Application of High-Throughput Sequencing to the Genomic Epidemiology of Bacterial Pathogens

Trabajo realizado por

Carlos Francés Cuesta

Para optar al título de Doctor por la Universidad de Valencia

Director:

Fernando González Candelas

VNIVERSITAT DE VALÈNCIA

Enero, 2021

D. **Fernando González Candelas**, Doctor en Ciencias Biológicas y Catedrático del Departamento de Genética de la Universidad de Valencia.

CERTIFICA

Que D. **Carlos Francés Cuesta**, Licenciado en Biología por la Universidad de Valencia, ha realizado bajo mi dirección el trabajo titulado:

"Application of High-Throughput Sequencing to the Genomic Epidemiology of Bacterial Pathogens", para optar al Grado de Doctor por la Universidad de Valencia.

Y para que conste, en el cumplimiento de la legislación vigente, firmo el presente certificado en Valencia, a 14 de enero de 2021.

Fdo.: Dr. Fernando González Candelas

A mi familia

DECLARACIÓN

El trabajo presentado en esta tesis ha sido desarrollado en la Unidad Mixta de Investigación “Infección y Salud Pública” FISABIO-Universidad de Valencia, Instituto de Biología Integrativa de Sistemas (CSIC-UV), ubicada en el área Genómica y Salud de FISABIO.

Este trabajo no hubiera sido posible sin la “ayuda para contratos predoctorales para la formación de doctores 2015” del Ministerio de Ciencia, Innovación y Universidades [BES-2015-074204].

Los fondos proceden de los proyectos [BFU2014-58656-R] y [BFU2017-89594-R], del Ministerio de Ciencia, Innovación y Universidades, y del proyecto [Prometeo2016/122] de la Generalitat Valenciana. Los fondos para las infraestructuras e instrumental para el Servicio de Secuenciación de FISABIO proceden de los fondos europeos FEDER.

AGRADECIMIENTOS

Parece que fue ayer cuando inicié esta aventura, con la ilusión de quien por fin logra lo que quería y el temor a la novedad. Porque todo era nuevo para mí, desde los conceptos más básicos de la genómica, hasta la negra terminal de los sistemas Linux (y lo que me costó pillar que un directorio no era más que las dichosas carpetas de toda la vida, así era el nivel). Desde entonces han pasado unos cuantos años y muchas personas a las que se les debe un espacio en este libro.

En primer lugar, gracias al grupo EpiMol, quien ha sido mi segunda familia durante todos estos años. Gracias a Paula, por su ayuda incondicional desde que entré en el equipo, muchas de las cosas aprendidas te las debo a ti, gracias también por tu amistad. Gracias a “las niñas” (ya no tan niñas), Neris, Marta y Bea, por vuestra simpatía y vuestra inestimable ayuda. Gracias también a Iván, por las conversaciones a veces surrealistas que teníamos en las comidas, y por estar siempre dispuesto a ayudar en los muchos problemas que he tenido a la hora de “programar” (sigo pensando que haces magia). Gracias también al resto del grupo, por haber sido parte de mi familia científica.

Gracias especialmente a Fernando por haberme dado la oportunidad de desarrollarme como científico abriéndome las puertas de su grupo. Por su paciencia y por el hecho de tener siempre la puerta de su despacho abierta a pesar de la cantidad de faena.

Retrocediendo un poco en el tiempo, he de agradecer también al grupo de Tuberculosis, quien fue mi primera familia científica en FISABIO. Especialmente a Iñaki, por darme la oportunidad de entrar de pleno en el mundo de la genómica microbiana y por su amabilidad siempre que he requerido algo de él. También a sus chicos, por ayudarme a dar mis primeros pasos en el mundillo de la terminal.

Gracias a todos los compañeros de FISABIO, quienes habéis hecho más ameno el transcurso de esta aventura. Si me pusiera a escribir nombres no terminaría. Tanto a los que ya acabasteis vuestras aventuras particulares como a los que las estáis empezando, gracias por los momentos que me habéis regalado.

Agradecer profundamente a todas las personas que han colaborado en el desarrollo de esta tesis. A algunos os he llegado a conocer en persona y a otros no, pero el agradecimiento es igual de sincero. Sin vosotros no hay muestras y sin muestras no hay trabajo. Gracias.

Gracias a Leo, quien ha estado dispuesta a echarme una mano desde el principio y desde la distancia, aunque ahora ya no es tanta (welcome back).

Gracias a la Prof. Verena Schuenemann, por acogerme en su grupo y permitirme volver de nuevo, temporalmente, al wet lab. Pero, sobre todo, gracias a Natasha, por haberse esforzado al máximo en hacer mi estancia en Zúrich lo más agradable posible, en unos pocos meses te convertiste en una gran amistad.

Gracias a mis amigos, por estar siempre ahí y por todos los buenos momentos, que han sido y son muchos.

Un millón de gracias a mis padres y a mi hermano, porque a vosotros os lo debo todo. Gracias por permitir vivir la vida a mi manera, por respetar mis decisiones y por animarme en todos mis proyectos. Os quiero.

Finalmente, gracias a Lidia, por tu amor, por tu apoyo incondicional todos estos años. Sin tu forma de ver la vida, y, en ocasiones, sin tu testarudez, muchas cosas un hubieran sido lo mismo. Gracias por permanecer siempre a mi lado. Sigamos avanzando juntos en nuestra aventura particular.

CONTENTS

RESUMEN

1. Introducción

1.1. Los patógenos bacterianos como amenaza para la Salud Pública

Las bacterias causantes de enfermedades infecciosas pueden ser patógenos obligados, bacterias que han evolucionado para adaptarse a un medio —tan nutritivo como hostil— como es el organismo hospedador, o pueden ser patógenos oportunistas, bacterias comensales o ambientales que pueden provocar una enfermedad infecciosa cuando el sistema inmune del hospedador se ve comprometido por otros factores.

Las enfermedades infecciosas suponen una enorme carga para los sistemas de Salud Pública, siendo causa del 26% de las muertes a nivel global en 2019, según datos de la Organización Mundial de la Salud (OMS). Esto pone de manifiesto que este tipo de enfermedades constituye una de las principales preocupaciones para las agencias y sistemas de Salud Pública.

Algunas de las principales amenazas para la Salud Pública debidas a agentes infecciosos son:

- **Infecciones de transmisión sexual (ITS).** Son infecciones adquiridas a través del contacto sexual con una persona infectada. En este grupo también están incluidas las infecciones congénitas. Según la OMS, cada día se adquiere más de 1 millón de ITS a nivel global, siendo mayoritarias la tricomoniasis, clamidiasis, gonorrea y sífilis, estas tres últimas causadas por bacterias. Estas infecciones pueden llegar a causar la muerte, infertilidad, lesiones que incrementan el riesgo de una infección por VIH, afecciones en el feto que pueden incluso llevar al aborto, desarrollo de algunos tipos de cáncer, y un fuerte estigma social. Todo el mundo está expuesto a este tipo de infecciones, aunque hay poblaciones más vulnerables debido a que están vinculadas a

factores de riesgo que favorecen la transmisión de ITS. La prevención de las ITS pasa por la educación sexual, la reducción de parejas sexuales y de actividades de riesgo y el uso de preservativos.

- **Infecciones nosocomiales.** Son infecciones adquiridas durante procedimientos de cuidado de la salud en hospitales, unidades de cuidados intensivos (UCIs) u otras instalaciones sanitarias. El espectro clínico es amplio, incluyendo infecciones urinarias, neumonías y sepsis, entre otras. Aunque no hay recuentos globales, se estima que millones de personas se ven afectadas por estas infecciones, especialmente en países subdesarrollados. Todo paciente es susceptible de adquirir una infección nosocomial, pero ancianos, neonatos y pacientes inmunodeprimidos tienen mayor riesgo. Los principales agentes causantes —aunque no todos— son patógenos oportunistas que colonizan instrumental médico y otras superficies. Las medidas de prevención incluyen la higiene de manos de los trabajadores sanitarios, la limpieza de superficies y la reducción del número de trabajadores en contacto con el paciente.
- **Enfermedades infecciosas emergentes.** Son aquellas infecciones que han aparecido recientemente en una población o bien existían previamente pero su incidencia se ha incrementado rápidamente. Muchas de estas infecciones son zoonosis o son transmitidas por un vector —generalmente un artrópodo—. Algunos de los factores que contribuyen a la emergencia o re-emergencia de una enfermedad infecciosa son los cambios ecológicos, la demografía humana y globalización, la adaptación y cambios en los microorganismos o la relajación de las medidas implementadas en Salud Pública. Los brotes producidos por este tipo de infecciones suelen producir una elevada mortalidad; muchas de estas infecciones no tienen cura y a menudo los sanitarios son víctimas de estas enfermedades.
- **Resistencia a los antibióticos.** Sucede cuando los microorganismos evolucionan y dejan de responder a los fármacos, lo que dificulta o imposibilita el tratamiento de las infecciones que producen. El uso abusivo y descontrolado de antibióticos, tanto en medicina como en agricultura y ganadería, favorecen esta evolución y la consecuente dispersión de las cepas microbianas resistentes en el medio ambiente

o en las comunidades. La emergencia de la resistencia a antibióticos constituye uno de los problemas más serios para la Salud Pública en la actualidad.

1.2. Secuenciación de alto rendimiento aplicada a la vigilancia de patógenos: Epidemiología Genómica

Aunque la revolución genómica es un fenómeno relativamente reciente, su desarrollo empezó hace décadas, en los años 70, con el método por terminación de la cadena de Sanger —también conocido como secuenciación Sanger—, con el que se secuenció el genoma del bacteriófago ΦX174. En los años 90 se desarrolló un método consistente en secuenciar numerosas secuencias que se superponían, permitiendo la secuenciación del primer genoma bacteriano, *Haemophilus influenzae*. Ya en la primera década del siglo XXI, se desarrollaron tecnologías de segunda generación que permiten paralelizar reacciones de secuenciación de forma masiva, aumentando el rendimiento y reduciendo los costes en comparación con las tecnologías anteriores. Estas tecnologías trabajan sobre un genoma fragmentado y generan millones de secuencias de 100-150 nucleótidos que serán empleadas para reconstruir el genoma secuenciado. Sin embargo, presentan algunas limitaciones como, por ejemplo, una alta tasa de error en los extremos de las secuencias. Recientemente, se han desarrollado tecnologías de tercera generación que generan secuencias más largas que las de segunda generación, pero que presentan tasas más elevadas de error, por lo que estas tecnologías requieren un mayor desarrollo.

Las tecnologías de secuenciación de segunda y tercera generación se conocen como secuenciación de alto rendimiento —o secuenciación masiva— y han alcanzado una gran implementación en el campo de la genómica microbiana. Sin embargo, aún está en etapas tempranas de implementación en la práctica clínica y epidemiológica. Su empleo en el análisis de brotes y en vigilancia epidemiológica y de resistencias a antibióticos tiene un gran potencial. Sin embargo, su implementación dependerá de la resolución de los obstáculos que presenta y del beneficio que aporten.

1.3. Patógenos bacterianos estudiados en esta tesis

En esta tesis se van a explorar los problemas para la Salud Pública introducidos anteriormente, mediante cuatro ejemplos de implementación de la tecnología de secuenciación de alto rendimiento al estudio de tres patógenos bacterianos.

- ***Neisseria gonorrhoeae.*** Es un patógeno obligado cuyo único hospedador natural es el ser humano. Posee un genoma de unas 2.22 Mb y puede contener distintos tipos de plásmidos. El plásmido críptico no confiere fenotipo de virulencia o resistencia, el plásmido asiático y sus variantes generadas por inserción o deleción de fragmentos de ADN contienen el gen bla_{TEM} que codifica una β-lactamasa que confiere resistencia a penicilinas y cefalosporinas, y el plásmido conjugativo cuyas variantes americana y holandesa poseen el gen *tetM* que confiere resistencia a las tetraciclinas. Esta bacteria es el agente causal de la gonorrea, la segunda ITS bacteriana más común. Sus síntomas incluyen uretritis, cervicitis, gonorrea orofaríngea y rectal, y diversas complicaciones, como endometritis, enfermedad inflamatoria pélvica, epididimitis, *ophthalmia neonatorum*, artritis, endocarditis y meningitis. El tratamiento recomendado es la terapia dual con una cefalosporina de espectro extendido y azitromicina. Sin embargo, esta bacteria ha ido desarrollando resistencia a cada antibiótico que se ha usado para tratar sus infecciones. Por otro lado, su incidencia se ha ido incrementando a nivel global en los últimos años, por lo que, junto a su creciente adquisición de resistencias a distintos antibióticos, ha sido declarada como una amenaza para la Salud Pública y un patógeno prioritario con elevada necesidad de desarrollo de nuevos antibióticos frente a él.
- ***Serratia marcescens.*** Es una enterobacteria de vida libre con un genoma de unas 5.3 Mb. Es muy ubicua, pudiéndose encontrar en suelos, agua, alimentos y animales. Originalmente no se la consideraba patógena, pero desde la década de los 50 se ha asociado a diversos brotes nosocomiales, por lo que se le considera un patógeno oportunista. Las infecciones por esta bacteria tienen un espectro

clínico muy amplio, pudiendo producir infecciones urinarias, en piel y tejidos blandos, conjuntivitis, neumonía, sepsis, endocarditis y meningitis. La mayoría de los brotes declarados implican UCIs neonatales, debido a los múltiples factores de riesgo que poseen los pacientes ingresados. El tratamiento habitual incluye varios antibióticos en función del lugar donde se localice la infección y del perfil de sensibilidad a antibióticos. Esta bacteria pertenece a un grupo de bacterias que poseen una β-lactamasa cromosómica inducible que les confiere resistencia a múltiples antibióticos. Sin embargo, *S. marcescens* no suele causar brotes complicados.

- ***Lactococcus garvieae.*** Es una bacteria de vida libre, muy ubicua, que se puede encontrar en agua, suelo, vegetales, piel de animales, leche y otros productos lácteos. Su genoma posee unas 2.07 Mb. *L. garvieae* es un patógeno de peces que causa importantes pérdidas en acuicultura, aunque también se ha aislado en infecciones en otros animales de ganadería. *L. garvieae* es un patógeno emergente en humanos con un amplio espectro clínico, sin embargo, el más frecuente es la endocarditis infecciosa. La fuente principal de contagio es a través de la ingesta de pescado o productos lácteos contaminados. El tratamiento puede incluir diversos fármacos y dependerá del lugar de la infección y del resultado del antibiograma, aunque las endocarditis se suelen tratar con amoxicilina y gentamicina.

2. Objetivos

El objetivo principal es estudiar la epidemiología genómica de diversos patógenos bacterianos mediante la obtención de sus genomas completos empleando la secuenciación de alto rendimiento, evaluando la utilidad de esta herramienta para su aplicación en epidemiología y microbiología forense.

De este objetivo se derivan una serie de objetivos secundarios:

- Estudiar la variabilidad genómica de *Neisseria gonorrhoeae* en muestras clínicas procedentes de la Comunidad Valenciana, Cataluña y Madrid.
- Definir la estructura poblacional y las dinámicas de transmisión espacio-temporal de *N. gonorrhoeae* en dichas regiones.
- Comparar la secuenciación de alto rendimiento y los esquemas de tipado de *N. gonorrhoeae* evaluando los resultados de ambos tipos de datos.
- Analizar los eventos de recombinación entre los distintos grupos estructurales de *N. gonorrhoeae*.
- Detectar y evaluar los determinantes de resistencia a antibióticos contrastando con los datos fenotípicos en *N. gonorrhoeae*.
- Aplicar el análisis de secuencias obtenidas por secuenciación de alto rendimiento para la resolución de un caso forense de transmisión de gonorrea a una menor en un supuesto caso de abuso sexual.
- Analizar dos brotes nosocomiales de *Serratia marcescens* en las UCIs neonatales de dos hospitales de la Comunidad Valenciana.
- Evaluar el impacto de elegir distintos genomas de referencia para *S. marcescens* en la interpretación de resultados de los análisis de brotes.
- Estudiar el pangenoma de *Lactococcus garvieae* y detectar genes recombinantes en una cepa clínica procedente de un caso fatal de endocarditis a tres niveles taxonómicos —especie, género y clase—.

3. Métodos

3.1. Procesamiento de muestras y generación de metadatos

El muestreo fue realizado por los hospitales colaboradores. Las muestras de gonococo se cultivaron en agar chocolate PoliVyteX o en agar Thayer-Martin a 35-37ºC en una atmósfera enriquecida en CO_2 al 5%. Las muestras de *Serratia marcescens* se cultivaron en agar chocolate PoliVyteX o en CHROMagar Orientation Medium a 37ºC. La cepa de *Lactococcus garvieae* se aisló post-mortem en caldo de cultivo *brain heart infussion* (BHI). En los tres casos, los aislados fueron identificados mediante espectrometría de masas (MALDI-TOF). Las tres especies bacterianas fueron sometidas a antibiograma para distintos antibióticos dependiendo de la especie y del hospital colaborador. Para gonococo y *S. marcescens* se usó el sistema Etest y los resultados se interpretaron siguiendo las recomendaciones de EUCAST, mientras que para *L. garvieae* se empleó el método de difusión en disco y se usaron los puntos de corte recomendados por la Sociedad Francesa de Microbiología para testar Gram-positivos.

El ADN se extrajo por diversos métodos —por golpe térmico, mediante kits de extracción o empleando sistemas automáticos— dependiendo del hospital colaborador. El ADN fue enviado a nuestro laboratorio, donde se procedió a la cuantificación por fluorometría.

Las muestras de gonococo y *S. marcescens* se secuenciaron empleando la plataforma Illumina NextSeq 500 que genera secuencias emparejadas de 150 nucleótidos. La cepa de *L. garvieae* se secuenció empleando la plataforma PacBio RS II que genera secuencias de entre 15 y 25 Kb.

3.2. Análisis primario

El análisis primario implica el control de calidad de las secuencias obtenidas por secuenciación de alto rendimiento. Primero se inspeccionó la calidad de dichas secuencias empleando los programas FastQC y MultiQC. Posteriormente se filtraron las secuencias de baja calidad empleando el programa PRINSEQ-lite.

En el caso de *S. marcescens* se realizó la identificación *in silico* de la especie empleando los programas Kraken y Mash Screen. Este paso adicional se llevó a cabo porque dos de los aislados no cubrían el genoma de referencia, lo que nos llevó a sospechar que no pertenecían a la especie de estudio.

3.3. Análisis secundario

Este análisis reordena las secuencias obtenidas con el fin de reconstruir el genoma de las cepas de estudio. Hay dos estrategias: mapeo y ensamblado *de novo*.

Los gonococos del Capítulo 1 se mapearon frente a la cepa de referencia FA1090, que es la cepa representativa para esta especie, mientras que los del Capítulo 2 se mapearon frente a la cepa WHO P porque se identificó mediante el programa kmerID como la más cercana genéticamente a los aislados de estudio. En el caso de *S. marcescens*, los aislados se mapearon inicialmente frente a la cepa más cercana genéticamente (UMH9) y posteriormente frente a la cepa representativa de la especie (Db11). En todos los casos, el mapeo se realizó empleando el programa BWA-MEM. Tras el mapeo, se empleó SAMtools y BCFtools para detectar los polimorfismos y generar un fichero de alineamiento con todos los genomas de estudio y la referencia empleada.

Adicionalmente, los genomas de los gonococos del Capítulo 1 se ensamblaron *de novo* utilizando SPAdes. Los *contigs* resultantes se emplearon para curar los resultados ambiguos del tipado y para la detección de determinantes de resistencia.

En el caso de los gonococos del Capítulo 2 y los aislados de *S. marcescens* se extrajeron las secuencias no mapeadas mediante SAMtools y se ensamblaron con SPAdes, de modo que también se pudo analizar la fracción genómica no mapeada con el fin de detectar diferencias entre los aislados.

Finalmente, el genoma de la cepa de *L. garvieae* únicamente fue ensamblado empleando Celera WGS-assembler. Los dos *contigs* resultantes correspondieron con el cromosoma y el plásmido, y fueron circularizados mediante Circlator.

3.4. Análisis terciario

Comprende todos los análisis que dan significado biológico a los datos.

- **Tipado.** Los gonococos (Capítulos 1 y 2) poseen tres esquemas de genotipado —MLST, NG-MAST y NG-STAR—. Los tres esquemas fueron empleados para tipar los gonococos del Capítulo 1, mientras que los del Capítulo 2 se tiparon con el esquema MLST. Para ello se empleó el programa SRST2. Los resultados ambiguos se confirmaron utilizando BLAST sobre los *contigs* ensamblados. Como *S. marcescens* (Capítulo 3) y *L. garvieae* (Capítulo 4) no poseen esquemas de genotipado, este paso no se realizó con dichas especies.
- **Análisis filogenético.** Los alineamientos resultantes tras el mapeo y la detección de polimorfismos se emplearon para la reconstrucción filogenética mediante IQ-TREE. En los Capítulos 2 y 3, se requirió computar la matriz de distancias genéticas empleando el paquete APE de R.
- **Estructura genética poblacional.** En el Capítulo 1 se analizó la estructura de la población gonocócica empleando STRUCTURE. También se analizó la estructura geográfica mediante un test AMOVA empleando el paquete poppr de R.
- **Análisis de recombinación.** En los Capítulos 1 y 4 se detectaron los genes recombinantes mediante tests de congruencia topológica. Primero se evaluó la señal filogenética mediante un test de mapeo de

verosimilitudes, que compara las verosimilitudes de las tres topologías posibles para cada cuarteto de secuencias tomadas al azar del conjunto de datos. Los genes con suficiente señal filogenética fueron sometidos a tests de congruencia topológica, comparando las topologías de los árboles de cada gen frente a la del árbol de referencia. Se tuvieron en cuenta los resultados de los tests SH y ELW. Los genes para los que ambos tests rechazaron la congruencia entre topologías se seleccionaron como recombinantes. Todos estos análisis se realizaron con IQ-TREE. Para detectar los movimientos de los genes entre aislados, los árboles de los genes recombinantes se inspeccionaron visualmente comparándolos con el árbol de referencia empleando Phylo.io. La diferencia entre ambos capítulos es que en el 1 se partía del alineamiento generado tras el mapeo, mientras que en el 4 se partía de genomas ensamblados, por lo que se tuvo que detectar genes ortólogos y analizar el pangenoma y genoma *core* de *L. garvieae*.

- **Análisis del genoma accesorio no mapeado.** En el Capítulo 1, los *contigs* ensamblados se emplearon también para detectar plásmidos y la presencia de la isla genética gonocócica empleando BLAST. En el Capítulo 2, la fracción no mapeada se ensambló, se detectó la presencia de plásmidos utilizando BLAST y se anotaron los *contigs* restantes con Prokka. Las secuencias codificantes de cada aislado se compararon frente a las de los demás empleando Proteinortho. En el Capítulo 3 se realizó el mismo procedimiento que en el 2 salvo que los *contigs* no se anotaron, sino que directamente se compararon frente a los de los demás aislados utilizando BLAST.
- **Análisis de los determinantes de resistencia a antibióticos en gonococo.** En el Capítulo 1 se detectaron las mutaciones descritas en la literatura que confieren resistencia a distintos antibióticos utilizando BLAST sobre los *contigs* ensamblados y distintas bases de datos. Estos resultados genotípicos se compararon con los resultados fenotípicos. Adicionalmente, se realizó una comparación entre los resultados genotípicos y los resultados obtenidos con ARIBA, con el fin de evaluar la fiabilidad de esta herramienta.
- **Análisis de datación.** Para estimar el ancestro común más reciente de la población gonocócica en el Capítulo 1, se evaluó la señal temporal

con TempEst y se estimó el tiempo hasta el ancestro común más reciente con LSD2 empleando un modelo de reloj molecular relajado.

4. Capítulo 1 | Epidemiología genómica y vigilancia de la resistencia a antibióticos de *Neisseria gonorrhoeae* en España

Un total de 342 aislados de *N. gonorrhoeae* procedentes de la Comunidad Valenciana, Cataluña y Madrid fueron secuenciados mediante la plataforma Illumina NextSeq 500. El período de muestreo fue de 5 años, desde noviembre de 2012 a noviembre de 2017. La mayoría de pacientes (92.1%) fueron hombres, con edades comprendidas entre los 16 y los 67 años. La mayoría de las muestras fueron de origen uretral (76.9%) o rectal (13.2%).

La secuenciación produjo una media de 2,015,226 de secuencias de 150 nucleótidos por aislado, número que descendió a 1,782,226 tras el filtrado de las secuencias de baja calidad. El mapeo alcanzó una cobertura promedio de 80.1 X, cubriendo el 93.4% del genoma de la cepa de referencia. El alineamiento generado tuvo una longitud de 2.15 Mb conteniendo 34,907 polimorfismos, número que descendió a 30,042 tras eliminar las regiones recombinantes.

Previo a la detección de recombinación, se analizó la estructura genética de la población con STRUCTURE, obteniendo 6 grupos poblacionales, los cuales presentaron un elevado nivel de mezcla genética. Por ello, se procedió a la identificación de genes recombinantes tal y como se describió en la sección Métodos. Muchos de estos genes recombinantes codifican proteínas de membrana o citoplasmáticas, implicadas en procesos biosintéticos, redox, de traducción y de transporte, aunque también hay proteínas implicadas en el ensamblado de *pilus* de tipo IV y en patogénesis. Tras la eliminación de los genes recombinantes en el alineamiento, se repitió el análisis de la estructura genética poblacional, obteniéndose 8 grupos. Sin embargo, el nivel de mezcla genética en dichos grupos aún fue elevado, lo que nos lleva a pensar que no todos los eventos de recombinación presentes en la población pudieron ser detectados.

El análisis de la estructura geográfica mediante el test AMOVA reveló que no hay diferencias significativas entre los aislados de las regiones estudiadas.

El genotipado de los aislados reveló una gran variabilidad de secuenciotipos (STs). Siguiendo el esquema MLST, se detectaron 47 STs diferentes, siendo mayoritarios el 7363 (16.4%), 1901 (15.8%) y 9363 (9.9%). Respecto al sistema NG-MAST, se encontraron 111 STs distintos englobados en 75 genogrupos, siendo mayoritarios G3378 (9.9%), G2400 (8.5%) y G2992 (8.2%). Respecto al esquema NG-STAR, se detectaron 81 STs distintos, siendo mayoritarios los STs 90 (8.5%), 158 (7.9%) y 63 (7%). Al analizar la evolución de los STs a través de los años comprendidos en el período de estudio, se observa en los tres esquemas una sustitución de STs por otros nuevos.

Respecto a la detección de determinantes de resistencia a antibióticos, se realizó una búsqueda exhaustiva de mutaciones descritas que confieren resistencia a distintos fármacos, así como la detección de genes plasmídicos que confieren resistencia a antibióticos β-lactámicos o a tetraciclinas. Estos resultados genotípicos se compararon con los resultados fenotípicos proporcionados por los hospitales, pudiéndose observar algunas incongruencias que nos llevan a pensar que, en determinados casos, la aparición de un fenotipo de resistencia está ligada a una acumulación de mutaciones en genes y *loci* concretos y no a una única mutación. Estos resultados también se compararon con la herramienta ARIBA, enfocada a la detección de determinantes de resistencia, y, aunque con algunas diferencias respecto a los resultados de nuestro análisis, los resultados fueron bastante congruentes, lo que nos lleva a concluir que esta herramienta es suficientemente fiable, rápida y relativamente sencilla de emplear para este tipo de análisis.

Finalmente, se trató de datar el ancestro común más reciente de la población de gonococos, pero la señal temporal evaluada por TempEst no fue suficientemente robusta para realizar un análisis fiable.

5. Capítulo 2 | Secuenciación de genomas completos de gonococo en un caso forense de transmisión

En este capítulo se exploró la aplicación de la secuenciación de alto rendimiento a un caso forense de una transmisión de gonorrea por posible abuso sexual a una menor.

El hospital colaborador detectó un caso de gonorrea en una niña y dos parientes suyos y activó un protocolo de investigación en el entorno familiar de la niña. Se aislaron los gonococos de la víctima y el sospechoso. La otra pariente infectada ya había empezado el tratamiento antibiótico y no se pudo aislar la bacteria. El laboratorio del hospital realizó una serie de análisis moleculares, como el genotipado por MLST y una electroforesis de campos pulsados (PFGE) junto a una serie de muestras control procedentes de pacientes con gonorrea no relacionados con el caso.

Los resultados de MLST no permitieron discriminar entre los aislados, ya que todos pertenecían al mismo ST 9363. Los resultados del PFGE sí permitieron discriminar los aislados del caso frente a los controles salvo por uno, por lo que no se pudo relacionar claramente el aislado del sospechoso y el de la víctima. Con estos resultados, enviaron muestras de ADN a nuestro laboratorio para realizar un análisis genómico.

Las muestras del sospechoso, la víctima y tres controles —incluyendo al control indistinguible de los aislados implicados en el caso— fueron secuenciadas por Illumina NextSeq 500, produciendo un promedio de 1,750,335 de secuencias. Para los análisis subsiguientes se añadieron otros 26 controles más procedentes del dataset del Capítulo 1, los cuales pertenecían al ST 9363 y procedían de la Comunidad Valenciana y Cataluña.

Las secuencias se mapearon frente a la referencia más cercana a las muestras del caso, la cepa WHO P. Se alcanzó una cobertura promedio de 85.57 X, cubriendo un 94.78% del genoma de referencia. El alineamiento alcanzó una longitud de 2.17 Mb y contenía 6,792 polimorfismos. La matriz de distancias genéticas reveló que no existían diferencias entre el aislado del

sospechoso y el de la víctima, mientras que el control que no podía ser distinguido de estas muestras difería en 2 SNPs respecto a los aislados del caso. Esto hace pensar que el sospechoso y el paciente control contrajeron la infección en el mismo lugar. Los otros controles locales difirieron en 66 y 1,096 SNPs respecto a las muestras del caso. Los controles procedentes de la Comunidad Valenciana y Cataluña diferían entre 1,060 y 1,661 SNPs respecto a los aislados del caso, por lo que claramente no guardan relación con estos aislados.

Para realizar un análisis lo más completo posible, se ensamblaron las secuencias que no mapearon con la referencia de los aislados del caso y los dos controles locales más próximos. Se identificó en todos ellos el mismo plásmido —plásmido críptico de la cepa WHO P—. El resto de *contigs* se anotaron y los genes se analizaron con Proteinortho, mostrando que todos los aislados compartían 67 genes idénticos y el control más alejado poseía dos genes adicionales. En consecuencia, se pudo determinar que los aislados del caso eran completamente idénticos, tanto a nivel de cromosoma como de plásmido.

Estos resultados mostraron el superior poder de resolución de la secuenciación de alto rendimiento en comparación con otras técnicas moleculares como MLST o PFGE, y el potencial de esta tecnología para ayudar a la resolución de casos de microbiología forense.

6. Capítulo 3 | Análisis genómico de dos brotes de *Serratia marcescens* en hospitales de la Comunidad Valenciana

En este capítulo se estudió la relación genómica entre aislados de *S. marcescens* procedentes de dos brotes producidos en las UCIs neonatales de dos hospitales de la Comunidad Valenciana.

El brote A afectó a 7 neonatos ingresados en la UCI, mientras que el brote B afectó a 6 pacientes. Ningún caso de ambos brotes resultó fatal. Las medidas de control e higiene fueron efectivas para contener ambos brotes.

Se aislaron 21 muestras del brote A y 6 muestras más un control externo del brote B. El ADN de los aislados fue secuenciado por Illumina NextSeq 500. El brote A produjo una media de 1,383,705 secuencias por muestra, mientras que en el brote B el promedio fue de 2,454,515. La identificación *in silico* de la especie fue positiva para *S. marcescens* en todas las muestras salvo en dos del brote A, que correspondieron a *S. liquefaciens* y fueron excluidas de los restantes análisis.

Los aislados del brote A se mapearon frente al genoma de la cepa de referencia UMH9, que era la más próxima genéticamente a estos aislados. Se obtuvo una cobertura promedio de 19 X, con el 99.9% de las secuencias cubriendo al 94.4% del genoma de referencia. El 5.6% restante probablemente no fue cubierto porque las secuencias no pasaron los filtros de calidad. El alineamiento generado tras el mapeo y la detección de polimorfismos tuvo una longitud de 5.02 Mb y contenía 27,775 polimorfismos. Este número de variantes descendía a 157 si se excluía la cepa control. La matriz de distancia genética entre pares de aislados reveló que había 0-2 SNPs de diferencia entre las cepas del brote y entre 0-1 SNPs entre las cepas del brote y la referencia, lo que la hacía indistinguible respecto a los aislados del brote. Las secuencias no mapeadas fueron tan escasas que no se pudieron ensamblar, excepto la cepa control, cuya fracción genómica no mapeada tuvo una longitud de 655,961 nucleótidos distribuidos en 321 *contigs*.

A la vista de estos resultados, no pudimos definir a nuestros aislados como parte de un brote, ya que una muestra ajena entraba dentro de dicho brote. Por ello, repetimos el análisis empleando una referencia más distante genéticamente, la cepa Db11. La cobertura promedio fue de 16.3 X, con un 88.7% de las secuencias siendo mapeadas y cubriendo el 79.5% del genoma de referencia. El alineamiento alcanzó las 5.11 Mb y contuvo 162,590 polimorfismos. La matriz de distancias reveló de nuevo diferencias entre 0-2 SNPs entre los aislados del brote y entre 5,535-5,536 SNPs entre los aislados del brote y la cepa de referencia, por lo que en esta ocasión sí pudimos definir claramente el brote como tal. El genoma accesorio tuvo una longitud promedio de 532,084 nucleótidos entre las cepas del brote y de 318,456 nucleótidos para la cepa control. No se detectaron muchas diferencias entre los genomas accesorios de los aislados del brote, aunque sí entre estos y el aislado control. Tampoco se detectaron plásmidos en ninguna de las cepas.

Para determinar si la cepa UMH9 pudiera ser un clon que se estuviera dispersando por la Comunidad Valenciana, repetimos el análisis con el brote B. Mapeamos los aislados de este brote frente a UMH9 y frente a Db11, aunque en esta ocasión los resultados de ambos mapeos fueron muy similares, ya que ninguna de las dos cepas de referencia era genéticamente cercana a los aislados del brote. En ambos casos se pudo ensamblar la fracción genómica no mapeada, donde no se observaron diferencias significativas entre los aislados implicados en el brote y sí entre estos y la cepa control. En este brote tampoco se identificaron plásmidos.

Los resultados de este capítulo pusieron de manifiesto una limitación de este tipo de análisis. Si al analizar un brote, una cepa externa resulta estar dentro del brote, no podemos estar seguros, basándonos únicamente en los datos genéticos, de que los aislados analizados constituyan un brote. Por ello es importante tanto elegir una referencia apropiada a la hora de analizar un brote, así como disponer de datos microbiológicos y epidemiológicos que nos ayuden a interpretar correctamente los resultados.

7. Capítulo 4 | Un enfoque desde el pangenoma para detectar transferencia horizontal de genes en una cepa de *Lactococcus garvieae*

En este capítulo obtuvimos, a través de una colaboración con la Universidad Complutense de Madrid, una cepa de *L. garvieae* que había sido secuenciada mediante PacBio RS II, ensamblada y anotada en un laboratorio externo, de modo que consiguieron cerrar el cromosoma y el plásmido. Nuestro objetivo fue detectar recombinación entre esta cepa y otras de la misma especie, y extender este análisis a niveles taxonómicos superiores —género y clase—.

Para el análisis a nivel de especie, descargamos todos los genomas disponibles en NCBI hasta la fecha, sumando 24 genomas si contamos nuestra cepa de estudio —Lg-Granada—. A continuación, se detectaron los genes ortólogos con Proteinortho y se determinó el genoma *core* estricto, que comprende los genes compartidos por todas las cepas, y el genoma *core* relajado, que comprende los genes compartidos por al menos el 80% de las cepas. Estos genomas *core* contienen 1,157 y 1,556 genes para el estricto y el relajado, respectivamente. Por su parte, el pangenoma de *L. garvieae* alcanzó los 5,031 genes. Un total de 592 genes del *core* relajado fueron identificados como recombinantes, la mayoría codificando proteínas de membrana y citoplasmáticas con funciones catalíticas y de unión, envueltas en procesos metabólicos y de transporte, principalmente.

A nivel de género, Lg-Granada se comparó con 6 genomas de *Lactococcus*. La reconstrucción filogenética mostró que *L. garvieae* comparte ancestro común con *L. lactis*, quedando *L. piscium* y *L. raffinolactis* más alejadas. El genoma *core* estricto para el género contiene 924 genes, mientras que el relajado contiene 1,462. Se detectaron 97 genes recombinantes, aunque sólo 27 de ellos implicaban transferencia entre Lg-Granada y las otras especies. La mayoría de los genes transferidos a Lg-Granada procedían de alguna especie no identificada de *Lactococcus,* que nombramos *Lactococcus* spp. Los genes recombinantes detectados estaban implicados en transporte, regulación de la transcripción y división celular, principalmente.

A nivel de clase, Lg-Granada fue comparada frente a 19 especies de la clase *Bacilli*. La reconstrucción filogenética mostró que *L. garvieae* está estrechamente relacionado con el género *Streptococcus*. El *core* estricto de la clase *Bacilli* contiene 409 genes, mientras que el relajado contiene 775. Un total de 135 genes fueron identificados como recombinantes, de los cuales solo 34 implicaban transferencia entre Lg-Granada y las otras especies. Lg-Granada donó mayor número de genes al género *Lactobacillus*, mientras que recibió mayor número de genes del género *Streptococcus*. La mayoría de los genes recombinantes estaban implicados en procesos metabólicos y de transporte, así como en traducción y respuesta a estímulos, entre otros.

El estudio de los eventos de recombinación en bacterias patógenas es importante porque constituye el principal mecanismo de adquisición de virulencia y patogenicidad. En este capítulo hemos comprobado que, a nivel intraespecífico, *L. garvieae* es una bacteria muy recombinogénica, por lo que sería interesante controlar su presencia en aquellos ambientes desde los que puede alcanzar a los humanos, como ganado y agricultura.

8. Discusión

La secuenciación de alto rendimiento es una tecnología que cada vez tiene mayor implantación en la investigación epidemiológica de patógenos. Se está desarrollando cada vez más la transición de la epidemiología molecular a la epidemiología genómica. En esta tesis se ha mostrado, a través de cuatro estudios distintos, cómo esta tecnología puede proporcionar información útil para estudiar bacterias de relevancia clínica.

Gracias al acceso a los genomas completos de gonococos hemos podido investigar la estructura poblacional de este patógeno en España, poniendo de manifiesto el elevado nivel de mezcla genética que tienen los aislados de estudio. También nos permitió identificar fácilmente todas las mutaciones presentes en genes que confieren resistencia a antibióticos. Esto es interesante desde el punto de vista clínico, porque conocer las mutaciones circulantes en la región permitiría adaptar el tratamiento con el fin de que

fuese lo más efectivo posible y, a su vez, evitar favorecer la aparición de más mutaciones que condujesen al desarrollo de resistencia a los fármacos.

La secuenciación de alto rendimiento posee el poder de resolución más elevado entre las técnicas existentes de biología molecular para detectar clonalidad entre aislados en casos de transmisión. Este potencial se ha explotado en el Capítulo 2, donde se exploró la aplicación de esta tecnología en un caso de investigación forense de transmisión de gonorrea en el contexto de un posible caso de abuso sexual a una menor, y en el Capítulo 3, donde se analizaron dos brotes nosocomiales y evaluamos el impacto de elegir un determinado genoma de referencia para comparar los aislados de un brote. El efecto de la elección de una referencia u otra se ha discutido en otros trabajos, donde se ve afectada la filogenia, pero también la estimación de diversos parámetros evolutivos. Sin embargo, en este capítulo de la tesis se proporciona un ejemplo en el que se pone de manifiesto que la secuenciación de alto rendimiento es una técnica más de biología molecular, la cual tiene un elevado potencial para convertirse en un estándar en epidemiología y clínica, pero cuyos resultados deben ser complementados con los resultados de otros análisis, como los microbiológicos.

También hemos aprovechado la capacidad de reconstruir un genoma completo, con el cromosoma y el plásmido cerrados, gracias al uso de la secuenciación de alto rendimiento de tercera generación, que genera secuencias más largas, permitiendo resolver regiones de baja complejidad. Gracias al acceso a la secuencia completa, no perdemos información genética y se pudo estudiar los eventos de recombinación en una cepa de *L. garvieae*.

Aunque la implantación de esta tecnología es amplia en investigación, su utilización en clínica aún no está generalizada, especialmente por la complejidad de los análisis y la falta de estandarización y recursos, por lo que el desarrollo de herramientas que faciliten los análisis bioinformáticos y su interpretación será una de los factores clave para alcanzar esta implementación.

9. Conclusiones

1. La aplicación de la secuenciación de alto rendimiento tiene un gran potencial en los estudios epidemiológicos.
2. Gracias a la implementación de esta herramienta, hemos sido capaces de analizar la estructura de la población gonocócica en España, verificando el alto grado de mezcla genética presente en los aislados muestreados.
3. La disponibilidad de secuencias genómicas completas obtenidas por secuenciación de alto rendimiento facilitó la detección de mutaciones clave en ciertos genes que pueden conducir al desarrollo de resistencia a antibióticos. De este modo, algunas de estas mutaciones fueron detectadas incluso en aislados cuyo fenotipo era sensible. Esta información puede ser interesante en clínica para desarrollar protocolos de tratamiento apropiados para no favorecer la acumulación de mutaciones en esos genes y la consecuente emergencia de fenotipos resistentes.
4. La secuenciación de alto rendimiento ofrece mayor resolución que cualquier otra técnica molecular. Esto evidencia la utilidad de esta técnica para la detección de aislados estrechamente relacionados.
5. En base a la conclusión anterior, la secuenciación de alto rendimiento es muy útil en investigaciones de transmisión en contextos forenses. Fuimos capaces de discriminar entre aislados de gonococo procedentes del sospechoso y la víctima de un posible caso de abuso sexual a una menor y los aislados de los pacientes control, cuando las otras técnicas moleculares fracasaron.
6. Los análisis de brotes basados en el mapeo frente a una referencia están limitados por la elección de dicha referencia. Si elegimos un genoma de referencia que es demasiado cercano genéticamente a los aislados que queremos analizar, corremos el riesgo de perder la resolución filogenética hasta el punto de ser incapaces de definir al brote como tal, como hemos podido observar en el análisis de los brotes de *Serratia marcescens*.

7. La tecnología de secuenciación de alto rendimiento que produce secuencias más largas tiene la ventaja de resolver los ensamblados de genomas mejor que la tecnología de secuenciación que produce secuencias cortas. Esto nos permite obtener genomas completos, incluyendo cromosomas y plásmidos cerrados. Gracias a esta tecnología, pudimos analizar los eventos de recombinación presentes en el patógeno emergente *Lactococcus garvieae* a nivel de especie y la subsiguiente extensión a niveles taxonómicos superiores.

8. En resumen, las tecnologías de secuenciación de alto rendimiento son muy útiles para el análisis de patógenos de relevancia clínica, así como para la vigilancia de la emergencia de resistencias a antibióticos. Con la maduración de esta tecnología en el campo de la investigación básica, donde los obstáculos que esta tecnología puede ir presentando deben ser resueltos, podrá ser transferida a la práctica clínica a un nivel más generalizado.

ABBREVIATIONS

AMR: Antimicrobial Resistance.

BAM: Binary Alignment/Map.

BIC: Bayesian Information Criterion.

BLAST: Basic Local Alignment Search Tool.

bp: base pair.

BWA: Burrows-Wheeler Aligner.

cgMLST: core-genome multilocus sequence typing.

CV: Comunidad Valenciana.

DNA: deoxyribonucleic acid.

ECDC: European Centre for Disease Prevention and Control.

ENA: European Nucleotide Archive.

ESC: extended-spectrum cephalosporin.

ESCHAPPM: *Enterobacter* spp., *Serratia marcescens, Citrobacter freundii, Hafnia* spp., *Aeromonas* spp., *Providencia* spp., *Proteus vulgaris*, and *Morganella morganii*.

EUCAST: European Committee on Antimicrobial Susceptibility Testing.

GB: Gigabyte.

GC: Guanine-Cytosine.

GFF: General Feature Format.

GGI: gonococcal genetic island.

GTR: General Time Reversible.

HAI: Healthcare-associated infection.

HCV: Hepatitis C Virus.

HIV: Human Immunodeficiency Virus.

HTS: high-throughput sequencing.

ICU: intensive care unit.

IDU: injection drug user.

Kb: Kilobase.

LM: likelihood mapping.

MALDI-TOF: Matrix-Assisted Laser Desorption/Ionization - Time-of-Flight.

Mb: Megabase.

MCMC: Markov chain Monte Carlo.

ML: maximum likelihood.

MLST: multilocus sequence typing.

MRCA: most recent common ancestor.

MSM: men who have sex with men.

N: unknown nucleotide.

NCBI: National Center for Biotechnology Information.

NGS: next-generation sequencing.

NICU: neonatal intensive care unit.

OG: orthologous gene.

PCR: polymerase chain reaction.

PFGE: pulsed-field gel electrophoresis.

PPNG: penicillinase-producing *Neisseria gonorrhoeae*.

RNA: ribonucleic acid.

SAM: Sequence Alignment/Map.

SNP: single nucleotide polymorphism.

SSI: surgical site infection.

ST: sequence type.

STI: sexually-transmitted infection.

TRNG: tetracycline-resistant *Neisseria gonorrhoeae*.

TVM: transversion model.

T4SS: type IV secretion system.

UTI: urinary tract infection.

UV: ultraviolet.

VCF: Variant Call Format.

WGS: whole-genome shotgun.

WHO: World Health Organization.

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

1. Bacterial pathogens as a threat for Public Health

Bacterial pathogens are bacteria that can cause infectious diseases. Many of them are obligate pathogen that evolved from environmental ancestors[1]. The genetic mechanisms involved in the short- and long-term evolution of bacteria include point mutations, DNA rearrangements and gene transfer, and all of these mechanisms can lead to acquisition of pathogenicity and/or virulence[2]. From the pathogenic bacteria perspective, the host body constitutes a complex ecosystem with numerous niches that can be colonized and that offer numerous nutrient sources[3], but it is also an extreme environment for microorganisms due to the development by the host of systems for avoiding infections[1]. Several studies highlight the fact that most pathogenic bacteria has adapted to this environment by reducing their genome size, that is, the balance between the gain of virulence genes and the loss of genes involved in metabolic processes is inclined in favor of the latter[4–7].

