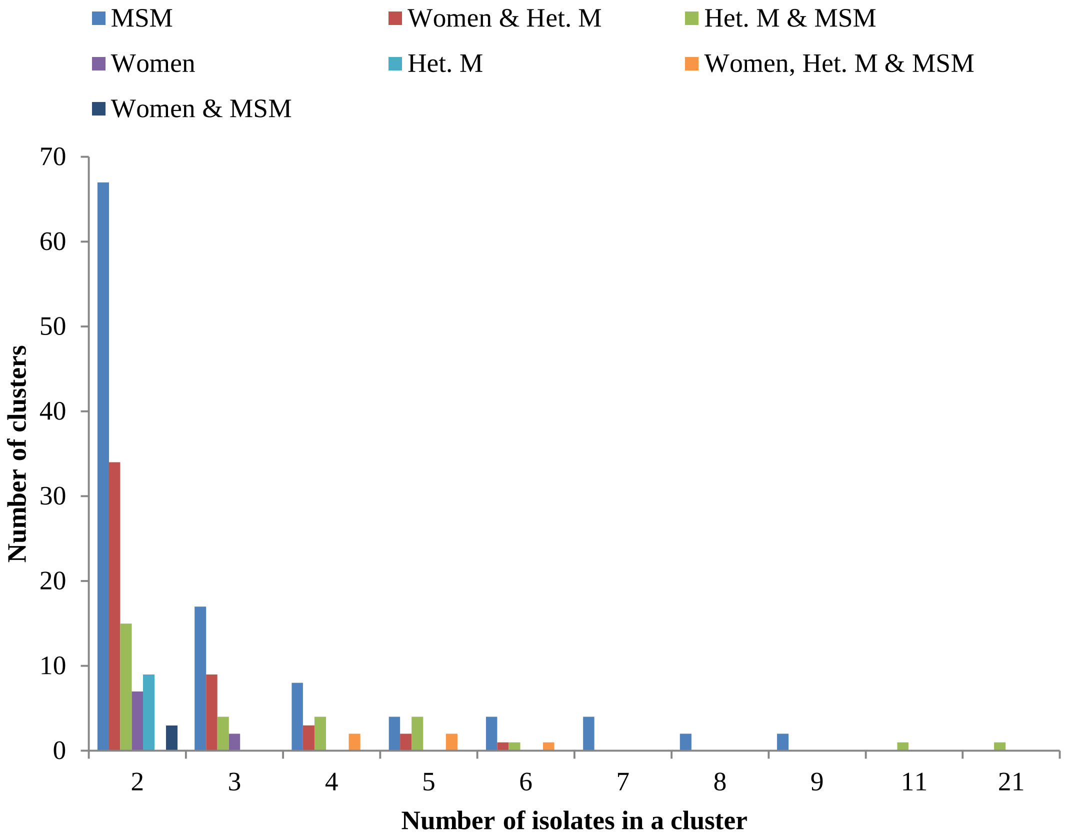

Figure S3 Size and frequency of *N. gonorrhoeae* clusters by sexual risk

Clusters defined by SNP difference threshold ≤5. Het. M = heterosexual men (men who reported sexual activity exclusively with women), MSM = men who reported sex with men