However, there is another wide class of bacterial pathogens that are not obligated parasites. These are the opportunistic pathogens, which can be part of the host's commensal microbiota, or be environmentally acquired, and can become pathogenic when the immune system of the host is compromised —e.g. previous infection or disease, medication, ageing, etc.—[8].

With some exceptions[9–11], both obligate and opportunistic pathogens evolve rapidly because bacteria have short generation times. Selective pressures within the host, such as immune system interactions, antibiotics, and competition with commensal microbiota, favor the evolution of pathogenic bacteria[12]. On the other hand, selective pressures by bacterial pathogens favors the evolution of host proteins, such as immune system receptors[13]. These facts highlight the existence of an evolutionary "arms race" between the pathogenic bacteria and the host.

According to the World Health Organization (WHO), infectious diseases accounted for 26% of deaths globally in 2019[14]. With this scenario, it is clear that communicable diseases are one of the main concerns for humankind.

Next, some of the main Public Health threats due to infectious agents will be introduced.

1.1. Sexually transmitted infections

A sexually transmitted infection (STI) is an infection acquired from an infected person through sexual intercourse, including vaginal, oral and anal sex, as well as congenital infections, which are infections from infected pregnant mother to child[15]. Causal agents of STIs include more than 30 species of bacteria, viruses, parasites —protozoa and insects—, and fungi[16].

According to WHO, more than 1 million of STIs are acquired every day worldwide. In 2016, 376 million new infections with one of the following STIs were estimated to have occurred: trichomoniasis (156 million), chlamydiasis (127 million), gonorrhea (87 million), and syphilis (6.3 million)[17]. The first one is caused by a protozoon, while the others are caused by bacteria.

STIs can have serious consequences beyond the impact of the infection itself[18]:

- Can lead to infertility, especially in women.
- Skin and mucosal lesions can increase the risk of human immunodeficiency virus (HIV) acquisition.
- Mother-to-child transmission of STIs can result in stillbirth, prematurity, sepsis, pneumonia, neonatal conjunctivitis, and congenital deformities.
- Some infections can lead to the development of cervical cancer and, in most cases, to death.
- People affected by some STI are often stigmatized by the society.

Although anyone can be exposed to STIs, there are vulnerable populations such as sex workers and their clients, men who have sex with men (MSM), injection drug users (IDUs), teenagers, and rape victims[19–21]. These populations are specially linked to a series of risk factors which favor the STIs transmission, such as having multiple sexual partners, being part of

marginal groups with limited access to education, having unprotected sex, or abusing of substances —alcohol and drugs— that reduce risk perception.

The most reliable method to avoid contagion by STIs is sexual abstinence[22]. However, Prevention of STIs is sought through the sexual education of citizens, the reduction of the number of sexual partners as well as risky activities, and the use of barrier methods, such as condoms[18].

Rapid and accurate diagnosis of STIs is key because, in general, there are antibiotic or antiviral treatments for most of these infections, although antimicrobial resistance (AMR) is an emerging threat of which we will talk about later. In this sense, controlling STIs require well-stablished surveillance systems that monitor epidemic trends and detect outbreaks, as well as the development of strategies to direct resources for prevention, treatment and control[23].

1.2. Healthcare-associated infections

Healthcare-associated infections (HAIs) are infections in which the disease is a result from exposure to infectious agents during healthcare procedures[24]. These infections can be acquired in hospitals, in intensive care units (ICUs), or in long-term care facilities. The infections may be caused by bacteria, viruses, fungi, parasites, or prions. The clinical spectrum is broad, including urinary tract infections (UTIs), surgical site infections (SSIs), pneumonia, bloodstream infections, gastrointestinal and systemic infections, or skin and soft tissue infections[25].

According to the WHO, although global estimates of HAIs are not available, the published studies provide evidence that hundreds of millions of people are affected every year worldwide, especially in the developing countries[26,27]. All the patients are susceptible to acquire a HAI, but the elderly, neonates —specially premature—, and patients with a compromised immune system are the ones who have the highest risk[28–30]. SSIs are one of the most common types of HAIs, as they involve skin incision and foreign implants, but the most frequent ones are UTIs, associated with the use of catheters, and pneumonia,

associated with motorized automatic ventilation. The major —but not the only ones— causative agents of HAIs are opportunistic pathogens which colonize surfaces and medical instruments[31].

The measures of the prevention and control of HAIs include hand hygiene of healthcare workers[32] and cleaning of surfaces, and reducing the number of workers that come into contact with the patient. The lack of well-established HAI surveillance systems and the emergence of AMR pathogens in healthcare settings, partly due to the incorrect administration of antibiotics, require improvements in reporting and surveillance systems at the national and institutional levels, improvement in the staff education on control and hygiene measures, and adaptation of the surveillance protocols to the reality of each country[33].

1.3. Emerging infectious diseases

Emerging infections are those that have recently appeared in a population or have existed previously but their incidence or geographical range are rapidly increasing[34].

Most of emerging infections are zoonoses —infectious agents from animals that have jumped to human hosts— or vector-borne infections —infections transmitted from an animal, usually an arthropod, to humans—, but they can also be caused by microorganisms that persist in the environment, or by known microbes causing new diseases[35,36].

There are several factors that contribute to infectious disease emergence or re-emergence[34,35]:

- **Ecological changes:** deforestation, climate change, agriculture and livestock.
- **Human demographics and globalization:** population growth, migration from rural areas to cities, sexual behavior, worldwide movements of goods and people.

- **Technology, industry and medicine:** changes in food processing, widespread use of antibiotics, organ transplantation.
- **Microbial adaptation and change:** evolution and adaptation to environment pressures —e.g. antimicrobial resistant bacteria—.
- **Breakdown in Public Health measures:** reduction in prevention programs, inadequate sanitation and vector control measures.

Emerging infections are a significant threat to Public Health. Outbreaks of these infections may cause high numbers of deaths as they spread, and have a potential impact in society and economics in the context of an interconnected world. Most of these diseases do not have a cure and, often, healthcare workers are also victims of these infections. It is mandatory to establish appropriated surveillance systems, with effective risk communication and management of preventive and control measures[37].

1.4. Antimicrobial resistance

Antimicrobial resistance (AMR) occurs when microorganisms evolve and no longer respond to drugs making infections increasingly difficult or impossible to treat, which increases the risk of disease spread, severe illness and death[38].

AMR is a natural phenomenon result of a Darwinian competition for survival in a hostile environment mediated by antimicrobial molecules produced by other microorganisms[39]. However, the use, misuse, or overuse of antimicrobial drugs in medicine has become a major driving force of antimicrobial resistance[40,41]. But the antibiotic abuse is not exclusive of healthcare settings: the same problem is also found in food-producing animals and in agriculture, where drugs are used for animal growth and to prevent infections in animals and plants, and resulting AMR bacteria end up spreading in the environment or the community (Figure 1)[42–44].

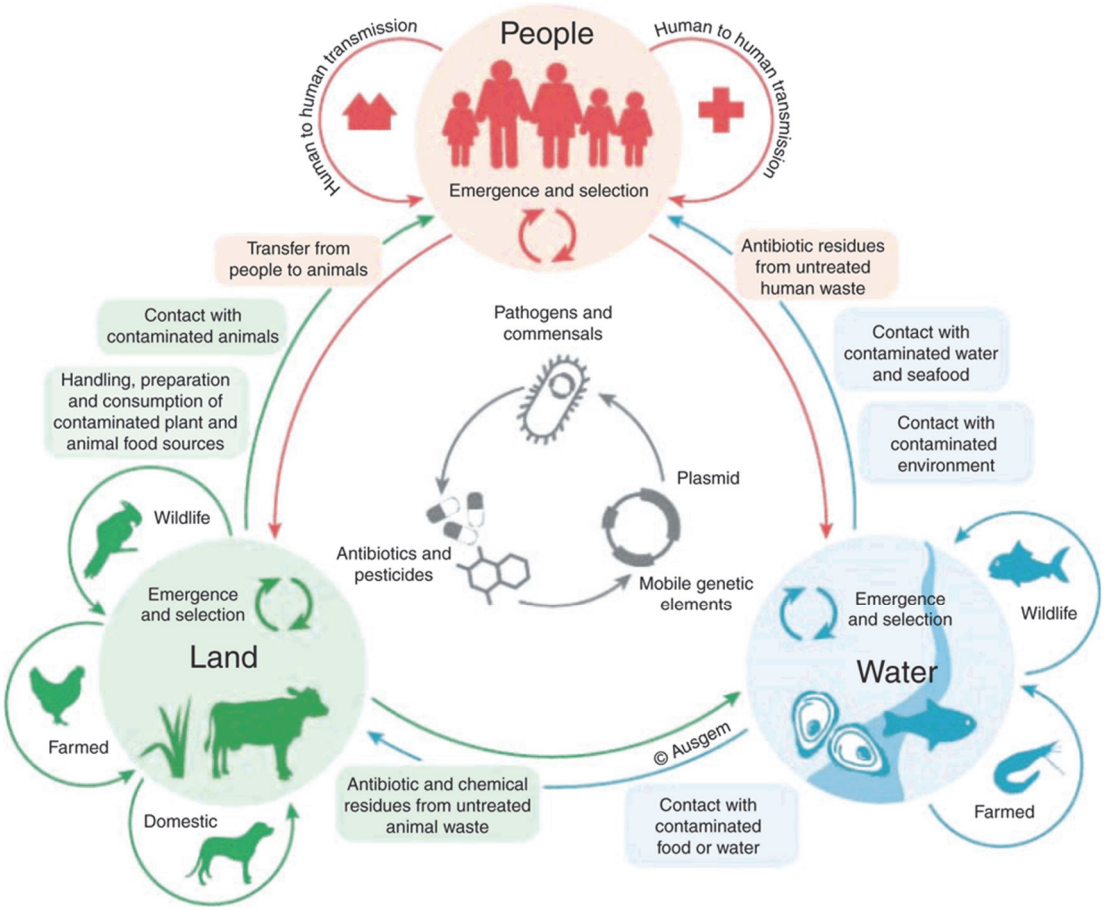


Figure 1 | Spread of AMR through human community and land and water ecosystems. Source: Djordjevic & Morgan, 2019[44].

AMR determinants are transmitted between pathogens through recombination, by three main routes[39] (Figure 2):

- **Transformation:** bacteria take exogenous DNA from the environment and incorporate it into their chromosome.
- **Transduction:** a phage-mediated transfer of DNA between bacteria.
- **Conjugation:** transfer of DNA between bacteria through cell-to-cell contact. This is the way of plasmid transference.

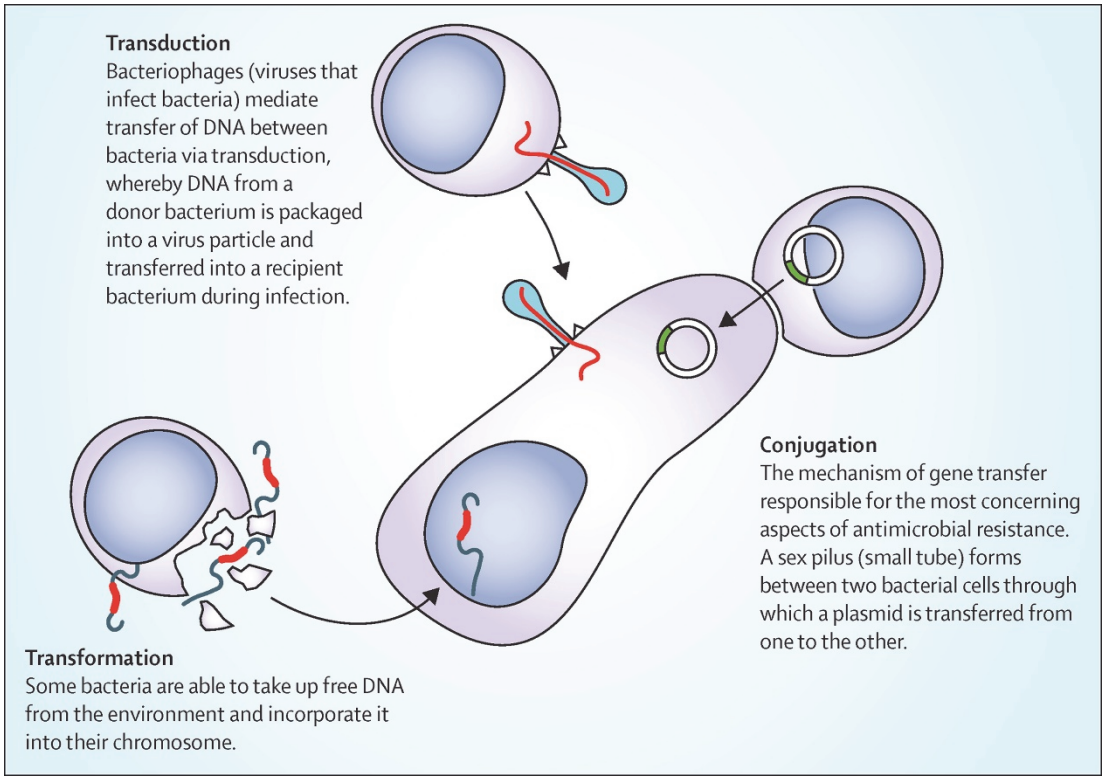


Figure 2 | Mechanisms of DNA acquisition in bacteria. These are potential ways of AMR transfer between bacteria. Source: Holmes *et al.*, 2016[39].

Public Health agencies consider the emergence of AMR pathogens as one of the most serious public health threats nowadays, that makes it difficult the effective prevention and treatment of infections by pathogens no longer susceptible to the drugs used in their treatment[42]. The role of drugs use in the emergence of AMR may be specific to each drug and to each microorganism, as it is the effect of changes in this use, so policies need to be aware of this complexity in addressing AMR and must adopt an integrated approach across both the community —environment, agriculture and livestock— and healthcare settings[39]. In this sense, in 2015 the WHO endorsed a Global Action Plan to tackle AMR by improving awareness of AMR, increasing knowledge through surveillance and research, optimizing the use of drugs, reducing the incidence of infections, and developing economic budgets for investment in resources to fight against AMR attending the needs of all countries[45].

2. High-throughput sequencing applied to pathogen surveillance: Genomic Epidemiology

The genomics revolution is a relatively recent phenomenon, but its development commenced many decades ago. The first DNA sequencing technique was the chain termination method, established by Frederick Sanger and associates during the 1970s, which is known as Sanger sequencing[46]. Using this method, the first sequenced genome was that of bacteriophage ΦX174[47]. Two decades later, strategies of shotgun sequencing —which use numerous overlapping sequences to improve the efficiency— were developed[48], allowing the sequencing of the first bacterial genome, *Haemophilus influenzae*[49].

The high demand for low-cost genomic sequencing led to the development, in the first decade of 2000s, of second-generation sequencing —also known as next-generation sequencing (NGS)—, with technologies allowing the massive parallelization of sequencing reactions, enhancing throughput and reducing costs in comparison to the first-generation technologies[50]. These technologies work on a fragmented genome and generate an output of millions of sequences of about 100-150 bp —short reads— that are used to reconstruct the sequenced genome. Illumina is one of the leading companies developing this technology[51]. But this technology is not exempt of limitations, including the high error rates near the terminal ends of the reads, the short length of the reads —which causes problems with low-complexity genomic regions—, and the source of bias and sequencing errors because they are based on PCR amplification to generate enough DNA template[52–54].

Recently, the so-called third-generation sequencing methodology has been developed. Pacific Biosciences (PacBio) has developed the single molecule real-time (SMRT) sequencing technology, where the DNA synthesis is done using a single DNA molecule without amplification step, and it occurs uninterrupted generating long reads[55]. On the other hand, Oxford Nanopore has developed the nanopore sequencing technology, which differs from the previously commented technologies in that it infers nucleotide identities through variations in electric current rather than detecting fluorescent

signals. Nanopore technology produce longer reads than PacBio technology[56,57]. Both third-generation sequencing technologies require further development and maturity, as they have higher error rates than those of second-generation sequencing technologies[56,58].

High-throughput sequencing (HTS) —which encompasses both second- and third-generation sequencing technologies— has reached a wide implementation in the research field of microbial genomics[59], but it is still in the early stages of implementation in clinical and epidemiological practice. Only a few countries have been working to adopt such technologies into their Public Health systems, as it is the case of the United Kingdom[60]. Public agencies such as the European Centre for Disease Prevention and Control (ECDC) have concluded that the application of HTS for typing pathogens provides higher resolution and accuracy than classical molecular methods, but the costs and lack of expertise limit its use by public health laboratories[61]. In clinical microbiology, although its implementation is uneven depending on the disease, HTS has an increasingly predominant role[62,63].

The use of whole genomes to analyze epidemic outbreaks is a relatively recent methodology[64,65], but it has proved its potential to analyze and elucidate the origins of outbreaks[66,67]. Another potential application of HTS is AMR surveillance[68,69]. In this sense, HTS provides an enormous amount of information and a high resolution for pathogen typing, so its application for global surveillance can provide information on the emergence and spread of AMR, and complements the information from phenotypic methods.

In addition to these applications in Public Health, the results of comparative genomics studies can shed light on the epidemiology, evolution, and population dynamics of pathogens[70–73].

In summary, HTS is at the same level as the quantitative PCR (qPCR) technique was in its beginnings —it was difficult to imagine it as a diagnostic tool, since it lacked standardization, ease of use, and trained staff—[74]. HTS implementation will depend on the resolution of the existing obstacles and the benefits obtained[62,75].

3. Bacterial pathogens studied in this thesis

In this thesis, all the issues addressed in the previous sections will be explored through four examples of application of HTS to the genomic study of three pathogens.

3.1. Neisseria gonorrhoeae*: a sexually-transmitted bacterium resistant to antimicrobial drugs*

Neisseria gonorrhoeae —also known as gonococcus— is a Gram-negative bacterium belonging to the *Betaproteobacteria* class. It is non-motile, non-spore forming, and characteristically grows in pairs as diplococci. Gonococcus is an obligate pathogen whose only natural host are humans, being unable to survive for long periods outside the human body. The genus *Neisseria* includes several non-pathogenic species, but there are two species of clinical relevance for humans, *N. gonorrhoeae* and *Neisseria meningitidis* —meningococcus—, the etiological agent of one form of bacterial meningitis[76]. Both species are evolutionarily very close each other as it can be seen in Figure 3.

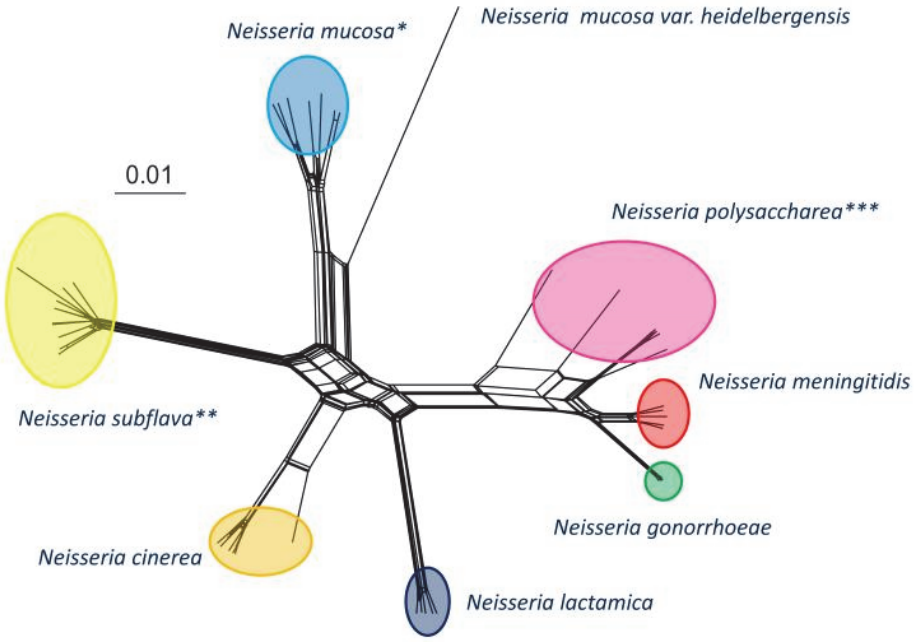


Figure 3 | Phylogenetic network of the *Neisseria* genus based on 53 ribosomal genes. Source: Bratcher *et al.*, 2012[77].

As of January 1, 2021, there were 746 *N. gonorrhoeae* genomes deposited at the NCBI database, 61 of which were completely closed[78]. According to data from those 61 complete genomes, the size of its genome is around 2.22 Mb (range 2.15-2.29 Mb), with 2,064 coding sequences on average (range 1,973-2,203), and a GC content around 52.43% (range 52.10-52.70%).

Gonococci can contain plasmids of diverse nature. The most common one is the cryptic plasmid, a small plasmid of 4.2 Kb that does not confer AMR or virulence phenotype, but which is present in 96% of gonococcal strains[79]. There are several plasmids that carry the bla_{TEM} gene, which encodes a β-lactamase that confers resistance to penicillin. The main plasmids of this type are Asian (7.4 Kb), African (5.6 Kb) and Toronto/Rio (5.2 Kb)[80]. There are other minor types of β-lactamase plasmids, such as Nimes (6.8 Kb), New Zealand (9.3 Kb), Johannesburg (4.8 Kb), and Australian (3.2 Kb)[81]. It has been suggested that the original plasmid is the Asian, while the remainder have originated from insertions or deletions, and/or rearrangements in fragments of the Asian plasmid sequence[82]. Finally, the other type of plasmid that can be found in gonococcal isolates is the conjugative, that can be markerless (39 Kb) or one of the two types of 42 Kb plasmids carrying the *tetM* gene, which confers resistance to tetracyclines —Dutch or American types—[81,83]. Isolates carrying a β-lactamase plasmid are known as PPNG (penicillinase-producing *N. gonorrhoeae*), while those carrying a *tetM*-conjugative plasmid are TRNG (tetracycline-resistant *N. gonorrhoeae*)[84].

Another characteristic element of the mobilome in gonococci is the 57 Kb gonococcal genetic island (GGI), which encodes a type IV secretion system (T4SS). T4SS is found in most gonococcal isolates, and its function is to secrete chromosomal fragments outside the cell, having a direct role in gonococci transformation[85,86].

Gonococcus is the etiological agent of gonorrhea, the second most common bacterial STI[87]. Its symptoms include urethritis in men and cervicitis in women. Rectal and oropharyngeal gonorrhea is more frequent in men who have sex with men (MSM), but it can be found in both sexes. Complications include endometritis, pelvic inflammatory disease or epididymitis —resulting

in infertility—, ophthalmia neonatorum —which leads to blindness in newborns—, and disseminated gonococcal infection —causing arthritis, endocarditis or meningitis—[88].

Gonococcal infection (Figure 4) starts when the bacterium adheres to the host's epithelial cells, through type IV pili and other surface proteins such as opacity proteins (Opa), the major outer membrane protein (PorB), and the lipooligosaccharide (LOS). Gonococci colonize the surface of the point of infection forming a biofilm and competing with local microbiota. When colonizing the epithelium, gonococci can invade the host cells and go through them via transcytosis. Gonococcal cells release fragments of peptidoglycan, LOS and vesicles, activating immune signaling in epithelial and dendritic cells, and in macrophages, which activate inflammatory response attracting numerous polymorphonuclear leukocytes —or neutrophils— to the point of infection that engulf gonococci. However, gonococci survive in phagosomes. Finally, the influx of neutrophils gives rise to a purulent exudate that facilitates transmission[89].

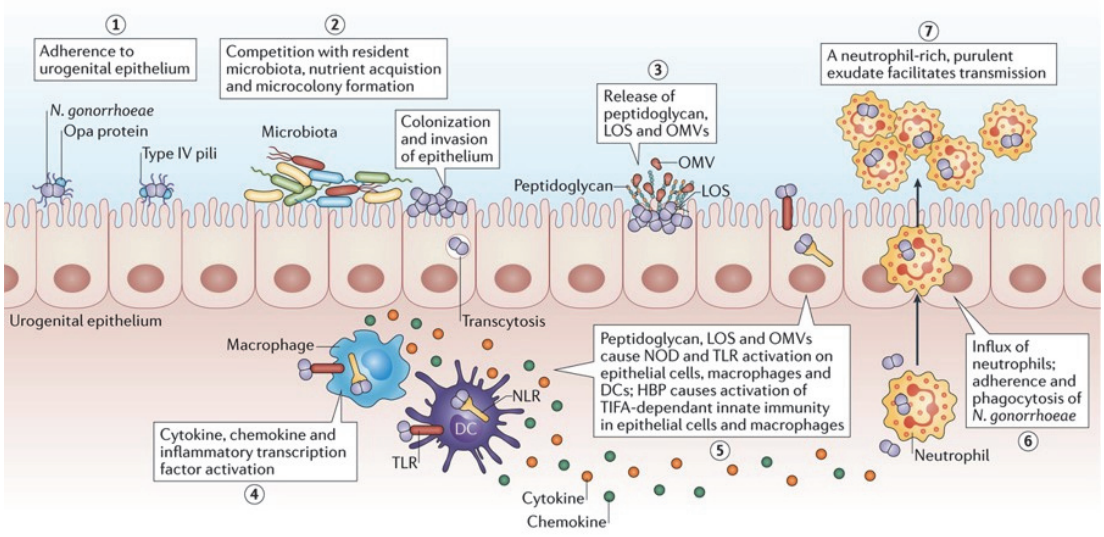


Figure 4 | Stages of gonococcal infection. Source: Quillin & Seifert, 2018[89]

The recommended treatment for gonorrhea is a dual therapy with an extended-spectrum cephalosporin (ESC) —ceftriaxone or cefixime— along with the macrolide azithromycin. For neonates with gonococcal conjunctivitis the recommendation is a single dose of ceftriaxone or an aminoglycoside, such as kanamycin or spectinomycin[90]. However, gonococci are developed AMR to every antibiotic that has been used to treat their infections (Figure 5):

- **Sulfonamides:** these antibiotics target the dihydropteroate synthase enzyme, encoded by the *folP* gene, to inhibit folic acid synthesis in the bacterium, which is essential for bacterial growth. Mutations in this gene reduce target affinity[91].
- **Penicillines:** these antibiotics interfere with the synthesis of peptidoglycan from the cell wall through binding of the β-lactam ring to penicillin-binding proteins (PBPs). Mutations in the *ponA* gene, which encodes PBP1, reduce penicillin activity in this target[92], mutations in the *porB* gene —encoding the major outer membrane porin, PorB— reduce penicillin influx[93,94], mutations in the *pilQ* gene —encoding PilQ secretin of type IV pili— increase the penicillin efflux[95], and the presence of a plasmid carrying a bla_{TEM} gene —encoding mainly the TEM-1 or TEM-135 alleles of β-lactamase— inactivate penicillin[96].
- **ESCs:** like penicillin, interfere with the synthesis of peptidoglycan from the cell wall through binding of the β-lactam ring to PBPs. Therefore, in principle, mutations in resistance determinants that affect susceptibility to penicillin apply to ESCs. However, in the case of ESCs, the most significant mutations are those in the *penA* gene —encoding PBP2—, and some mosaic forms of this gene, such as types X and XXXIV, associated with higher levels of resistance to ESCs[88,97].
- **Fluoroquinolones:** this family of antibiotics act by inhibiting the DNA gyrase and topoisomerase IV, which are essential for the metabolism of DNA in the bacterial cell. Mutations in the *gyrA* gene —encoding subunit A of DNA gyrase—, and/or in *parC* and *parE* genes —encoding subunits of topoisomerase IV— reduce the binding affinity of fluoroquinolones to such enzymes[88,94]. Also, the NorM efflux pump contributes to resistance to fluoroquinolones; point mutations within

the -35 region of the *norM* promoter increase resistance to such antibiotics[98].

- **Macrolides:** this family of antibiotics block protein synthesis by binding to the 50S ribosomal subunit, blocking the peptide exit channel by interacting with 23S rRNA. Azithromycin is the macrolide of election for treating gonorrhea. Mutations in 23S rRNA block or reduce the target affinity for the drug[99,100]. Also, the MacA-MacB efflux system exports macrolides. A SNP in the -10 region of *macAB* promoter enhances resistance to macrolides[101].
- **Tetracyclines:** these antibiotics inhibit protein synthesis by binding to the 30S ribosomal subunit. A point mutation in the *rpsJ* gene, which encodes the ribosomal protein S10, reduces the affinity of the drug to its target[102]. As penicillin, mutations in the *porB* and *pilQ* determinants reduce susceptibility to tetracyclines. Also, the presence of TetM-encoding plasmids confers resistance to these drugs, as TetM binds to the tetracycline-binding region in the 30S ribosomal subunit[103].
- **Spectinomycin:** this drug interacts with the 16S rRNA, blocking protein elongation during translation. SNPs in the 16S rRNA reduce the affinity of spectinomycin to its target[104]. Also, mutations in *rpsE* —encoding the 30S ribosomal protein S5— prevent binding of the antibiotic to the ribosomal subunit[105,106].
- **Multiple antibiotics:** The MtrCDE efflux pump can recognize several antibiotics, such as penicillin, macrolides, ESCs, and tetracyclines. Mutations in the efflux repressor MtrR or in the -35 region of its promoter can lead to the efflux pump overexpression, which increases resistance to those antibiotics[107,108]. Also, mosaic forms of *mtrR* and *mtrD* occurring together enhance resistance to macrolides[109].

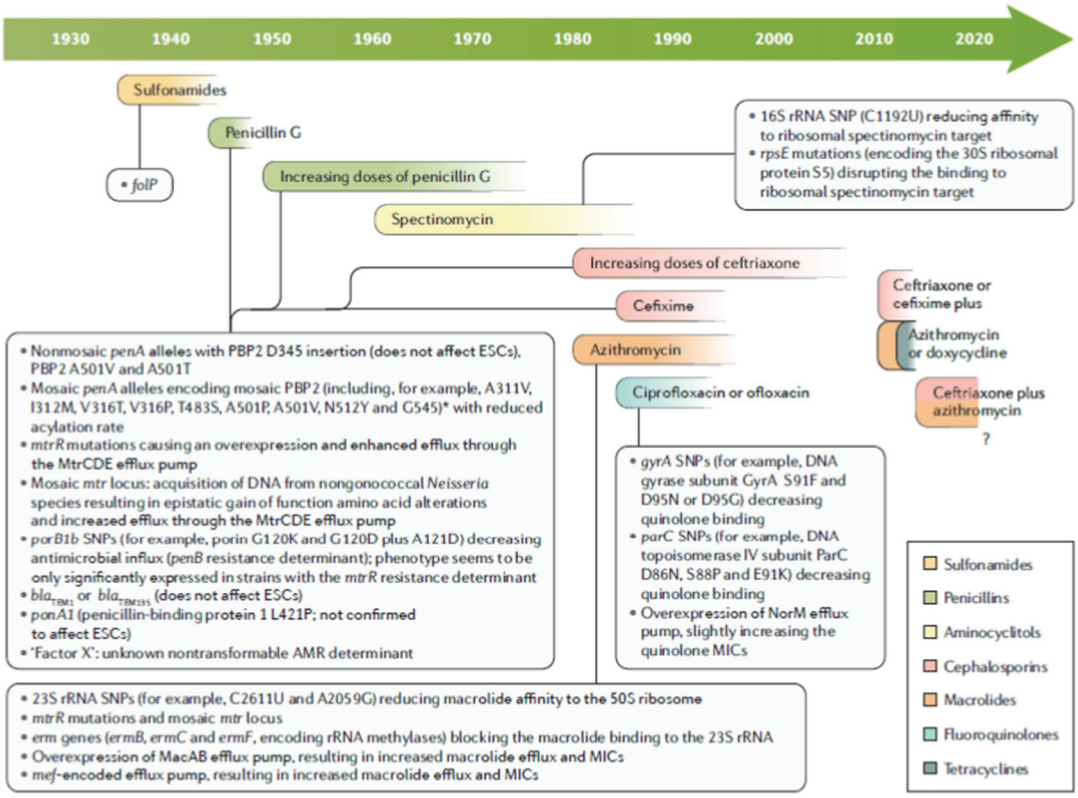


Figure 5 | Timeline of antimicrobial therapy for gonorrhea and acquisition of resistance determinants to such antibiotics. Source: Unemo *et al.*, 2019[110].

The incidence of gonorrhea has increased worldwide in the last years[23,111,112]. In 2016, the WHO estimated a global incidence of 86.9 million cases[17] (Figure 6). In 2017, there were 89,624 cases in Europe[113] (Figure 7). The same year, there were 8,732 reported cases in Spain, and 11,044 cases in 2018[114] (Figure 8). This, along with the increasingly common acquisition of resistance to antibiotics[88,115–119], has led to the declare *N. gonorrhoeae* as a threat to public health and it has been included in the WHO list of priority pathogens that highly need the development of new antibiotics[120].

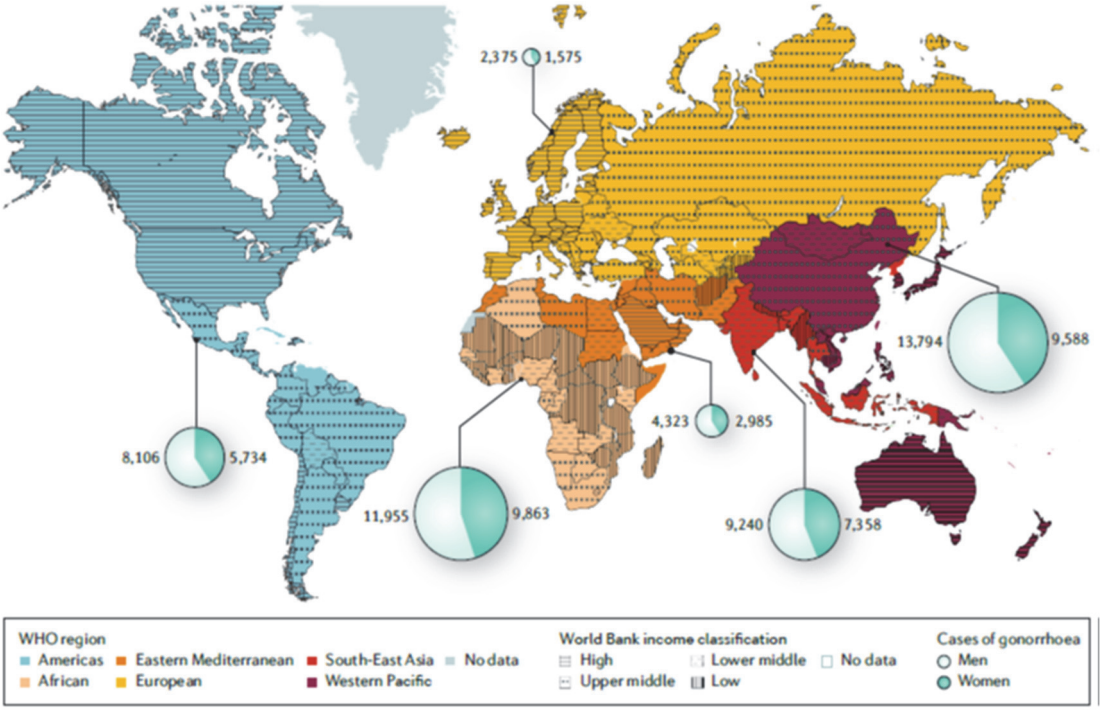


Figure 6 | Global incidence of gonorrhea cases, 2016. The number of cases are in millions. Source: Unemo *et al.*, 2019[110].

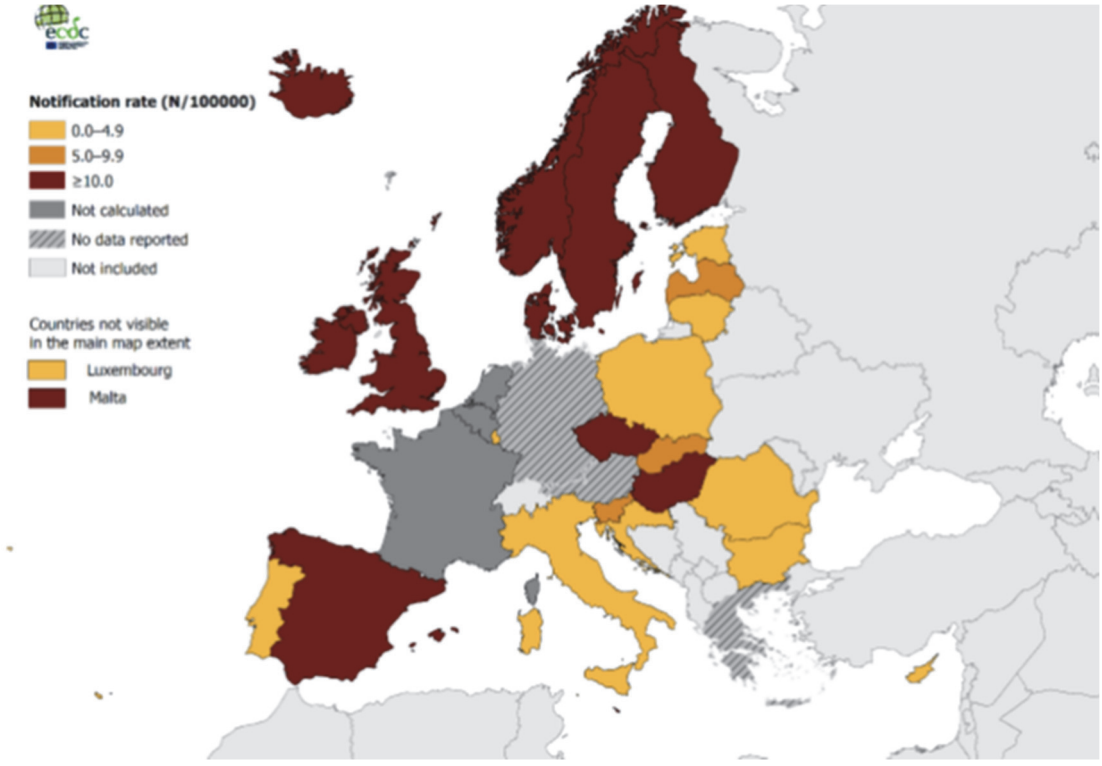


Figure 7 | Incidence of gonorrhea cases in Europe, 2017. Source: ECDC, 2019[113].

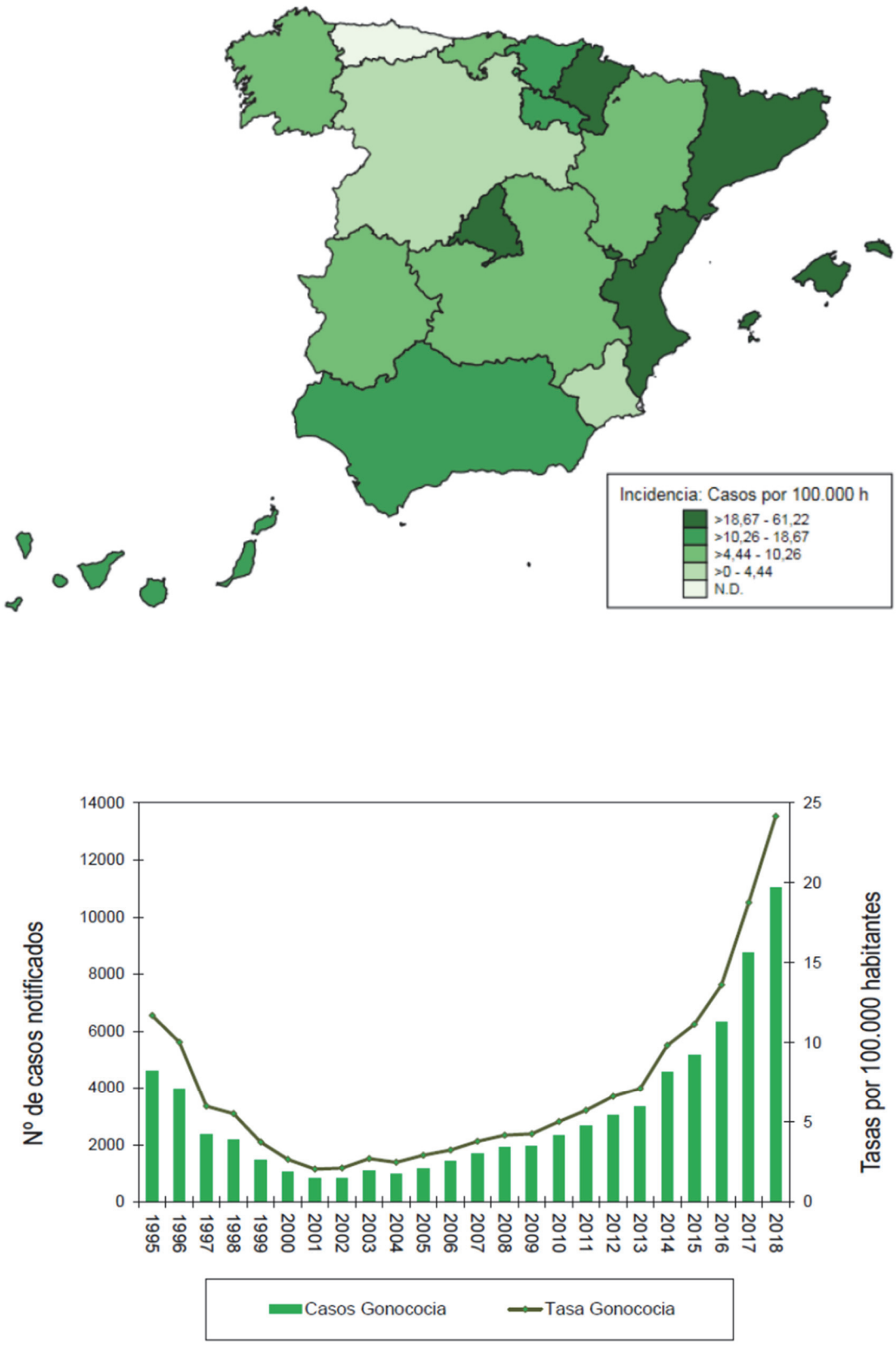


Figure 8 | Incidence of global cases of gonorrhea in Spain in 2018 by region (top), and incidence of cases in Spain for period 1995-2018 (bottom). Source: ISCIII, 2020[114].

3.2. Serratia marcescens*: an opportunistic pathogen causing healthcare-associated outbreaks*

Serratia marcescens is a Gram-negative bacterium belonging to the *Gammaproteobacteria* class. It is a rod-shaped, motile, and non-spore forming enterobacterium. Traditionally, it has been included in the *Enterobacteriaceae* family, but a genomic study reviewed the taxonomy of enterobacteria and now it is included in the *Yersiniaceae* family[121] (Figure 9).

As of January 1, 2021, there were 701 *S. marcescens* genomes deposited at the NCBI database, 72 of which were closed[122]. According to data from these 72 complete genomes, the size of its genome is around 5.30 Mb (range 4.93-5.90 Mb), with 4,846 coding sequences on average (range 3,998-5,430), and a GC content around 59.50% (range 58.45–60.20%).

S. marcescens is ubiquitous in the environment and can be found in soil, water, food, and animals. It is characterized by its ability to produce prodigiosin, a red pigment which is linked to medieval miracles such as holy bread and statues bleeding[123]. In the past, it was considered as a non-pathogenic, saprophytic microorganism. Because of the consideration of *S. marcescens* as a harmless microorganism along with the characteristic production of prodigiosin, this bacterium was used in medical and military experiments to track infections[124]. However, around the middle of the 20th century, *S. marcescens* was associated with nosocomial infections, being considered as an opportunistic pathogen, especially in healthcare settings[125,126]. Since the 1950s, the reports of *S. marcescens* outbreaks have been increased, with an incidence of 5 out of 39 outbreaks in neonatal intensive care units being caused by this microorganism[127].

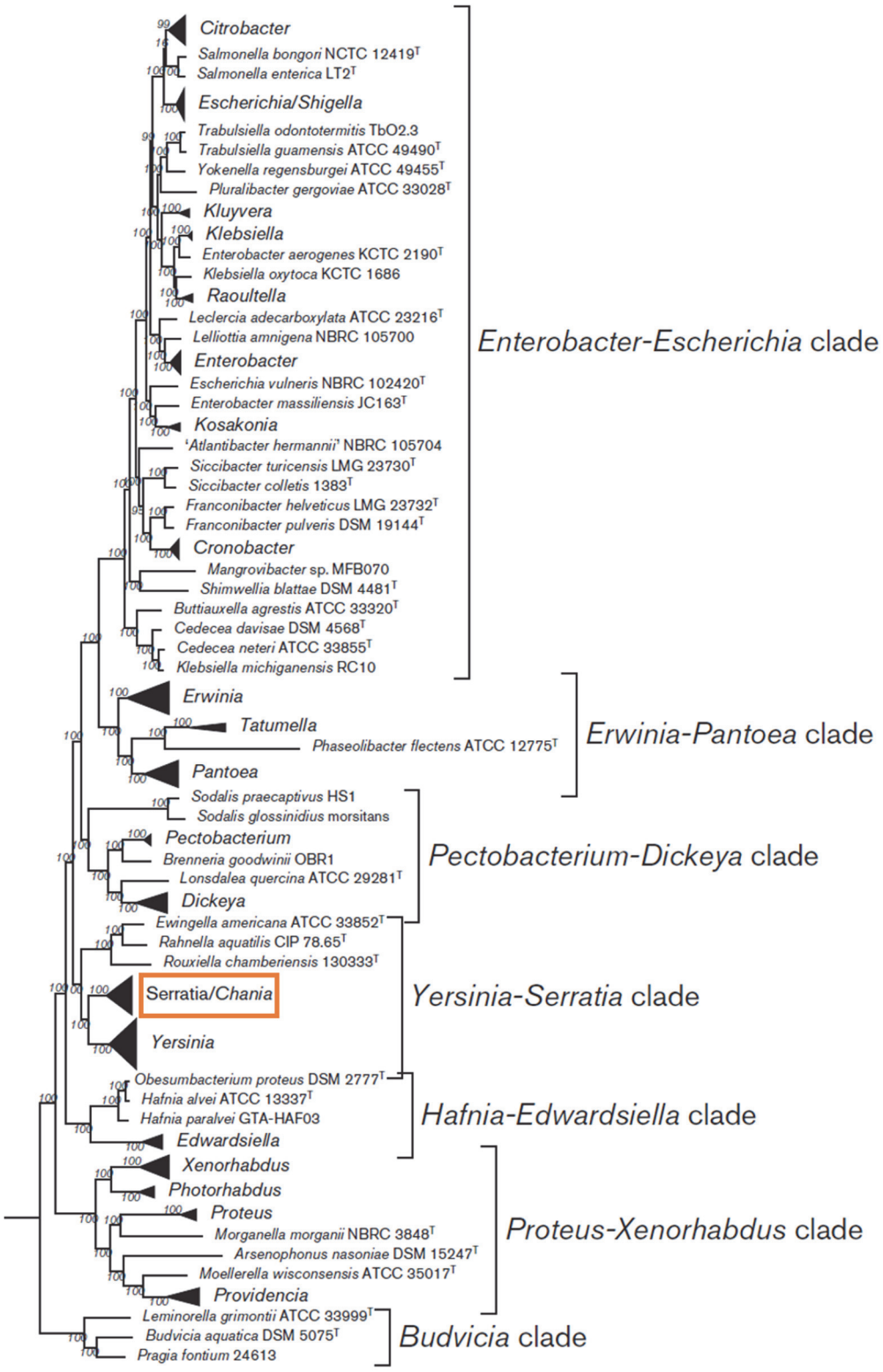


Figure 9 | Maximum-likelihood phylogenetic tree for the order *Enterobacteriales*. The tree was constructed using 1,548 core proteins. *Serratia* clade was highlighted. Source: adapted from Adeolu *et al.*, 2016[121].

S. marcescens produces different hemolysins that are toxic to different host cells[124], and infections by this microorganism have a broad clinical spectrum, including urinary tract infections, skin and soft tissues infections, conjunctivitis, pneumonia, septicemia, endocarditis, and meningitis[128–138]. Curiously, prodigiosin non-producing strains seem to be more significant in clinical specimens[139].

Most *S. marcescens* outbreaks involve intensive care units (ICUs), especially neonatal ICUs (NICUs), due to the multiple risk factors of the newborns admitted in them, such as prematurity, low weight, immature or compromised immune system, or antibiotic treatment[130,132,140].

The usual treatment may include piperacillin-tazobactam, fluoroquinolones, aminoglycosides, or carbapenems[124]. Antibiotic choice depends on the site of infection and the susceptibility testing results, as it is possible that AMR emerges during the treatment course. *S. marcescens* belongs to the ESCHAPPM group —*Enterobacter* spp, *S. marcescens*, *Citrobacter freundii*, *Hafnia* spp, *Aeromonas* spp, *Providencia* spp, *Proteus vulgaris*, and *Morganella morganii*—, which have an inducible, chromosomal AmpC β-lactamase that confers intrinsic resistance to multiple antibiotics, such as ampicillin, amoxicillin, first- and second-generation cephalosporins, macrolides, tetracycline, nitrofurantoin, and colistin[141,142]. However, although some cases of multidrug-resistant *S. marcescens* have been described, it seldom causes complicated outbreaks[143,144].

3.3. Lactococcus garvieae*: an emerging zoonotic pathogen*

Lactococcus garvieae is a Gram-positive bacterium belonging to the *Bacilli* class. It is a non-motile, non-spore forming coccus occurring in short chains. *L. garvieae* is an ubiquitous bacterium that can be found in water, soil, vegetables, animal skin, milk, and dairy products[145–147].

As of January 1, 2021, there were 42 *L. garvieae* genomes deposited at the NCBI database, 9 of which were closed[148]. According to the data from these 9 complete genomes, the size of its genome is around 2.07 Mb (range 1.95-2.29 Mb), with 1,924 coding sequences on average (range 1,792-2,130), and a GC content around 38.36% (range 37.79-38.80%).

The genus *Lactococcus* comprises several species being *Lactococcus lactis* and *L. garvieae* the most significant ones. *L. lactis* is considered safe for humans and animals[149], and it is commonly used in the dairy industry. However, *L. lactis* has also been associated sporadically with animal and human infections[150–153]. In contrast, *L. garvieae* is an important pathogen for freshwater and marine fish that causes important economic losses in aquaculture[145,154]. It has also been frequently isolated from infections in other animals, such as cattle or pigs, and humans[146,147,155–157].

L. garvieae is the causative agent of lactococcosis in fishes, a sepsis whose clinical symptoms include anorexia, lethargy, loss of orientation and erratic swimming, multiple hemorrhages, splenomegaly, enteritis, and a high rate of deaths[145,158]. In addition to causing disease in fish, it is also associated with mastitis in ruminants, and pneumonic processes and pericarditis in other animals[146,147,155,159–162]. However, in the last few years this pathogen has been isolated in an increasing number of human infections[157,163], and is currently considered as an emerging zoonotic agent.

As an emerging human pathogen, *L. garvieae* is associated with different clinical manifestations, such as urinary infections, meningitis, bacteremia, peritonitis, liver abscess, osteomyelitis, or endophthalmitis but, most notably with endocarditis, which represents more than 50% of the total human infections associated with this pathogen[157,164,165]. The source of human

infections seems to be the ingestion of contaminated foods, mainly raw fish, seafood or crude dairy products, and subsequent passage into the bloodstream and spread to diverse organs[157,163,166]. The existence of defects such as alterations of the heart valves has been described as one of the main risk factors for developing the infection[167,168]. Treatment depends on the antibiogram results, being the dual therapy of amoxicillin and gentamicin the most used to treat endocarditis, but ampicillin, ceftriaxone or vancomycin are also recommended, and there are no remarkable AMR in this pathogen[165,169].

OBJECTIVES

The main aim of this thesis is to study the genomic epidemiology of several bacterial pathogens through obtaining complete genomes of some isolates using high-throughput sequencing (HTS) and evaluating the usefulness of this tool for its application in epidemiology and forensic settings.

A series of secondary objectives are derived from the main objective:

1. To study the genomic variability of *Neisseria gonorrhoeae* in clinical samples from populations of the Comunidad Valenciana, Catalonia and Madrid.

2. To define the population structure and the dynamics of spatial and temporal transmission of *N. gonorrhoeae* in these regions.

3. To compare HTS and the bacterial typing schemes of *N. gonorrhoeae* by evaluating the correspondence between the results of both types of data.

4. To analyze the recombination events between the different structural groups of *N. gonorrhoeae*.

5. To detect and evaluate antimicrobial resistance determinants contrasting with treatment response data (phenotype) in *N. gonorrhoeae*.

6. To apply the analysis of sequences obtained by HTS to resolve a forensic case of gonorrhea transmission to a minor in an alleged case of sexual abuse.

7. To analyze two nosocomial outbreaks of *Serratia marcescens* in the neonatal intensive care units from two hospitals of the Comunidad Valenciana.

8. To evaluate the impact of choosing different reference genomes of *S. marcescens* in the interpretation of the results from outbreak analyses.

9. To study the pangenome of *Lactococcus garvieae* and to detect recombinant genes in a clinical strain from a fatal endocarditis at three taxonomic levels —species, genus, and class—.

METHODS

1. Sample processing and metadata generation

Data associated to our analyses, such as phenotypic traits or demographic information, were obtained by the collaborating hospitals staff. We received the extracted DNA at FISABIO were we proceed to quantify and verify its integrity. Next, samples were handled over to the FISABIO sequencing service staff who proceeded to prepare libraries and sequencing as detailed below. *Lactococcus garvieae* strain was sequenced by Life-Sequencing company.

1.1. Sampling, isolation and identification of Neisseria gonorrhoeae

Sampling was carried out as recommended by the national and international guidelines[170–172]. Specimens collected were urethral, rectal, oropharyngeal, balanopreputial, and endocervical swabs, as well as a synovial fluid puncture.

Depending on the needs of each hospital, the specimens were cultured in one or both of the following media, as recommended by the guidelines[173]:

- Chocolate agar PoliVyteX™ (bioMérieux, Marcy-l'Étoile, France). A non-selective medium with heated equine or bovine blood and a commercial mix of supplements to provide growth factors for exigent species.
- Thayer–Martin medium[174]. A selective medium for *Neisseria* species, which is basically a chocolate agar medium containing antimicrobial agents —vancomycin, colistin, and nystatin or another antifungal agent— to inhibit the growth of other bacteria, and fungi.

In both cases, plates were incubated at 35–37°C in a 5% CO_2 atmosphere, which was supplied by a CO_2 incubator or CO_2-generating tablets. Plates were examined at 24 and 48 hours, and suspected *N. gonorrhoeae* colonies were identified by Gram stain, the cytochrome oxidase test (Remel Pathotec™, Termo Fisher Scientific, MA, US), the catalase test using 3% hydrogen peroxide, and mass spectrometry (MALDI-TOF) using the VITEK® MS system (bioMérieux).

1.2. Sampling, isolation and identification of Serratia marcescens

Isolates of *Serratia marcescens* were obtained from two outbreaks in two Valencian hospitals. Clinical specimens collected were conjunctival, perineal, blood, pharyngeal, and perianal. Environmental specimens were also collected, being sampled from medical instruments —incubator tubing and hood, peripheral catheter, and endotracheal tube—, mother's milk —corresponding to breast pump contamination or contamination by the mothers' manipulation during milk extraction—, milk waste container, sinks, and healthcare workers' hands.

Collected specimens were cultured by duplicate in both media:

- Chocolate agar supplemented with PolyViteX™ (bioMérieux). As described for gonococci.
- CHROMagar™ Orientation Medium (BD, NJ, US)[175]. A non-selective medium focused especially in urinary tract pathogens. It contains chromogenic substrates which allow to distinguish between different genus and/or species by the color and appearance of the colonies. For *Serratia* species, the colonies are medium-sized and dark blue.

Plates were incubated at 37ºC without atmospheric requirements, as *S. marcescens* is a facultative anaerobic bacterium. Plates were examined at 24 hours and suspected colonies were identified by mass spectrometry (MALDI-TOF) using the VITEK® MS system (bioMérieux).

1.3. Sampling, isolation and identification of Lactococcus garvieae

Lactococcus garvieae strain Lg-Granada was isolated from a fatal case of infective endocarditis in the city of Granada (Spain). As explained in the case report study[176], a blood specimen was taken from the patient and cultured in blood agar and CHROMagar in aerobic conditions. Plates were examined at 24 hours and colonies were identified by mass spectrometry (MALDI-TOF) using the VITEK® MS system (bioMérieux). After the patient's demise, a sample of the heart valve tissue was taken to analysis by PCR and sequencing of 16S rRNA, confirming the species.

To isolate the Lg-Granada strain in the Madrid laboratory for high-throughput sequencing, the strain was cultured using log-phase cultures (OD600, ~1) in brain heart infusion (BHI) broth (Panreac AppliChem, Barcelona, Spain).

1.4. Antimicrobial susceptibility testing

The three species studied here were subjected to antimicrobial susceptibility tests as part of the working algorithm of each hospital. The susceptibility testing was performed using the quantitative method Etest (bioMérieux). Etest consists of a strip with a scale and a gradient of antibiotic concentration that diffuses to agar medium, creating an inhibition ellipse. The point where the ellipse intersects the strip indicates the minimum inhibitory concentration (MIC) for the antibiotic tested, which is the lowest concentration that prevents microorganism growth. The interpretation of the results as susceptible, intermediate resistance —or decreased susceptibility—, or resistant (SIR system) was performed using the values established by the European Committee on Antimicrobial Susceptibility Testing (EUCAST)[177].

Isolates of *N. gonorrhoeae* in Chapter 1 were tested for 9 antibiotics, with variations depending on the hospitals:

- **Penicillin G.**
- **Tetracycline:** doxycycline.
- **Macrolide:** azithromycin.
- **ESCs:** cefixime, cefotaxime, and/or ceftriaxone.
- **Fluoroquinolones:** ciprofloxacin or levofloxacin.
- **Aminoglycoside:** spectinomycin.

Additionally, the chromogenic test of nitrocefin was performed to detect PPNG isolates, which had a plasmidic β-lactamase. This was done using BD BBL™ Cefinase™ paper discs (BD).

There was no report of antimicrobial susceptibility testing of gonococci in the forensic analysis of a transmission case (Chapter 2).

Isolates of *S. marcescens* in Chapter 3 were tested for 17 antibiotics, but only the outbreak B isolates —excluding the control—. There was no report of antimicrobial susceptibility testing of outbreak A isolates. The antibiotics tested were:

- **Penicillins:** ampicillin, amoxicillin/clavulanic acid, piperacillin/ tazobactam.
- **Carbapenems:** imipenem, ertapenem.
- **ESCs:** ceftazidime, cefotaxime, cefuroxime, cefepime, cefoxitin, cefuroxime axetil.
- **Fluoroquinolones:** ciprofloxacin, nalidixic acid.
- **Aminoglycosides:** amikacin, gentamicin.
- **Sulfonamide:** co-trimoxazole (trimethoprim/sulfamethozaxole).
- **Tetracycline:** tigecycline.

However, these data were not relevant to the objectives covered in this thesis (see Chapter 3), so they are included as supplementary information in appendix V.

The *L. garvieae* isolate in Chapter 4 was tested for 7 antibiotics during patient stay at the hospital as described in the case report study[176]. The antibiotics tested were:

- **Penicillin.**
- **ESC:** cefotaxime.
- **Fluoroquinolone:** levofloxacin.
- **Macrolide:** erythromycin.
- **Lincosamide:** clindamycin.
- **Vancomycin.**
- **Daptomycin.**

After isolation of the strain Lg-Granada in the Madrid laboratory, the antimicrobial susceptibility test was replicated using the disc diffusion method using Oxoid™ antibiotic discs (Thermo Fisher Scientific) in Mueller Hinton blood (MHS) agar (bioMérieux). As no specific inhibition zone diameter (IZD) breakpoints for *L. garvieae* are available, the IZD breakpoints were those recommended by the French Society of Microbiology for testing Gram-positive bacteria[178]. The AMR profile was identical in both cases, and the results are attached as supplementary information in appendix VI.

1.5. DNA extraction and quality checking

Bacterial DNA was extracted using by different methods depending on the hospital:

- Gonococci in Chapter 1: Comunidad Valenciana (CV) isolates DNA was extracted using QIAamp DNA Mini Kit (Qiagen, Hilden, Germany), Catalonia isolates DNA was extracted using a heat shock method —colonies were resuspended in 100 µL of Milli-Q water, heated at 95°C for 10 minutes, and centrifuged at 13,000 rpm for 3 minutes; the supernatant was diluted at 1:2 with Milli-Q water—, and Madrid isolates DNA was extracted using NucliSENS® easyMAG® (bioMérieux).
- Gonococci in Chapter 2: DNA was extracted using QIAamp DNA Mini Kit.
- *Serratia marcescens* isolates in Chapter 3: DNA was extracted using NucliSENS® easyMAG®.

- *Lactococcus garvieae* in Chapter 4: DNA was extracted using Blood & Cell Culture DNA Maxi Kit (Qiagen).

DNA concentration was quantified in all cases by fluorometric methods using Invitrogen Qubit® 3.0 fluorometer (Thermo Fisher Scientific), or Picogreen® fluorometer (Promega, WI, US) in the case of *L. garvieae*, to ensure an optimal DNA concentration as an input for HTS platforms.

1.6. High-throughput sequencing

1.6.1. Short-reads sequencing strategy

High-throughput sequencing (HTS) producing short reads was the main methodology used in this thesis to obtain genomic information, and it was used in isolates of the Chapters 1 to 3, i.e. gonococci and *S. marcescens*.

Prior to the sequencing step, DNA libraries need to be prepared. Libraries allow both samples to be identified and DNA to adhere to the sequencing flowcell. Libraries were constructed using Nextera® XT Library Preparation Kit (Illumina, San Diego, CA, US). This kit performs the enzymatic fragmentation of the genomic DNA and adapters ligation in a single step called "tagmentation". Next, the indexes are attached to adapters in a PCR step (Figure 10).

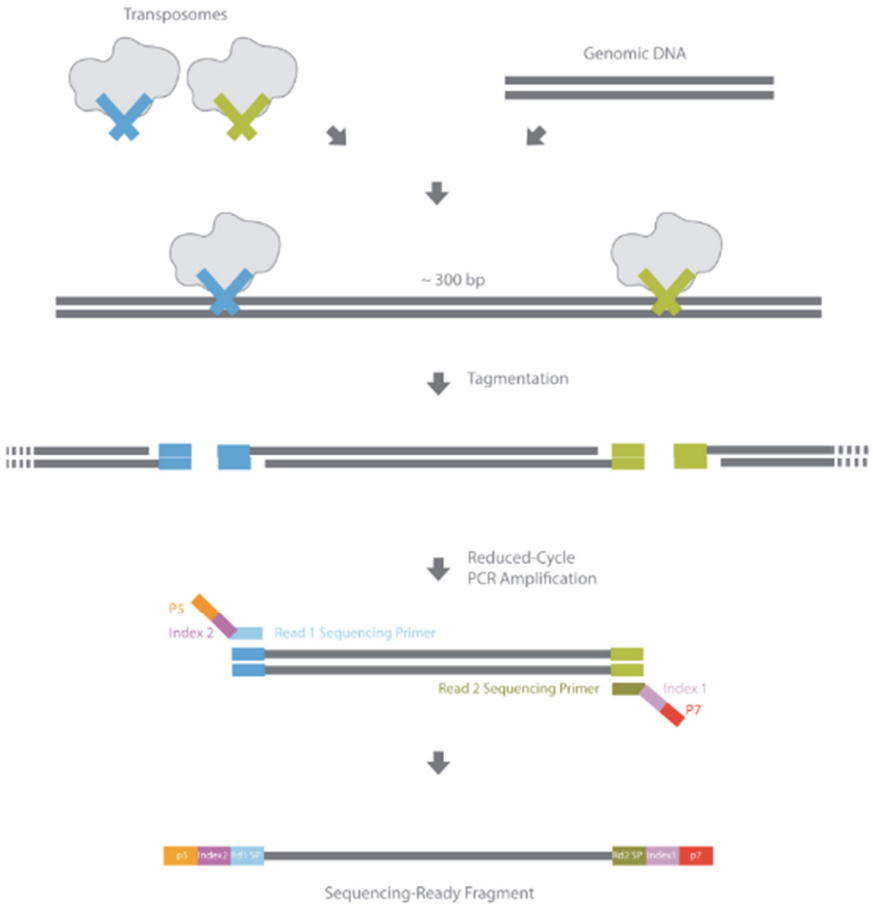


Figure 10 | Library preparation using Nextera XT kit. Source: Illumina.

Adaptors are short DNA sequences with several functions (Figure 11). P5 and P7 edges are regions that bind to oligonucleotides on the flowcell surface. Indexes are unique sequences —or "barcodes"— which are used to differentiate samples during data analysis, and allow to save resources, as multiple libraries can be pooled together and sequenced in the same run —a process known as multiplexing—. Up to 384 uniquely indexed samples can be pooled and sequenced together, with an input of 1 ng of DNA by sample. Finally, adaptors have binding regions for the primers thus enabling forward and reverse reads —in paired-end reads, as is the case—.

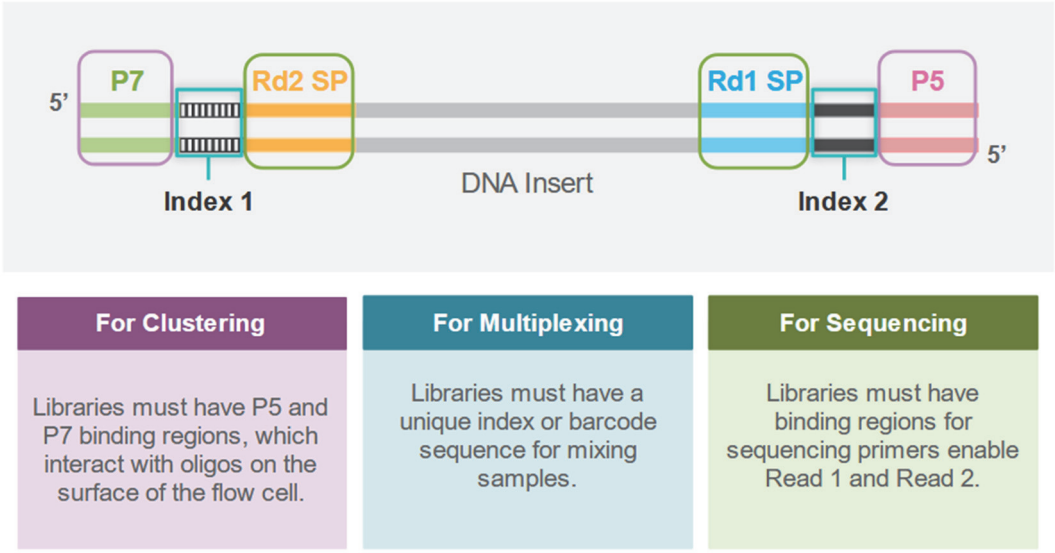


Figure 11 | Structure and functions of library adaptors. Source: Illumina.

To control the size of DNA fragments we performed a capillary electrophoresis using a Fragment Analyzer System (Agilent Technologies, Santa Clara, CA, US). Controlling the fragment size is crucial for sequencing because shorter fragments amplify more efficiently than longer sequences, while longer fragments generate more diffuse clusters than shorter sequences. In order to balance these features, the target size is around 300 bp, but successfully sequenced libraries of up to 1.5 Kb can be obtained[179].

Once the libraries were generated, they were attached to the flowcell surface by means of the oligonucleotides previously fixed onto it. Then, clusters are generated as the fragments are attached to the flowcell and a polymerase amplifies them generating millions of copies, which form localized clusters on the flowcell. Finally, the sequencing step takes place in a process called sequencing by synthesis (SBS), by union of complementary nucleotides to the DNA strand. These nucleotides are marked with a fluorescent tag and bind one at a time releasing the fluorescent signal which is detected by optic sensors (Figure 12). The sequencing platform used was Illumina® NextSeq™ 500, which generated 150 bp paired-end reads.

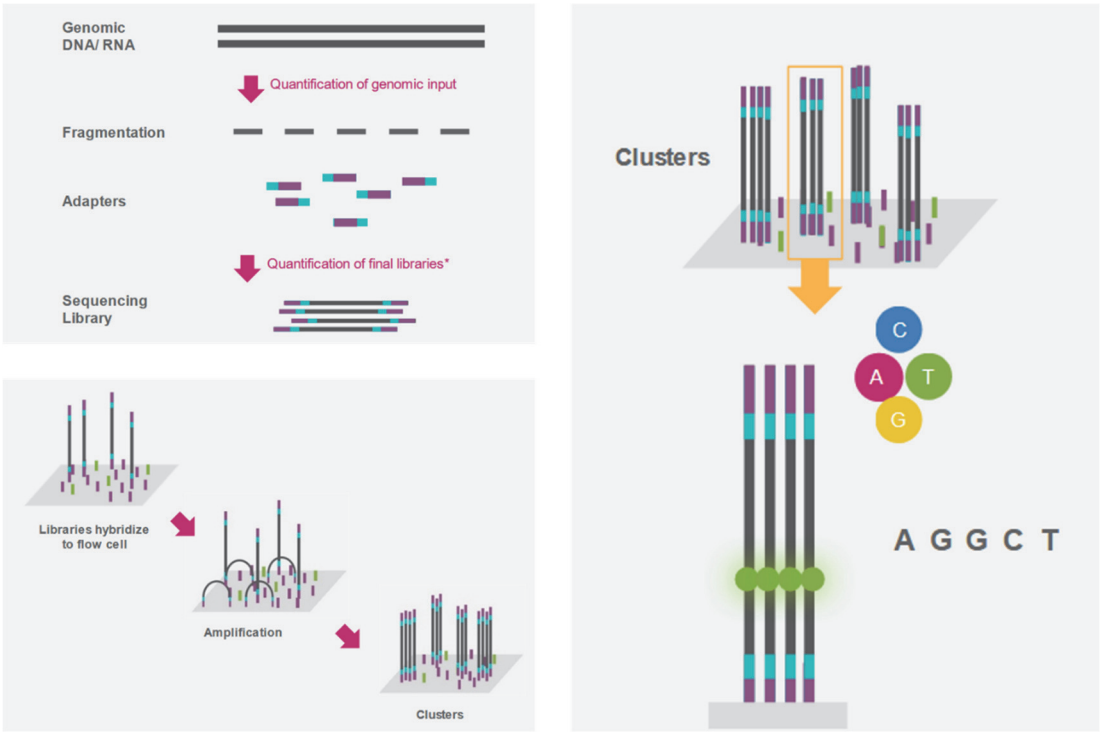


Figure 12 | Steps of Illumina high-throughput sequencing by synthesis. Top left summarizes the genomic DNA fragmentation and library preparation step. Bottom left depicts the binding of the libraries to the flowcell and the generation of clusters step. Right represents the fluorescent nucleotides binding and the detection of the fluorescent signal. Source: Illumina.

1.6.2. Long-reads sequencing strategy

The *Lactococcus garvieae* isolate analyzed in Chapter 4 was sequenced using a different HTS method, the PacBio single molecule real-time (SMRT) sequencing methodology.

The initial steps are very similar to those described above. First, DNA is fragmented and fragments' size is controlled using BluePippin™ (Sage Science, Beverly, MA, US), a DNA size selector which uses pulsed-field electrophoresis for collecting fragments of the optimal size. In this case, the optimal DNA length was around 15 to 25 Kb. Libraries are generated by ligation of hairpin adapters, to which the sequencing primers attach (Figure 13). Samples may be barcoded to multiplex in order to increase the throughput.

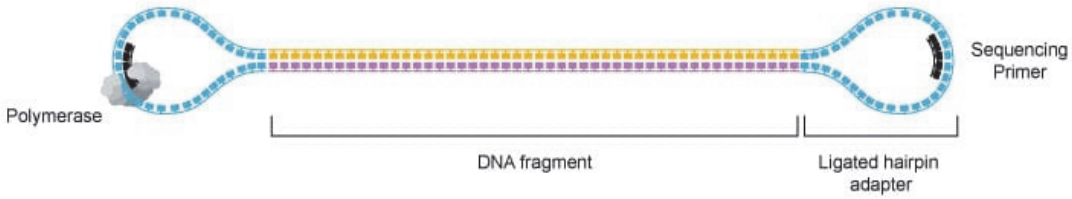


Figure 13 | PacBio sequencing library with primer and polymerase attached. Source: PacBio.

L. garvieae strain Lg-Granada was sequenced in a PacBio® RS II platform (Pacific Biosciences, Menlo Park, CA, US) using the P6-C4 polymerase chemistry combination, a PacBio proprietary technology that allows obtaining longer reads than previous versions. The libraries are immobilized in the wells of the SMRT cell and the polymerase incorporates fluorescent nucleotides that release a luminous signal that is measured in real time. The shape of the library allows the polymerase to move through the insert several times, improving the reliability of the signals (Figure 14).

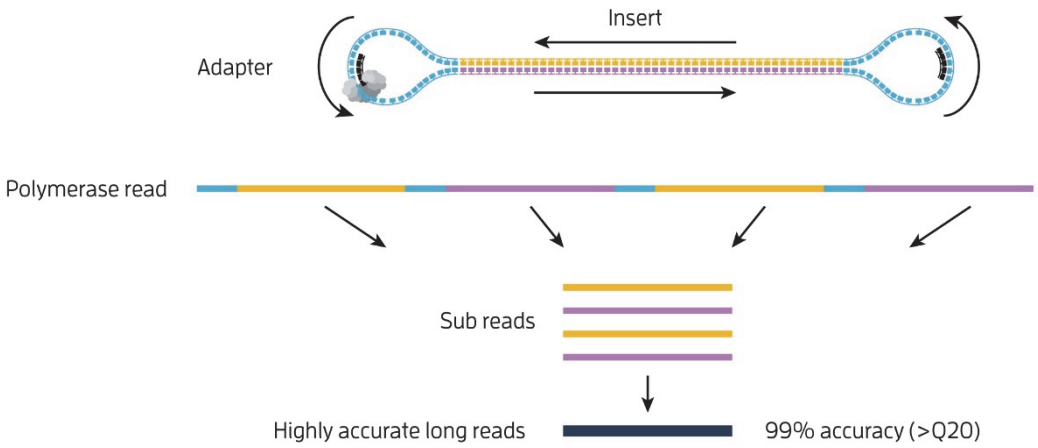


Figure 14 | Sequencing steps in PacBio SMRT technique. Source: PacBio.

2. Primary analysis

Bioinformatics analyses of HTS data can be divided into three stages. The primary analysis is focused on quality control and assurance of the reads generated by the sequencing platform. Different versions of each program were used in the different chapters and they are detailed in appendix I.

The first step is to generate a report with basic statistics of the raw reads, such as the number of reads, their length, and several plots with parameters indicative of their quality. This task is done with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). As we obtain a report for each fastq file, we can generate a summary report using MultiQC[180].

The next stage is to clean the raw reads to remove low-quality sequences and adaptors. To do this, PRINSEQ-lite[181] was used to filter sequencing reads shorter than 50 bp, and trimming those sequences with an average *Phred* quality score (*Q*) below 30, calculated using a sliding window of 20 bp. The *Q* score is a measure of the quality assigned to each nucleotide that is logarithmically linked to the probability of error (P_e) in nucleotide calling[182,183]. This relationship is defined by the following formula:

$$Q = -10 \log_{10} P_e$$

Therefore, a Q of 30 implies a probability of error in the nucleotide calling of 1 in 1,000 or, what is the same, an accuracy in the nucleotide calling of 99.9%.

After the read cleaning step, we run again FastQC and MultiQC to generate the corresponding quality reports.

To check the identity of the isolates, we used Kraken[184] and the pre-built 8 GB MiniKraken database, which was constructed from complete prokaryotic and viral genomes in RefSeq. Kraken splits the reads in shorter sequences of *k* nucleotides (*k*-mers) and maps them against the most recent common ancestor of all the genomes containing the same *k*-mer. So, it assigns taxonomic labels to the reads, allowing their identification. This step was double checked using Mash Screen[185], which also splits reads in *k*-mers and screens them against a pre-built RefSeq database (available at https://gembox.cbcb.umd.edu/mash/refseq.genomes.k21s1000.msh), thus assigning an identity score between the set of reads and the genomes in the database. Both programs are designed to classify genomes in metagenomic samples and to detect contamination in sets of reads. However, they are useful for our goals because they inform about the species present in highest proportion among the reads, thus allowing the identification of the species. It should be noted that this step was carried out only in *Serratia marcescens* in Chapter 3, because two isolates did not cover the reference genome during the mapping step, which led us to suspect that they were possibly not *S. marcescens*.

3. Secondary analysis

After the quality evaluation and filtering of the reads, we can proceed with the secondary analyses that reorder the reads to reconstruct the genome under study. There are two ways to do this: mapping and assembly.

Assembly is computationally more demanding than mapping —in RAM memory and CPU time—[186], it usually has more sources of error than mapping[187], and the reads' length is a critical parameter for assembly while we can map short reads with high accuracy[188].

The strategy followed in this dissertation is a combination of both methodologies.

3.1. Mapping

The procedures described in this section were employed in Chapters 1 to 3.

3.1.1. Election of a reference genome

Before mapping the cleaned reads, a genome must be selected to act as a reference for alignment. The election of the reference depended on the study as follows:

- **Chapter 1:** the reference genome used was *N. gonorrhoeae* FA1090 strain (NCBI accession NC_002946.2), because it is the representative genome for the species.
- **Chapter 2:** the reference was *N. gonorrhoeae* WHO P strain (NCBI accession LT592157.1), which is the closest reference genome to the isolates of the study. To identify the closest reference genome, the tool kmerID (https://github.com/phe-bioinformatics/kmerid) was used. This program computes a similarity index between the raw reads of each sample and each of the 23 complete *N. gonorrhoeae* genomes selected from NCBI database as of the starting date of the study

(August, 2017) by calculating the percentage of 18-nucleotide *k*-mers (18-mers) in the reference that were also present in the reads.

- **Chapter 3:** two references were used, *S. marcescens* UMH9 strain (NCBI accession NZ_CP018923.1), which is the closest reference genome to the isolates of the first outbreak, and *S. marcescens* Db11 strain (NCBI accession NZ_HG326223.1), which is the representative genome for the species. The choice of UMH9 strain as the closest reference genome was done using kmerID and the 21 complete *S. marcescens* genomes available in the NCBI database as of the starting date of the analyses (March, 2018).

3.1.2. Mapping and processing of the alignment files

Once the reference genomes were chosen, the reads were aligned against them using BWA-MEM[189], which was selected over other mappers for its excellent relationship between alignment accuracy and running time[190]. The algorithm of this mapper seeds local alignments that match exactly and then extends the seed over the entire length of the read, discarding putative misalignments based on a score threshold. The resulting alignment was written in SAM (Sequence Alignment/Map) format, which was converted to binary format —BAM (Binary Alignment/Map) format— using SAMtools[191] to save computational resources.

Next, the BAM file was processed to remove artifacts that could bias the results in the next step. Firstly, duplicates were removed using the *MarkDuplicates* tool from the Picard suite (http://broadinstitute.github.io/picard). Duplicates are reads generated from the same DNA fragment. They can be originated by errors during the construction of libraries, or because the optical sensor of the sequencer has detected two amplification clusters that are actually only one —optical duplicates—. Duplicates could bias the results of the variant calling step if they carry a wrong polymorphism, since they would support this false SNP.

Finally, the *IndelRealigner* tool from GATK[192] was used to detect insertions/deletions (*indels*) and to realign the reads against those regions. As mappers can only consider each read independently, they can, for example, favor alignments with mismatches instead of opening gaps in the read or the reference, generating a mapping error. *Indel* realignment improves the mapping of reads in these regions.

3.1.3. Variant calling

Subsequent to mapping the reads, the polymorphisms present in the reads with respect to the reference genome must be identified in a step known as "variant calling".

Variant calling was performed with the SAMtools/BCFtools[191] suite. Here, we used strict quality filters to ensure accuracy of the data:

- Positions with base and/or mapping qualities below a Phred score of 30 were filtered.
- The minimum depth coverage for calling a variant position was adjusted to 10 and to 5 in the case of a non-variant position, with at least 90% of the reads coincident with the corresponding call.

When these conditions were not met, an undetermined base (N) was called in that position. The main disadvantage is that if the reads do not have a good level of quality, we may lose a lot of information. In contrast, these filters ensure the reliability of the data obtained.

A BCFtools script —VCFutils— together with the program seqtk (https://github.com/lh3/seqtk) were used to generate quality draft genomes that were joined into a single multiple alignment file along with the corresponding reference genome.

3.2. Assembly

We approach the genome assembly from different perspectives depending on the problem analyzed in each chapter.

3.2.1. Assembly of short-reads (Chapter 1)

We use SPAdes[193] to reconstruct the gonococcal genomes from the cleaned reads. This assembler uses a *k*-mer-based algorithm in which several *k*-mer sizes are used simultaneously. It constructs a de Bruijn graph which detects and corrects problematic regions, estimates the distances between graph edges —i.e. between *k*-mers in the genome—, and constructs the contigs. Furthermore, we used the option *careful* which activates an additional step of mapping with BWA to detect and correct mismatches. Also, we provided an additional set of auxiliary contigs to improve the quality of the results. These auxiliary contigs were created using IDBA-UD[194], another assembler based on the de Bruijn graph algorithm. The quality of the assembly was checked with QUAST[195], which reports useful parameters such as the number of contigs —indicating how fragmented your assemblies are—, the N_{50} —the median of contigs length, indicating that the 50% of the assembled genome is contained in contigs of this length or longer—, the number of undetermined nucleotides (N), or the length of the assembled genome.

In this chapter, contigs were used to detect plasmids, the gonococcal genetic island (GGI), AMR determinants, and to curate typing results (see tertiary analysis section).

3.2.2. Assembly of unmapped reads (Chapters 2 and 3)

The mapping step only allows analyzing the mapped fraction of the genome, which is usually almost the complete genome. However, we can lose part of important information contained in the unmapped fraction, as the reference genome may lack some genes or plasmids that the studied isolates may have.

To analyze the unmapped genome, the unmapped reads were extracted from the BAM files using SAMtools and then they were assembled as explained above, firstly using IDBA-UD to construct auxiliary contigs and then with SPAdes. The quality of assemblies was checked using QUAST.

3.2.3. Assembly of long-reads (Chapter 4)

The case of *Lactococcus garvieae* isolate was different, and the steps of quality checking of reads, the *de novo* assembly and posterior genome annotation were done in the Life-sequencing company facilities.

As commented before, *L. garvieae* was sequenced using the PacBio platform which generates long reads output. Quality checking of the reads and assembly were performed using the PacBio-focused tools.

First, a pre-assembly step was performed using HGAP —implemented within SMRT analysis[196]— with default parameters and minimum seed read length set at 6,000 bp. This step generated long and highly accurate sequences by mapping single pass reads to long reads (seeds). Reads were filtered using minimal polymerase read length and minimal subread length of 500 bp, and minimal polymerase read quality of 0.8.

The second step of the approach was the assembly using the Overlap Layout Consensus (OLC) algorithm implemented by the Celera® WGS-assembler[197]. This algorithm finds overlapping regions in subsequent contigs and try to resolve mismatches by generating a consensus based on most probably nucleotides. This step was polished using the Quiver algorithm —implemented in the GenomicConsensus package available at

https://github.com/PacificBiosciences/GenomicConsensus— to reduce the remaining indel and base errors in the assembly.

Finally, the resulting two contigs —corresponding to the chromosome and a plasmid— were circularized using Circlator[198]. This program tries to rearrange the contigs by finding their *dnaA* gene, usually close to the origin of replication.

4. Tertiary analysis

Finally, the tertiary analysis encompasses all the analyses that give biological meaning to the data, helping in the interpretation of the results.

4.1. Typing

Molecular typing is a method widely used in molecular epidemiology to identify strains of microorganisms based on their genetic material.

In *Neisseria gonorrhoeae* there are three main schemes for molecular typing:

- **Multilocus Sequence Typing (MLST)**[199]**.** This scheme is based on 7 housekeeping genes —*abcZ, adk, aroE, fumC, gdh, pdhC,* and *pgm*—. Alleles of each gene are assigned with a number following the order in which they were incorporated to the database. The 7-allele combination corresponds to a sequence type (ST) profile. The MLST database of *Neisseria* spp. is available at PubMLST website (https://pubmlst.org/organisms/neisseria-spp/), and as of January 1, 2021 it contained 15,734 ST profiles.
- ***Neisseria gonorrhoeae* Multi-Antigen Sequence Typing (NG-MAST)**[200]**.** This scheme is based on 2 highly polymorphic genes —*porB,* and *tbpB*—. The database is available at http://www.ng-mast.net/, and as of January 1, 2021 it contained 22,036 ST profiles. NG-MAST STs can be clustered in genogroups if they share one of the alleles and the other has an identity ≥99% —≤5 SNPs for *porB* gene, or

≤4 SNPs for *tbpB* gene—. Each genogroup will be numbered with the major ST of the group[201].

- ***Neisseria gonorrhoeae* Sequence Typing for Antimicrobial Resistance (NG-STAR)[202].** This scheme is based on 7 well-characterized genes linked with antimicrobial resistance to cephalosporines, macrolides and fluoroquinolones —*penA, mtrR, porB, ponA, gyrA, parC,* and 23S rRNA—. The database is available at https://ngstar.canada.ca, and as of January 1, 2021 it contained 3,397 ST profiles. This scheme has the advantage of directly linking the ST profiles with AMR profiles.

All three schemes were used for typing gonococcal isolates using SRST2[203]. This program aligns the reads against the alleles of the databases and calculate scores that determine the results in the output. The alleles are assigned to each isolate and, using the table of ST profiles, the corresponding STs are also assigned.

Additionally, we used the assembled contigs to curate the results that were not very clear. To do this, we used BLAST[204] to align the contigs to the alleles database. Results with 100% identity and coverage confirm the assigned allele.

As there were no typing schemes for *S. marcescens* nor *L. garvieae,* typing was only performed with gonococci.

4.2. Phylogenetic analysis

4.2.1. Phylogenetic reconstruction

The alignment file generated after mapping and variant calling steps was used for phylogenetic reconstruction. First, we removed the problematic regions, that is:

- **Repetitive regions:** these regions can be problematic because they attract reads during the mapping step that may correspond to another identical or very similar region in the chromosome, falsely increasing the coverage of this region. This can bias the results if an erroneous SNP is found, thus being called in the variant calling step —false SNP—. Repetitive regions are detected using the *repeat-match* script implemented in MUMMER.
- **Phages, high-density of SNPs regions, and recombinant regions:** these regions are problematic as they contain an increased number of contiguous SNPs that can cause false positives during the variant calling. In Chapter 1, the phages of the FA1090 strain had been identified in a previous study[205], high-density of SNPs regions were detected using Gubbins[206], and recombinant regions were detected by likelihood mapping and topological congruence testing, as will be explained later. In Chapters 2 and 3, only repetitive regions and phages were removed, as the objective in these chapters was to compare the largest portion of the genomes of the isolates as possible. Phages were detected using PHASTER[207].

The problematic regions detected were masked using the script *remove_blocks_from_aln* (available from: https://github.com/sanger-pathogens/remove_blocks_from_aln), and then were removed together with other poorly aligned regions using GBlocks[208]. This program was run by allowing gaps in up to 50% of the sequences, with a minimum block length of 10 bp, and a maximum number of contiguous non-conserved positions of 8.

We extracted the variant positions of the cleaned alignment using SNP-sites[209] in order to reduce the computational cost for the phylogenetic reconstructions.

Phylogenetic reconstructions were performed by maximum likelihood (ML) using IQ-TREE[210]. First, we used the ModelFinder[211] tool implemented in IQ-TREE to select the most appropriate model of nucleotide substitution, based on Bayesian Information Criterion (BIC) parameter. BIC is dependent on the likelihood values, such that the model with the lowest BIC is preferred.

Once the model of nucleotide substitution is selected, the phylogenetic reconstruction starts. We used the ultrafast bootstrap (UFboot)[212] option with 2,000 replicates in Chapter 2, and 10,000 replicates in Chapters 1 and 3 to evaluate the statistical support for the groupings.

The visualization of all the phylogenetic trees —except by tanglegram in Chapter 1, which was done with the dendextend[213] package in R[214]— was performed using iTol[215].

4.2.2. Calculation of genetic distances between isolates

In Chapters 2 and 3, we needed to compute the genetic distances between isolates to establish whether such isolates were clonal. We used the cleaned alignment and pairwise nucleotide distances between the isolates were computed using the *dist.dna* function of the APE[216] package in R.

5. Population genetic structure

We analyzed the population structure of the gonococcal isolates in Chapter 1. We cluster isolates using STRUCTURE[217,218]. This program assumes a model in which there are *K* populations characterized by the allele frequencies at each locus. The isolates are assigned probabilistically to one or more populations, depending if they are admixed genetically. STRUCTURE applies the Markov chain Monte Carlo (MCMC) estimation, a Bayesian method that assigns isolates to a predetermined number of groups at random, estimates the variant frequencies in each group and then re-assigns the isolates based on those estimates. This process is repeated many times.

We run STRUCTURE for *K* values from 1 to 10 populations, with 100,000 iterations of the MCMC sampling, and a burn-in length of 10,000 —i.e. the simulation runs 10,000 times before starting to collect data to minimize the effect of the starting configuration—. For each *K* value, we performed 20 runs.

The results were collected using STRUCTURE harvester[219], which reformats data for use in downstream steps, and performs the Δ*K* method[220] to select the optimal *K* value of the dataset. This method calculates the average value of log likelihood, L(*K*), at each step of the MCMC, then calculates the absolute value of the difference between successive values of L(*K*), and finally, estimates Δ*K* by dividing these differences by the standard deviation. The optimal *K* value of populations corresponds to a peak in the Δ*K* value when plotted against *K*.

Finally, we run CLUMMP[221] with all the STRUCTURE runs corresponding to best estimate of *K* to obtain a consensus file from the results of each of the 20 runs.

We run STRUCTURE twice, before and after detection and removal of recombinant genes in the population.

To analyze the geographical structure of the *N. gonorrhoeae* samples we performed an analysis of the molecular variance (AMOVA) on the non-recombinant fraction of the genome. To do that, we stratified the

populations in two hierarchical levels, area —Mediterranean or Central areas— and region —CV, Catalonia, and Madrid—, to calculate the variance within and between groups. We used the poppr[222] package in R.

6. Recombination analysis

6.1. Recombination detection in *Neisseria gonorrhoeae* (Chapter 1)

To detect recombination events in the 342 gonococcal isolates, first we needed to reduce the dataset to eliminate clonal isolates. For this, we run CD-HIT-EST[223], which clusters isolates into groups by a similarity threshold that was set to 98%.

Once the dataset was reduced to representative isolates, we extracted the coding sequences (CDS) from the alignment file using the coordinates of the strain FA1090 in its GFF file and the *extractalign* tool in EMBOSS[224] suite. CDS in the complementary strain were reversed using the *revseq* tool in EMBOSS.

We performed a likelihood mapping[225] (LM) test, as implemented in IQ-TREE, to evaluate the phylogenetic signal in each gene. This test determines the likelihood of the three possible topologies of each quartet of sequences drawn at random from the whole dataset. The likelihoods are plotted in a triangle whose vortexes correspond with the 3 topologies. So, the likelihoods plotted in the areas near the vortexes correspond to quartets with a completely resolved topology, likelihoods plotted in areas between two vortexes correspond to quartets with topologies partially solved, and likelihoods plotted in the central area of the triangle correspond to quartets with unresolved topologies. We set the number of quartets to 10,000 random draws for each gene. The proportion of quartets in the three vortex areas indicates if there is phylogenetic signal[226] for the gene. We selected genes with a strong signal, i.e. with ≥70% of quartets completely solved, to the next stage of the analysis.

Next, genes that met the above requirement were tested for topology congruence. To do this, a ML phylogenetic tree was constructed for each gene using the same parameters as explained in section 4.1. Then, the topologies of the tree derived from each gene and that from the complete genome were compared using the Shimodaira-Hasegawa (SH)[227] and Expected-Likelihood Weights (ELW)[228] tests implemented in IQ-TREE. Those genes for which both tests rejected the congruence between the topologies of the two trees, were selected as recombinants. The p-values of the SH test were corrected using the false discovery rate (FDR) method implemented in the *p.adjust* function in R.

The recombinant gene trees were visually compared with the complete genome tree using Phylo.io[229] to analyze the movement of genes between the STRUCTURE groups. Genome positions of the recombinant genes were masked for phylogenetic and population structure analyses.