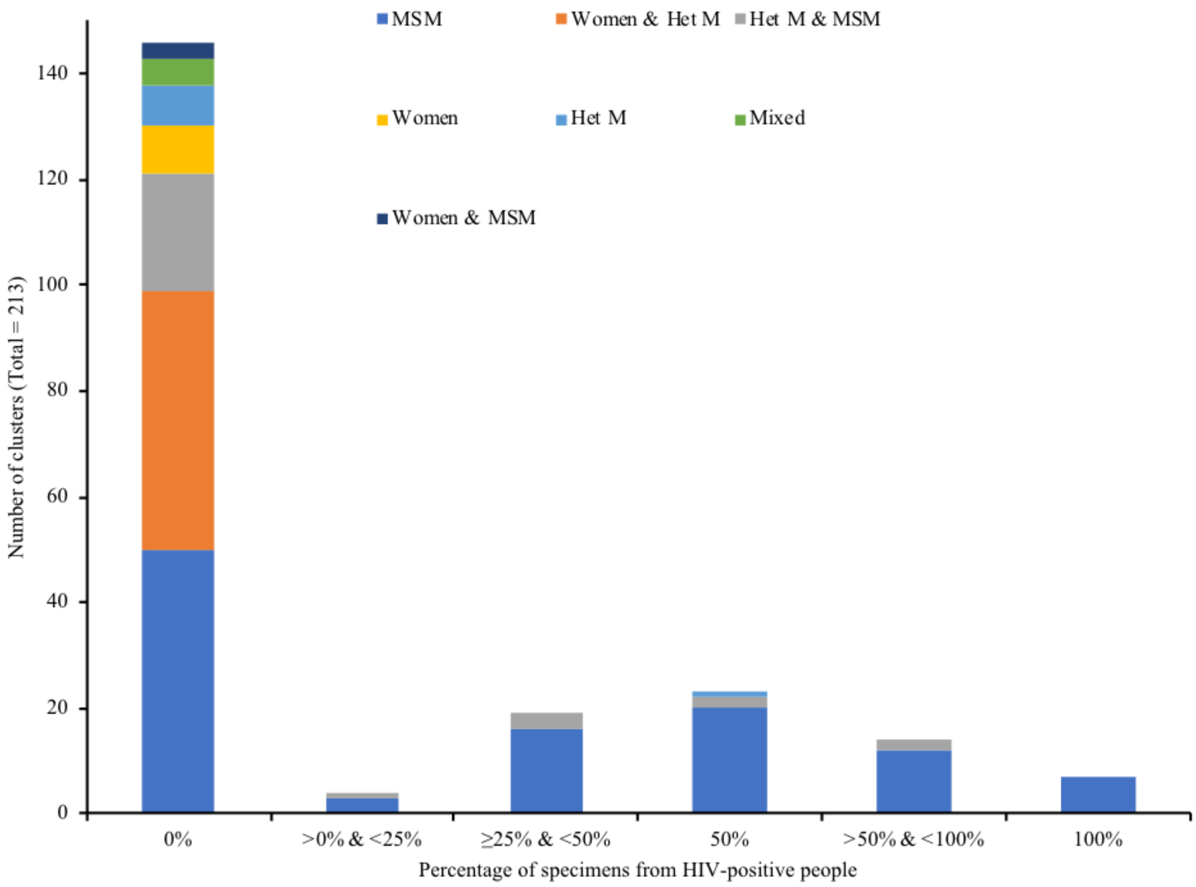

Figure S4 Number of clusters by percentage of isolates from HIV-positive people in each cluster and sexual risk of cluster

Het. men = heterosexual men (men who reported sexual activity exclusively with women), MSM = men who reported sex with men

STROME-ID Checklist

Manuscript section	**Item Number**	**STROME-ID item**	**Manuscript Page**
Title and abstract			
Introduction	1.1	The term molecular epidemiology should be applied to the study in the title or abstract and the keywords when molecular and epidemiological methods contribute substantially to the study	4
Background rationale	2.1	Provide background information about the pathogen population and the distribution of pathogen strains within the host population at risk	5, 6, 9
Objectives	3.1	State the epidemiological objectives of using molecular typing	5
Methods			
Molecular terminology	4.1	Define or cite definitions for key molecular terms used within the study (eg, strain, isolate, and clone)	6, 7
Molecular markers	4.2	Clearly define the molecular markers that were used with a standard nomenclature	7
Infectious disease case definition	4.3	Clearly state the infectious-disease case definitions	6
Laboratory methodology	4.4	Describe sample collection and laboratory methods, including any methods used to minimise and measure cross-contamination, and give the criteria used to interpret strain classification	6, appendix page 1
Setting	5.1	Clearly state the timeframe of the study; consider and appropriately reference the molecular clock of markers if known, and the natural history of the infection	6
Participants	6.1	State the source of participants and clinical specimens, and clearly describe sampling frame and strategy	6
Multiple-strain infections	8.1	Describe any methods used to detect multiple-strain infections and measure their effect on the study findings	N/A
Bias	9.1	Describe any efforts made to address discovery or ascertainment bias	6, 7
Study size	10.1	Describe any unique restrictions placed on the study sample size	6
Statistical methods	12.1	State how the study took account of the non- independence of sample data, if appropriate	N/A
	12.2	State how the study dealt with missing data	9
Results			
Participants	13.1	Report numbers of participants and samples at each stage of the study, including the number of samples obtained, the number typed, and the number yielding data	8, 9, 10
	13.2	If the study investigates groups of genetically indistinguishable pathogens (molecular clusters), state the sampling fraction, the distribution of cluster sizes, and the study population turnover, if known	8, 9
Descriptive data	14.1	Give information by strain type if appropriate, with use of standardised nomenclature	N/A
Main results	16.1	Consider showing molecular relatedness of strain types by means of a dendrogram or phylogenetic tree	Figure 1
Discussion			
Limitations	19.1	Consider alternative explanations for findings when transmission chains are being investigated, and report the consistency between molecular and epidemiological evidence	12, 13
Other information			
Ethics	23.1	Report any ethical considerations with specific implications for infectious-disease molecular epidemiology	6, 14

N/A – not applicable

References

1. Illumina. An introduction to next-generation sequencing technology.
2. Illumina. Specification Sheet: Sequencing - HiSeq X Series of Sequencing Systems. 2018.
3. Page AJ, De Silva N, Hunt M, et al. Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. *Microb Genom* 2016; 2(8): e000083.
4. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014; 15(3): R46.
5. NCBI. Genome Assembly and Annotation report: *Neisseria gonorrhoeae* FA 1090. https://www.ncbi.nlm.nih.gov/genome/864?genome_assembly_id=300414 (accessed 20/11/2018 2018).
6. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* 2013; 1303(3997).
7. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; 25(16): 2078-9.
8. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010; 20(9): 1297-303.
9. Croucher NJ, Page AJ, Connor TR, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* 2015; 43(3): e15.
10. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014; 30(9): 1312-3.
11. Piekarowicz A, Klyz A, Majchrzak M, Adamczyk-Poplawska M, Maugel TK, Stein DC. Characterization of the dsDNA prophage sequences in the genome of Neisseria gonorrhoeae and visualization of productive bacteriophage. *BMC Microbiol* 2007; 7: 66.
12. Harrison OB, Clemence M, Dillard JP, et al. Genomic analyses of Neisseria gonorrhoeae reveal an association of the gonococcal genetic island with antimicrobial resistance. *J Infect* 2016; 73(6): 578-87.
13. Page AJ, Taylor B, Delaney AJ, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microbial Genomics*. 2016 2(4)