6.2. Recombination detection in *Lactococcus garvieae* (Chapter 4)

Essentially the same methodology as in gonococci was applied to *L. garvieae* with some differences:

- In this case, we started from an assembled genome and not from a genome resulting from mapping to a reference genome.
- The objective was to detect recombinant genes between this isolate and other isolates of the same species, and then to extend this level of analysis to higher taxonomic levels —genus and class—.

So, the first step was to download the genomes for the analyses. For the intra-species level, we downloaded 23 genomes of *L. garvieae*. For the genus level, we downloaded 6 genomes of *Lactococcus*, and for the class level, we downloaded 19 genomes of the class *Bacilli*. The average nucleotide identity (ANI) between pairs of isolates was calculated using pyani[230] to have a preliminary evaluation of the variation among the genomes. The orthologous genes (OGs) between the genomes in each taxon were detected using the BLAST-based program Proteinortho[231]. We used the default parameters

except for the minimum similarity, which was adjusted to 80% for the species level, and 70% for genus and class levels. The core genome was determined as strict core —common orthologues for all the isolates of the corresponding level— and as relaxed core —orthologues shared by ≥80% isolates—.

OGs included in the strict and relaxed cores were aligned using MACSE[232], and concatenated using a custom python script (https://drive.google.com/file/d/1acMucj7byGDVLlxE0j3gznVAEttRbymy/view?usp=sharing). The resulting multiple alignments were used for phylogenetic reconstruction by ML using IQ-TREE as described previously. For the three taxonomic levels both the strict and relaxed core phylogenetic trees were topologically identical.

The methodology for the detection of recombination was identical as that described for gonococci, with the only difference that, in this case, we included an additional filter in the genes passing the LM test. The proportion of pairwise-informative SNPs of each alignment was also computed. So, only the OGs with ≥70% of quartets completely resolved in the LM test and with a proportion of informative sites of at least 10% from the total length of the alignment were eligible to topology congruence testing.

7. Analysis of the accessory, unmapped genome

The gonococcal isolates in Chapter 1 were assembled as described. We separated suspected plasmidic contigs —i.e., contigs with a coverage value much higher than the average— and looked for plasmids using BLAST against the complete online nucleotide database of NCBI. We retrieved only results matching *Neisseria gonorrhoeae*. For conjugative plasmid typing, we checked the percentage identity to sequences with accession numbers L12241 and L12242 for American and Dutch plasmids, respectively. For β-lactamase plasmids, it was not possible to type them because of the level of fragmentation in these contigs and the similarity of the sequences as they derive from the Asian type; however, we were able to detect the presence of these plasmids. Additionally, we run a local version of BLAST to find the GGI,

which is not present in the reference genome FA1090. We used the GGI from strain MS11 (NCBI accession AY803022.1) as reference.

Unmapped reads of gonococcal isolates in Chapter 2 were assembled and suspected plasmid contigs were separated and identified using BLAST as described above. The remaining contigs were annotated using Prokka. The annotated CDS from the different isolates in this dataset were compared to each other using Proteinortho.

Unmapped reads of *Serratia marcescens* isolates in Chapter 3 were assembled as described before. BLAST was used with default settings to compare the resulting contigs among the strains of each outbreak. Additionally, the presence of plasmids was determined using BLAST with default settings and a custom database containing all the *S. marcescens* plasmids published in NCBI database as of December, 2019.

8. Analysis of antimicrobial resistance determinants in gonococci

For the analysis of mutations conferring resistance to antibiotic in gonococci (see Chapter 1) we analyzed the assembled contigs using several approaches:

- AMR determinants included in the NG-STAR typing scheme —*penA, mtrR, porB, ponA, gyrA, parC*, and 23S rDNA— were detected using BLAST+ against the NG-STAR database, allowing only results with 100% identity and coverage. As each allele has associated a metadata informing about the known mutations which confer resistance to the corresponding antimicrobial drug, or if the allele corresponds with a mosaic form of the gene, we retrieved such information. For the *penA* gene we further inspected the alignment using MEGA X[233], which allows translation of the nucleotide alignment to amino acids and checks for all the known variants, as the metadata from NG-STAR do not contain all the described mutations.
- For the AMR determinants not included in the NG-STAR scheme —*pilQ, parE, rpsJ,* and *rpsE*—, we extracted the translated sequences from the reference genome FA1090. Then, we ran tBLASTn to compare

the translated genes and the contigs, and we looked for the presence of mutations in the regions of interest. In the case of the 16S rRNA and the promoters of *macAB* and *norM,* it was not necessary to extract the translated sequences, only the nucleotide sequences.

- A database of mosaic forms of *mtrD* containing 20 alleles was found in NCBI (Bioproject accession PRJEB36607). We used BLAST+ with all the contigs to find mosaic alleles of this gene.
- The presence of the plasmidic determinants —bla_{TEM} and *tetM*— was determined by the presence of a β-lactamase plasmid, or a Dutch or American conjugative plasmid, detected as described before.

Additionally, we used ARIBA[234] with the cleaned reads to detect the main AMR determinants, and to check if there existed differences between this tool and our methodology. ARIBA makes clusters of the sequences in its databases by similarity using CD-HIT. The reads mapped in a cluster were assembled using FERMI-lite[235]. The assembled sequence is aligned to the sequences of the cluster in the database using NUCMER —a script from MUMMER— to identify the variants, and re-confirmed these variants by mapping the reads to the assembled sequence and identifying the variants.

ARIBA was run using a pre-computed database for gonococci available at https://github.com/martinghunt/ariba-publication/. We compared the results of our method to those of ARIBA in the genes and mutations shared by the two analyses.

Genotypic and phenotypic data were compared, in the cases where phenotype was available to check the presence of incongruences.

9. Dating analysis

To estimate the most recent common ancestor (MRCA) of gonococcal populations in Chapter 1, we first used TempEst[236], a tool for evaluating the temporal signal in asynchronous sequences —sampled at different times— along a phylogeny. To do that, the program calculates, through a regression analysis, if the sampling dates of the isolates are relevant for the phylogeny. When there is a significant association between the divergences and the sampling dates, this indicates that there is a temporal signal, and the sampling dates will be useful to estimate the time to the MRCA (tMRCA). To evaluate the temporal signal, we used the coefficient of determination (R^2); the higher value the value of R^2, the better temporal signal we will have.

Next, we proceeded to estimate the tMRCA using LSD2[237] which calculates the dates of all the ancestral nodes and the substitution rate in a phylogenetic tree based on a Gaussian model and exploiting the tree structure, such as branch lengths. When an unrooted tree is used as input, as in this case, LSD2 infers the best position of the root. We run LSD2 with the 30,042 non-recombinant SNPs alignment and the corresponding phylogenetic tree, and a lognormal relaxed clock with mean 1 and standard deviation 0.2.

CHAPTER 1

GENOMIC EPIDEMIOLOGY AND ANTIMICROBIAL RESISTANCE SURVEILLANCE OF *Neisseria gonorrhoeae* IN SPAIN

1. Results

1.1. Patient demographics

The dataset consists of 342 *Neisseria gonorrhoeae* isolates from 3 Spanish regions —150 isolates from Comunidad Valenciana (CV), 93 from Catalonia, and 98 from Madrid—. In geographical terms, CV and Catalonia represent the Mediterranean region of the country, while Madrid constitutes the central region. According to 2020 data from the Spanish National Statistics Institute (https://www.ine.es), these 3 regions are the most populated in the country. Of the 47,329,981 inhabitants in Spain, the CV concentrates 5,028,650 (10.6%), Catalonia concentrates 7,652,069 (16.2%), and Madrid concentrates 6,747,425 (14.3%), so that the sum of the 3 regions represents 41.1% of the total population in Spain. However, the distribution of the samples is different depending on the region. The samples from Catalonia come from the city of Barcelona, those from Madrid region come from the city of Madrid; however, those from CV come from 12 hospitals that cover the entire Valencian region (Figure 15).

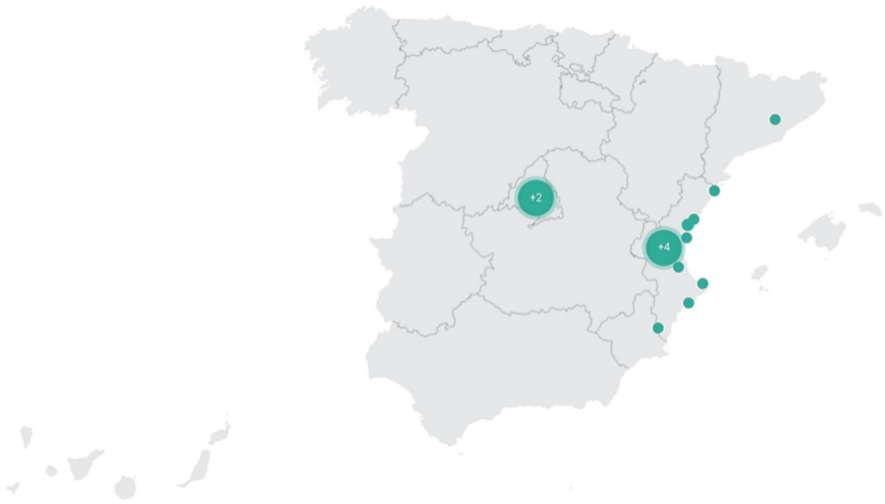


Figure 15 | Areas of study. Points mark the collaborating hospitals.

Regarding patients, 92.1% (n=315) were males, 4.4% (n=15) were females, and 3.5% (n=12) did not report their gender. The age of the patients ranged from 16 to 67 years (quartiles Q1=25 years, Q2=31 years, and Q3=35.5 years), with peaks at 24-25 and 31-32 years. The age of 119 patients was not reported (Figure 16).

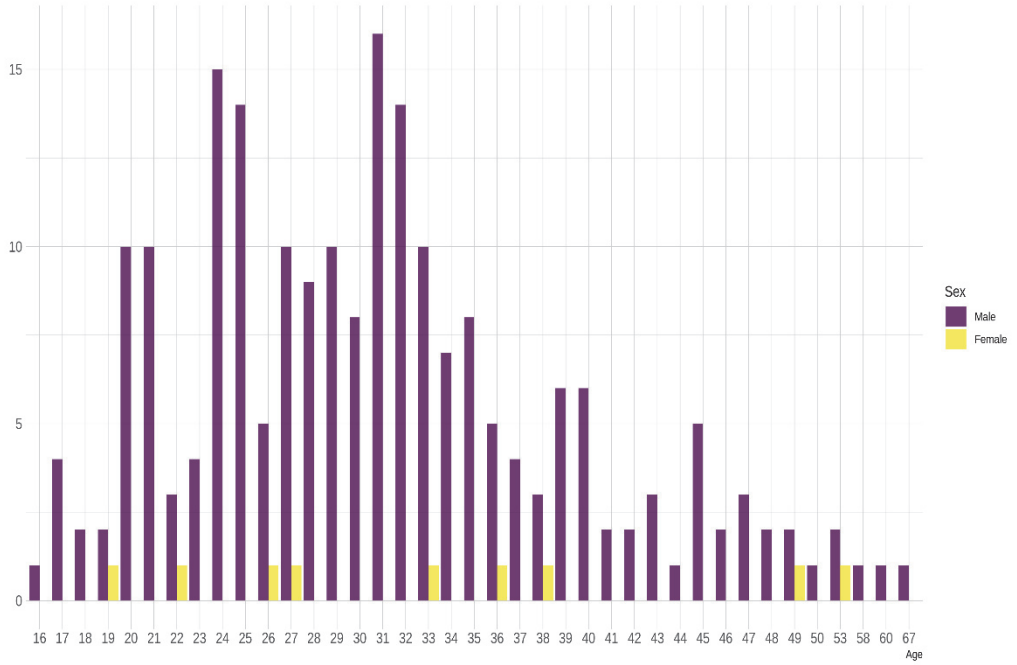


Figure 16 | Patients age distribution by sex. 106 males and 6 females of unknown age, as well as 12 individuals without reported sex, were excluded.

Specimens were mainly urethral (76.9%, n=263), followed by rectal (13.2%, n=45), vaginal (2.9%, n=10), endocervical (0.9%, n=3), oropharyngeal (0.6%, n=2), balanopreputial (0.3%, n=1), and a joint fluid puncture (0.3%, n=1). Specimens for 17 patients were not reported (Figure 17).

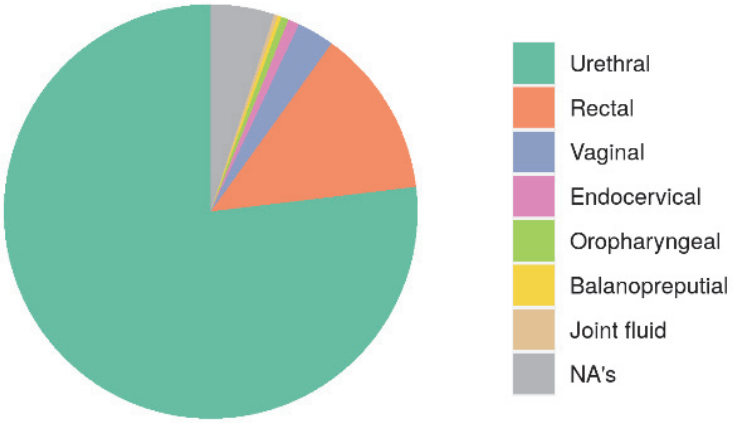


Figure 17 | Distribution of specimens by their type.

The sampling period spanned 5 years, from November 2012 to November 2017 (Supplementary Table 1).

1.2. Sequencing and mapping quality

The 342 gonococcal isolates were sequenced using Illumina NextSeq platform in 3 different sequencing runs, generating an average of 2,015,226 paired-end reads of 150 bp (ranging 249,622 to 34,304,514). After applying the quality filters (Supplementary Figure 1), the average number of reads was 1,782,226 (ranging 230,224 to 30,339,826). Because the average number of reads differs for the different sequencing runs, the data per run is shown in Table 1. The wide range in number of reads may be caused by technical difficulties when pooling, as the samples in each run were multiplexed together with other bacterial species[238].

Table 1 | Average number of reads by sequencing run.

Run	N. of isolates	Raw reads	Cleaned reads
1	220	1,299,592 (249,622-3,145,614)	1,170,270 (230,224-2,833,148)
2	49	6,698,479 (319,386-34,304,514)	5,995,826 (287,402-30,339,826)
3	73	1,028,377 (489,004-2,079,080)	798,171 (362,502-1,671,232)

The mapping step yielded an average depth coverage of 80.1X and an average breadth coverage of 93.4% (range 78–96.2%) (Supplementary Table 2).

1.3. Alignment and phylogenetics

The resulting alignment of mapped positions spanned 2,153,922 bp and contained 34,907 variant positions (SNPs) after removing repetitive regions, phages and high-density SNP regions, and cleaning the poorly aligned regions. Recombinant genes were also removed (see next section), leaving a 30,042 SNPs alignment.

This 30,042 bp alignment was used to reconstruct a phylogenetic tree by maximum-likelihood (ML). The nucleotide substitution model was inferred by ModelFinder, which selects the best-fit model according to Bayesian Information Criterion (BIC). This model was K3Pu+F+ASC+G4, that means the following:

- K3Pu: three substitution types model and unequal base frequencies[239].
- F: empirical base frequencies.
- ASC: ascertainment bias correction[240]. This should be applied if the alignment does not contain constant sites (e.g. SNP alignment.
- G4: discrete Gamma model with four rate categories[241].

The ML phylogenetic tree was reconstructed using 10,000 ultrafast bootstrap replicates.

1.4. Recombination detection

Before proceeding to the detection of recombination, a ML phylogenetic tree was reconstructed using the 34,907 SNPs alignment. This alignment was used to determine genetic clusters using STRUCTURE. The number of hypothetical ancestral populations (*K*) was estimated in 6 by Δ*K* method[220] (Supplementary Figure 2). However, we found a high level of genetic admixture in all 6 populations except in population III (Figure 18), so we proceeded to detect recombinant regions.

To avoid redundant sequences, we subset the dataset in 105 isolates using CD-HIT-EST. Then, we extract the coding regions using the coordinates in the reference genome FA1090 strain, and subjected them to likelihood mapping (LM) analysis. 77 genes passed the LM test and were eligible for congruence testing against the topology of the whole-genome SNPs alignment. A total of 51 genes were statistically significant in rejecting the whole-genome SNPs tree topology for both SH and ELW tests, after applying the FDR correction, being detected as recombinant genes. There were 5 recombination events spanning 2 genes, and one event spanning 3 genes. The remaining 38 recombination events encompassed only one gene.

Most recombinant genes encode membrane or cytoplasm proteins, with catalytic and binding activities, mainly. Proteins encoded by recombinant genes are involved in a wide variety of biological processes, the most common are biosynthetic and oxidation-reduction processes, translation, and transport. There are also some proteins involved in the type IV pilus assembly, and others involved in pathogenesis (Supplementary Table 3).

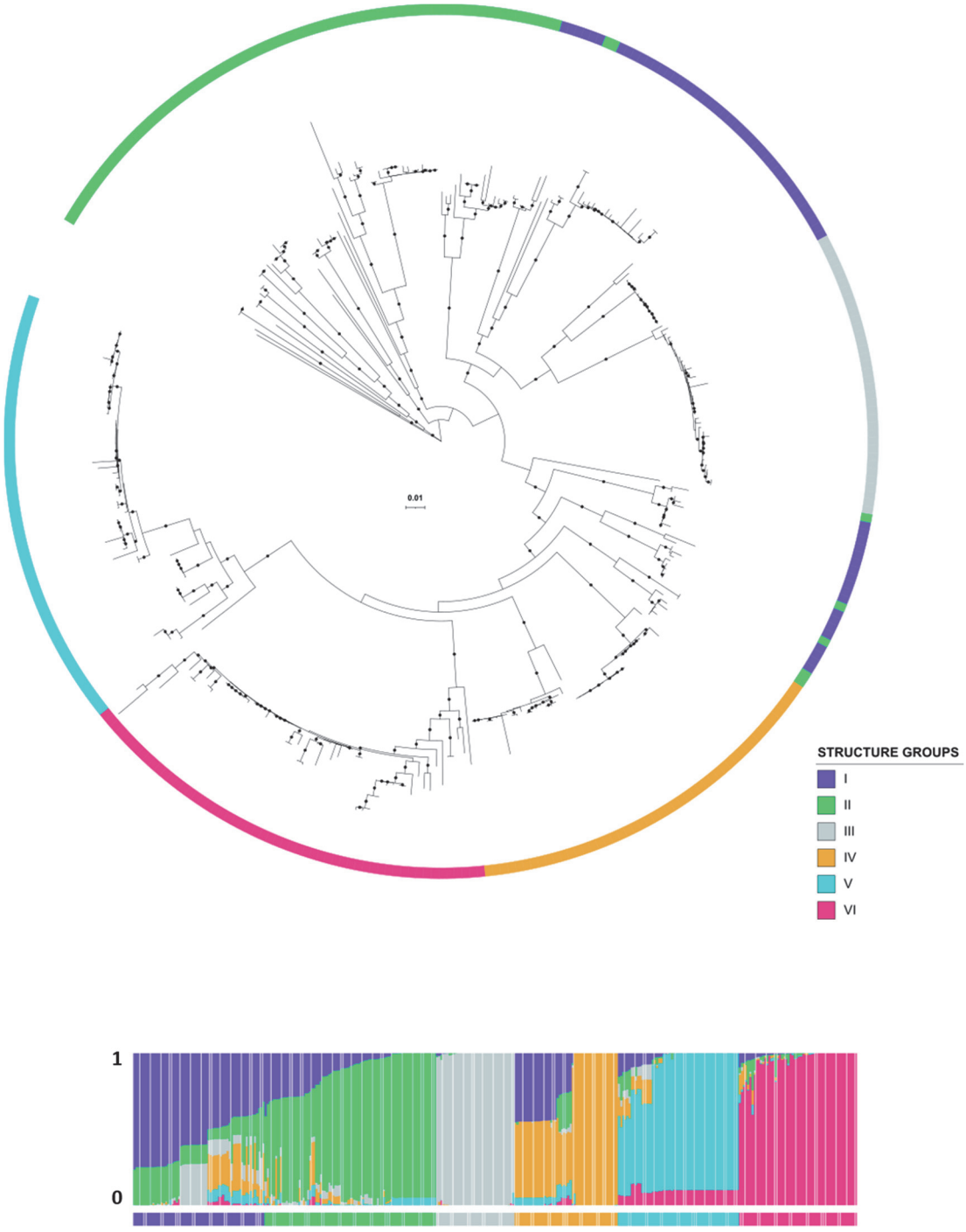


Figure 18 | Maximum-likelihood tree prior recombinant genes removal. The 34,907 SNPs alignment was used in the phylogenetic reconstruction. Black dots in branches represent bootstrap support values above 90%. The six STRUCTURE groups are highlighted. The plot below shows the genetic admixture of these six groups.

Based on the membership of the isolates to the six STRUCTURE groups, gene transfer between these groups was detected, as represented in the Figure 19. Group III was the least active in gene transfer, which is consistent with its low level of genetic admixture. The main donors were groups I, II and V, but they received approximately the same proportion of genes from all the other groups.

Figure 19 | Genetic transfer between the six STRUCTURE groups.

After recombinant genes were removed from the alignment, we obtained a 30,042 SNPs alignment, that we used to reconstruct the phylogenetic tree as explained in the previous section. Both trees were compared (Figure 20), and

the topological congruence tests were statistically significant in rejecting the reciprocal tree topologies (Table 2).

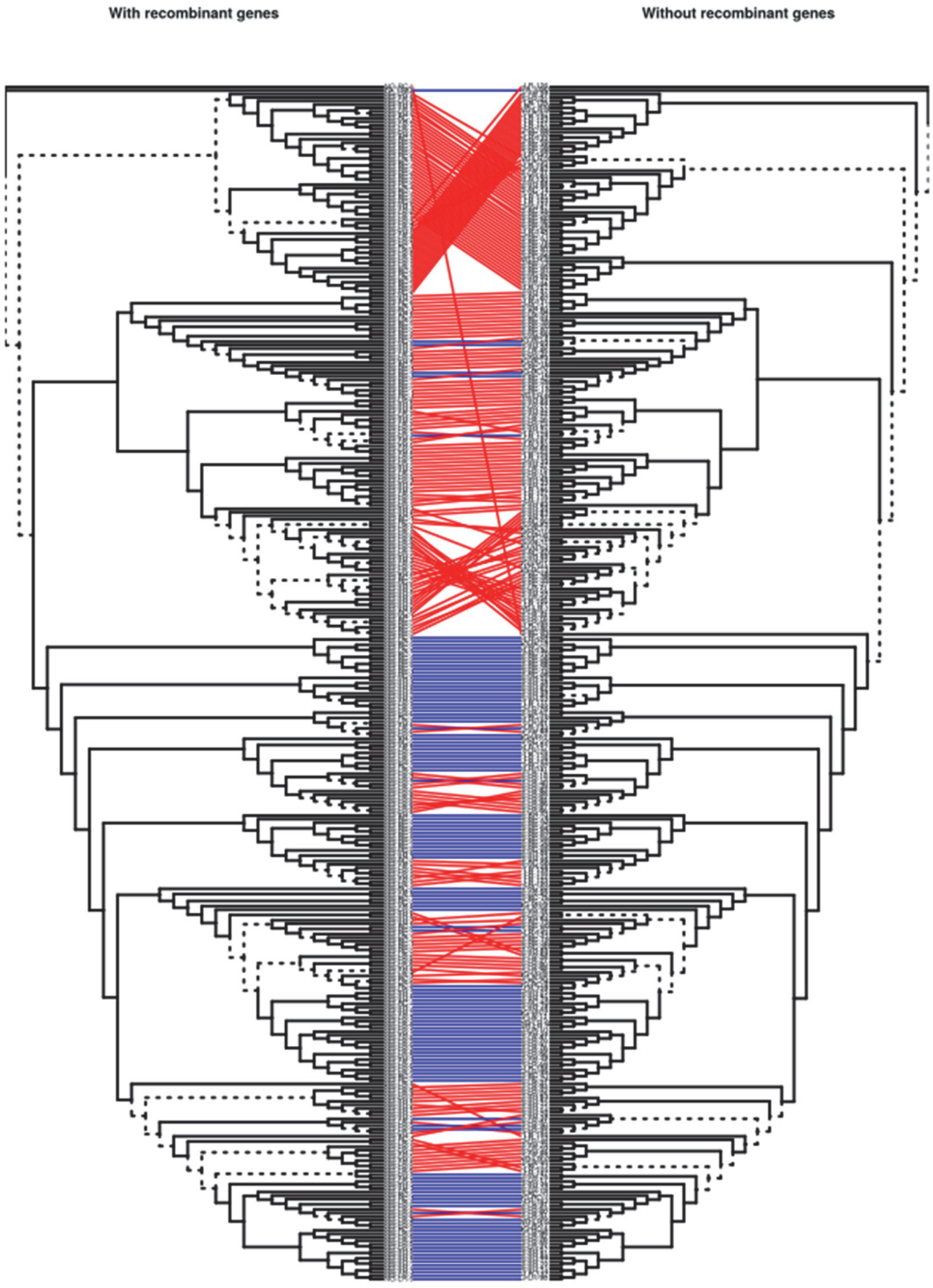


Figure 20 | Tanglegram comparing both phylogenetic trees. On the left, the tree with recombinant genes (34,907 SNPs alignment); on the right, the tree without them (30,042 SNPs alignment). Red links and dashed branches highlight the incongruences between both trees.

Table 2 | P-values for SH test and posterior weights for ELW test when comparing the topologies of tree with and without recombinant genes.

	With recombinant genes		Without recombinant genes	
	SH	ELW	SH	ELW
With recombinant genes	1.0000	1.0000	0.0000	0.0000
Without recombinant genes	0.0000	0.0000	1.0000	1.0000

1.5. Population structure

The 30,042 SNPs alignment was used to determine genetic clusters using STRUCTURE. The number of hypothetical ancestral populations (*K*) was estimated in 8 by Δ*K* method[220] (Supplementary Figure 3). Again, high levels of genetic admixture were found between the estimated groups, although this time there was a higher proportion of isolates without admixture than before removing the recombinant genes (Figure 21). The fact that there are still significant levels of genetic admixture —especially in groups I, IV and VI, where no "pure" isolates were found— makes us suspect that not all the recombination events actually existing in the gonococcal population could be detected.

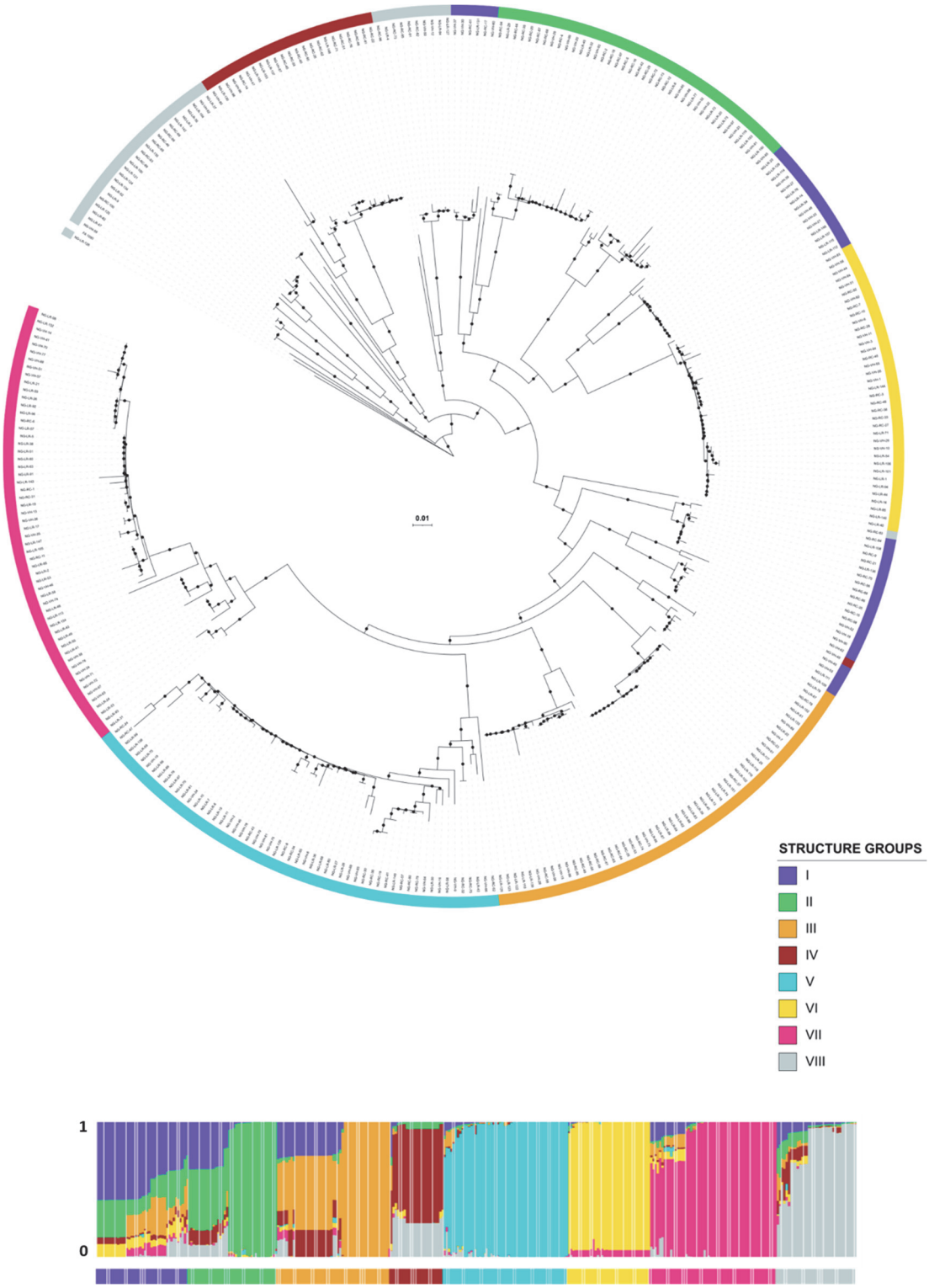


Figure 21 | Structure populations of gonococcal isolates. Black dots in branches highlight bootstrap support values above 90%. The plot depicts the genetic admixture in the eight groups.

To study the population structure from the geographical origin of the isolates, we performed an AMOVA test with two geographical hierarchies: area —Mediterranean (encompassing the CV and Catalonia regions' isolates) and Central (which includes the isolates from Madrid)— and region. Results revealed that, at level of country, there was no significant differentiation between Mediterranean and Central areas (Table 3).

Table 3 | AMOVA test results.

				Components of variance		
	Df	Sum Sq	Mean Sq	Sigma	%	p-value
Between areas	1	4704.172	4704.172	-67.97807	-2.354924	0.6723
Between regions within area	1	14203.523	14203.523	100.86795	3.832587	0.0009
Within regions	339	879013.554	2592.960	2592.96034	98.522337	0.0009
Total	341	897921.249	2633.200	26.85022	100.0000	

1.6. Typing

There are currently three typing systems for gonococcus. Using the MLST system, 47 different sequence types (STs) were obtained. For the complete dataset, the main STs were 7363 (16.4%, n=56), 1901 (15.8%, n=54), 9363 (9.9%, n=34), 8143 (6.1%, n=21), 8156 (4.3%, n=15), 1588 (4.1%, n=14), 1599 (4.1%, n=14), 7822 (3.2%, n=11), 7827 (3.2%, n=11), and 1579 (2.9%, n=10). There were 21 isolates whose ST could not be identified either because it was a new ST or because some of its alleles could not be identified.

Inspecting the distribution of the main STs by year, an evolution is observed in which some of the most abundant STs at the beginning of the period of study are being replaced by other STs, some of which were not even at the beginning. The ST 1901 was the most abundant in 2013 —28% of all STs in that year—, but its presence was decreasing until it almost disappeared (1.4%) in 2017. In contrast, STs that were not represented during the first year of the period studied ended up becoming the main STs in 2017. This was the case of STs 8143, 8156, 1599 —all appeared in 2014, the last of which was established as the most abundant in 2017—, and 7827 —this one appeared in 2015—. However, some main STs in 2013 remained present as one of the main STs throughout the period of study, as was the case of ST 7363 (Figure 22).

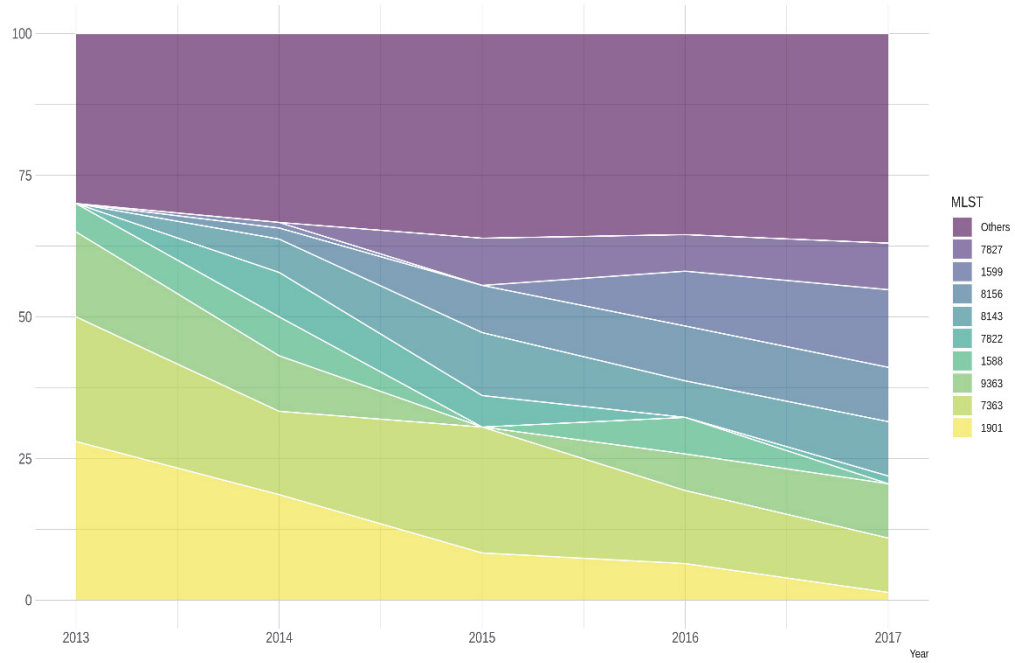


Figure 22 | Proportion of main STs (MLST scheme) by year. STs represented as main are those that had at least 5 individuals in any of the years included in the period of study. The *"Others"* category includes the remaining 38 STs. Because there were only 3 isolates from 2012, they were included in 2013.

Using the NG-MAST system, 111 different STs were obtained, encompassed in 75 different genogroups (G). For the complete dataset, the main STs were 2992 (6.7%, n=23), 2400 (4.7%, n=16), 5624 (4.7%, n=16), 5441 (4.1%, n=14), 3378 (3.5%, n=12), and 13288 (3.5%, n=12). However, the main genogroups were G3378 (9.9%, n=34), G2400 (8.5%, n=29), G2992 (8.2%, n=28), G5441 (4.7%, n=16), G5624 (4.7%, n=16), G21 (4.1%, n=14), G13288 (3.5%, n=12), G11461 (3.2%, n=11), and G437 (2.9%, n=10) (Supplementary Table 4). There were 36 isolates whose ST could not be identified.

As in the MLST scheme, when inspecting the distribution of the main genogroups by year, some of the most abundant genogroups at the beginning of the period of study are being replaced by others. G3378 was the most abundant in 2013 —18% of all genogroups in that year—, but its presence was decreasing until it almost disappeared (1.4%) in 2017. This was the same for G2992, which went from 15% in 2013 to 1.4% in 2017. G13288 was the third most abundant in 2013 (12%), but in the following years it disappeared. G2400 was one of the main genogroups in both 2013 (7%) and 2017 (9.6%). On the other hand, genogroups that were not represented during the first year of the period studied became the main genogroups in 2017. This was the case of G5441 —appeared in 2014, and present in 9.6% of the isolates in 2017—, G10386 —appeared in 2015, present in 6.8% of isolates—, or G11461 and G4186 —both appeared in 2016, and present in 12.3% and 6.8% of isolates, respectively— (Figure 23).

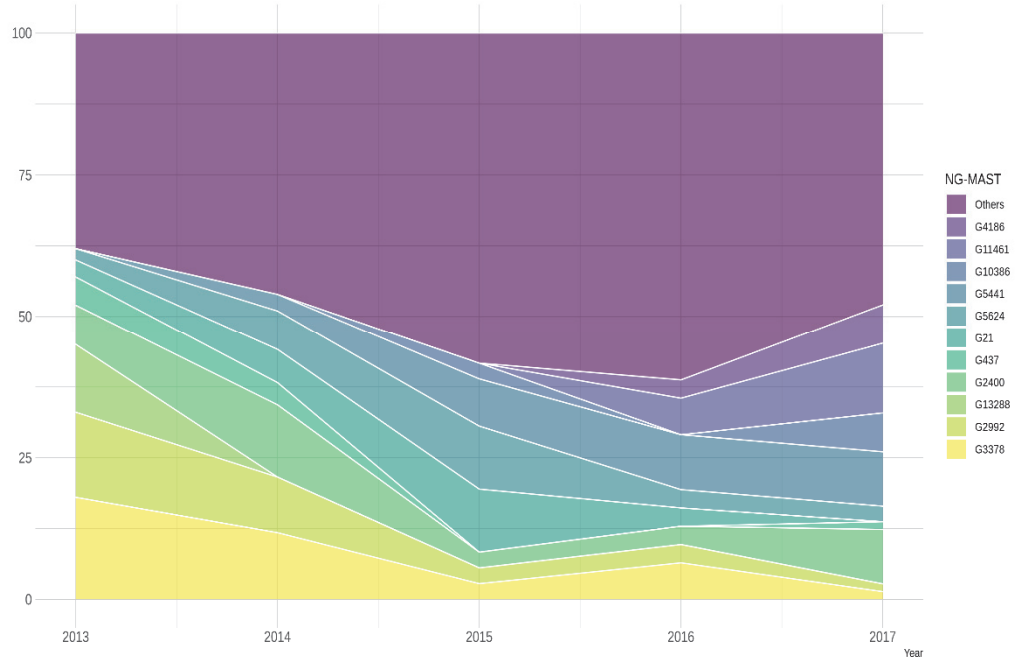


Figure 23 | Proportion of main genogroups (NG-MAST scheme) by year. Genogroups represented as main are those that had at least 5 individuals in any of the years included in the period of study. The *"Others"* category includes the remaining 64 genogroups. Because there were only 3 isolates from 2012, they were included in 2013.

Using the NG-STAR system, 81 different STs were obtained. For the complete dataset, the main STs were 90 (8.5%, n=29), 158 (7.9%, n=27), 63 (7%, n=24), 442 (4.7%, n=16), 426 (4.4%, n=15), 520 (4.4%, n=15), 348 (3.5%, n=12), 139 (3.2%, n=11), and 38 (2.9%, n=10). There were 38 isolates whose ST could not be identified.

Similar to what happened in the MLST and NG-MAST schemes, inspecting the distribution of the main STs by year, the most abundant STs at the beginning of the period of study are being replaced by others. STs 90, 63 and 348 were the most abundant in 2013 —17%, 13% and 12% of all STs in that year, respectively—, but they decreased until they disappeared in 2017 —ST 63 not disappeared, but it was reduced to 2.7% in 2017—. ST 38 appeared in the dataset in 2015, representing 8.3% of the isolates in that year, and it was the fourth most abundant in 2017 (6.8%). STs 442 and 520 appeared in

the dataset in 2014 (2% and 1%, respectively), and they became the second and first most abundant STs in 2017 (12% the ST 442, and 15.1% the ST 520). So we can talk about a ST replacement during the period of study (Figure 24).

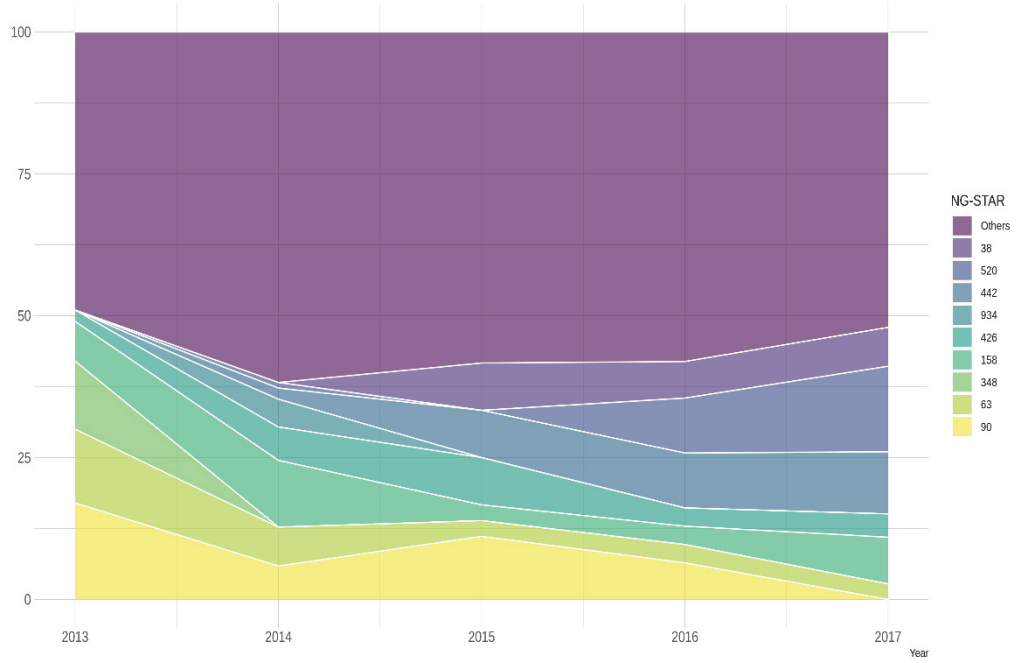


Figure 24 | Proportion of main STs (NG-STAR scheme) by year. STs represented as main are those that had at least 5 individuals in any of the years included in the period of study. The *'Others'* category includes the remaining 72 STs. Because there were only 3 isolates from 2012, they were included in 2013.

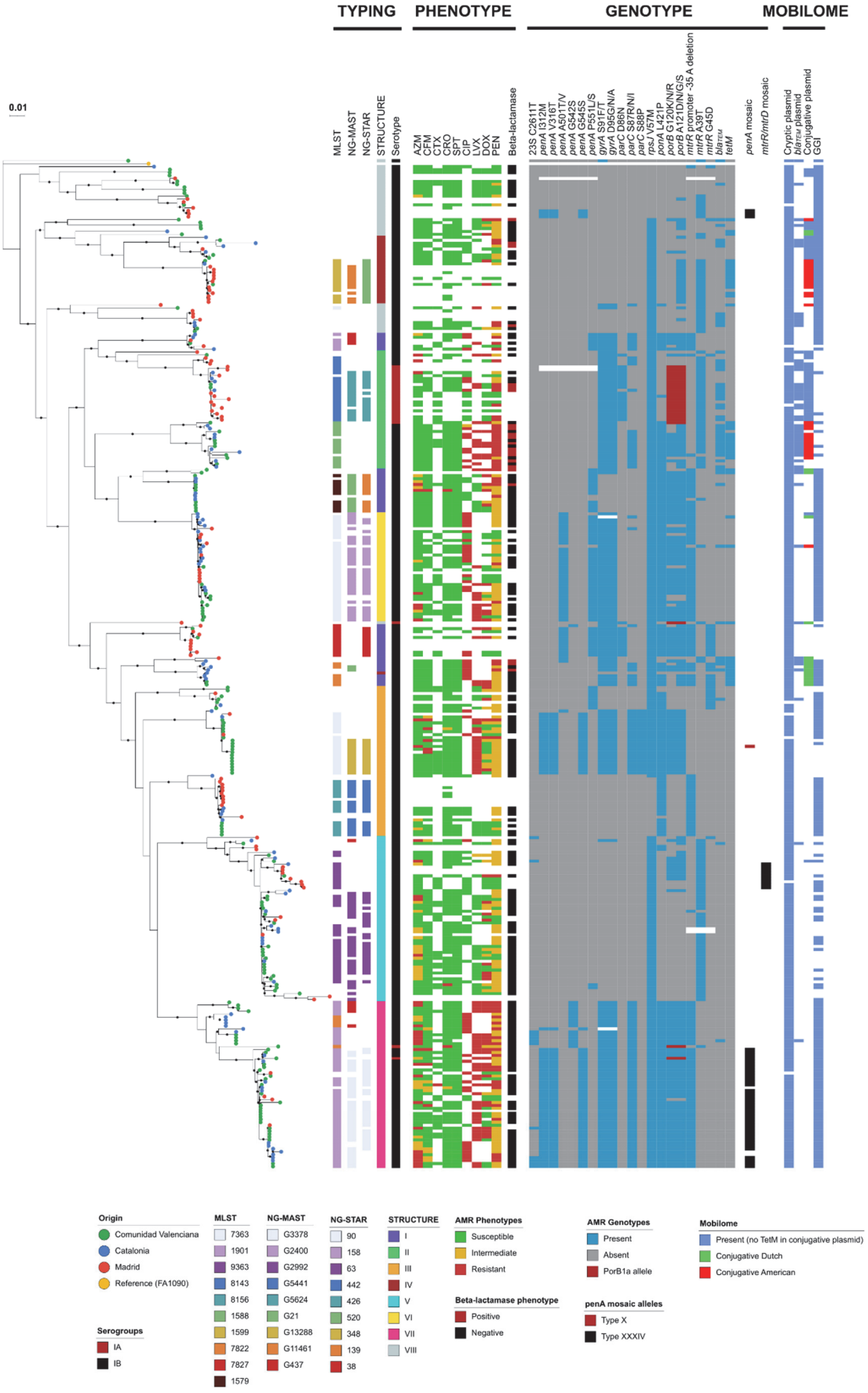

TYPING
PHENOTYPE
GENOTYPE
MOBILOME
0.01
MLST
NG-MAST
NG-STAR
STRUCTURE
Serotype
AZM
CFM
CTX
CRO
SPT
CIP
LVX
DOX
PEN
Beta-lactamase
23S C2611T
penA I312M
penA V316T
penA A501T/V
penA G542S
penA G545S
penA P551L/S
gyrA S91F/Y
gyrA D95G/N/A
parC D86N
parC S87R/N/I
parC S88P
rpsJ V57M
porB L421P
porB G120K/N/R
porB A121D/N/G/S
mtrR promoter -35 A deletion
mtrR A39T
mtrR G45D
blaTEM
tetM
penA mosaic
mtrR/mtrD mosaic
Cryptic plasmid
blaTEM plasmid
Conjugative plasmid
GGI
Origin
Comunidad Valenciana
Catalonia
Madrid
Reference (FA1090)
Serogroups
IA
IB
MLST
7363
1901
9363
8143
8156
1588
1599
7822
7827
1579
NG-MAST
G3378
G2400
G2992
G5441
G5624
G21
G13288
G11461
G437
NG-STAR
90
158
63
442
426
520
348
139
38
STRUCTURE
I
II
III
IV
V
VI
VII
VIII
AMR Phenotypes
Susceptible
Intermediate
Resistant
Beta-lactamase phenotype
Positive
Negative
AMR Genotypes
Present
Absent
PorB1a allele
penA mosaic alleles
Type X
Type XXXIV
Mobilome
Present (no TetM in conjugative plasmid)
Conjugative Dutch
Conjugative American

Figure 25 | Maximum likelihood phylogenetic tree of the Spanish gonococcal isolates including the metadata about typing schemes, AMR phenotype and genotype, and mobilome. For more resolution visit: https://drive.google.com/file/d/10xZ29HOfEAr1XLmRswbZWYixRa80U3tm/view?usp=sharing.

1.6. Antimicrobial resistance

The collaborating hospitals provided phenotypic information on the antibiotic resistance profile that they had available. The information gaps observed were due either to the fact that not all hospitals test the same antibiotics, or that some strains were in storage and could not be cultured again —despite the latter, we were able to sequence them—. Table 4 summarizes the phenotypic results of the isolates in our dataset (see Supplementary Table 1).

Table 4 | Number of isolates for each antimicrobial susceptibility phenotype.

Antibiotic	Susceptible	Intermediate	Resistant	No data
Azithromycin	197	34	35	76
Cefixime	212	10	9	111
Ceftriaxone	242	2	3	95
Cefotaxime	96	0	1	245
Spectinomycin	231	0	0	111
Ciprofloxacin	63	1	63	215
Levofloxacin	64	1	79	198
Doxycycline	86	25	60	171
Penicillin	56	153	63	70

Given the tested isolates for each antibiotic —excluding those for which we do not have phenotypic data—, the results presented in the table show that

79.4% of the gonococci studied present resistance or reduced susceptibility to penicillin, 50.4% to ciprofloxacin, 55.6% to levofloxacin, 49.7% to doxycycline, and 25.9% to azithromycin. On the other hand, they present greater susceptibility to extended spectrum cephalosporins (ESCs) —8.2% of the isolates presented resistance or reduced susceptibility to cefixime, 2% to ceftriaxone, and 1% to cefotaxime—, and to spectinomycin, to which all isolates tested to it were susceptible.

Table 5 | Genetic determinants of AMR present in the isolates.

Antibiotic[a]	**Determinant**	**Mutations**	**N. isolates**
AZM	23S rDNA	A2059G	0
		C2611T	12
		C2611G	0
	mtrR promoter	-35A deletion	155
	mtrR	A39T	132
		G45D	30
	Both mosaic *mtrR* and *mtrD*		9
	macAB promoter	-10 hexamer TAGAAT→TATAAT	0
ESCs	*penA*	A311V	0
		I312M	66
		V316P	0
		V316T	66
		T483S	0
		A501P	0
		A501T	36
		A501V	13
		G542S	16

Continued on next page.

Antibiotic[a]	Determinant	Mutations	N. isolates
ESCs	*penA*	G545S	66
		P551S	18
PEN	*ponA*	L421P	201
	porB	G120D	0
		G120K	145
		G120R	1
		G120N	15
		G120Q	0
		G120E	0
		G120T	0
		A121D	87
		A121S	39
		A121N	50
		A121G	24
		A121V	0
	pilQ	E666K	0
	mtrR promoter		
	mtrR		
	bla$_{TEM}$		52
FQs	*gyrA*	S91F	186
		S91Y	0
		S91I	0
		S91T	1
		D95N	21
		D95G	113

Continued on next page.

Antibiotic[a]	Determinant	Mutations	N. isolates
FQs	*gyrA*	D95A	52
		D95Y	0
	parC	D86N	30
		S87I	2
		S87N	14
		S87R	97
		S88P	22
		E91K/G/Q/A	0
	parE	G410V	0
	norM promoter	-35 hexamer CTGACG→TTGACG	0
DOX	*rpsJ*	V57M	303
	porB		
	pilQ		
	tetM		43
SPT	16S rRNA	C1192U	0
	rpsE	T24P	0
		V27 deletion	0
		K28E	0

a. AZM: azithromycin; ESCs: extended spectrum cephalosporins (CFM: cefixime; CTX: cefotaxime; CRO: ceftriaxone); FQs: fluoroquinolones (CIP: ciprofloxacin; LEV: levofloxacin); DOX: doxycycline; SPT: spectinomycin.

An exhaustive search of known mutations that confer resistance or decreased susceptibility to the antimicrobial agents tested was done (Table 5; Supplementary Table 1).

1.6.1. Resistance to macrolides

For macrolides, such as azithromycin, only mutation C2611T in 23S rDNA was found in 12 isolates. This mutation is associated to low-level resistance to azithromycin[242]. The remaining resistant isolates to azithromycin could be due to the presence of mutations in the *mtrR* promoter and/or in *mtrR* gene, which also causes resistance to penicillin. Of all isolates in Table 5 with mutations in *mtrR* and *mtrR* promoter, 20 isolates had both the deletion in the promoter and the G45D mutation in the gene. Is also remarkable the presence of a cluster of 9 isolates with mosaic forms both *mtrR* and *mtrD* genes, which is also associated with resistance to azithromycin. There are 125 isolates with mosaic *mtrD* but without the presence of the mosaic *mtrR* it cannot causes resistance to this drug[109].

Of the 35 isolates with phenotype of resistance to azithromycin, 31.4% (n=11) had the C2611T mutation in the 23S rDNA, 74.3% (n=26) had the -35 A deletion in the *mtrR* promoter, and 22.9% (n=8) had A39T or G45D mutations in the *mtrR* gene. While mutations in 23S were found only in resistant isolates, we found variable percentages of the others in both intermediate resistant and susceptible isolates. Of the 34 isolates with decreased susceptibility to azithromycin —or intermediate resistant isolates— phenotype, 55.9% (n=19) had the deletion in the *mtrR* promoter, 38.2% (n=13) had the A39T mutation in the *mtrR* gene, and 2.9% (n=1) had both *mtrR* and *mtrD* mosaic alleles. Of the 197 isolates with phenotypic susceptibility to azithromycin, 43.1% (n=85) had the deletion in the *mtrR* promoter, 45.2% (n=89) had the A39T and/or G45D mutations in the *mtrR* gene, and 2.5% (n=5) had both *mtrR* and *mtrD* mosaic genes.

1.6.2. Resistance to extended spectrum cephalosporins

For ESCs, such as cefixime, cefotaxime, and ceftriaxone, a few mutations in the *penA* gene were found. Some mutations appear together, but not always, except for I312M, V316T and G545S, which appeared always together in 66 isolates. 22 isolates had the *penA* mosaic type X, and 41 had the *penA* mosaic type XXXIV, both linked to high-level resistance to ESCs[243].

Regarding cefixime, of the 9 isolates phenotypically resistant to this antibiotic, 88.9% (n=8) had the I312M, V316T and G545S mutations in the *penA* gene, 11.1% (n=1) had the P551L mutation, the mosaic allele *penA* type X was present in 55.6% (n=5) of the resistant isolates, and the mosaic allele *penA* type XXXIV was present in 33.3% (n=3) of the resistant isolates. Of the 10 isolates with phenotype of intermediate resistance to cefixime, 90% (n=9) had the I312M, V316T and G545S mutations in the *penA* gene, 20% (n=2) had the mosaic *penA* type X, and 50% (n=5) had the mosaic *penA* type XXXIV. Of the 212 isolates phenotypically susceptible to this drug, 17.9% (n=38) had the I312M, V316T and G545S mutations in the *penA* gene, 13.2% (n=28) had the A501T/V mutation in the *penA* gene, 7.1% (n=15) had the G542S mutation in the *penA* gene, 22.2% (n=47) had the P551L/S mutation in the *penA* gene, 6.1% (n=13) had the mosaic *penA* type X, and 11.3% (n=24) had the mosaic *penA* type XXXIV.

For cefotaxime phenotypes, only 1 isolate had resistance to this drug, having the three I312M, V316T and G545S mutations in the *penA* gene, and the mosaic *penA* type XXXIV. Of the 96 isolates with phenotypic susceptibility to cefotaxime, 30.2% (n=29) had the I312M, V316T and G545S mutations in the *penA* gene, 15.6% (n=15) had the A501T/V mutation in the *penA* gene, 3.1% (n=3) had the G542S mutation in the *penA* gene, 22.9% (n=22) had the P551L/S mutation in the *penA* gene, 12.5% (n=12) had the mosaic *penA* type X, and 17.7% (n=17) had the mosaic *penA* type XXXIV. There was no phenotype of intermediate resistance to this antimicrobial agent.

Regarding ceftriaxone phenotypes, of the 3 isolates with phenotype of resistance to this antibiotic, 33.3% (n=1) had the G542S mutation in the *penA* gene, and 66.7% (n=2) had the P551L/S mutation in the *penA* gene. Of the 2

isolates with intermediate resistance to this drug, both of them had the I312M, V316T and G545S mutations in the *penA* gene, and one of them had the mosaic *penA* type X. Of the 242 isolates phenotypically susceptible to ceftriaxone, 22.3% (n=54) had the I312M, V316T and G545S mutations in the *penA* gene, 12% (n=29) had the A501T/V mutation in the *penA* gene, 5.8% (n=14) had the G542S mutation in the *penA* gene, 20.2% (n=49) had the P551L/S mutation in the *penA* gene, 7.9% (n=19) had the mosaic *penA* type X, and 13.6% (n=33) had the mosaic *penA* type XXXIV.

1.6.3. Resistance to penicillin

Regarding penicillin resistance, 58.8% of the isolates had the L421P mutation in the *ponA* gene. Mutations in the *porB* were also found. 39 isolates had only the A121S mutation, while 161 isolates had both G120K/R/N and A121D/N/G mutations. All these mutations affect the *porB1b* allele, leading to acquisition of penicillin resistance. This allele represents 93.3% of the isolates, but there are 23 isolates with *porB1a* allele, not linked to penicillin resistance. These two alleles allow us to classify the isolates in two serogroups, IA (allele *porB1a*) and IB (allele *porB1b*). Determinant *pilQ* mutation E666K, that confers resistance both to penicillin and tetracyclines, such as doxycycline, was not found in any isolate. Finally, 52 isolates carry a plasmid with the bla_{TEM} determinant.

Of the 63 isolates phenotypically resistant to penicillin, 77.8% (n=49) had the L421P mutation in the *ponA* gene, 71.4% (n=45) had the G120K/N/R mutation in the *porB* gene, 81% (n=51) had the A121D/N/G/S mutation in the *porB* gene, 54% (n=34) had the -35 A deletion in the *mtrR* promoter, 42.9% (n=27) had the A39T or G45D mutation in the *mtrR* gene, and 47.6% (n=30) had the plasmidic bla_{TEM} gene. Of the 153 isolates with phenotype of intermediate resistance to penicillin, 64.7% (n=99) had the L421P mutation in the *ponA* gene, 53.6% (n=82) had the G120K/N mutation in the *porB* gene, 62.7% (n=96) had the A121D/N/G/S mutation in the *porB* gene, 49% (n=75) had the -35 A deletion in the *mtrR* promoter, 37.9% (n=58) had the A39T or G45D mutation in the *mtrR* gene, and 4.6% (n=7) had the plasmidic bla_{TEM}

gene. Of 56 isolates phenotypically susceptible to penicillin, 46.4% (n=26) had the L421P mutation in the *ponA* gene, 35.7% (n=20) had the G120K/N mutation in the *porB* gene, 42.9% (n=24) had the A121D/N/G/S mutation in the *porB* gene, 42.9% (n=24) had the -35 A deletion in the *mtrR* promoter, 46.4% (n=26) had the A39T or G45D mutation in the *mtrR* gene, and 3.6% (n=2) had the plasmidic bla_{TEM} gene.

1.6.4. Resistance to fluoroquinolones

Resistance to fluoroquinolones, such as ciprofloxacin or levofloxacin, is mainly caused by mutations in the *gyrA* and *parC* determinants, but also by mutations in *parE,* and *norM* promoter determinants. Gonococci in this study had no mutations in the last two, but 186 isolates had both *gyrA* S91F and D95A/N/G mutations. 22 isolates had both *parC* S97R and S88P mutations, and the remaining 121 had the *parC* mutations shown in Table 5. 142 isolates had both *gyrA* and *parC* genes with some of the described mutations.

For ciprofloxacin AMR phenotypes, of the 63 isolates with resistance to this drug, 95.2% (n=60) had the S91F mutation in the *gyrA* gene, 96.8% (n=61) had the D95G/A/N mutation in the *gyrA* gene, 6.3% (n=4) had the D86N mutation in the *parC* gene, 61.9% (n=39) had the S87R/N/I mutation in the *parC* gene, and 4.8% (n=3) had the S88P mutation in the *parC* gene. Only 1 isolate had a phenotype of intermediate resistance to ciprofloxacin, having the S91F and D95A mutations in the *gyrA* gene. Of the 63 isolates phenotypically susceptible to ciprofloxacin, 33.3% (n=21) had the S91F mutation in the *gyrA* gene, 33.3% (n=21) had the D95G/A mutation in the *gyrA* gene, 15.9% (n=10) had the D86N mutation in the *parC* gene, and 7.9% (n=5) had the S87R mutation in the *parC* gene. 66.7% (n=42) of the isolates with phenotype of susceptibility to ciprofloxacin had no mutations conferring resistance in any of the *gyrA* and *parC* genes.

Regarding levofloxacin AMR phenotypes, of the 79 isolates with phenotype of resistance to this drug, 100% of them had the S91F and D95G/A/N mutations in the *gyrA* gene, 6.3% (n=5) had the D86N mutation in the *parC*

gene, 75.9% (n=60) had the S87R/N mutation in the *parC* gene, and 21.5% (n=17) had the S88P mutation. The unique isolate with phenotype of intermediate resistance to levofloxacin had no detected mutations conferring resistance. Of the 64 isolates phenotypically susceptible to this antibiotic, 1.6% (n=1) had the S91F and D95N mutations in the *gyrA* gene and the S87R and S88P mutations in the *parC* gene, the remaining 98.4% (n=63) had no mutations conferring resistance in *gyrA* or *parC*.

1.6.5. Resistance to tetracyclines

Resistance to tetracyclines, such as doxycycline, is mainly caused by mutation V57M in *rpsJ* or by the presence of a conjugative plasmid carrying the *tetM* gene. Less frequently, it can be caused by the same mutations in *porB* and *pilQ* as in the case of penicillin. The mutation V57M in *rpsJ* was found in 88.6% of the isolates, while the *tetM* gene was found in 43 isolates.

Of the 60 isolates with phenotype of resistance to doxycycline, 100% of them had the V57M mutation in the *rpsJ* gene, 20% (n=12) had the plasmidic determinant *tetM*, 66.7% (n=40) had the G120K/N/R mutation in the *porB* gene, and 76.7% (n=46) had the A121D/N/G/S mutation in the *porB* gene. Of the 25 isolates with phenotype of intermediate resistance to doxycycline, 100% of them had the V57M mutation in the *rpsJ* gene, 48% (n=12) had the G120K/N/R mutation in the *porB* gene, and 56% (n=14) had the A121D/N/G/S mutation in the *porB* gene. Of the 86 isolates phenotypically susceptible to doxycycline, 83.7% (n=72) had the V57M mutation in the *rpsJ* gene, 1.2% (n=1) had the plasmidic determinant *tetM*, 7% (n=6) had the G120K/N mutation in the *porB* gene, and 45.3 (n=39) had the A121D/N/G/S mutation in the *porB* gene.

1.6.6. Resistance to spectinomycin

Mutations in the 16S rRNA or in *rpsE* gene cause resistance to spectinomycin, but no isolates with these mutations were found, and all the isolates with information about AMR phenotype to this drug (n=231) were susceptible.

1.6.7. Comparison with ARIBA results

We run ARIBA with the cleaned reads to double-check our results and to compare both methods of identification of AMR determinants (see Supplementary Table 5). Most of the results were identical for both methods, but there were some differences:

- Our BLAST-based method lacked information about the *penA* gene in 3 isolates and ARIBA found them, although they had no mutations conferring resistance. The BLAST method detected 1 isolate with the mutation P551S in *penA* and ARIBA did not; however, ARIBA detected 1 isolate with the mutation A501V that was not detected with the BLAST method.
- The BLAST method did not find the *gyrA* gene in 2 isolates but ARIBA found them, and they had the S91F and D95G mutations.
- The BLAST method detected 2 isolates with the D86N mutation in the *parC* gene that ARIBA did not detect. The BLAST method detected 10 isolates with the S87R mutation in the *parC* gene that ARIBA did not detect. Both methods detected a S87I mutation in the *parC* gene that the other method did not detect.
- ARIBA detected 31 isolates with the G120K mutation in the *porB* gene that were not detected using the BLAST method, and also detected 30 isolates with the A121D mutation in the *porB* gene that the BLAST method did not detect. The BLAST method detected 1 isolate with the A121D mutation in the *porB* gene that ARIBA did not detect.
- The BLAST method detected 3 isolates with the L421D mutation in the *ponA* gene that ARIBA did not detect.
- The BLAST method failed to detect the *mtrR* gene and its promoter in 3 isolates; ARIBA detected them but they did not have mutations

conferring resistance to antimicrobials. ARIBA detected 1 isolate with the G45D mutation in the *mtrR* gene that the BLAST method did not detect.

- The BLAST method detected 5 isolates with the bla_{TEM} plasmidic determinant that ARIBA did not detect.
- The BLAST method detected 5 isolates with the *tetM* plasmidic determinant that ARIBA did not detect.

1.7. Mobilome

The cryptic plasmid was present in 97.4% (n=333) of the isolates. The conjugative plasmid was found in 24% (n=82) of the isolates. Of these 82 isolates, 47.6% (n=39) did not carry the *tetM* gene that confers resistance to tetracycline, 32.9% (n=27) carried the American-type plasmid and 19.5% (n=16) carried the Dutch-type plasmid, both with the *tetM* gene. The β-lactamase producing plasmid was present in 15.2% (n=52) of the isolates.

The Gonococcal Genetic Island (GGI) was found in 72.3% (n=248) of isolates (Figure 25; Supplementary Table1).

1.8. Dating analysis

The analysis of the correlation between the sampling dates of the isolates and their divergence with their ancestor was done with TempEst both for the complete dataset and for each of the 8 STRUCTURE groups. We obtained low values of the coefficient of determination (R^2) in the regression line, which indicates a poor temporal signal (Table 6).

Despite this, we wanted to explore the dating of the ancestor of our dataset with LSD2. Unsurprisingly, the results between the complete dataset and the different STRUCTURE groups were, in some cases, inconsistent as the time of the most recent common ancestor (tMRCA) for the complete dataset was estimated in year 1756 and tMRCAs for STRUCTURE groups I and VIII were estimated in 1734 and 1200, respectively, i.e. decades and centuries earlier

than the complete dataset. The higher R^2 values corresponded to STRUCTURE groups I to V, which estimated their tMRCAs between 1734 and 1973 (Table 6). That would lead us to think that the current gonococcal population in Spain is not as old as previously thought, nor as old as has been estimated globally[244]; in fact, some of the STRUCTURE groups would have very recent ancestors, well into the 20th century. However, knowing that the temporal signal for our dataset is poor, we cannot take these results as valid. The estimated evolutionary rates were around 10^{-6} substitutions/site/year (s/s/y). This was consistent with the results of other studies which used isolates from longer periods of time than ours[244,245].

Table 6 | TempEst and LSD2 results for the complete dataset and the 8 STRUCTURE groups.

	N	R^2	Rate	Rate CI	tMRCA	tMRCA CI
All	342	7.57E-2	5.76E-6	4.93E-6; 6.37E-6	1756/02/13	1706/12/03; 1785/10/01
I	41	0.4241	4.37E-6	2.64E-6; 6.16E-6	1734/12/03	1544/12/08; 1820/09/14
II	40	0.2492	1.25E-5	8.47E-6; 1.66E-5	1922/01/26	1871/06/23; 1947/05/25
III	51	0.3403	3.09E-6	2.00E-6; 4.12E-6	1762/02/21	1613/06/21; 1833/05/19
IV	24	0.6414	1.19E-5	6.59E-6; 1.61E-5	1878/09/19	1756/01/18; 1922/03/02
V	56	0.4490	1.08E-5	7.49E-6; 1.38E-5	1973/09/09	1954/10/18; 1984/01/09
VI	37	0.1446	2.97E-6	2.14E-6; 3.78E-6	1990/10/11	1977/12/17; 1998/01/14
VII	57	5.79E-2	4.62E-6	2.95E-6; 6.01E-6	1920/01/14	1868/04/15; 1943/02/02
VIII	36	0.1786	1.77E-6	1.39E-12; 7.32E-6	1200/08/12	-1.12E9; 1818/01/21

2. Discussion

Gonococcal infection is an increasingly worrying concern for Public Health worldwide due to its increasing incidence in the last decades and the capacity of *Neisseria gonorrhoeae* for acquiring resistance to every antibiotic that has been used to treat its infections, which is causing this bacterium to evolve towards a superbug state[88]. This is also favored because, apparently, the acquisition of many of the AMR determinants has no cost on the biological fitness of the gonococcus[110].

As a result, it is of high importance to establish solid surveillance programs that report the incidence of this pathogen, its transmission dynamics, and the acquisition and evolution of mutations that confer resistance to treatment. In this regard, there are different programs at different scales to monitor gonococci. For example, the gonococcal isolate surveillance project (GISP) is a program that monitors gonococcal trends in the United States since 1986[246,247], the gonococcal antimicrobial surveillance program (GASP) is a worldwide network of laboratories that monitors the trends of AMR patterns in gonococci[248], and the Euro-GASP program does the same focused on the European Union countries[249].

High-throughput sequencing technologies have facilitated the study of the genomic epidemiology of many pathogens[250–253] and *N. gonorrhoeae* is not an exception, with most studies focusing on gonococcal epidemiology and AMR surveillance[70,244,254–261]. The increasing use of whole-genome sequencing to study the epidemiology is reaching the Public Health systems and agencies. The World Health Organization has published a report discussing the benefits and limits of HTS for AMR surveillance[69].

Here, we present a study of the gonococcal genomic epidemiology and AMR determinants surveillance in Spain. To our knowledge, this study represents the first in-depth HTS-based analysis of *Neisseria gonorrhoeae* epidemiology and AMR in this country.

The population structure analysis showed that gonococci has no significant differences between the individuals from the Mediterranean areas

—Catalonia and the Comunidad Valenciana (CV)— and those from the central area, represented by Madrid isolates. In fact, gonococcal isolates showed a high level of genetic admixture even when the recombinant genes were removed from the analysis. Our method for detecting recombination was very conservative, which is translated into some level of undetected recombinant events, but it was surprising that the admixture levels were so high after the removal of recombination. This is evident when we observe that STRUCTURE groups have isolates within other groups in the phylogenetic tree —this also occurs in other works using BAPS approach to cluster isolates of the population[244,258]—, or in the presence of some polyphyletic STs or genogroups from the different typing schemes —e.g. ST 7363 from MLST scheme, or G437 from NG-MAST scheme—. However, the fact that these polyphyletic STs appear in our whole-genome SNPs phylogeny is another proof of the higher resolution of this methodology in contrast to using the classical typing schemes.

Despite the above, in genomic studies it is very useful to inform about the STs detected, given their extensive use in the clinical and epidemiological fields. For this reason, many studies that already use HTS to study the epidemiology of a pathogen report the STs of their isolates[262,263], even some genomic studies focus on a specific ST[264]. In this sense, we have explored the dynamics of the STs inferred from the three available typing schemes in gonococcal isolates during the period of study. For example, we observed a replacement of MLST ST 1901 by other STs, such as ST 1599. The same holds for NG-MAST genogroups, where G3378 or G2992 were replaced by G11461, and NG-STAR STs, where STs 90 or 163 were replaced by 442 or 520. Similar genogroups were found in molecular studies of other Spanish regions[265,266]. Also, ST profiles can be associated with population risk groups, but we lack this information.

Lack of information about population risk groups and of the sexual orientation of the patients are the main limitations of this study, because we cannot relate genotypes with risk groups. In addition, the bias in the sex of the patients is evident, a fact that is very common in other studies[267–270]. We can speculate that the majority of males included in the study are MSM, but

without the real data it is only an assumption. Another limitation derives from the sampling dates of the isolates, which were insufficient to show a temporal signal and it was impossible for us to estimate the date of the most recent common ancestor for the Spanish gonococcal population represented by this dataset.

In this work, we have shown that resistance to penicillin, fluoroquinolones and tetracyclines is highly prevalent in Spain, and resistance to macrolides started to be of concern during the period of this study. However, the phenotypes show high rates of susceptibility to extended-spectrum cephalosporins (ESCs) and spectinomycin. Since the emergence of resistance to ESCs in gonococci is the main concern to Public Health systems, as ESCs are the last-line treatment for gonorrhea[119], we can think that the situation in Spain is not so worrisome. However, when we inspected the genotypes we found a high number of known AMR-conferring mutations in the *gyrA, ponA, porB,* and *rpsJ* genes —causing resistance to fluoroquinolones, penicillin, and tetracyclines—, as well as in the *mtrR* gene and its promoter —which confer resistance to multiple antibiotics—. Consistent with the phenotype, the number of isolates with mutations in the *penA* gene —conferring resistance to ESCs— was not too high. However, the presence of *penA* mosaic alleles type X and, above all, type XXXIV, which are associated with increased resistance to ESCs[243], was notable.

There were discrepancies between genotypic and phenotypic results, and this is a matter that has been discussed in several studies[271–273] and the arguments go from the expression, or not, of the genes that contain these mutations, errors in the interpretation of results —unlikely at these levels—, or the different levels of resistance that these mutations confer and whether they require the presence of other mutations so that the corresponding phenotype appears. In any case, the level of resolution offered by HTS for the monitoring of resistance determinants is clear, and although the resistant phenotype is not observed, the presence of relevant mutations should be taken into account when taking decisions for treatment to prevent favoring the appearance of more mutations and, finally, the dreaded resistance phenotype. In addition, computer programs, existing and under development,

facilitate the identification of resistance determinants directly from the reads, which facilitates and accelerates genotypic identification. We have compared our results with ARIBA and, although there are discrepancies due to limitations in both methods —generally related to the quality of the sequencing reads—, the results are generally comparable.

Summarizing, in this work we have exploited the enormous amount of data derived from HTS to study the epidemiology of gonococci in Spain and monitor the AMR determinants present in such isolates. This is the first study on using whole-genome data to explore the Spanish gonococcal population, adding useful information to the Public Health system.

CHAPTER 2

WHOLE-GENOME SEQUENCING OF GONOCOCCI IN A FORENSIC CASE OF TRANSMISSION

This Chapter has been published as:

Francés-Cuesta C, de la Caba I, Idigoras P, Fernández-Rodríguez A, Del Valle Pérez D, Marimón JM, González-Candelas F. Whole-genome sequencing of *Neisseria gonorrhoeae* in a forensic transmission case. *Forensic Sci Int Genet* 2019; **42**: 141-6. DOI: 10.1016/j.fsigen.2019.07.003

1. Background

In the last decades, molecular epidemiology analyses have been applied to the study of virus transmission cases and outbreaks in forensic settings[274–279]. This methodology was introduced in the forensic field by Ou and colleagues[274], who demonstrated the transmission of human immunodeficiency virus (HIV) by a dentist to some of his patients. The first case in a criminal court was presented by Metzker and colleagues[277], whose evidence contributed to the conviction of a physician for the attempted homicide of his former lover by deliberate injection of HIV. There are other cases involving more transmission events during a longer period of time, such as the case of a Spanish anesthesiologist who infected hundreds of his patients with hepatitis C virus (HCV) during 10 years[279]. Once again, the phylogenetic evidence provided by the researchers contributed to show the common ancestry of the viruses found in the source and those in the outbreak victims, ultimately leading to the conviction of the suspect.

Viruses, especially those with an RNA genome such as HIV and HCV, have very fast evolutionary rates which result in the fast accruing of new genetic variants in very short —even days— periods of time in their small genomes —typically around 10 Kb for these two viruses—. Bacteria are cellular organisms with DNA genomes several times larger than those of RNA viruses —typically between 1 and 7 Mb— and with much slower evolutionary rates, similar to those of eukaryotes. In order to evaluate the genealogical relationships among bacterial isolates, it is necessary to evaluate longer portions of the corresponding genome sequences, ideally their complete genomes. The application of complete genome sequencing to the anthrax mail attacks in the United States in 2001[280–282] eventually led to the establishment of Microbial Forensics as a discipline[283]. Currently, the increasing availability and low economic costs associated with high-throughput sequencing (HTS) technologies have facilitated the sequencing of complete bacterial genomes, thousands of which are already deposited in publicly accessible databases. These techniques allow a fast analysis of genome sequences although the processing of the raw data they generate is much more computationally and technically demanding[284]. However, to our knowledge, no forensic analysis of

bacterial genome sequences presented to courts, civil or penal, has been published to date, much less so using HTS techniques.

In this chapter, the biological results of the genetic and genomic analyses of the isolates involved in this case are presented. For this, a variety of techniques of increasing sensibility and resolution power were used: from pulsed-field gel electrophoresis (PFGE) along with multi-locus sequence typing (MLST) to complete genome sequencing. Given the absence of literature on bacterial whole-genome sequencing (WGS) related to a forensic case, our methodology could establish the bases of a new way of expert analysis in criminal cases.

2. Description of the case and specific methods

2.1. Case description

In the first half of 2017, *N. gonorrhoeae* was isolated from the vaginal exudate of a young girl who visited her ambulatory general pediatrician in Donostia (Basque Country, Spain) because of a vaginal purulent secretion. Two adult family members, a male and a female, also tested positive for gonorrhea, leading to the suspicion of a case of child abuse. The gonococci from the girl and her male-relative could be isolated by culture unlike those from the female-relative, which was detected only by PCR. This person had started antibiotic treatment a few days earlier because of abdominal pain. A genetic comparison between the *N. gonorrhoeae* isolates from the male-relative —the suspect— and the girl —the victim— was requested to help in establishing a possible case of sexual abuse.

2.2. Gonococcal isolation and detection

The first case detected was a young girl who attended her ambulatory general pediatrician because of purulent vaginal secretions. In the Gram stain of the vaginal exudate, Gram-negative cocci compatible with gonococcus were observed and cultivated after 48 hours in blood and chocolate agar plates (bioMérieux, Marcy-l'Étoile, France) in a 5% CO_2 enriched atmosphere. Species identification was performed by MALDI-TOF. *Candida albicans* was also isolated in the vaginal exudate. The vaginal swab was also tested with the commercial multiplex-PCR for Sexually Transmitted Infections (STI) targeting *N. gonorrhoeae, Chlamydia trachomatis, Trichomonas vaginalis, Mycoplasma genitalium, Mycoplasma hominis, Ureaplasma urealyticum*, and *Ureaplasma parvum* (Allplex™ STI Essential Assay, Seegene, Seoul, South Korea), giving a positive result for *N. gonorrhoeae* and negative for the other pathogens.

After this finding of possible child abuse, an investigation in the familiar setting was initiated three days after the victim's examination. At the onset of the investigation a female adult relative, admitted to hospital four days before because of abdominal pain, was in treatment with piperacillin-tazobactam. In the urine collected 4 days after the onset of antibiotic treatment, *N. gonorrhoeae* was detected by the commercial PCR GeneXpert CT/NG test (Cepheid, Sunnyvale, CA, USA). However, it was not possible to culture the gonococcus. An elder sister of the child was also tested for possible sexually transmitted pathogens, obtaining a negative result for all of them. A direct male adult relative was also examined, showing dysuria and a clear urethral exudate in which *N. gonorrhoeae* was detected by PCR and by culture. All the PCRs for *N. gonorrhoeae* were performed in duplicate.

The chain of custody was secured with the use of forms for the transportation and during the analysis of the samples that guarantee the samples identity and quality. Samples and subsamples followed a unique identification label code that guaranteed their identification.

2.3. Molecular epidemiological studies at the hospital

A preliminary analysis to establish the genetic relatedness between the isolates from the suspect and victim was performed by MLST and PFGE. The *N. gonorrhoeae* isolates of another twelve unrelated individuals —living in other villages as the one where the family lived— cultured in the same dates as the ones from the family were also studied as negative controls. PFGE was performed according to the protocol used for *N. meningitidis*[285] using *Nhe*I and *Spe*I separately. Fragments were subjected to 23-hour electrophoresis with a pulse angle of 120°, switching times increasing from 0.5 to 25 seconds.

For MLST, bacterial DNA was automatically extracted using the Nuclisens easyMag platform (bioMérieux, Marcy-l'Étoile, France) and housekeeping gene fragments amplified using the primers and conditions described at the *Neisseria* Sequence Typing webpage (https://pubmlst.org/neisseria/). After sequencing, allelic numbers and ST were assigned using the software and database available at the webpage.

Molecular analyses were performed according to a separate-areas workflow, with dedicated pre-PCR —extraction and PCR-set up zones— and post-PCR areas in a DNA-free environment —UV-irradiated workstation— using certified sterile DNA-free plastic consumables to avoid contamination. In the extraction, amplification, and sequencing steps negative controls were included.

2.4. DNA sequencing and bioinformatics analyses

DNA extraction, quantification, and whole-genome sequencing have been explained in the Methods section.

To verify the ST assignment, *in silico* genotyping was performed using SRST2[203] and the *Neisseria* spp. MLST database available at https://pubmlst.org/neisseria/. 26 additional controls were added from the Chapter 1 gonococcal dataset. They were selected because they had the same ST (ST 9363) as the isolates from the hospital. These additional controls were

derived from two different Spanish regions, Catalonia and Comunidad Valenciana (CV), geographically apart from the Basque Country, where the case samples had been obtained.

The reference selection, mapping, SNP calling, phylogenetic reconstruction, generation of SNP distances matrix, and accessory genome identification steps were detailed in the Methods section.

3. Results

3.1. Molecular analyses at hospital

The *Spe*I and *Nhe*I PFGE profiles of the suspect and the victim were indistinguishable (Table 7, Supplementary Figure 4). Surprisingly, of the 12 control isolates studied, one apparently epidemiologically unrelated isolates showed a PFGE pattern with 100% similarity after restriction with both enzymes —isolate identified as local control 3, LC3—. So, these three isolates along with two additional local control isolates showing high similarity with the PFGE patterns of the isolates under investigation (Table 7) were subjected to MLST analysis.

Table 7 | MLST and PFGE profiles of the 5 *N. gonorrhoeae* isolates from Donostia, Spain (LC = Local Control).

ISOLATE	**MLST**	**PFGE *Spe*I**	**PFGE *Nhe*I**
Victim	ST 9363	A	A
Suspect	ST 9363	A	A
LC1	ST 9363	C (88%)[1]	A
LC2	ST 9363	B (94%)[1]	B (90%)[1]
LC3	ST 9363	A	A

1. % similarity to profile A.

The MLST analysis revealed that the five isolates were identical and corresponded to ST 9363. In consequence, these results did not provide enough resolution to ascertain whether the isolates of the victim and the suspect were closely related because of direct transmission, or had been independently acquired and were indistinguishable because this particular strain was common and circulating in this geographical area, as revealed by their similarity to other unrelated isolates. Hence, we proceeded to obtain complete genome sequences of these five isolates.

3.2. Genomic epidemiology analyses

The *N. gonorrhoeae* genomes from the suspect, the victim, and the three local controls were sequenced using the Illumina NextSeq platform. The average number of reads obtained per strain was 1,750,335 (range 935,938-2,860,862). After the trimming and cleaning steps, this number was reduced to 1,611,630 reads (range 856,274-2,613,098). The mapping step yielded an average depth coverage of 85.57 X and an average mapping percentage of the positions of the reference genome of 94.78% (range 94.15–95.09%) (Supplementary Table 6). In order to increase the number of control samples, raw reads from 26 gonococci isolates of the same ST from an independent study with samples from Catalonia and CV were added to the analyses (Supplementary Table 6). The resulting alignment of core genome positions spanned 2,173,861 bp and contained 6,792 variant positions (SNPs).

The matrix of pairwise distances in the core genome (Supplementary Table 7) revealed that there were no differences between the suspect and victim isolates, which suggests that both isolates are clonal or share a very close most recent ancestor. One of the local controls (LC3) was very close to these isolates, from which it differs by only 2 SNPs. Further investigation revealed that there was no relationship between this control and the case family, but also revealed that the source of infection of this control patient was located in a local brothel. The other two local controls differed from the suspect/victim isolates by 66 (LC1) and 1,096 (LC2) SNPs (Supplementary Table 7). The first value indicates a relatively close relationship between the

four isolates —suspect, victim, LC3 and LC1—. The number of SNPs between the case isolates and the additional controls from distant Spanish regions ranged between 1,060 and 1,661, indicating lack of close relationships. This also holds true for LC2 and the case isolates.

A similar range of differences was observed among the additional controls (0-1,541 SNPs). There were 4 cases of perfect identity between pairs of additional controls (AC8-AC12, AC9-AC13, AC11-AC15, AC18-AC20), 3 cases with only 1 SNP (AC18-AC23, AC20-AC23, AC24-AC25) and 5 pairs which differed by 2 SNPs (Supplementary Table 7). Previous studies have suggested that isolates with a temporal difference of isolation of up to 3 months approximately, and with 0–6 SNPs can be considered as a direct transmission[254]. Due to the anonymous character of the data, no additional information could be obtained, but these patients had similar ages, dates of isolation, and attended the same hospital, so they likely represented sexual partners.

The multiple alignment of 6,792 SNPs was used to obtain the maximum likelihood tree shown in Figure 26. This tree summarizes the information reflected in Supplementary Table 7 and reveals that the isolates included in the analysis —cases and controls— fall in well-supported clusters with a certain level of geographical structure. In all the cases of close relatedness among isolates, the corresponding samples were from the same geographical area.

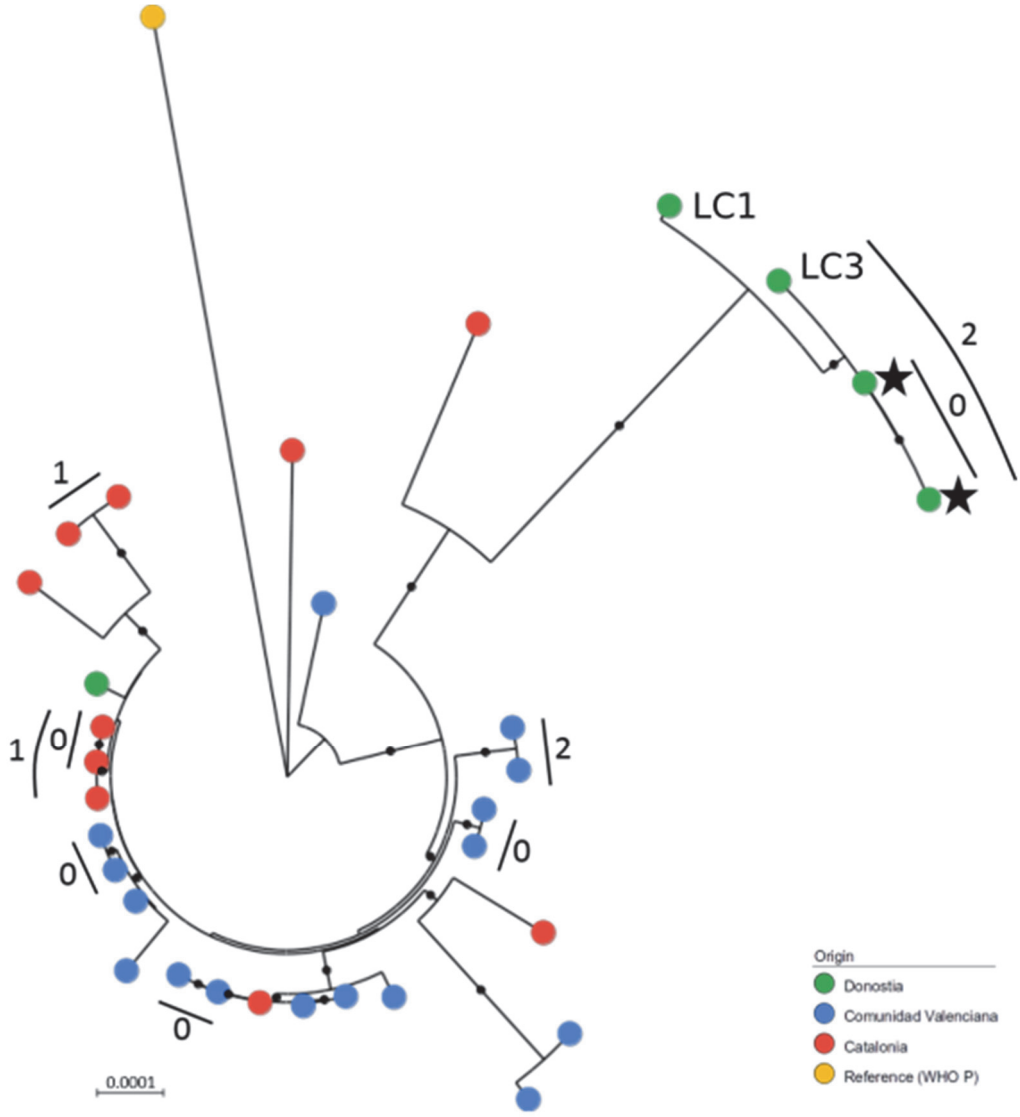


Figure 26 | Unrooted maximum-likelihood tree showing the suspect and victim isolates (black stars) and the unrelated control isolates. Black dots in branches indicate bootstrap support values higher than 90%. Numbers correspond to SNPs between some isolates.

3.3. Accessory genome analysis

The previous analyses had been performed with the mapped genome, which includes the nucleotide positions common to all the studied isolates, but not the accessory genome, represented by positions not included in the reference and not necessarily shared by all the isolates. To further evaluate the genetic similarity among these, we proceeded to analyze the accessory genes, and to

identify and characterize the plasmids from the victim, suspect, and the two closest control isolates. For this, the unmapped reads were extracted and assembled. The number of contigs ranged from 17 to 25, but this range reduced to 10–13 when the contigs shorter than 250 bp (Supplementary Table 8) were removed.

Contigs suspected to contain plasmid-related sequences were separated, and the remaining contigs were annotated. Clustering of orthologous genes resulted in the 4 isolates sharing 67 identical genes, and one of the control strains (LC1) had 2 additional genes (Supplementary Table 9). Among the contigs containing plasmid-related sequences, the WHO P strain plasmid (GenBank accession LT592158.1) was identified in the 4 isolates. The plasmid corresponds to the gonococcal cryptic plasmid[286]. This is a small plasmid with 4,207 bp encoding 8 proteins with no major biological or clinical relevance. The sequence of the reference plasmid matched with an identity score of 100% to those derived from the 4 isolates (Figure 27, Supplementary Table 10). In these evaluations of similarity, we considered not only mismatches but also indels. In consequence, the genetic similarity between the suspect and victim isolates was 100% in both the bacterial chromosome and the plasmid.

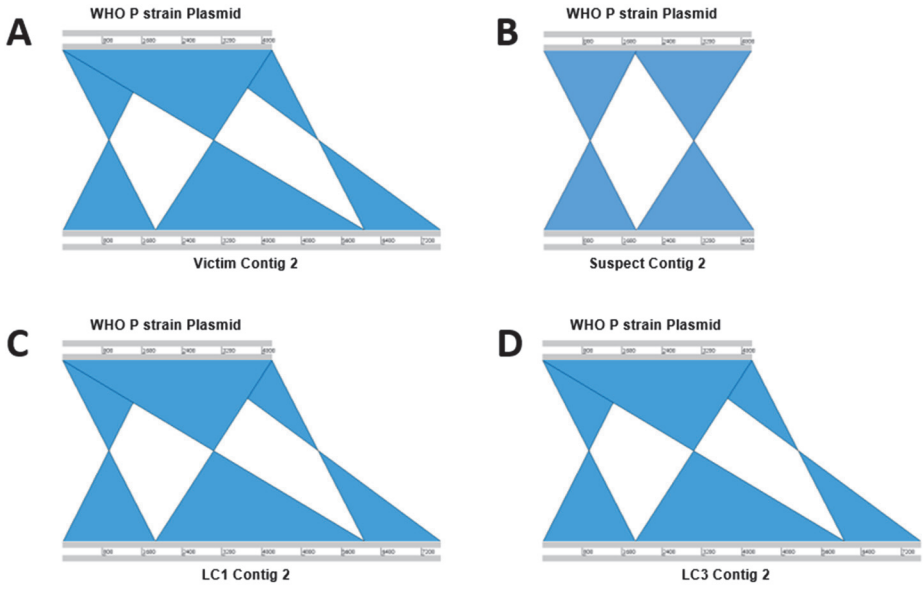


Figure 27 | Location of the *N. gonorrhoeae* WHO P strain plasmid within the plasmid contigs of the 4 isolates. In the case of the victim and local controls

(A, C, D) the sequence of the plasmid is found almost twice because the sequence recircularized during the assembly step. In the case of the suspect (B) the plasmid is found once and it is split in half but the same plasmid matched perfectly once both halves were reordered, with an overlapping region of 55 nucleotides. In all cases, the sequences of the isolates were reverse-complementary to the reference plasmid sequence, but the match with this plasmid was 100% (see Supplementary Table 10).

4. Discussion

The use of molecular epidemiology to investigate cases of transmission in a forensic context has become increasingly frequent. Since the pioneering study by Ou and colleagues[274], the analyses have become more complex, both in the molecular technologies used to obtain the sequence information and in the ensuing bioinformatics, evolutionary and statistical methods applied to that information. Although this type of studies usually deals with deliberate or unintentional transmission of viruses, mainly HIV or HCV, the increasing accessibility to complete genome sequences of cellular organisms, especially bacteria, opens new possibilities for the application of the same methods to cases of bacterial transmissions.

A technique used in recent decades as a universal method for the characterization of bacteria is the MLST[287]. This technique involves the sequencing of several loci —7 loci in the case of *Neisseria* spp.— to create an allelic profile that corresponds to a specific genotype. MLST is widely used in molecular epidemiology studies. In a similar way, in the cases of transmission of a virus, genes encoding key proteins —from the envelope, the matrix or the polymerase— are sequenced.

For studying the short-term epidemiology of *N. gonorrhoeae*, PFGE seems to be more discriminative than MLST[288]. In fact, PFGE has already been used in child sexual abuse cases to *N. gonorrohoeae* typing[289]. In this study, PFGE was able to discriminate between some isolates with the same ST 9363, although no differences were obtained between the isolates of the suspect, victim and one of the control isolates whose genome only differed in 2

nucleotides. In this context, PFGE as well as MLST could be useful if they showed different PFGE patterns or different ST to exclude a relationship between isolates.

However, the resolution offered by sequencing a few genes (MLST) or mapping the genome with rare-cutting endonucleases (PFGE) cannot be compared to that offered by the sequencing of whole genomes for bacterial epidemiological studies[64,290–292]. For this reason, this type of sequencing is frequently applied to the analysis of outbreaks[293]. For the present study, we wanted to take advantage of this power of resolution to elucidate a bacterial transmission in a criminal context.

Bacteria do not evolve as quickly as viruses, and gonorrhea is an infection that usually does not become chronic. Therefore, by means of phylogenetic methods the direction of transmission cannot be determined. Also, the time of detecting the bacteria will not help in determining the direction of transmission as alleged victims will be commonly detected before the suspect is investigated. However, the case that concerned us was an alleged sexual child abuse, so the direction of the transmission seemed quite obvious. It was necessary to establish whether the gonococci of the suspect and the victim were clonal. Our analyses were based on the study of the complete genomes, including both the core and the accessory genomes, and determined not only that the suspect and victim isolates were identical, but that they also had the same plasmid.

Medical-forensic investigations of sex-related crimes usually include an evaluation of biological, criminal and toxicological analyses. Most biological studies are aimed at comparing the DNA from samples of the suspected aggressor with that found in very specific cells that might have been left on the victim, mainly spermatozoa and epithelial cells. Nevertheless, these studies are usually limited by different unavoidable elements such as the time elapsed between the aggression and sampling, washing, sample degradation, scarceness of the biological fluids of the aggressor, external contamination, etc.

However, when a sexual assault results in a sexually-transmitted infection it is possible to compare the bacterial DNA from the victim with that of the presumed aggressor. This approach can overcome some of the above limitations and, eventually, provide additional evidence within a strong, well-established scientific domain. Hence, the analysis of complete genome sequences from microorganisms opens new avenues for forensic investigations within the framework of microbial forensics[283,294]. Nevertheless, there is still a pressing need for standardization and validation of the procedures, especially when the sequences are obtained using high-throughput (HTS) sequencing methodologies[284,295].

In this particular case, the genome analysis results revealed the existence of a transmission through contact, which was included in the judicial process as an additional element along with other evidences. The court, after evaluation of all the evidences, concluded that there was not conclusive evidence that the contact was exclusively related to sexual abuse and that other types of contact could not be discarded. This led to the dismissal of the case but, at least, it allowed institutional control and follow-up of the minor and her family.

Sexual abuses in children, particularly those occurring in the family setting, are a matter of concern for judicial authorities. The presence of a sexually transmitted infection can be used to support allegations of sexual abuse, but the particular significance of the identification of a sexually transmitted agent as an evidence of possible child sexual abuse varies by the type of pathogen. Postnatal acquired gonorrhea usually suggests some kind of sexual contact[296,297]. Although traditional methods such as PFGE[289] have been used to compare strains in judicial cases, more accurate methods based on HTS such as the one reported in this study are required. Efforts should be made to go in depth in the epidemiology of the pathogen —in this case, *N. gonorrhoeae*— and to enlarge and improve the existing databases, which should provide data of the different geographical populations of the species.

In summary, the development of HTS technologies and bioinformatics techniques and tools for information processing has allowed multiple studies

in the field of microbial genomics and molecular epidemiology[293,298]. Despite the great progress in these fields, no studies applying whole-genome sequences from HTS to shed light on a case of forensic microbiology have been published yet. Here, we have provided a methodology for determining the clonality of two isolates involved in a transmission event related to a criminal case taking advantage of the high resolution of HTS and involving both the analysis of the core and the accessory genome. We hope that the methodology used in this study helps in establishing the bases for the development of bacterial transmission in forensic context.

CHAPTER 3

GENOMIC ANALYSIS OF TWO *Serratia marcescens* OUTBREAKS IN HOSPITALS FROM COMUNIDAD VALENCIANA

This Chapter has been published as:

Francés-Cuesta C, Sánchez-Hellín V, Gomila B, González-Candelas F. Is there a widespread clone of *Serratia marcescens* producing outbreaks worldwide? *J Hosp Infect* 2021; **108**: 7-14. DOI: 10.1016/j.jhin.2020.10.029

1. Specific methods

1.1. Outbreaks investigation and measures of control

In September 2017 an outbreak of *S. marcescens* —outbreak A— was declared at the NICU of a 492-bed tertiary care hospital in Comunidad Valenciana (CV). The NICU had a capacity of 9 cubicles. Seven patients were affected, and 15 specimens were collected from clinical samples. Three environmental samples and three controls were also collected (Supplementary Table 11).

While the first outbreak was being analyzed, a second outbreak —outbreak B— was declared in April 2018 at the NICU of a 519-bed tertiary hospital in a different city in the CV, about 260 km apart. This NICU also had a capacity of 9 cubicles but expandable up to 12 if necessary. Six patients were affected, and a specimen was collected from each of them. A control isolate was also included (Supplementary Table 12).

Collected specimens were cultured and bacterial colonies suspected of being *S. marcescens* were identified as explained in the Methods section. Isolates were kept for genomic analysis.

Commonly for both hospitals, multiple infection-control measures were implemented to prevent additional transmissions. All the affected patients were isolated and were attended by exclusive healthcare staff, reducing the influx of people in the affected areas. Frequent hand hygiene with either soap and water or hydroalcoholic solution was heavily emphasized, as well as the use of gloves, both in healthcare workers and newborns' parents.

To control environmental contamination, NICU's cubicles were cleaned using 1% sodium hypochlorite and disinfected with pulsed UV light (Xenex®, San Antonio, TX, USA), and clinical surfaces were disinfected with specific Surfa' Safe Premium spray product (Instrunet®, Barcelona, Spain).

1.2. DNA extraction and genome sequencing

Genomic DNA was extracted, quantified and sequenced as explained in the Methods section. Raw sequence data generated in this study were deposited in the European Nucleotide Archive (ENA) under project PRJEB36342.

1.3. Genomic epidemiology analysis

The genomic analyses were very similar to those in Chapter 2 and were described in the Methods section. A list of the genomes and plasmids used in this Chapter is provided in Supplementary Tables 13 and 14.

2. Results

2.1. Description of the outbreaks

The first case of outbreak A was detected on 6 September 2017 in a newborn admitted to the NICU of the hospital. From this date through 14 November 2017 (weeks 36-46), a total of seven patients —five males and two females— admitted to this NICU were affected (Figure 28a). The microbiological analyses of the samples were positive for *S. marcescens*. Only patient 1 developed a *S. marcescens* bacteremia, which was treated with meropenem. The remaining patients showed colonization by this bacterium but they did not develop an infection and treatment was not required.

The first case of outbreak B was detected on 17 April 2018 in a newborn admitted to the NICU of the second hospital. From this date through 20 July 2018 (weeks 16-29), a total of six patients —two males and four females— admitted to this NICU were affected (Figure 28b). The microbiological analyses were positive for *S. marcescens*. Patient 2 developed a *S. marcescens* bacteremia, which was treated with ceftazidime/amikacin, and patient 6 developed conjunctivitis, which was treated with tobramycin. The remaining patients showed colonization but not infection.

The infection-control measures as well as disinfection of environmental contamination and treatment of infected patients were effective in containing both outbreaks, and none of them was fatal.

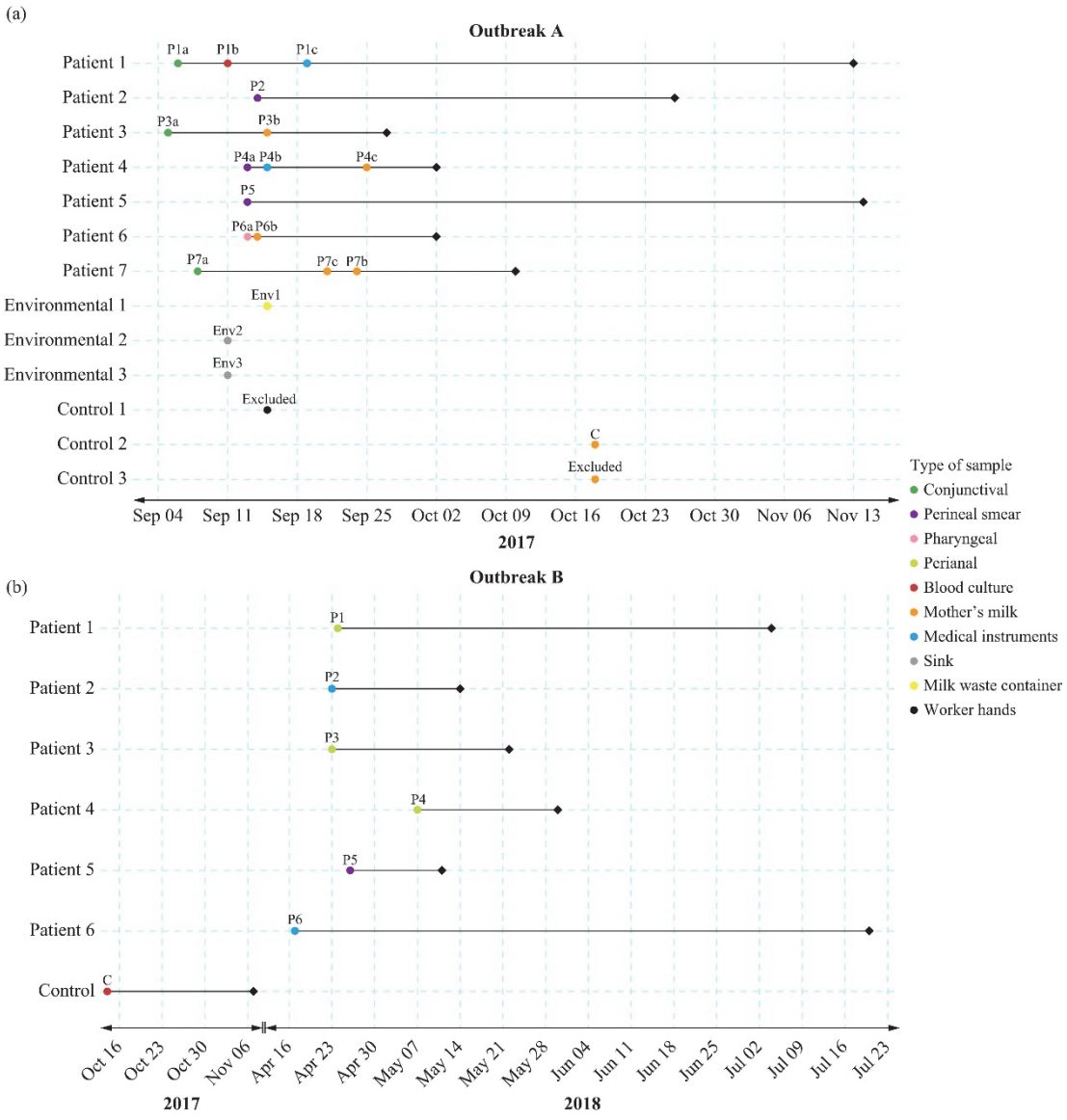


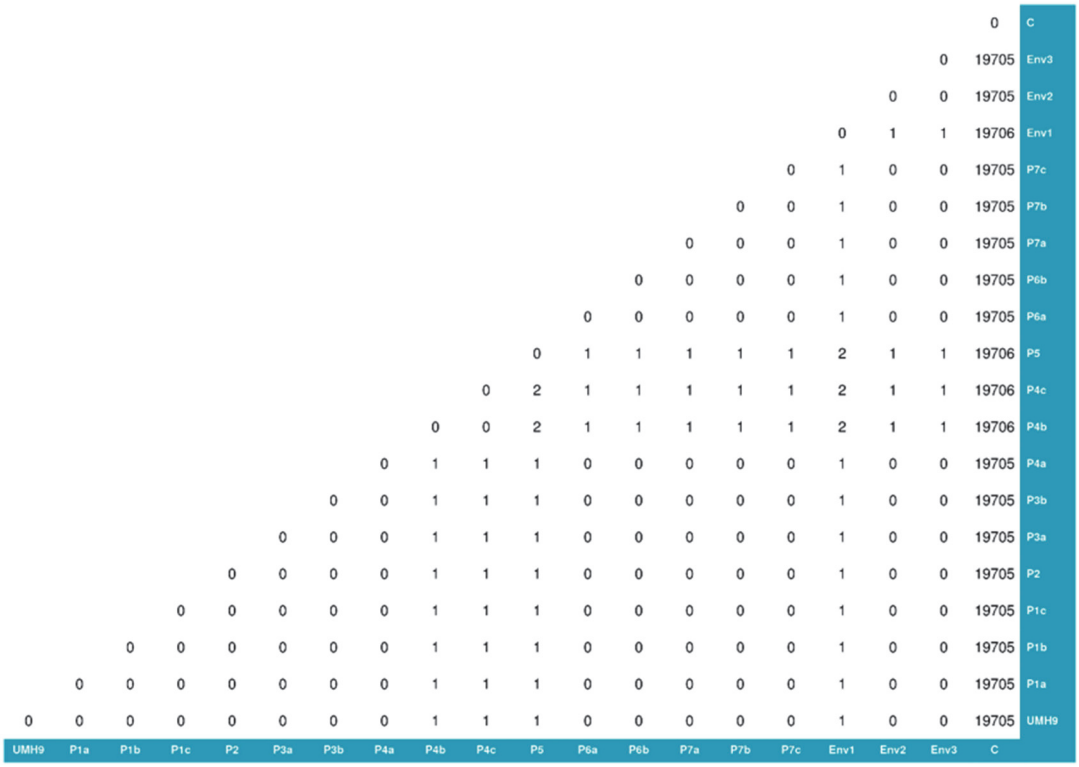
Figure 28 | Outbreaks timelines. Outbreak A timeline (A) shows the different samples taken from each patient during their stay at the NICU and the two excluded controls, while outbreak B timeline (B) shows only one sample from each patient. The dates of discharge are represented by black diamonds. See Supplementary Tables 11 and 12 for further details about the samples.

2.2. HTS and in silico *verification of the bacterial species*

The *S. marcescens* genomes from both outbreaks were sequenced by HTS in two separate runs. The average number of reads obtained for samples from outbreak A was 1,383,705 (range 767,334-2,643,420) and it was 2,454,515 (range 1,716,034-3,410,598) for samples from outbreak B. After applying the quality filters, these values decreased to 1,028,338 (range 557,366-1,988,044) and 2,144,567 (range 1,483,848-2,990,070), respectively (Supplementary Table 15). We confirmed *in silico* that all the strains were *S. marcescens* except two of the three controls from outbreak A which were identified as *S. liquefaciens* and were removed from the subsequent analyses.

2.3. Genomic analysis of outbreak A

Isolates from outbreak A were mapped against the NCBI closest reference genome —strain UMH9—, yielding an average depth coverage of 19.0 X. The samples involved in the outbreak had an almost perfect identity with the reference, with an average of 99.9% of reads being mapped, which covered 94.4% of the UMH9 strain genome. The control sample was far apart from this reference, with 85.2% of its reads being mapped, covering 66.0% of the reference genome (Supplementary Table 16). The full-genome alignment spanned 5,024,591 bp but, after removing repetitive regions, phages, and poorly aligned regions, the alignment was reduced to 4,770,885 bp, which included 27,775 variant positions. The number of variant positions decreased to 157 when the control strain was excluded. The pairwise distance matrix revealed that there were 0-2 single nucleotide polymorphisms (SNPs) between strains in the outbreak (Figure 29). Remarkably, only 0-1 SNPs were detected between samples in the outbreak and the reference genome, making the UMH9 strain indistinguishable from the outbreak strains (Figure 29).

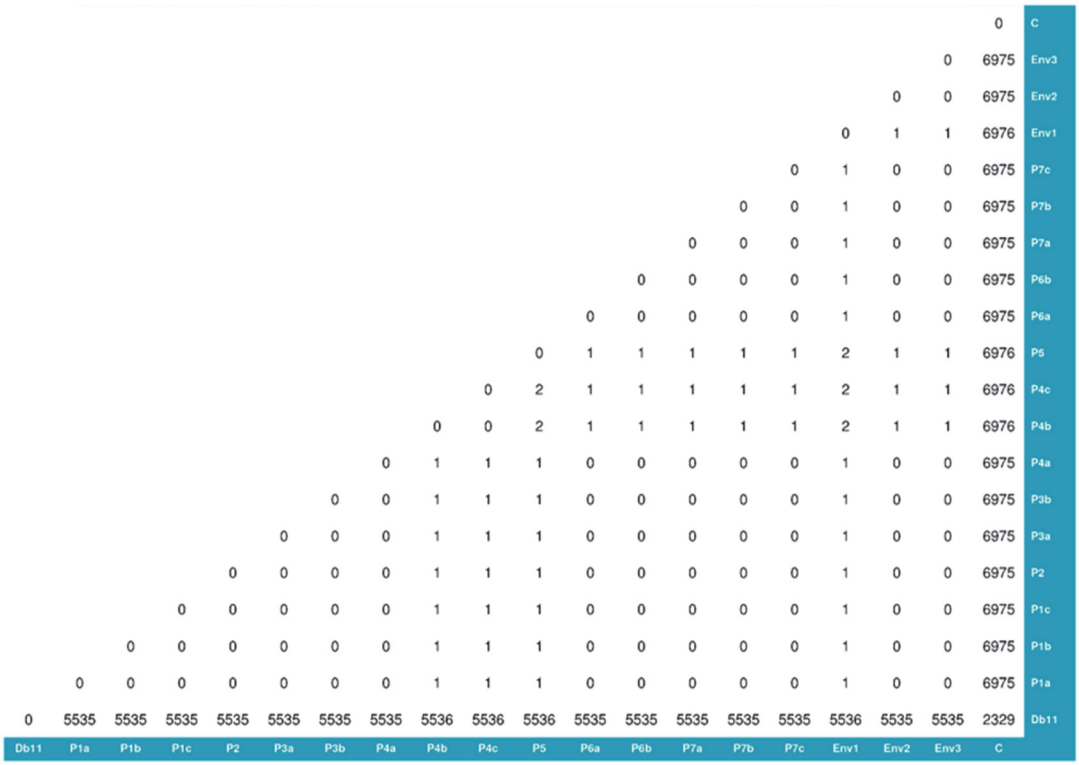


	UMH9	P1a	P1b	P1c	P2	P3a	P3b	P4a	P4b	P4c	P5	P6a	P6b	P7a	P7b	P7c	Env1	Env2	Env3	C
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Env3	19705	19705	19705	19705	19705	19705	19705	19705	19705	19705	19706	19705	19705	19705	19705	19705	19706	19705		
Env2	19705	19705	19705	19705	19705	19705	19705	19705	19706	19706	19706	19705	19705	19705	19705	19705	19706			
Env1	19706	19706	19706	19706	19706	19706	19706	19706	19706	19706	19706	19706	19706	19706	19706	19706				
P7c	19705	19705	19705	19705	19705	19705	19705	19705	19705	19705	19705	19705	19705	19705	19705					
P7b	19705	19705	19705	19705	19705	19705	19705	19705	19705	19705	19705	19705	19705	19705						
P7a	19705	19705	19705	19705	19705	19705	19705	19705	19705	19705	19705	19705	19705							
P6b	19705	19705	19705	19705	19705	19705	19705	19705	19705	19705	19705	19705								
P6a	19705	19705	19705	19705	19705	19705	19705	19705	19705	19705	19705									
P5	19706	19706	19706	19706	19706	19706	19706	19706	19706	19706										
P4c	19706	19706	19706	19706	19706	19706	19706	19706	19706											
P4b	19706	19706	19706	19706	19706	19706	19706	19706												
P4a	19705	19705	19705	19705	19705	19705	19705													
P3b	19705	19705	19705	19705	19705	19705														
P3a	19705	19705	19705	19705	19705															
P2	19705	19705	19705	19705																
P1c	19705	19705	19705																	
P1b	19705	19705																		
P1a	19705																			
UMH9																				

Figure 29 | Pairwise distance matrix for outbreak A when the UMH9 strain is used as reference for mapping, and the corresponding ML phylogenetic tree. The reference strain is highlighted. Black dots in branches represent a bootstrap support value equal or higher than 90%.

Because of the almost perfect identity between outbreak A isolates and the reference, there was no accessory genome except in the control isolate. There was, on average, 5.6% of the UMH9 genome not covered by reads from outbreak isolates. This was probably due to the fact that the coverage and quality of the bases in these regions of the genome did not pass the filters. The unmapped reads of these isolates were only 0.01% and no accessory genome could be assembled. The accessory genome of the control isolate was 655,961 bp, distributed in 321 contigs (Supplementary Table 17).

The mapping step was repeated using the classical reference strain for *S. marcescens* —strain Db11—, which was clearly more distant from the outbreak strains than UMH9. The average depth coverage was 16.3 X, and the breadth coverage for this reference was 79.5% on average for samples involved in the outbreak with 88.7% of the reads being mapped, and 74.8% for control sample with 93.2% mapped reads (Supplementary Table 18). The alignment length was 5,113,802 bp, which was reduced to 4,134,629 bp after cleaning, and contained 162,590 variant positions. If the control strain was excluded, the number of variant positions was 160,012. Again, 0-2 SNPs of difference were found among outbreak samples, with the control strain at a distance of 6,975-6,976 SNPs, and this time with the reference clearly separated from the outbreak by 5,535-5,536 SNPs (Figure 30).



	Db11	P1a	P1b	P1c	P2	P3a	P3b	P4a	P4b	P4c	P5	P6a	P6b	P7a	P7b	P7c	Env1	Env2	Env3	C	
Db11	0	5535	5535	5535	5535	5535	5535	5535	5536	5536	5536	5535	5535	5535	5535	5535	5536	5535	5535	2329	Db11
P1a		0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	1	0	0	6975	P1a
P1b			0	0	0	0	0	0	1	1	1	0	0	0	0	0	1	0	0	6975	P1b
P1c				0	0	0	0	0	1	1	1	0	0	0	0	0	1	0	0	6975	P1c
P2					0	0	0	0	1	1	1	0	0	0	0	0	1	0	0	6975	P2
P3a						0	0	0	1	1	1	0	0	0	0	0	1	0	0	6975	P3a
P3b							0	0	1	1	1	0	0	0	0	0	1	0	0	6975	P3b
P4a								0	1	1	1	0	0	0	0	0	1	0	0	6975	P4a
P4b									0	0	2	1	1	1	1	1	2	1	1	6976	P4b
P4c										0	2	1	1	1	1	1	2	1	1	6976	P4c
P5											0	1	1	1	1	1	2	1	1	6976	P5
P6a												0	0	0	0	0	1	0	0	6975	P6a
P6b													0	0	0	0	1	0	0	6975	P6b
P7a														0	0	0	1	0	0	6975	P7a
P7b															0	0	1	0	0	6975	P7b
P7c																0	1	0	0	6975	P7c
Env1																	0	1	1	6976	Env1
Env2																		0	0	6975	Env2
Env3																			0	6975	Env3
C																				0	C

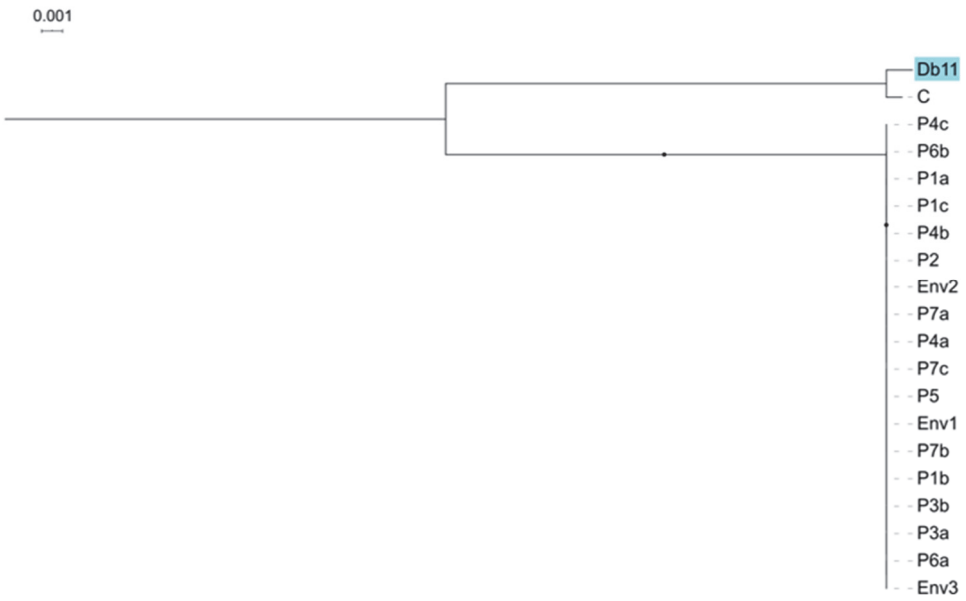
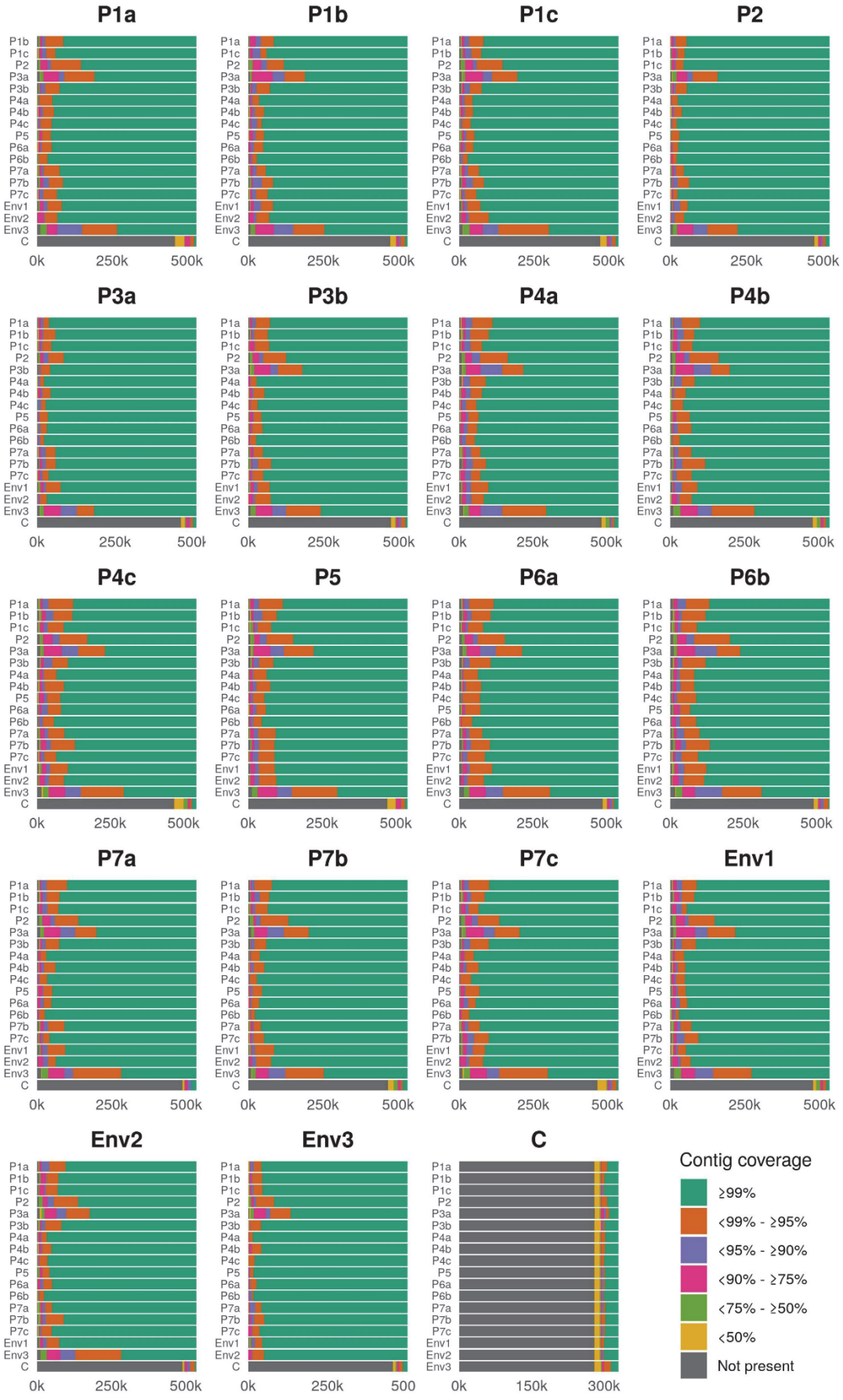


Figure 30 | Pairwise distance matrix for outbreak A when the UMH9 strain is used as reference for mapping, and the corresponding ML phylogenetic tree. The reference strain is highlighted. Black dots in branches represent a bootstrap support value equal or higher than 90%.

The average length of the accessory genome was 532,084 bp distributed in 199 contigs, on average. The accessory genome of the control isolate had 318,456 bp distributed in 124 contigs (Supplementary Table 19). There were very few differences in the accessory genome of the isolates involved in the outbreak, being the Env3 isolate the one that accumulated more differences with respect to the others (Figure 31). Consistent with the mapping results, the control isolate was the most divergent one.

No plasmids were found in any of the isolates involved in this outbreak, including the control isolate.

Figure 31 (next page) | Comparison between accessory genomes of all the isolates from outbreak A when the reference for mapping was Db11 strain. The outbreak-related isolates share the same accessory genome with minor differences, while the control isolate remains completely different.


P1a
P1b
P1c
P2
P3a
P3b
P4a
P4b
P4c
P5
P6a
P6b
P7a
P7b
P7c
Env1
Env2
Env3
C
0k
250k
500k
150k
300k
Contig coverage
≥99%
<99% - ≥95%
<95% - ≥90%
<90% - ≥75%
<75% - ≥50%
<50%
Not present

2.4. Genomic analysis of outbreak B

To check whether the UMH9 strain was a widespread clone in CV, reads from isolates of outbreak B were mapped against that genome. The average depth coverage was 45.1 X. For outbreak-related isolates, 88.6% of the reads were mapped, covering 83.0% of the UMH9 genome. These statistics were similar to those of the control isolate, with 84.1% mapped reads, covering 83.9% of the reference (Supplementary Table 20). The alignment length was 5,024,591 bp, which was reduced to 4,420,688 bp after filtering noisy regions, and contained 248,654 variant positions. There were no differences between samples from outbreak B, but the reference had 158,883 SNPs compared to the outbreak samples, revealing no genetic relationship with this outbreak. The control isolate was also very different from the outbreak isolates and the reference —171,776 and 129,295 SNPs, respectively— (Figure 32).

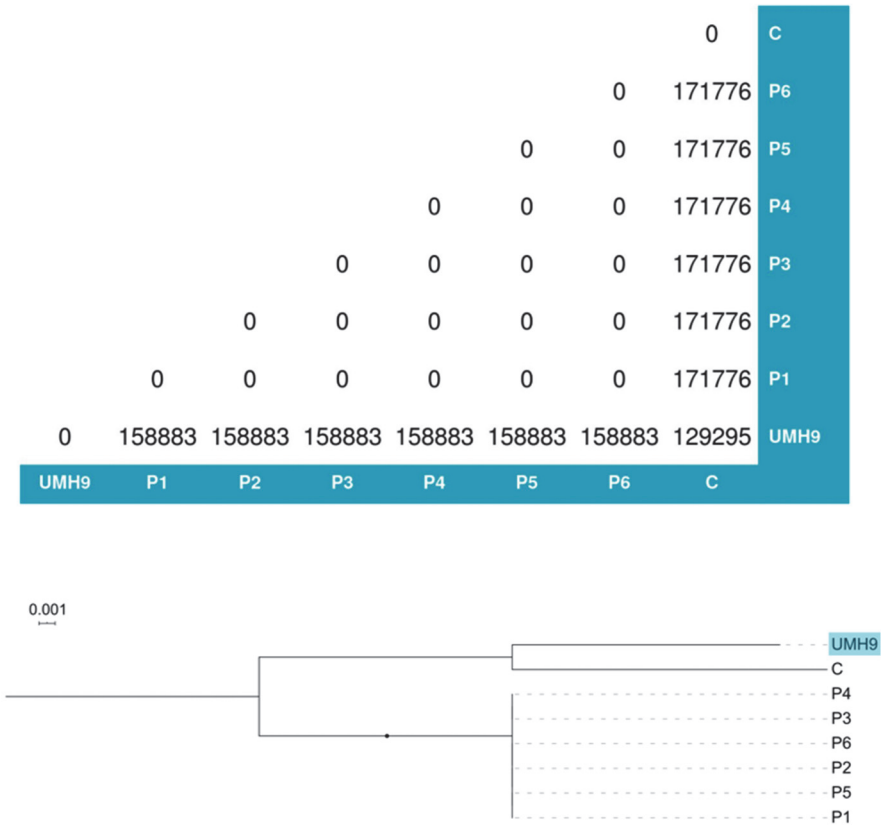


Figure 32 | Pairwise distance matrix for outbreak B when the UMH9 strain is used as reference for mapping, and the corresponding ML phylogenetic tree. The reference strain is highlighted. Black dots in branches represent a bootstrap support value equal or higher than 90%.

The accessory genome had 528,360 bp on average distributed in about 184 contigs. The accessory portion of the control isolate was longer, with 720,630 bp distributed in 205 contigs (Supplementary Table 21). Among outbreak-related isolates the accessory genome was very similar and different from that of the control isolate (Figure 33).

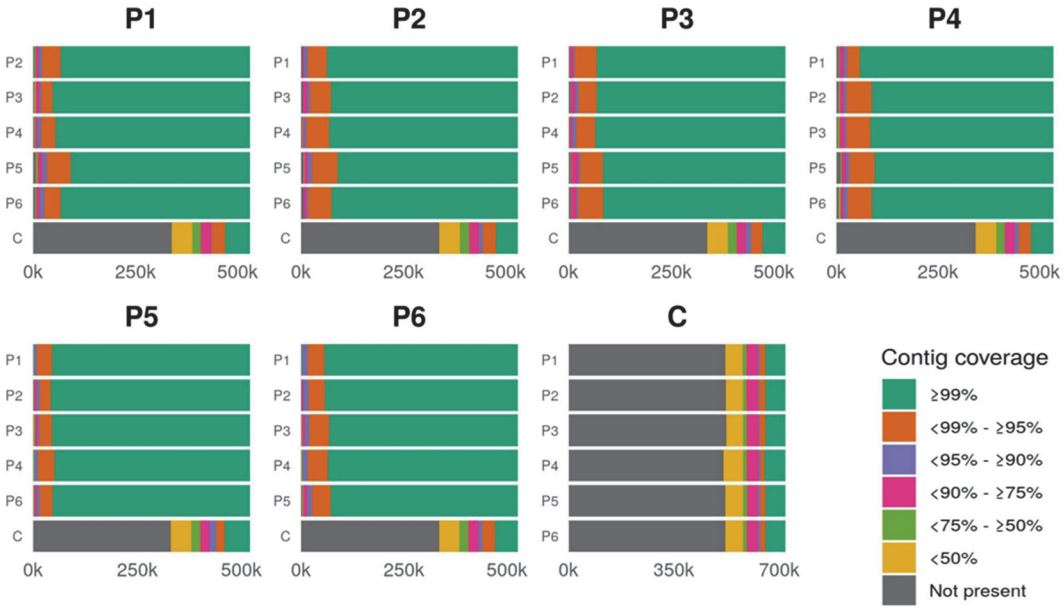


Figure 33 | Comparison between accessory genomes of all the isolates from outbreak B when the reference for mapping was UMH9 strain. The outbreak-related isolates share the same accessory genome with minor differences, while the control isolate remains completely different.

Similar results were obtained when samples from outbreak B were mapped against the Db11 reference strain. On average, the depth coverage was 45.2 X and 90.4% of the reads from strains involved in the outbreak were mapped, covering 84.9% of the reference. For the control sample, 82.9% of the reads were mapped, which covered 80.6% of the Db11 genome (Supplementary Table 22). The whole-genome alignment spanned 5,113,802 bp, which reduced to 4,544,972 bp after cleaning, containing 264,973 variant positions. Consistently with the previous results, there were no differences between samples from the outbreak, with a genetic distance of 151,705 SNPs to the

Db11 strain, and 175,335 SNPs to the control strain. The distance between the control and reference strains was 169,750 SNPs (Figure 34).

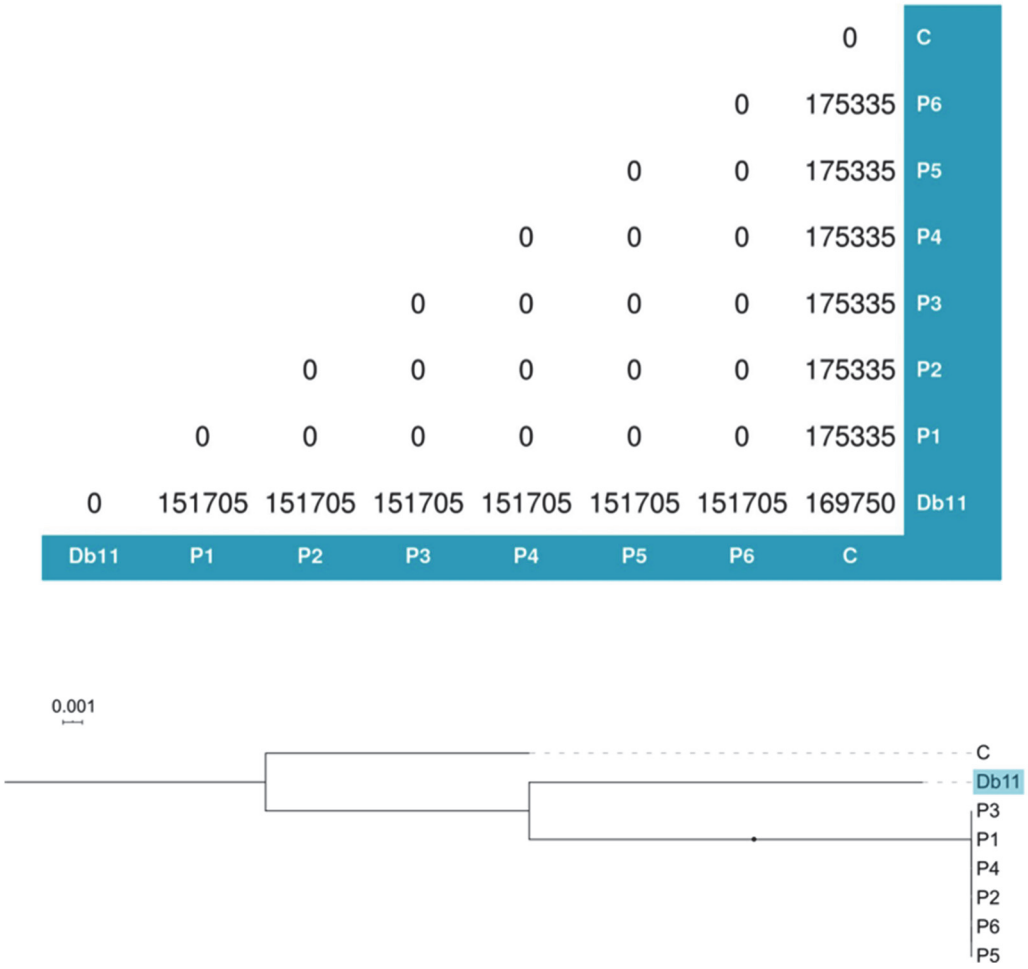


Figure 34 | Pairwise distance matrix for outbreak B when the Db11 strain is used as reference for mapping, and the corresponding ML phylogenetic tree. The reference strain is highlighted. Black dots in branches represent a bootstrap support value equal or higher than 90%.

The accessory genome of outbreak-related isolates had an average of 415,749 bp distributed in 123 contigs, and 774,488 bp distributed in 194 contigs in the case of the control isolate (Supplementary Table 23). Again, the accessory genome was very similar between outbreak-related isolates, unlike the control isolate (Figure 35).

As with outbreak A, no plasmids were found in any isolate of outbreak B, including the control isolate.

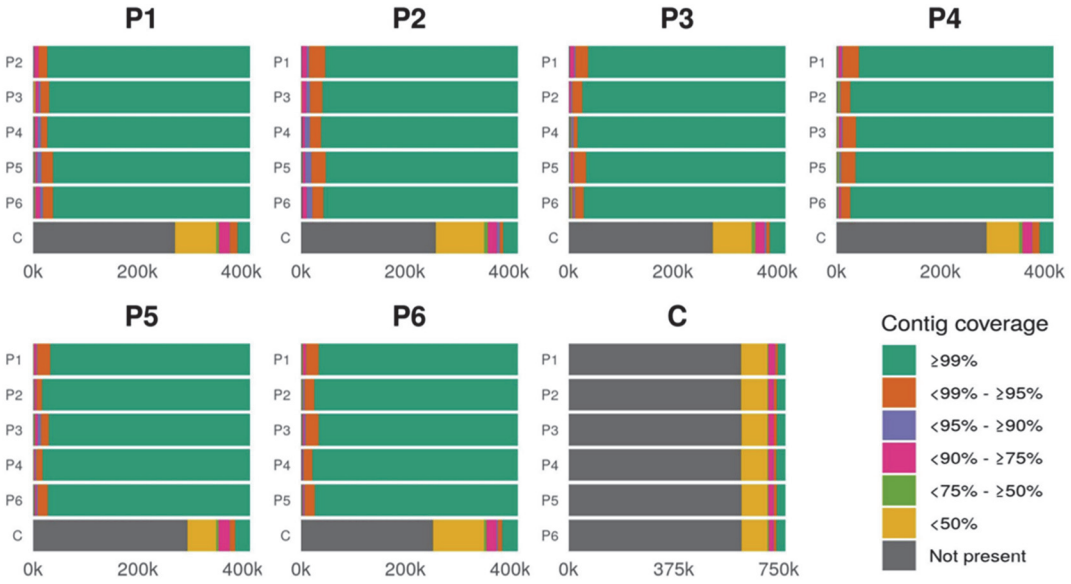


Figure 35 | Comparison between accessory genomes of all the isolates from outbreak B when the reference for mapping was Db11 strain. The outbreak-related isolates share the same accessory genome with minor differences, while the control isolate remains completely different.

2.5. Phylogenetic relationship between the two outbreaks

The joint reconstruction of a phylogenetic tree including isolates from both outbreaks and other available genomes of *S. marcescens* showed that they were genetically very distant and that there was no close relationship between them. Outbreak A isolates were very close to the UMH9 strain, as detailed previously, whereas the control isolate of this outbreak was closer to the Db11 strain. On the other hand, outbreak B isolates were close to strain UMH8, and its control isolate was closer to strain UMH5 (Figure 36). This phylogenetic reconstruction also shows that there is no clear separation between the strains of clinical origin and environmental strains.

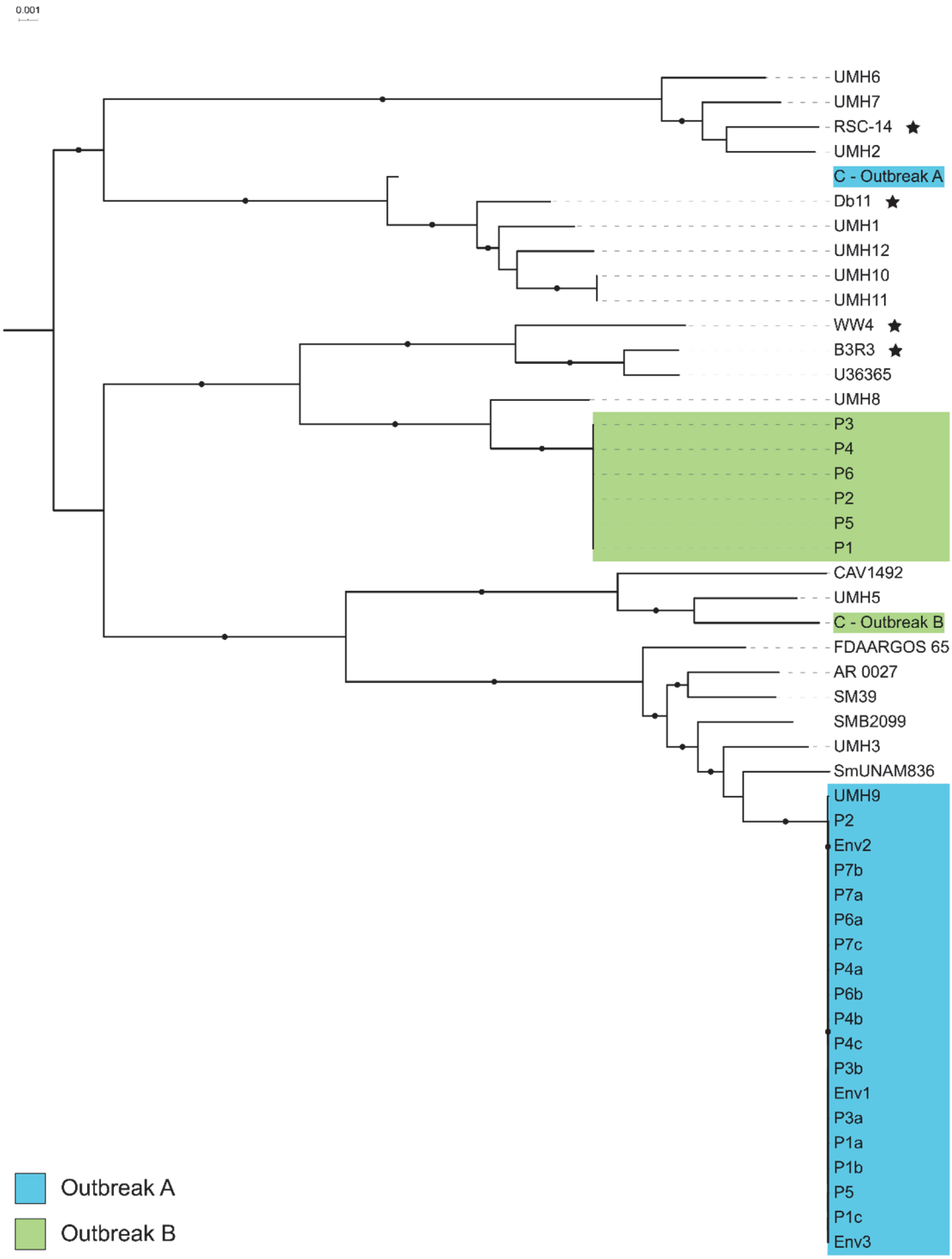


Figure 36 | ML phylogenetic tree of both outbreaks along with all *S. marcescens* genomes used in this study (see Supplementary Table 13). The black dots in branches represent bootstrap support values equal or higher than 90%. The joint reconstruction of the phylogeny of both outbreaks showed that there was no genetic relationship between them. It could also be seen that the UMH9 strain falls inside the outbreak A cluster. Black stars highlight the non-clinical strains.

3. Discussion

In this chapter, the genetic relationships between isolates of two different outbreaks of *Serratia marcescens* in two hospitals of the Comunidad Valenciana (Spain) have been analyzed through the complete genome sequencing of clinical and environmental isolates using high-throughput sequencing.

S. marcescens belongs to the ESCHAPPM group —*Enterobacter* spp*, S. marcescens, Citrobacter freundii, Hafnia* spp*, Aeromonas* spp*, Providencia* spp*, Proteus vulgaris,* and *Morganella morganii*—, which have a chromosomal inducible AmpC beta-lactamase that confers intrinsic resistance to multiple antibiotics[141,142]. However, although some cases of multidrug-resistant *S. marcescens* have been described[143,144], it seldom causes complicated outbreaks, and the usual treatment may include piperacillin-tazobactam, fluoroquinolones, aminoglycosides, or carbapenems[124].

In terms of evolutionary parameters, the nucleotide substitution rate of *S. marcescens* was estimated at around 10^{-7} substitutions per site per year, and its recombination rate was estimated between 0.01 and 0.07 events per genome per year[143]. Probably, the moderate value of the *S. marcescens* evolutionary rate in comparison with that of other bacteria[299] is related to the low-acquisition of factors that enhance its virulence and/or pathogenicity.

The use of HTS for the study of bacterial outbreaks is relatively recent[64,65], but it has proven to be more discriminative than other molecular techniques[287,288]. Despite lacking a standard for the experimental and analytical procedures and ease of use, some countries are already implementing HTS in their Public Health systems. So, studies that show both the power and limitations of this technology in a clinical context are still necessary.

In this study, we have followed the strategy of mapping the sequences against a reference genome because it has fewer sources of error than the *de novo*

assembly strategy[187]. Nevertheless, the mapping strategy was combined with the assembly of the unmapped fraction of the genome —in this context, the accessory genome— in order to obtain an as complete and accurate analysis as possible. In outbreak analyses, in which isolates are expected to have a very close genetic relationship with each other, it is not expected to find large differences in this accessory genome. Indeed, the results showed closely related isolates at the genetic level, with a maximum of 2 SNPs of difference in the mapped fraction of their genome. The small differences that could be found in the accessory genome do not affect the result, because the unmapped fraction barely represents about 10% of the whole genome of *S. marcescens*.

Nevertheless, the most interesting finding in this study stems from mapping the isolates of outbreak A against the UMH9 reference strain. An almost perfect identity between outbreak isolates and the reference strain was obtained. The similarity was such that it was impossible to separate the UMH9 strain from the outbreak isolates. This represents a serious limitation to define outbreaks on the sole basis of genomic information. Based on clinical, microbiological, and epidemiological data, the results of the genomic analyses for the samples of outbreak A confirmed that they matched the usual definition of outbreak[293]. In fact, if a genetically more distant reference genome —such as the strain Db11— had been used initially, it would have been concluded that it was a genomically clear *S. marcescens* outbreak with possible source in the NICU's sinks, which are a potential source of transmission[300–302], because the environmental isolates from these were very closely related to the clinical isolates from the outbreak.

After the detection of a second outbreak with similar characteristics but in a hospital located in another city of CV, we were interested in checking whether the UMH9 strain was spreading throughout this Spanish region. This was not the case but, in light of the identity of the complete genome sequences from outbreak A and that of UMH9 and in the absence of genomic information from other *S. marcescens* outbreaks, we cannot reject the possibility that UMH9 is a clone spreading throughout the world.

Other similar studies[303,304] opted for *de novo* assembly strategy and construction of their own core-genome (cg)MLST schemes, despite having a lower resolution than using whole-genome SNPs[305]. However, another study[306] opted for mapping against the UMH9 strain although their isolates were not as closely related to this strain as the isolates from outbreak A, but they found differences of 0-5 SNPs among their isolates. All these studies involved *S. marcescens* outbreaks in NICUs except one[303], and all concluded that HTS is a powerful tool to identify clonality among outbreak isolates.

In summary, although the impact of choosing different reference sequences in epidemiological analyses has already been discussed[307], this case dramatically illustrates how relevant this impact can be, even affecting the definition of an outbreak. HTS is a very powerful tool for the clinical setting and it will be increasingly easier and cheaper, allowing its implementation on a regular basis. However, this study highlights that, like other methodologies, it also has limitations, being always necessary to use not only the genomic information, but also the information based on microbiological and epidemiological data.

CHAPTER 4

A PANGENOME APPROACH TO DETECT HORIZONTAL GENE TRANFER IN A *Lactococcus garvieae* STRAIN

1. Results

1.1. Main features of L. garvieae *Lg-Granada*

The complete genome of Lg-Granada was obtained by HTS using PacBio long-read technology. Two contigs of 2,099,060 bp and 50,557 bp were obtained, corresponding to the chromosome and the plasmid (pGL50), respectively. The genome contains a total of 2,167 CDS —2,101 in the chromosome and 66 in the plasmid—, and 81 structural RNAs —16 rRNAs and 65 tRNAs—.

When compared with other strains of the species used in this study, Lg-Granada has 67 unique genes, some of which encode phage proteins and others are involved in metabolic and transport processes (Supplementary Table 24). If we separate the genes that Lg-Granada shares exclusively with the other strains according to their isolation source, we can see that it shares 34 genes with one or more strains of animal origin —some phage proteins, or proteins with catalytic activity—, 21 genes with strains isolated from food —some proteins involved in metabolism, but also plasmid proteins involved in pathogenesis and defense against other bacteria—, 2 genes of unknown function with the strain isolated from the soil, and 1 gene of unknown function with the strains of human origin (Supplementary Table 25).

1.2. Phylogenetics, pangenome, and core genome of L. garvieae

The phylogenetic reconstruction of the 24 strains of *L. garvieae* showed that they group in 4 clusters (Figure 37, Supplementary Table 26). Lg-Granada strain is very close to strains Tac2, UBA5784, UBA11300, and IPLA 31405 (ANI values above 99%; Supplementary Table 27). Curiously, the Lg-Granada strain is in a different cluster than the other strains isolated from human infections, 21881 and Lg-ilsanpaik-gs201105 (ANI of 98.71%), with which it shares an ANI of around 94%. Strains A1 and DCC43 are very divergent, clearly distant from the rest, with which they share an ANI of no more than 82.03% with respect to the others. In fact, the average ANI value for the 24 strains is 93.14% (ranging from 80.87 to 99.99%) but, after removal of these

two strains, it rises to 95.6% (ranging from 90.67 to 99.99%). These values suggest that these strains may be a subspecies within *L. garvieae* or even a novel species of *Lactococcus* very close to *L. garvieae*.

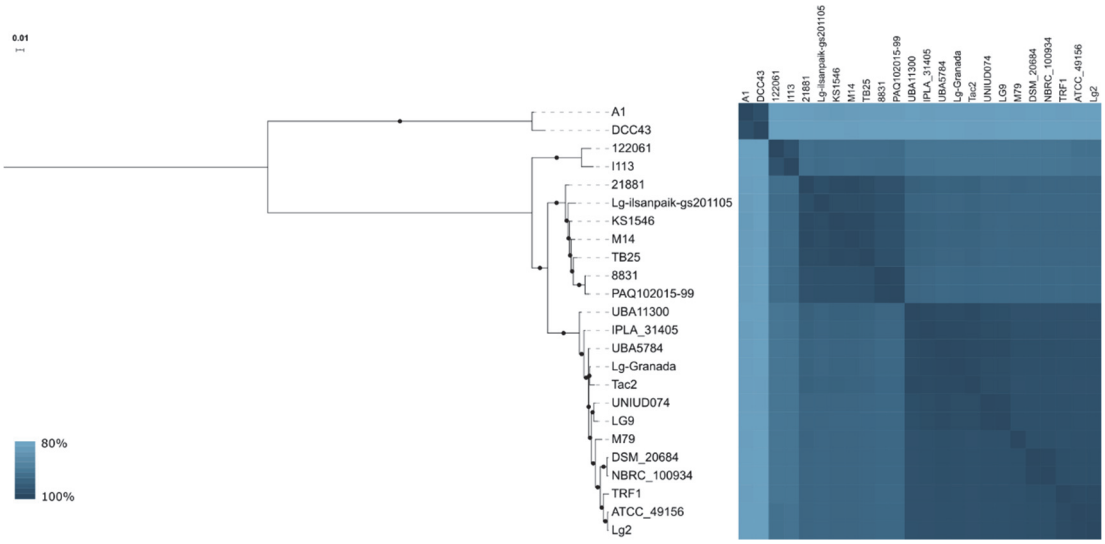


Figure 37 | Phylogenetic tree of *L. garvieae* strains used in this study. Black dots in branches represent a supporting bootstrap value equal or higher than 90%. The associated heatmap shows the pairwise ANI values.

The average number of coding sequences (CDS) in the *L. garvieae* genomes is 1,975 (ranging from 1,781 to 2,167; Supplementary Table 26). The strict core genome has 1,157 genes, spanning 1,098,087 bp of which 277,438 are variant positions. That is, the 24 strains used in this study share 58.6% of the average number of CDS, covering approximately half the length of the *L. garvieae* genome. These values increased when the core was relaxed to include genes shared by at least 80% of the genomes (19 strains or more). In this case, the relaxed core genome included 1,556 shared genes, spanning 1,488,588 bp with 391,094 variant positions.

The pangenome of *L. garvieae* reaches 5,031 genes. This number is based on these 24 genomes, but it would probably be higher if more genomes were available, as suggested by the rarefaction curve shown in Figure 38. In contrast, the core genome appears to reach a plateau from the 23 genomes. Regarding gene frequencies (Figure 39), unique genes, i.e. the genes present

in a single genome, represent the 34% of the pangenome (1,708 genes), an amount immediately followed by the strict core (23 % of the pangenome).

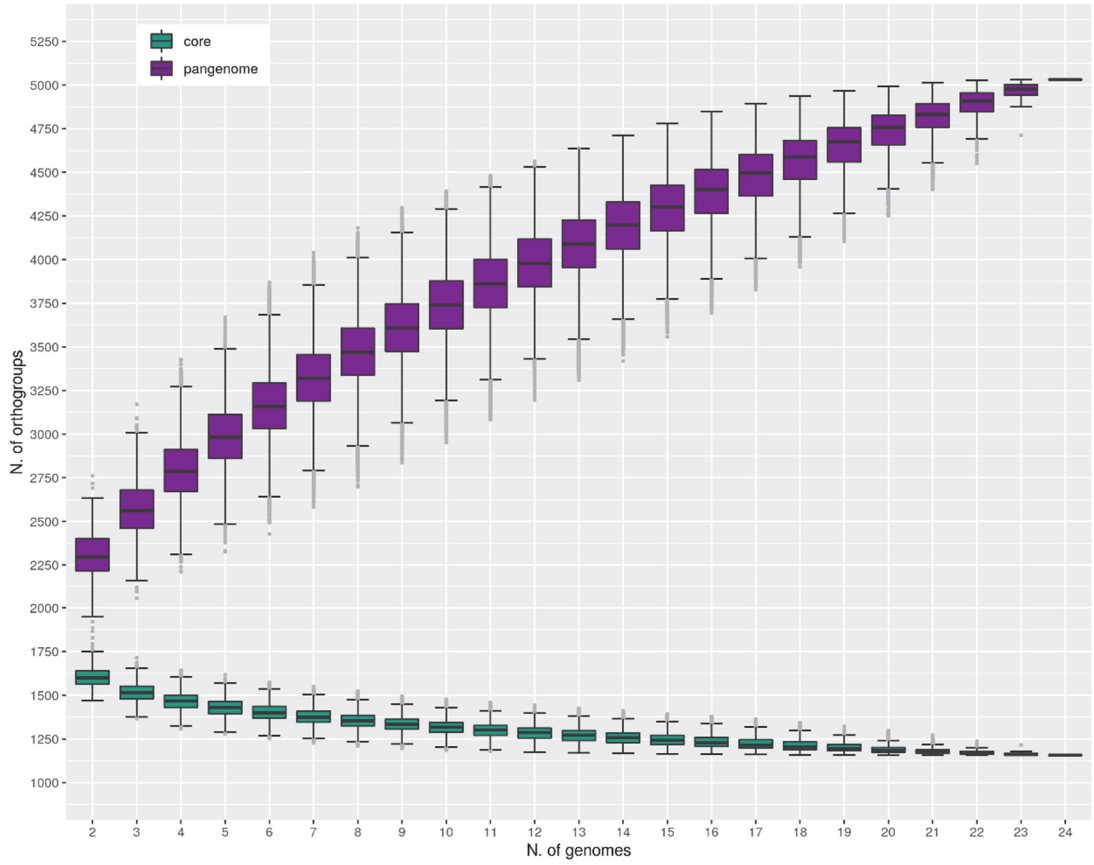


Figure 38 | Rarefaction curve for the total number of genes (purple) and the genes in the strict core (green) given a number of genomes of *L. garvieae* used.

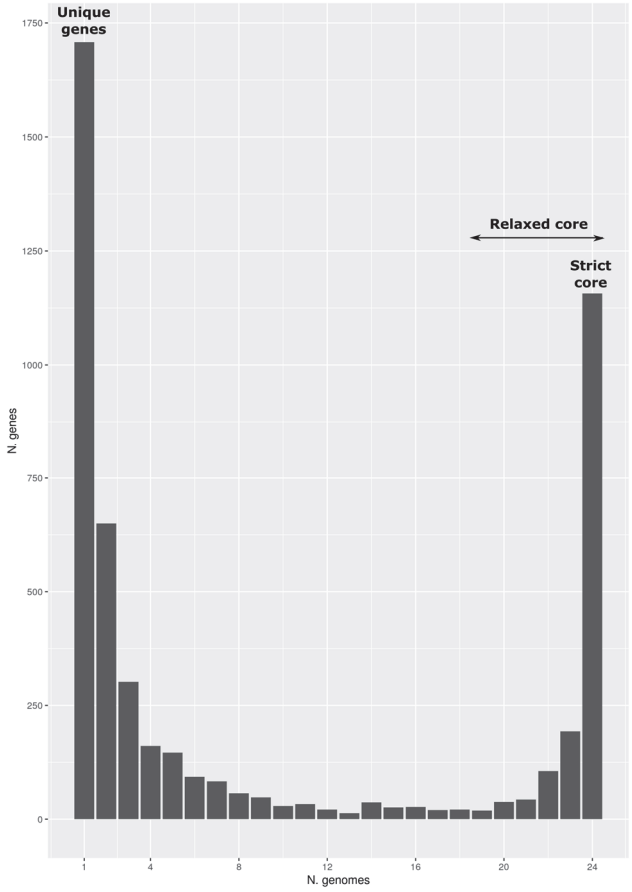


Figure 39 | Frequency of genes within the 24 genomes of *L. garvieae* included in this analysis. Unique genes and both strict and relaxed cores are marked.

1.3. Intraspecies recombination in L. garvieae

A total of 739 genes from the relaxed core passed the LM test, although 698 of them also had an adequate proportion of informative sites to be subjected to the topological congruence tests. Finally, 592 genes were statistically significant in rejecting the reference tree topology for both SH and ELW tests, after applying the FDR correction. Thus, they were identified as recombinant (Supplementary Table 28). Most of the recombinant genes corresponded to events encompassing only one gene, but there 57 genes were detected in recombination events spanning 2 genes, 13 spanned 3 genes, 12 spanned 4 genes, 3 spanned 5 genes, and the largest event encompassed 10 genes (Supplementary Table 29).

The comparison of the functions of recombinant genes to those in the Lg-Granada strain chromosome revealed that intraspecific recombinant genes are enriched in membrane and cytoplasmic proteins, with catalytic and binding activity, which are involved in metabolic and transport processes. There is also an enrichment of recombinant genes involved in response to external stimuli, cell division, and pathogenesis (Figure 40).

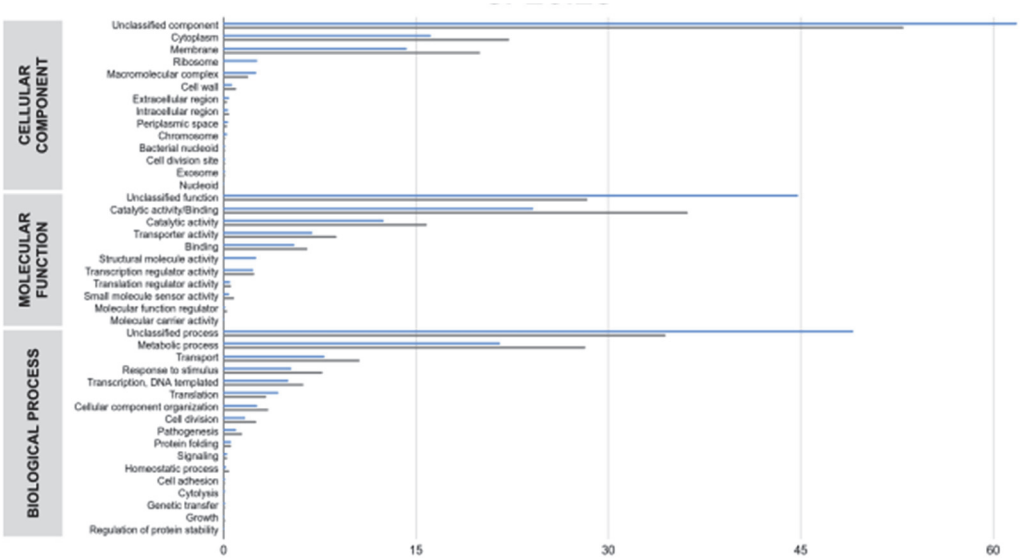


Figure 40 | Proportion of genes of Lg-Granada strain chromosome (blue) and recombinant genes (grey) at species level, classified by GO terms.

1.4. Recombination between L. garvieae *and other species of the* Lactococcus *genus*

The Lg-Granada strain was compared with 6 other *Lactococcus* genomes —which include 3 other species, *L. lactis, L. raffinolactis*, and *L. piscium*— (Supplementary Table 30). The phylogenetic reconstruction shows that *L. garvieae* shares an ancestor with *L. lactis*, being *L. piscium* and *L. raffinolactis* more distant (Figure 41). The ANI values ranged between 72.40% and 99.48%, with an average value of 78.74% (median of 74.40%). The highest values (87.14-99.48%) corresponded to comparisons between the subspecies of *L. lactis*, while the rest of the relationships identity ranged 72.40-77.12% (Supplementary Table 31).

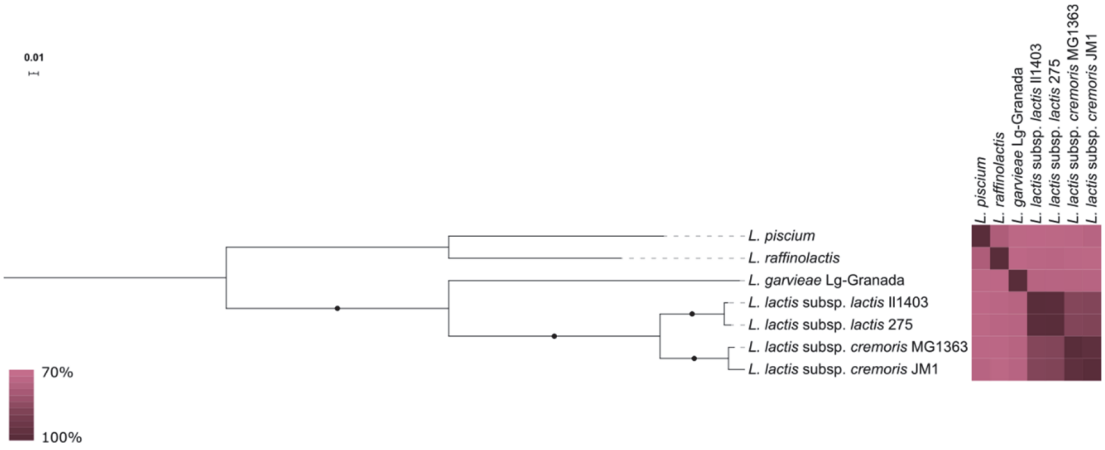


Figure 41 | Phylogenetic tree of *Lactococcus* strains used in this study. Black dots in branches represent a supporting bootstrap value equal or higher than 90%. The associated heatmap shows the pairwise ANI values.

The strict core genome of the genus has 924 genes, spanning 927,996 bp with 468,861 variant positions. The relaxed core (5 strains or more) encompassed 1,462 genes, spanning 1,434,816 bp with 723,149 variant positions. 103 orthologous genes (OGs) were eligible for topological congruence testing 97 of which were recombinant (Supplementary Table 32), but only 27 of them implied transfers between *L. garvieae* and the other species. Two recombination events encompassing 2 genes and one encompassing 4 genes were detected (Supplementary Table 33). The remaining recombination events included only one gene each.

L. garvieae is balanced in the number of genes it donates to and receives from other species in the genus (Figure 42). The main donor to *L. garviae* is an external, yet unidentified species —*Lactococcus* spp.—, followed by the different subspecies of *L. lactis*. This does not occur in the other species of *Lactococcus*, which receive more genes than they donate. The largest flow of gene movements occurs among the *L. lactis* subspecies, being the main donors among them and to the other species of the genus. *Lactococcus* spp. is the main donor to *L. lactis*, however there is no donation from *Lactococcus* spp. to *L. piscium* and *L. raffinolactis*.

Figure 42 | Summary of the gene movements between the species of *Lactococcus* used in this study.

Again, when compared with the distribution of gene functions in the Lg-Granada strain chromosome, interspecific recombinant genes at genus level mostly encode membrane and macromolecular complex proteins — cytoplasmic proteins are also an important fraction, although they are not comparatively different from the content of the Lg-Granada chromosome—. The functional categories enriched among the recombinant genes, compared to Lg-Granada, are found in genes involved in transport, transcription regulation, and cell division. The proportion of genes involved in homeostatic process is also remarkable (Figure 43).

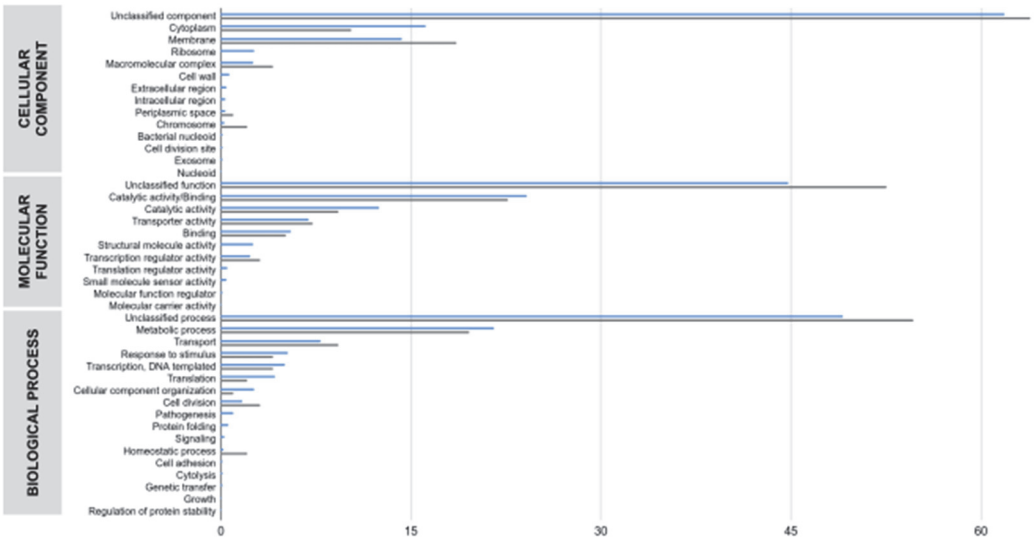


Figure 43 | Proportion of genes of Lg-Granada strain chromosome (blue) and recombinant genes (grey) at genus level, classified by GO terms.

1.5. Recombination between L. garvieae *and other species of the* Bacilli *class*

The Lg-Granada genome was compared with those of 19 additional species from the class *Bacilli*, grouped in 8 genera of the *Lactobacillales* order —*Aerococcoccus, Carnobacterium, Enterococcus, Tetragenococcoccus, Vagococcus, Lactobacillus, Leuconostoc*, and *Streptococcus*— and 2 genera of the *Bacillales* order —*Oceanobacillus* and *Listeria*— (Supplementary Table 34). The phylogenetic tree shows a close relationship between *L. garvieae* and *Streptococcus*, and they share a common ancestor with other species, some of clinical relevance, such as *E. faecalis, E. faecium*, or *L. monocytogenes* (Figure 44). The ANI values ranged 70.61-86.65%, with an average value of 72.34% (median of 71.92%). The highest values (82.81-86.65%) corresponded to the comparisons between the Listeria species, while the remaining pairwise comparisons ranged between 70.61 and 77.15%, very similar to the values found between *Lactococcus* species (Supplementary Table 35).

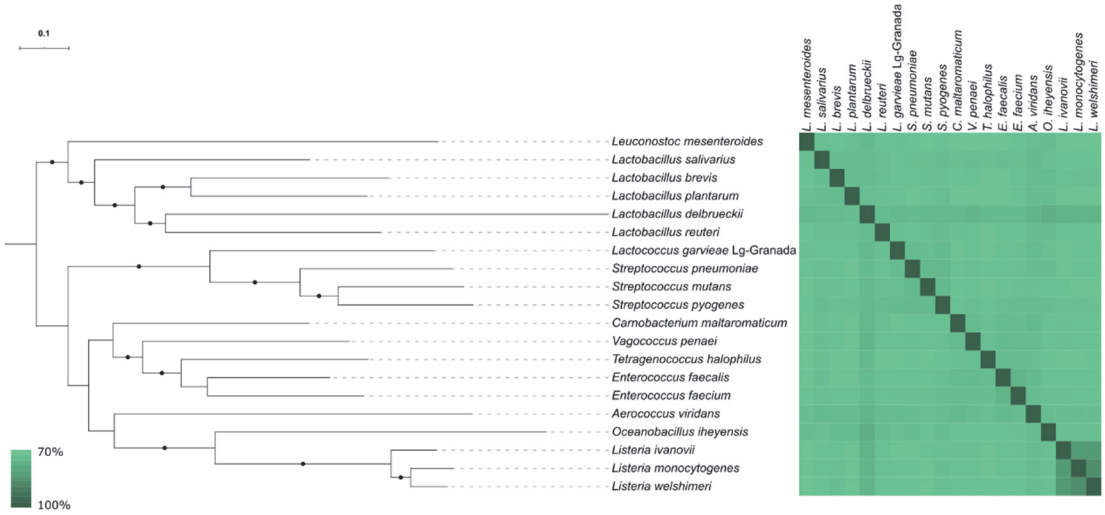


Figure 44 | Phylogenetic tree of Bacilli strains used in this study. Black dots in branches represent a supporting bootstrap value equal or higher than 90%. The associated heatmap shows the pairwise ANI values.

The strict core genome of the *Bacilli* class had 409 genes, spanning 466,956 bp with 344,201 variant positions. The relaxed core (16 strains or more) encompassed 775 genes, spanning 881,175 bp with 661,535 variant positions. 144 OGs were eligible for topological congruence testing of which 135 were recombinant (Supplementary Table 36), but only 34 of them implied movement between *L. garvieae* and another species. All the recombinant OGs were included in recombination events involving one single gene.

L. garvieae is less a donor than a recipient of genes to and from other genera of *Bacilli* (Figure 45). The main recipient from *L. garvieae* genes was *Lactobacillus*. The main donor of genes to *L. garvieae* was *Streptococcus*, followed by *Aerococcus* and *Enterococcus*. Genera not included in the analysis —marked as external, and probably included in the *Bacilli* class— were the main donors to *Streptococcus, Vagococcus, Enterococcus, Lactobacillus*, and *Listeria*.

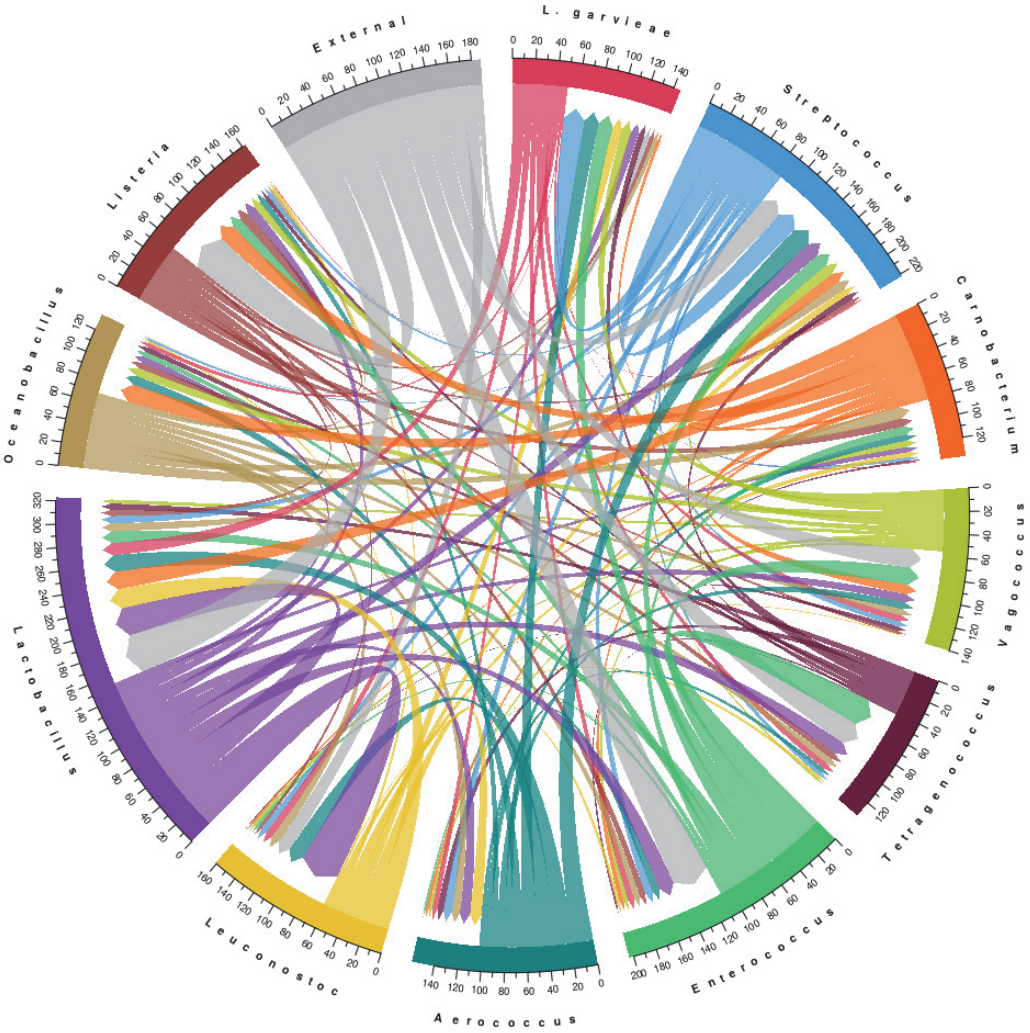


Figure 45 | Summary of the gene movements between the genus of *Bacilli* class used in this study.

In this case, when compared to the functional classes of genes in the Lg-Granada chromosome, the recombinant genes have a higher proportion of genes that encode cytoplasm, membrane and macromolecular complex proteins, but also ribosomal, chromosomal, and extracellular proteins. Most of these proteins have catalytic and binding activity, transporters, structural activity, or translation regulation, and they are involved in metabolic processes, transport, translation, response to stimulus, cellular organization, cell division, protein folding, homeostatic processes, and growth (Figure 46).

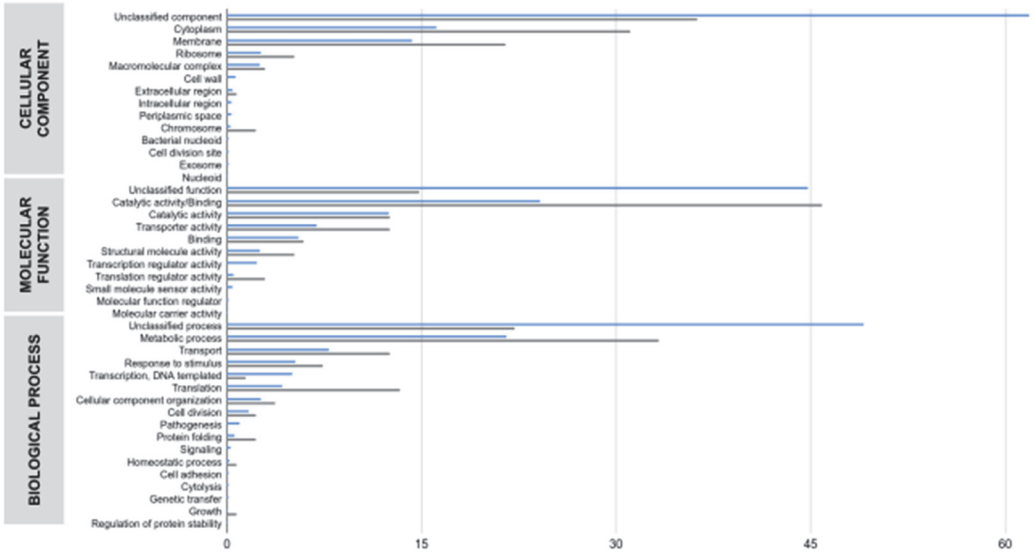


Figure 46 | Proportion of genes of Lg-Granada strain chromosome (blue) and recombinant genes (grey) at class level, classified by GO terms.

2. Discussion

Lactococcus garvieae is an ubiquitous bacterium that is the causative agent of lactococcosis in freshwater and marine fish, causing significant economic losses in aquaculture[154]. It has also been isolated from infections in mammals, such as cattle or pigs[146,147,150]. However, in the last few years this pathogen has been isolated in an increasing number of human infections[157,163], and is currently considered as an emerging zoonotic agent.

In this study, the genome of a new strain of *L. garvieae* —Lg-Granada— was completely sequenced and closed by PacBio sequencing technology, being the first clinical human strain of this species with such level of genomic completeness. We have estimated the core genome of *L. garvieae* at 1,157 genes, a value lower than that obtained by Ferrario and associates[308], which was estimated at 1,341 genes. However, our result is consistent with theirs if we take into account that this new study has used twice as many *L. garvieae* genomes. In fact, if we look at the rarefaction curve of the core genome (Figure 38), we can see that the value for 12 genomes is around 1,300 genes. We also

calculated the core genome of the *Lactococcus* genus at 924 genes, which is also consistent with the 949-genes core calculated in the Ferrario and associates study. Another recent study[309] estimated the core genome of the species at 1,850 genes based on only 8 genomes of *L. garvieae*. This value is not consistent with our results —around 1,350 genes for 8 genomes—, which may be explained by the use of less strict parameters in the orthologous identification step —e.g. lower percentages of identity and similarity—.

We found 67 genes exclusive of the Lg-Granada strain, but none of them encodes a protein with potential role in pathogenesis or virulence. Interestingly, it only shares one gene exclusively with the other human clinical strains, but its function is unknown. On the other hand, Lg-Granada shares more genes with strains of animal and food origin. It is very relevant that there are genes that encode proteins involved in pathogenicity and defense against other bacteria in food-borne strains. This may suggest that these strains acquired these genes from other bacteria in the food microbiota.

The study of recombination in bacteria has contributed to the understanding of their evolution, adaptability, and pathogenesis[310]. Detection of recombination could facilitate the identification of regions of interest in pathogen genomes[311]. However, despite its important role in the acquisition of virulence and pathogenic genes, recombination has not been studied in detail in *L. garvieae*. In this work, we used the Lg-Granada strain to explore the intra- and interspecific horizontal gene transfer to and from different taxa.

The results at the intraspecific level show that *L. garvieae* is a highly recombinogenic species. Of the 1,556 genes in the relaxed core genome, 592 genes are being transferred horizontally, representing approximately 38% of this relaxed core. If the average number of CDS in the genome of the species is 1,975, this means that almost 30% would be being transferred horizontally. A study similar to ours, focused on recombination in *Streptococcus* —a genus genetically very close to *Lactococcus*—, revealed that 35% of the *S. pyogenes* genome is recombinant, a value similar to that obtained for *L. garvieae*[312].

Of the 1,462 genes in the relaxed core genome of the genus *Lactococcus*, 97 genes were detected as recombinant, representing 6.6% of this core, a rate that indicates low recombination between species of the genus. Specifically for *L. garvieae*, 27 genes (1.8% of the relaxed core) were recombinant with other *Lactococcus* species. However, recombination between *Bacilli* species was superior to that at the genus level. Of the 775 genes in the relaxed core genome of the *Bacilli* class, 135 genes (17.4% of this core genome) were detected as recombinant. Specifically for *L. garvieae*, 34 genes (4.4% of the relaxed core) were recombinant with other *Bacilli* species. This indicates that *L. garvieae* has experienced more horizontal gene exchanges with species of the *Bacilli* than with closer species of its same genus. In general, at the interspecies level —both genus and class levels—, *L. garvieae* tends to import more genes than to export them (see Supplementary Tables 37 and 38 for further details).

Many of the genes detected as recombinant encode membrane proteins, many of them with transporter function. This has also been seen in other studies, both in Gram-positive and Gram-negative bacteria[312,313]. Many of these proteins will be on the cell surface, so it is very likely that they are involved in the interaction of *L. garvieae* with the environment —either the environment or a host—, and this medium exerts a selective pressure that favors both genetic exchange and fixation of these genes in the species.

In summary, recombination is the main system of bacteria to acquire virulence and pathogenesis genes, so it is important to analyze the ability of every potential pathogenic bacterium for catching DNA from their environment. Here, we have analyzed for first time the recombination events in the emerging pathogen *Lactococcus garvieae.* Knowing the potential of this bacterium for acquire and exchange big amounts of genes, we should control its presence in the environments where it could reach humans, such as the food-production animals and vegetables. Additionally, we calculated the ANI values between the different strains of this species and we found that there are pairs of strains with ANI values under 90%, far from the standard values established for consider two individuals as members of the same species[314], so it would be advisable to carry out a thorough review of the taxonomy of

this species, but that is something that is far from the objectives of this Chapter.

DISCUSSION

High-throughput sequencing (HTS) technology is being increasingly applied to the epidemiological investigation of pathogens. From the first applications in this field a decade ago[64,315] and with the maturation of massive sequencing technologies, more and more epidemiological and microbiological studies applying this technology are appearing[65,72,202,255,262,303]. We are thus witnessing the transition from molecular epidemiology to genomic epidemiology[316]. This thesis has tried to show, through four different studies, how HTS can provide useful insights for bacterial pathogens of clinical relevance.

In Chapter 1, we have used this technology to delve into the genomic epidemiology of *Neisseria gonorrhoeae*, a pathogen of high relevance to Public Health, given its elevated and increasing incidence worldwide, and its ability to acquire resistance against all the antibiotics used for the treatment of the infections it causes. These two factors highlight the importance of surveillance and monitoring of both the transmission dynamics of gonococci in human populations, and the accumulation of mutations in its genome that lead to the acquisition of antimicrobial resistance.

Our work has benefitted from the possibilities offered by the sequencing of complete genomes using HTS for the study of the population structure of gonococcus in Spain. Thanks to its high resolution, we were able to analyze how recombinogenic this bacterium is and the level of genetic admixture it possesses. Furthermore, we study the dynamics of the predominant sequence types (STs) during the period of study, being able to observe a substitution of some of the initial STs for others over time. However, the most interesting part of the study from a clinical perspective was the possibility of detecting mutations known to confer resistance to the antimicrobial agents used for its treatment. We also observed discrepancies with respect to the phenotype, whose monitoring is important to take precautions in the administration of antibiotics, as it could favor the appearance of more mutations that could lead to the development of the AMR phenotype[272,273]. There are similar works studying the genomic epidemiology of gonococci in many other countries[258–261,317], even worldwide[244]. However, the study presented in Chapter 1 represents the first one at genomic scale carried out in our country.

Another potential of HTS that is exploited in this thesis is its high resolution to detect clonality between isolates in cases of transmission. Chapters 2 and 3 make use of this feature. In Chapter 2, we applied this potential to solve a case of *Neisseria gonorrhoeae* transmission in a forensic context. Previous molecular investigations did not achieve enough resolution to distinguish between isolates from suspect and victim, and from one of the controls. Firstly, the typing by multilocus sequence typing (MLST) for all the isolates —including the controls— was performed, resulting in all isolates had the same ST. Then a pulsed-field gel electrophoresis (PFGE) was performed, a technique that allowed the differentiation between the isolates involved in the case and the controls, except for one, which was indistinguishable from the isolates of the case. In light of these results, we proceeded to sequence the complete genomes of the case isolates and the controls by HTS, and analyzed the differences between them. In this occasion we could see that, while the victim and suspect isolates were identical to each other, the control did differ with respect to them, although in only 2 SNPs, which suggests that the suspect and the patient from whom the control strain was extracted probably contracted infection from the same source and in a time interval close to each other. As of the date of publication of these results, we did not find any study that applied the sequencing of bacterial whole-genomes in a forensic context, so the methods and results obtained in our study could serve as a basis for the implementation of this technology in future forensic analyses that involve transmission of infectious agents.

We used the same methodology to analyze the genomic relationships between isolates of two nosocomial outbreaks of *Serratia marcescens* in Chapter 3. In order to obtain the maximum number of mapped reads as possible, we selected the closest reference genome available at the NCBI database. In this way, we wanted to ensure that we have mapped the maximum genome fraction as possible to compare the isolates with each other —although the unmapped fraction was also analyzed by assembly—. Surprisingly, we found a major limitation of our method: the reference genome —unrelated to the outbreak isolates being analyzed— was identical to the genomes of the isolates from the outbreak. With this result, we could not conclude that the isolates belonged to an epidemic outbreak. However, changing the reference

genome for a more distant one, allowed us to conclude that the isolates belonged to the same outbreak, as suggested from the epidemiological links and information. The results obtained in this chapter are a clear example of the effect that the choice of a reference genome can have on the results obtained in the analyses.

The impact of reference choice has been discussed in other studies. For example, a study by Lee & Behr 2016[307] about the genomic epidemiology of tuberculosis analyzed the effect of using different references on the phylogenies obtained, some of which lost completely their resolution, concluding that there exists a threshold in the mapping coverage of the reference genome beyond which the transmission of the pathogen cannot be discriminated. Another recent study from our group[318] also explores the impact of the election of a reference genome on subsequent analyses after the mapping step, which affects the results of the variant calling step, the estimation of recombination rates and *dN/dS* ratios, and the phylogenetic reconstruction.

These facts show that, although the sequencing of whole genomes by HTS is very useful for epidemiological and microbiological analyses, it must be taken into account that it is another molecular technique, such as PCR or other molecular biology techniques, and as such it provides the genetic information of the analyzed samples, so it is essential that this information is complemented with the information obtained from other microbiological and epidemiological analyses. If we had found this problem in the forensic case exposed in Chapter 2, we would not have been able to justify in any way the reliability of our conclusions. For this reason, it is crucial —in addition to the genomic information obtained— to put other factors into context, such as the origin of the samples and the references used.

Finally, in Chapter 4, we have analyzed the recombination events occurring between *Lactococcus garvieae* strains, and between *L. garvieae* and other species belonging to the same genus and to the same class. This work, derived from a collaboration with Dr. Alicia Gibello, from the Universidad Complutense de Madrid, started from a *L. garvieae* strain sequenced using long-reads HTS technology and it allowed us to detect the high number of intraspecific recombination events in this emerging pathogen. We were also

able to explore the recombination events present at the interspecific level. The analysis of recombination events is of interest in infectious pathogens because, in many cases, it is the origin of the acquisition of virulence and pathogenicity genes[319]. The long-reads HTS is not yet sufficiently developed for its implementation and its cost is still high compared to the short-reads HTS. However, it shows great potential for the evolution of genomic analyses, especially in the *de novo* assembly of genomes, since it solves shortcomings in the use of short reads, such as in assembling regions of complex variability or repetitive elements[320], and it has a great potential in AMR surveillance through the assembly of plasmids, which carry most of the AMR determinants in many bacterial pathogens[321].

In summary, the application of HTS technologies is well-established in research environments, especially —but not limited to— in microbial genomics studies, as evidenced by the huge and increasing number of studies that apply it to their analyses. However, its implementation in clinical settings and Public Health systems is still far from being generalized. Research works, as those presented in this thesis, evidence the enormous potential that this technology offers for epidemiological studies and the surveillance of health threats, such as the emergence of antimicrobial resistance. However, lack of standardization and the difficulty of the analyses derived from sequencing projects, which require the incorporation of experienced staff and training in clinical settings, is an obstacle that has to be solved before its implementation in these settings. The development of tools that facilitate bioinformatics analyses and the results interpretation will be one of the key factors to achieve this implementation.

CONCLUSIONS

1. The application of high-throughput sequencing (HTS) have a great potential in epidemiological studies.
2. Thanks to the implementation of this tool, we have been able to analyze the structure of gonococcal population in Spain and thus verifying the high degree of admixture present in the sampled isolates.
3. The availability of complete genomic sequences obtained by HTS facilitated the detection of key mutations in certain genes that can lead to the development of antimicrobial resistance (AMR). In this way, some of these mutations were detected even in gonococcal isolates whose phenotype was susceptible. This information could be interesting in clinical settings to develop appropriate treatment protocols that do not favor the accumulation of mutations in those genes and the consequent emergence of resistant phenotypes.
4. HTS offers higher resolution than any other molecular technology. This evidences the usefulness of this technique in the detection of closely related isolates.
5. Based on the previous conclusion, HTS is very useful in transmission investigations in forensic contexts. We were able to discriminate between gonococcal isolates from suspect and victim of an alleged case of abuse of a minor and the controls, when other molecular techniques failed.
6. Outbreak analyses based on reference mapping are limited in the choice of that reference. If we choose a reference genome that is too close to the isolates we want to analyze, we have the risk of losing the phylogenetic resolution to the point of not being able to define the outbreak itself, as observed in the analysis of *Serratia marcescens* outbreaks.
7. Long-reads-producing HTS technology has the advantage of resolving the assembly of genomes in the *de novo* assembly projects better than short-reads-producing HTS technology. This allows obtaining complete genomes, including closed chromosomes and plasmids, if any. Thanks to this technology, we were able to analyze the recombination events present in the emerging pathogen *Lactococcus garvieae* at the species level, and the subsequent extension to higher taxonomic levels.

8. In summary, HTS technologies are very useful for the analysis of pathogens of clinical relevance, and for the surveillance of AMR emergence. With the maturation of this technology in the basic research field, where obstacles that it may present must be resolved, it can be transferred to clinical settings at a more generalized level.

REFERENCES

1 Martínez JL. Bacterial pathogens: from natural ecosystems to human hosts. *Environ Microbiol* 2013; **15**: 325–33.

2 Ziebuhr W, Ohlsen K, Karch H, Korhonen T, Hacker J. Evolution of bacterial pathogenesis. *Cell Mol Life Sci* 1999; **56**: 719–28.

3 Eisenreich W, Dandekar T, Heesemann J, Goebel W. Carbon metabolism of intracellular bacterial pathogens and possible links to virulence. *Nat Rev Microbiol* 2010; **8**: 401–12.

4 Weinert LA, Welch JJ. Why might bacterial pathogens have small genomes? *Trends Ecol Evol* 2017; **32**: 936–47.

5 Georgiades K. Genomics of epidemic pathogens. *Clin Microbiol Infect* 2012; **18**: 213–7.

6 Arnold DL, Jackson RW. Bacterial genomes: evolution of pathogenicity. *Curr Opin Plant Biol* 2011; **14**: 385–91.

7 Merhej V, Georgiades K, Raoult D. Postgenomic analysis of bacterial pathogens repertoire reveals genome reduction rather than virulence factors. *Brief Funct Genomics* 2013; **12**: 291–304.

8 Brown SP, Cornforth DM, Mideo N. Evolution of virulence in opportunistic pathogens: generalism, plasticity, and control. *Trends Microbiol* 2012; **20**: 336–42.

9 Gibson B, Wilson DJ, Feil E, Eyre-Walker A. The distribution of bacterial doubling times in the wild. *Proc Biol Sci* 2018; **285**: 20180789.

10 An Wang R. Why is *Mycobacterium tuberculosis* hard to grow? The principle of biorelativity explains. *J Clin Exp Pathol* 2014; **04**: 176.

11 Kataoka N, Tokiwa Y, Tanaka Y, Takeda K, Suzuki T. Enrichment culture and isolation of slow-growing bacteria. *Appl Microbiol Biotechnol* 1996; **45**: 771–7.

12 Bliven KA, Maurelli AT. Evolution of bacterial pathogens within the human host. *Microbiol Spectr* 2016; **4**.

13 Adrian J, Bonsignore P, Hammer S, Frickey T, Hauck CR. Adaptation to host-specific bacterial pathogens drives rapid evolution of a human innate immune receptor. *Curr Biol* 2019; **29**: 616-630.e5.

14 WHO. Fact sheet: The top 10 causes of death. Geneva, 2020. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

15 Waugh M. History of Sexually Transmitted Infections. In: Gross GE, Tyring SK, eds. Sexually Transmitted Infections and Sexually Transmitted Diseases. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011: 3–11.

16 Casillas-Vega N, Morfín-Otero R, García S, Camacho-Ortiz A, Garza-González E. Causative agents, diseases, epidemiology and diagnosis of sexually transmitted infections. *Rev Med Microbiol* 2017; **28**: 9–18.

17 Rowley J, Vander Hoorn S, Korenromp E, *et al. Chlamydia*, gonorrhoea, trichomoniasis and syphilis: global prevalence and incidence estimates, 2016. *Bull World Health Organ* 2019; **97**: 548-562P.

18 WHO. Fact sheet: Sexually transmitted infections. Geneva, 2019. https://www.who.int/news-room/fact-sheets/detail/sexually-transmitted-infections-(stis)

19 Okigbo C, Eke A. Behavioural interventions to reduce the transmission of HIV infection among sex workers and their clients in low- and middle-income countries: RHL commentary (1 February 2013). The WHO Reproductive Health Library. Geneva. https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD005272.pub3/full

20 Shannon CL, Klausner JD. The growing epidemic of sexually transmitted infections in adolescents. *Curr Opin Pediatr* 2018; **30**: 137–43.

21 Estreich S, Forster GE, Robinson A. Sexually transmitted diseases in rape victims. *Sex Transm Infect* 1990; **66**: 433–8.

22 Workowski KA, Bolan GA, CDC. Sexually transmitted diseases treatment guidelines, 2015. *MMWR Recomm Reports* 2015; **64**: 1–137.

23 WHO. Report on global sexually transmitted infection surveillance 2018. Geneva, 2018. https://www.who.int/reproductivehealth/publications/stis-

surveillance-2018/en/

24 Grundmann H, Hellriegel B. Mathematical modelling: a tool for hospital infection control. *Lancet Infect Dis* 2006; **6**: 39–45.

25 ECDC. Point prevalence survey of healthcare-associated infections and antimicrobial use in European acute care hospitals. Stockholm, 2013.

26 WHO. Report on the burden of endemic health care-associated infection worldwide. Geneva, 2011. https://apps.who.int/iris/bitstream/handle/10665/80135/9789241501507_eng.pdf

27 Allegranzi B, Kilpatrick C, Storr J, Kelley E, Park BJ, Donaldson L. Global infection prevention and control priorities 2018–22: a call for action. *Lancet Glob Heal* 2017; **5**: e1178–80.

28 Institute of Medicine (US) Division of Health Promotion and Disease Prevention. Risk Factors for Infection in the Elderly. In: Berg RL, Cassells JS, eds. The Second Fifty Years: Promoting Health and Preventing Disability. Washington (DC): National Academies Press (US), 1992.

29 Ramasethu J. Prevention and treatment of neonatal nosocomial infections. *Matern Heal Neonatol Perinatol* 2017; **3**: 5.

30 Cabrera-Cancio MR. Infections and the compromised immune status in the chronically critically ill patient: Prevention strategies. *Respir Care* 2012; **57**: 979–92.

31 Haque M, Sartelli M, McKimm J, Abu Bakar M. Health care-associated infections - an overview. *Infect Drug Resist* 2018; **11**: 2321–33.

32 Creedon SA. Healthcare workers' hand decontamination practices: compliance with recommended guidelines. *J Adv Nurs* 2005; **51**: 208–16.

33 WHO. Guidelines on core components of infection prevention and control programmes at the national and acute health care facility level. Geneva, 2016. https://www.who.int/gpsc/ipc-components/en/

34 Morse SS. Factors in the emergence of infectious diseases. *Emerg Infect*

Dis 1995; **1**: 7–15.

35 Morens DM, Folkers GK, Fauci AS. The challenge of emerging and re-emerging infectious diseases. *Nature* 2004; **430**: 242–9.

36 Trent RJ. Infectious Diseases. In: Molecular Medicine, 3rd ed. Elsevier, 2005: 193–220.

37 WHO. A brief guide to emerging infectious diseases and zoonoses. Geneva, 2014. https://apps.who.int/iris/handle/10665/204722

38 WHO. Fact sheet: Antimicrobial resistance. Geneva, 2020. https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance

39 Holmes AH, Moore LSP, Sundsfjord A, *et al.* Understanding the mechanisms and drivers of antimicrobial resistance. *Lancet* 2016; **387**: 176–87.

40 Laxminarayan R, Duse A, Wattal C, *et al.* Antibiotic resistance—the need for global solutions. *Lancet Infect Dis* 2013; **13**: 1057–98.

41 WHO. The evolving threat of antimicrobial resistance: options for action. Geneva, 2012. https://apps.who.int/iris/handle/10665/44812

42 Prestinaci F, Pezzotti P, Pantosti A. Antimicrobial resistance: a global multifaceted phenomenon. *Pathog Glob Health* 2015; **109**: 309–18.

43 Landers TF, Cohen B, Wittum TE, Larson EL. A review of antibiotic use in food animals: Perspective, policy, and potential. *Public Health Rep* 2012; **127**: 4–22.

44 Djordjevic SP, Morgan BS. A One Health genomic approach to antimicrobial resistance is essential for generating relevant data for a holistic assessment of the biggest threat to public health. *Microbiol Aust* 2019; **40**: 73.

45 WHO. Global action plan on antimicrobial resistance. Geneva, 2015. https://www.who.int/antimicrobial-resistance/publications/global-action-plan/en/

46 Sanger F, Coulson AR. A rapid method for determining sequences in

DNA by primed synthesis with DNA polymerase. *J Mol Biol* 1975; **94**: 441–8.

47 Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci* 1977; **74**: 5463–7.

48 Green ED. Strategies for the systematic sequencing of complex genomes. *Nat Rev Genet* 2001; **2**: 573–83.

49 Fleischmann R, Adams M, White O, *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995; **269**: 496–512.

50 Hall N. Advanced sequencing technologies and their wider impact in microbiology. *J Exp Biol* 2007; **210**: 1518–25.

51 Quail M, Smith ME, Coupland P, *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 2012; **13**: 341.

52 Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res* 2015; **43**: e37–e37.

53 Aird D, Ross MG, Chen W-S, *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* 2011; **12**: R18.

54 Dozmorov MG, Adrianto I, Giles CB, *et al.* Detrimental effects of duplicate reads and low complexity regions on RNA- and ChIP-seq data. *BMC Bioinformatics* 2015; **16**: S10.

55 Eid J, Fehr A, Gray J, *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* 2009; **323**: 133–8.

56 Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol* 2016; **17**: 239.

57 Jain M, Koren S, Miga KH, *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* 2018; **36**: 338–45.

58 Koren S, Schatz MC, Walenz BP, *et al.* Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol* 2012; **30**: 693–700.

59 Loman NJ, Pallen MJ. Twenty years of bacterial genome sequencing. *Nat Rev Microbiol* 2015; **13**: 787–94.

60 Grant K, Jenkins C, Arnold C, Green J, Zambon M. Implementing pathogen genomics: a case study. London, 2018.

61 ECDC. Monitoring the use of whole-genome sequencing in infectious disease surveillance in Europe 2015–2017. Stockholm, 2018. https://www.ecdc.europa.eu/sites/default/files/documents/monitoring-WGS-infectious-disease-surveillance-in-Europe-2015-2017-updated-Dec-2018.pdf

62 Rossen JWA, Friedrich AW, Moran-Gilad J. Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. *Clin Microbiol Infect* 2018; **24**: 355–60.

63 Bertelli C, Greub G. Rapid bacterial genome sequencing: methods and applications in clinical microbiology. *Clin Microbiol Infect* 2013; **19**: 803–13.

64 Harris SR, Feil EJ, Holden MTG, *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 2010; **327**: 469–74.

65 Sherry NL, Porter JL, Seemann T, Watkins A, Stinear TP, Howden BP. Outbreak investigation using high-throughput genome sequencing within a diagnostic microbiology laboratory. *J Clin Microbiol* 2013; **51**: 1396–401.

66 Rasko DA, Webster DR, Sahl JW, *et al.* Origins of the *E. coli* strain causing an outbreak of hemolytic–uremic syndrome in Germany. *N Engl J Med* 2011; **365**: 709–17.

67 Chin C-S, Sorenson J, Harris JB, *et al.* The origin of the Haitian cholera outbreak strain. *N Engl J Med* 2011; **364**: 33–42.

68 Hendriksen RS, Bortolaia V, Tate H, Tyson GH, Aarestrup FM,

McDermott PF. Using genomics to track global antimicrobial resistance. *Front Public Health* 2019; **7**: 242.

69 Global Antimicrobial Resistance and Use Surveillance System (GLASS). GLASS whole-genome sequencing for surveillance of antimicrobial resistance. Geneva, 2020. https://www.who.int/publications/i/item/9789240011007

70 Grad YH, Kirkcaldy RD, Trees D, *et al.* Genomic epidemiology of *Neisseria gonorrhoeae* with reduced susceptibility to cefixime in the USA: a retrospective observational study. *Lancet Infect Dis* 2014; **14**: 220–6.

71 Price JR, Golubchik T, Cole K, *et al.* Whole-genome sequencing shows that patient-to-patient transmission rarely accounts for acquisition of *Staphylococcus aureus* in an intensive care unit. *Clin Infect Dis* 2014; **58**: 609–18.

72 Howden BP, Holt KE, Lam MMC, *et al.* Genomic insights to control the emergence of vancomycin-resistant enterococci. *MBio* 2013; **4**: e00412-3.

73 Walker TM, Kohl TA, Omar S V, *et al.* Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. *Lancet Infect Dis* 2015; **15**: 1193–202.

74 Comas I, Cancino-Muñoz I, Mariner-Llicer C, *et al.* Uso de las tecnologías de secuenciación masiva para el diagnóstico y epidemiología de enfermedades infecciosas. *Enferm Infecc Microbiol Clin* 2020; **38**: 32–8.

75 Köser CU, Ellington MJ, Cartwright EJP, *et al.* Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathog* 2012; **8**: e1002824.

76 Sparling PF. Biology of *Neisseria gonorrhoeae*. In: Holmes KK, Sparling PF, Stamm WE, *et al.*, eds. Sexually Transmitted Diseases, 4th ed. New York: McGraw-Hill Education, 2008: 607–26.

77 Bratcher HB, Bennett JS, Maiden MC. Evolutionary and genomic

insights into meningococcal biology. *Future Microbiol* 2012; **7**: 873–85.

78 NCBI Genome assembly and annotation report of *Neisseria gonorrhoeae*. https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/864/.

79 Korch C, Hagblom P, Ohman H, Göransson M, Normark S. Cryptic plasmid of *Neisseria gonorrhoeae*: complete nucleotide sequence and genetic organization. *J Bacteriol* 1985; **163**: 430–8.

80 Dillon JA, Yeung KH. Beta-lactamase plasmids and chromosomally mediated antibiotic resistance in pathogenic *Neisseria* species. *Clin Microbiol Rev* 1989; **2**: S125–33.

81 Cehovin A, Lewis SB. Mobile genetic elements in *Neisseria gonorrhoeae*: movement for change. *Pathog Dis* 2017; **75**.

82 Pagotto F, Aman A-T, Ng L-K, Yeung K-H, Brett M, Dillon J-AR. Sequence analysis of the family of penicillinase-producing plasmids of *Neisseria gonorrhoeae*. *Plasmid* 2000; **43**: 24–34.

83 Turner A, Gough KR, Leeming JP. Molecular epidemiology of *tetM* genes in *Neisseria gonorrhoeae*. *Sex Transm Infect* 1999; **75**: 60–6.

84 Lewis DA, Ison CA, Forster GE, Goh BT. Tetracycline-resistant *Neisseria gonorrhoeae*. *Sex Transm Dis* 1996; **23**: 378–83.

85 Ramsey ME, Woodhams KL, Dillard JP. The gonococcal genetic island and type IV secretion in the pathogenic *Neisseria*. *Front Microbiol* 2011; **2:** 61.

86 Woodhams KL, Benet ZL, Blonsky SE, Hackett KT, Dillard JP. Prevalence and detailed mapping of the Gonococcal Genetic Island in *Neisseria meningitidis*. *J Bacteriol* 2012; **194**: 2275–85.

87 Newman L, Rowley J, Vander Hoorn S, *et al.* Global estimates of the prevalence and incidence of four curable sexually transmitted infections in 2012 based on systematic review and global reporting. *PLoS One* 2015; **10**: e0143304.

88 Unemo M, Shafer WM. Antimicrobial resistance in *Neisseria gonorrhoeae* in the 21st century: past, evolution, and future. *Clin*

Microbiol Rev 2014; **27**: 587–613.

89 Quillin SJ, Seifert HS. *Neisseria gonorrhoeae* host adaptation and pathogenesis. *Nat Rev Microbiol* 2018; **16**: 226–40.

90 WHO. WHO guidelines for the treatment of *Neisseria gonorrhoeae*. Geneva, 2016. https://www.who.int/reproductivehealth/publications/rtis/gonorrhoea-treatment-guidelines/en/

91 Swedberg G, Fermér C, Sköld O. Point mutations in the dihydropteroate synthase gene causing sulfonamide resistance. *Adv Exp Med Biol.* 1993: 555–8.

92 Ropp PA, Hu M, Olesky M, Nicholas RA. Mutations in *ponA*, the gene encoding penicillin-binding protein 1, and a novel locus, *penC*, are required for high-level chromosomally mediated penicillin resistance in *Neisseria gonorrhoeae*. *Antimicrob Agents Chemother* 2002; **46**: 769–77.

93 Olesky M, Hobbs M, Nicholas RA. Identification and analysis of amino acid mutations in porin IB that mediate intermediate-level resistance to penicillin and tetracycline in *Neisseria gonorrhoeae*. *Antimicrob Agents Chemother* 2002; **46**: 2811–20.

94 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**: 578–87.

95 Nandi S, Swanson S, Tomberg J, Nicholas RA. Diffusion of antibiotics through the PilQ secretin in *Neisseria gonorrhoeae* occurs through the immature, sodium dodecyl sulfate-labile form. *J Bacteriol* 2015; **197**: 1308–21.

96 Muhammad I, Golparian D, Dillon J-AR, *et al.* Characterisation of blaTEM genes and types of β-lactamase plasmids in *Neisseria gonorrhoeae* – the prevalent and conserved blaTEM-135 has not recently evolved and existed in the Toronto plasmid from the origin. *BMC Infect Dis* 2014; **14**: 454.

97 Unemo M, Golparian D, Nicholas R, Ohnishi M, Gallay A, Sednaoui P. High-level cefixime- and ceftriaxone-resistant *Neisseria gonorrhoeae* in

France: novel *penA* mosaic allele in a successful international clone causes treatment failure. *Antimicrob Agents Chemother* 2012; **56**: 1273–80.

98 Rouquette-Loughlin C, Dunham SA, Kuhn M, Balthazar JT, Shafer WM. The NorM efflux pump of *Neisseria gonorrhoeae* and *Neisseria meningitidis* recognizes antimicrobial cationic compounds. *J Bacteriol* 2003; **185**: 1101–6.

99 Galarza PG, Abad R, Canigia LF, *et al.* New mutation in 23S rRNA gene associated with high level of azithromycin resistance in *Neisseria gonorrhoeae. Antimicrob Agents Chemother* 2010; **54**: 1652–3.

100 Chisholm SA, Dave J, Ison CA. High-level azithromycin resistance occurs in *Neisseria gonorrhoeae* as a result of a single point mutation in the 23S rRNA genes. *Antimicrob Agents Chemother* 2010; **54**: 3812–6.

101 Rouquette-Loughlin CE, Balthazar JT, Shafer WM. Characterization of the MacA–MacB efflux system in *Neisseria gonorrhoeae*. *J Antimicrob Chemother* 2005; **56**: 856–60.

102 Hu M, Nandi S, Davies C, Nicholas RA. High-level chromosomally mediated tetracycline resistance in *Neisseria gonorrhoeae* results from a point mutation in the *rpsJ* gene encoding ribosomal protein S10 in combination with the *mtrR* and *penB* resistance determinants. *Antimicrob Agents Chemother* 2005; **49**: 4327–34.

103 Morse SA, Johnson SR, Biddle JW, Roberts MC. High-level tetracycline resistance in *Neisseria gonorrhoeae* is result of acquisition of streptococcal *tetM* determinant. *Antimicrob Agents Chemother* 1986; **30**: 664–70.

104 Galimand M, Gerbaud G, Courvalin P. Spectinomycin resistance in *Neisseria* spp. due to mutations in 16S rRNA. *Antimicrob Agents Chemother* 2000; **44**: 1365–6.

105 Ilina EN, Malakhova M V., Bodoev IN, Oparina NY, Filimonova A V., Govorun VM. Mutation in ribosomal protein S5 leads to spectinomycin resistance in *Neisseria gonorrhoeae*. *Front Microbiol* 2013; **4**: 186.

106 Unemo M, Golparian D, Skogen V, *et al. Neisseria gonorrhoeae* strain with high-level resistance to spectinomycin due to a novel resistance mechanism (mutated ribosomal protein S5) verified in Norway. *Antimicrob Agents Chemother* 2013; **57**: 1057–61.

107 Shafer WM, Balthazar JT, Hagman KE, Morse SA. Missense mutations that alter the DNA-binding domain of the MtrR protein occur frequently in rectal isolates of *Neisseria gonorrhoeae* that are resistant to faecal lipids. *Microbiology* 1995; **141**: 907–11.

108 Zarantonelli L, Borthagaray G, Lee E-H, Shafer WM. Decreased azithromycin susceptibility of *Neisseria gonorrhoeae* due to *mtrR* mutations. *Antimicrob Agents Chemother* 1999; **43**: 2468–72.

109 Wadsworth CB, Arnold BJ, Sater MRA, Grad YH. Azithromycin resistance through interspecific acquisition of an epistasis-dependent efflux pump component and transcriptional regulator in *Neisseria gonorrhoeae*. *MBio* 2018; **9**: e01419-18.

110 Unemo M, Seifert HS, Hook EW, Hawkes S, Ndowa F, Dillon J-AR. Gonorrhoea. *Nat Rev Dis Prim* 2019; **5**: 79.

111 Van de Laar M, Spiteri G. Increasing trends of gonorrhoea and syphilis and the threat of drug-resistant gonorrhoea in Europe. *Euro Surveill* 2012; **17**: 20225.

112 ECDC. Sexually transmitted infections in Europe 2013. Stockholm, 2015. https://www.ecdc.europa.eu/sites/default/files/media/en/publications/Publications/sexual-transmitted-infections-europe-surveillance-report-2013.pdf

113 ECDC. Gonorrhoea - Annual Epidemiological Report for 2017. Stockholm, 2019. https://www.ecdc.europa.eu/en/publications-data/gonorrhoea-annual-epidemiological-report-2017

114 RENAVE. Vigilancia epidemiólogica de las infecciones de transmisión sexual en España, 2018. 2020. https://www.mscbs.gob.es/ciudadanos/enfLesiones/enfTransmisibles/sida/vigilancia/Vigilancia_ITS_1995_2018_def.pdf

115 Unemo M, del Rio C, Shafer WM. Antimicrobial resistance expressed by *Neisseria gonorrhoeae:* A major global public health problem in the 21st Century. *Microbiol Spectr* 2016: 213–37.

116 Unemo M. Current and future antimicrobial treatment of gonorrhoea – the rapidly evolving *Neisseria gonorrhoeae* continues to challenge. *BMC Infect Dis* 2015; **15**: 364.

117 ECDC. Gonococcal antimicrobial susceptibility surveillance in Europe, 2017. Stockholm, 2019. https://www.ecdc.europa.eu/en/publications-data/gonococcal-antimicrobial-susceptibility-surveillance-europe-2017

118 Unemo M, Jensen JS. Antimicrobial-resistant sexually transmitted infections: gonorrhoea and *Mycoplasma genitalium*. *Nat Rev Urol* 2017; **14**: 139–52.

119 Wi T, Lahra MM, Ndowa F, *et al.* Antimicrobial resistance in *Neisseria gonorrhoeae*: Global surveillance and a call for international collaborative action. *PLOS Med* 2017; **14**: e1002344.

120 WHO news. WHO publishes list of bacteria for which new antibiotics are urgently needed. 2017. https://www.who.int/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed

121 Adeolu M, Alnajar S, Naushad S, S Gupta R. Genome-based phylogeny and taxonomy of the 'Enterobacteriales': proposal for Enterobacterales ord. nov. divided into the families Enterobacteriaceae, Erwiniaceae fam. nov., Pectobacteriaceae fam. nov., Yersiniaceae fam. nov., Hafniaceae fam. nov., Morgane. *Int J Syst Evol Microbiol* 2016; **66**: 5575–99.

122 NCBI Genome assembly and annotation report of *Serratia marcescens*. https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/1112/

123 Bennett JW, Bentley R. Seeing red: The story of prodigiosin. In: Advances in Applied Microbiology. 2000: 1–32.

124 Mahlen SD. Serratia infections: from military experiments to current practice. *Clin Microbiol Rev* 2011; **24**: 755–91.

125 Yu VL. *Serratia marcescens*: historical perspective and clinical review. *N Engl J Med* 1979; **300**: 887–93.

126 Hejazi A, Falkiner FR. *Serratia marcescens*. J. Med. Microbiol. 1997; **46**: 903–12.

127 Johnson J, Quach C. Outbreaks in the neonatal ICU: a review of the literature. Curr. Opin. Infect. Dis. 2017; **30**: 395–403.

128 Lancaster LJ. Role of *Serratia* species in urinary tract infections. Arch. Intern. Med. 1962; **109**: 536.

129 Ishikawa K, Matsumoto T, Yasuda M, *et al.* The nationwide study of bacterial pathogens associated with urinary tract infections conducted by the Japanese Society of Chemotherapy. *J Infect Chemother* 2011; **17**: 126–38.

130 Campbell JR, Diacovo T, Baker CJ. *Serratia marcescens* meningitis in neonates. *Pediatr Infect Dis J* 1992; **11**: 881–6.

131 Passaro DJ, Waring L, Armstrong R, *et al.* Postoperative *Serratia marcescens* wound infections traced to an out-of-hospital source. *J Infect Dis* 1997; **175**: 992–5.

132 Casolari C, Pecorari M, Fabio G, *et al.* A simultaneous outbreak of *Serratia marcescens* and *Klebsiella pneumoniae* in a neonatal intensive care unit. *J Hosp Infect* 2005; **61**: 312–20.

133 Mills J. *Serratia marcescens* endocarditis: A regional illness associated with intravenous drug abuse. *Ann Intern Med* 1976; **84**: 29.

134 Hawe AJ, Hughes MH. Bacterial endocarditis due to *Chromobacterium prodigiosum*. *BMJ* 1954; **1**: 968–70.

135 Vano-Galvan S, Álvarez-Twose I, Moreno-Martín P, Jaén P. Fulminant necrotizing fasciitis caused by *Serratia marcescens* in an immunosuppressed host. *Int J Dermatol* 2014; **53**: e57–8.

136 Cohen AL, Ridpath A, Noble-Wang J, *et al.* Outbreak of *Serratia marcescens* bloodstream and central nervous system infections after interventional pain management procedures. *Clin J Pain* 2008; **24**: 374–80.

137 Das S. Association between cultures of contact lens and corneal scraping in contact lens–related microbial keratitis. *Arch Ophthalmol* 2007; **125**: 1182.

138 Ariel I, Arad H, Softer D. Autopsy findings of *Serratia* meningoencephalitis in infants. *Pediatr Pathol* 1986; **6**: 351–8.

139 Carbonell GV, Della Colleta HHM, Yano T, Darini ALC, Levy CE, Fonseca BAL. Clinical relevance and virulence factors of pigmented *Serratia marcescens*. *FEMS Immunol Med Microbiol* 2000; **28**: 143–9.

140 Cristina M, Sartini M, Spagnolo A. *Serratia marcescens* infections in neonatal intensive care inits (NICUs). *Int J Environ Res Public Health* 2019; **16**: 610.

141 Curtis N, Starr M, Connell T, Crawford N. Infectious diseases and immunisation. In: Gwee A, Rimer R, Marks M, eds. Paediatric Handbook, 9th edn. Oxford: John Wiley & Sons, 2015: 265–304.

142 EUCAST. EUCAST intrinsic resistance and exceptional phenotypes. Expert rules, version 3.1. Basel, Switzerland, 2016. https://www.eucast.org/fileadmin/src/media/PDFs/EUCAST_files/Expert_Rules/Expert_rules_intrinsic_exceptional_V3.1.pdf

143 Moradigaravand D, Boinett CJ, Martin V, Peacock SJ, Parkhill J. Recent independent emergence of multiple multidrug-resistant *Serratia marcescens* clones within the United Kingdom and Ireland. *Genome Res* 2016; **26**: 1101–9.

144 Šiširak M, Hukić M. An outbreak of multidrug-resistant *Serratia marcescens*: the importance of continuous monitoring of nosocomial infections. *Acta Med Acad* 2013; **42**: 25–31.

145 Vendrell D, Balcázar J, Ruiz-Zarzuela I, de Blas I, Gironés O, Múzquiz J. *Lactococcus garvieae* in fish: A review. *Comp Immunol Microbiol Infect Dis* 2006; **29**: 177–98.

146 Teixeira LM, Merquior VL, Vianni MC, *et al.* Phenotypic and genotypic characterization of atypical *Lactococcus garvieae* strains isolated from water buffalos with subclinical mastitis and confirmation of *L. garvieae* as a senior subjective synonym of *Enterococcus seriolicida*. *Int J Syst*

Bacteriol 1996; **46**: 664–8.

147 Tejedor JL, Vela AI, Gibello A, Casamayor A, Domínguez L, Fernández-Garayzábal JF. A genetic comparison of pig, cow and trout isolates of *Lactococcus garvieae* by PFGE analysis. *Lett Appl Microbiol* 2011; **53**: 614–9.

148 NCBI Genome assembly and annotation report of Lactococcus garvieae. https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/699/

149 Koutsoumanis K, Allende A, Alvarez-Ordóñez A, *et al.* Update of the list of QPS-recommended biological agents intentionally added to food or feed as notified to EFSA 11: suitability of taxonomic units notified to EFSA until September 2019. *EFSA J* 2020; **18**: e05965.

150 Plumed-Ferrer C, Uusikylä K, Korhonen J, von Wright A. Characterization of *Lactococcus lactis* isolates from bovine mastitis. *Vet Microbiol* 2013; **167**: 592–9.

151 Plumed-Ferrer C, Gazzola S, Fontana C, Bassi D, Cocconcelli P-S, von Wright A. Genome sequence of *Lactococcus lactis* subsp. cremoris Mast36, a strain isolated from bovine mastitis. *Genome Announc* 2015; **3**: e00449-15.

152 Chen F, Zhang Z, Chen J. Infective endocarditis caused by *Lactococcus lactis* subsp. lactis and *Pediococcus pentosaceus*. *Medicine (Baltimore)* 2018; **97**: e13658.

153 Shimizu A, Hase R, Suzuki D, *et al. Lactococcus lactis* cholangitis and bacteremia identified by MALDI-TOF mass spectrometry: A case report and review of the literature on *Lactococcus lactis* infection. *J Infect Chemother* 2019; **25**: 141–6.

154 Meyburgh C, Bragg R, Boucher C. *Lactococcus garvieae*: an emerging bacterial pathogen of fish. *Dis Aquat Organ* 2017; **123**: 67–79.

155 Plumed-Ferrer C, Barberio A, Franklin-Guild R, *et al.* Antimicrobial susceptibilities and random amplified polymorphic DNA-PCR fingerprint characterization of *Lactococcus lactis* ssp. *lactis* and *Lactococcus garvieae* isolated from bovine intramammary infections. *J Dairy Sci* 2015; **98**: 6216–25.

156 Kim JH, Go J, Cho CR, Kim J Il, Lee MS, Park SC. First Report of human acute acalculous cholecystitis caused by the fish pathogen *Lactococcus garvieae*. *J Clin Microbiol* 2013; **51**: 712–4.

157 Gibello A, Galán-Sánchez F, Blanco MM, Rodríguez-Iglesias M, Domínguez L, Fernández-Garayzábal JF. The zoonotic potential of *Lactococcus garvieae*: An overview on microbiology, epidemiology, virulence factors and relationship with its presence in foods. *Res Vet Sci* 2016; **109**: 59–70.

158 Eldar A, Ghittino C. *Lactococcus garvieae* and *Streptococcus iniae* infections in rainbow trout *Oncorhynchus mykiss*: similar, but different diseases. *Dis Aquat Organ* 1999; **36**: 227–31.

159 Carvalho MDGS, Vianni MDCE, Elliott JA, Reeves M, Facklam RR, Teixeira LM. Molecular analysis of *Lactococcus garvieae* and *Enterococcus gallinarum* isolated from water buffalos with subclinical mastitis. *Adv Exp Med Biol* 1997; **418**: 401–4.

160 Devriese L., Hommez J, Laevens H, Pot B, Vandamme P, Haesebrouck F. Identification of aesculin-hydrolyzing streptococci, lactococci, aerococci and enterococci from subclinical intramammary infections in dairy cows. *Vet Microbiol* 1999; **70**: 87–94.

161 Vela AI, Va□zquez J, Gibello A, *et al.* Phenotypic and genetic characterization of *Lactococcus garvieae* isolated in Spain from lactococcosis outbreaks and comparison with isolates of other countries and sources. *J Clin Microbiol* 2000; **38**: 3791–5.

162 Wyder AB, Boss R, Naskova J, Kaufmann T, Steiner A, Graber HU. *Streptococcus* spp. and related bacteria: Their identification and their pathogenic potential for chronic mastitis – A molecular approach. *Res Vet Sci* 2011; **91**: 349–57.

163 Reguera-Brito M, Galán-Sánchez F, Blanco MM, *et al.* Genetic analysis of human clinical isolates of *Lactococcus garvieae*: Relatedness with isolates from foods. *Infect Genet Evol* 2016; **37**: 185–91.

164 Lim SM, Wong B, Cross GB, Merchant R. *Lactobacillus garvieae* endocarditis presenting with leg cramps. *IDCases* 2018; **13**: e00427.

165 Malek A, De la Hoz A, Gomez-Villegas SI, Nowbakht C, Arias CA. *Lactococcus garvieae*, an unusual pathogen in infective endocarditis: case report and review of the literature. *BMC Infect Dis* 2019; **19**: 301.

166 Wang C-YC, Shie H-S, Chen S-C, *et al. Lactococcus garvieae* infections in humans: possible association with aquaculture outbreaks. *Int J Clin Pract* 2006; **61**: 68–73.

167 Chan JFW, Woo PCY, Teng JLL, *et al.* Primary infective spondylodiscitis caused by *Lactococcus garvieae* and a review of human *L. garvieae* infections. *Infection* 2011; **39**: 259–64.

168 Russo G, Ianetta M, D'Abramo A, *et al. Lactococcus garvieae* endocarditis in a patient with colonic diverticulosis: first case report in Italy and review of the literature. *New Microbiol* 2012; **35**: 495–501.

169 Wilbring M, Alexiou K, Reichenspurner H, Matschke K, Tugtekin SM. *Lactococcus garvieae* causing zoonotic prosthetic valve endocarditis. *Clin Res Cardiol* 2011; **100**: 545–6.

170 Galán Montemayor JC, Lepe Jiménez JA, Otero Guerra L, Serra Pladevall J, Vázquez Valdés F. 24a. Diagnóstico microbiológico de las infecciones de transmisión sexual y otras infecciones genitales. Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica (SEIMC), 2018.

171 Unemo M, Ison C. Gonorrhoea. In: Unemo M, Ballard R, Ison C, Lewis D, Ndowa F, Peeling R, eds. Laboratory diagnosis of sexually transmitted infections, including human immunodeficiency virus. Geneva: World Health Organization (WHO), 2013: 21–53.

172 Ng L-K, Martin IE. The laboratory diagnosis of *Neisseria gonorrhoeae*. *Can J Infect Dis Med Microbiol* 2005; **16**: 15–25.

173 Papp JR, Schachter J, Gaydos CA, Van Der Pol B. Recommendations for the laboratory-based detection of *Chlamydia trachomatis* and *Neisseria gonorrhoeae* - 2014. Atlanta, 2014.

174 Thayer JD, Martin JE. Improved medium selective for cultivation of *N. gonorrhoeae* and *N. meningitidis*. *Public Health Rep* 1966; **81**: 559–62.

175 Merlino J, Siarakas S, Robertson GJ, Funnell GR, Gottlieb T, Bradbury R. Evaluation of CHROMagar Orientation for differentiation and presumptive identification of gram-negative bacilli and *Enterococcus* species. *J Clin Microbiol* 1996; **34**: 1788–93.

176 Heras Cañas V, Pérez Ramirez MD, Bermudez Jiménez F, *et al.* *Lactococcus garvieae* endocarditis in a native valve identified by MALDI-TOF MS and PCR-based 16s rRNA in Spain: A case report. *New Microbes New Infect* 2015; **5**: 13–5.

177 EUCAST. Breakpoint tables for interpretation of MICs and zone diameters. Version 10.0. Basel, Switzerland, 2020. https://www.eucast.org/fileadmin/src/media/PDFs/EUCAST_files/Breakpoint_tables/v_10.0_Breakpoint_Tables.pdf

178 Soussy CJ, Carret G, Cavallo JD, *et al.* [Antibiogram Committee of the French Microbiology Society. Report 2000-2001] (Article in French). *Pathol Biol* 2000; **48**: 832–71.

179 Head SR, Komori HK, LaMere SA, *et al.* Library construction for next-generation sequencing: Overviews and challenges. *Biotechniques* 2014; **56**: 61-4.

180 Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016; **32**: 3047–8.

181 Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 2011; **27**: 863–4.

182 Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using Phred. I. Accuracy Assessment. *Genome Res* 1998; **8**: 175–85.

183 Ewing B, Green P. Base-calling of automated sequencer traces using Phred. II. Error Probabilities. *Genome Res* 1998; **8**: 186–94.

184 Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014; **15**: R46.

185 Ondov BD, Starrett GJ, Sappington A, *et al.* Mash Screen: high-

throughput sequence containment estimation for genome discovery. *Genome Biol* 2019; **20**: 232.

186 Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. *Nat Methods* 2009; **6**: S6–12.

187 Olson ND, Lund SP, Colman RE, *et al.* Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Front Genet* 2015; **6**: 235.

188 Palmieri N, Schlötterer C. Mapping accuracy of short reads from massively parallel sequencing and the implications for quantitative expression profiling. *PLoS One* 2009; **4**: e6323.

189 Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv.1303.3997v2* 2013.

190 Wang W, Wei Z, Lam T-W, Wang J. Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. *Sci Rep* 2011; **1**: 55.

191 Li H, Handsaker B, Wysoker A, *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25**: 2078–9.

192 Van der Auwera GA, Carneiro MO, Hartl C, *et al.* From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics* 2013; **43**: 11.10.1-11.10.33.

193 Bankevich A, Nurk S, Antipov D, *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012; **19**: 455–77.

194 Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 2012; **28**: 1420–8.

195 Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013; **29**: 1072–5.

196 Chin C-S, Alexander DH, Marks P, *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat*

Methods 2013; **10**: 563–9.

197 Myers EW, Sutton GG, Delcher AL, *et al.* A Whole-genome assembly of *Drosophila*. *Science* 2000; **287**: 2196–204.

198 Hunt M, Silva N De, Otto TD, Parkhill J, Keane JA, Harris SR. Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol* 2015; **16**: 294.

199 Maiden MCJ, Bygraves JA, Feil E, *et al.* Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci* 1998; **95**: 3140–5.

200 Martin IMC, Ison CA, Aanensen DM, Fenton KA, Spratt BG. Rapid sequence-based identification of gonococcal transmission clusters in a large metropolitan area. *J Infect Dis* 2004; **189**: 1497–505.

201 Chisholm SA, Unemo M, Quaye N, *et al.* Molecular epidemiological typing within the European Gonococcal Antimicrobial Resistance Surveillance Programme reveals predominance of a multidrug-resistant clone. *Euro Surveill* 2013; **18**: 20358.

202 Demczuk W, Sidhu S, Unemo M, *et al. Neisseria gonorrhoeae* sequence typing for antimicrobial resistance, a novel antimicrobial resistance multilocus typing scheme for tracking global dissemination of *N. gonorrhoeae* strains. *J Clin Microbiol* 2017; **55**: 1454–68.

203 Inouye M, Dashnow H, Raven L-A, *et al.* SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med* 2014; **6**: 90.

204 Camacho C, Coulouris G, Avagyan V, *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* 2009; **10**: 421.

205 Awska MA. Analysis of the filamentous bacteriophage genomes integrated into *Neisseria gonorrhoeae* FA1090 chromosome. *Polish J Microbiol* 2006; **55**: 251–60.

206 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**: e15–e15.

207 Arndt D, Grant JR, Marcu A, *et al.* PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* 2016; **44**: W16–21.

208 Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 2000; **17**: 540–52.

209 Page AJ, Taylor B, Delaney AJ, *et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genomics* 2016; **2**: e000056.

210 Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015; **32**: 268–74.

211 Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 2017; **14**: 587–9.

212 Minh BQ, Nguyen MAT, von Haeseler A. Ultrafast Approximation for Phylogenetic Bootstrap. *Mol Biol Evol* 2013; **30**: 1188–95.

213 Galili T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* 2015; **31**: 3718–20.

214 R Core Team. R: A language and environment for statistical computing. 2019.

215 Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 2019; **47**: W256–9.

216 Paradis E, Claude J, Strimmer K. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* 2004; **20**: 289–90.

217 Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* 2000; **155**: 945–59.

218 Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 2003; **164**: 1567–87.

219 Earl DA, VonHoldt BM. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the

Evanno method. *Conserv Genet Resour* 2012; **4**: 359–61.

220 Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol Ecol* 2005; **14**: 2611–20.

221 Jakobsson M, Rosenberg NA. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 2007; **23**: 1801–6.

222 Kamvar ZN, Tabima JF, Grünwald NJ. Poppr : an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* 2014; **2**: e281.

223 Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006; **22**: 1658–9.

224 Madeira F, Park Y mi, Lee J, *et al.* The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res* 2019; **47**: W636–41.

225 Strimmer K, von Haeseler A. Likelihood-mapping: A simple method to visualize phylogenetic content of a sequence alignment. *Proc Natl Acad Sci* 1997; **94**: 6815–9.

226 Münkemüller T, Lavergne S, Bzeznik B, *et al.* How to measure and test phylogenetic signal. *Methods Ecol Evol* 2012; **3**: 743–56.

227 Shimodaira H, Hasegawa M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol* 1999; **16**: 1114–6.

228 Strimmer K, Rambaut A. Inferring confidence sets of possibly misspecified gene trees. *Proc R Soc London Ser B Biol Sci* 2002; **269**: 137–42.

229 Robinson O, Dylus D, Dessimoz C. Phylo.io : Interactive viewing and comparison of large phylogenetic trees on the web. *Mol Biol Evol* 2016; **33**: 2163–6.

230 Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK.

Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal Methods* 2016; **8**: 12–24.

231 Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho: Detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics* 2011; **12**: 124.

232 Ranwez V, Harispe S, Delsuc F, Douzery EJP. MACSE: Multiple alignment of coding sequences accounting for frameshifts and stop codons. *PLoS One* 2011; **6**: e22594.

233 Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol* 2018; **35**: 1547–9.

234 Hunt M, Mather AE, Sánchez-Busó L, *et al.* ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb Genomics* 2017; **3**: e000131.

235 Li H. FermiKit: assembly-based variant calling for Illumina resequencing data. *Bioinformatics* 2015; **31**: 3694–6.

236 Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* 2016; **2**: vew007.

237 To T-H, Jung M, Lycett S, Gascuel O. Fast dating using least-squares criteria and algorithms. *Syst Biol* 2016; **65**: 82–97.

238 Muller BH, Mollon P, Santiago-Allexant E, Javerliat F, Kaneko G. In-depth comparison of library pooling strategies for multiplexing bacterial species in NGS. *Diagn Microbiol Infect Dis* 2019; **95**: 28–33.

239 Kimura M. Estimation of evolutionary distances between homologous nucleotide sequences. *Proc Natl Acad Sci* 1981; **78**: 454–8.

240 Lewis PO. A Likelihood approach to estimating phylogeny from discrete morphological character data. *Syst Biol* 2001; **50**: 913–25.

241 Yang Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J Mol Evol* 1994; **39**: 306–14.

242 Zhang J, van der Veen S. *Neisseria gonorrhoeae* 23S rRNA A2059G mutation is the only determinant necessary for high-level azithromycin resistance and improves in vivo biological fitness. *J Antimicrob Chemother* 2019; **74**: 407–15.

243 Shimuta K, Watanabe Y, Nakayama S ichi, *et al.* Emergence and evolution of internationally disseminated cephalosporin-resistant *Neisseria gonorrhoeae* clones from 1995 to 2005 in Japan. *BMC Infect Dis* 2015; **15**: 1–11.

244 Sánchez-Busó L, Golparian D, Corander J, *et al.* The impact of antimicrobials on gonococcal evolution. *Nat Microbiol* 2019; **4**: 1941–50.

245 Golparian D, Harris SR, Sánchez-Busó L, *et al.* Genomic evolution of *Neisseria gonorrhoeae* since the preantibiotic era (1928–2013): antimicrobial use/misuse selects for resistance and drives evolution. *BMC Genomics* 2020; **21**: 116.

246 Kirkcaldy RD, Zaidi A, Hook EW, *et al. Neisseria gonorrhoeae* antimicrobial resistance among men who have sex with men and men who have sex exclusively with women: the Gonococcal Isolate Surveillance Project, 2005–2010. *Ann Intern Med* 2013; **158**: 321.

247 Kirkcaldy RD, Kidd S, Weinstock HS, Papp JR, Bolan GA. Trends in antimicrobial resistance in *Neisseria gonorrhoeae* in the USA: the Gonococcal Isolate Surveillance Project (GISP), January 2006–June 2012. *Sex Transm Infect* 2013; **89**: iv5–10.

248 Unemo M, Lahra MM, Cole M, *et al.* World Health Organization Global Gonococcal Antimicrobial Surveillance Program (WHO GASP): review of new data and evidence to inform international collaborative actions and research efforts. *Sex Health* 2019; **16**: 412.

249 Cole MJ, Quinten C, Jacobsson S, *et al.* The European gonococcal antimicrobial surveillance programme (Euro-GASP) appropriately reflects the antimicrobial resistance situation for *Neisseria gonorrhoeae* in the European Union/European Economic Area. *BMC Infect Dis* 2019; **19**: 1040.

250 Moura A, Criscuolo A, Pouseele H, *et al.* Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes*. *Nat Microbiol* 2017; **2**: 16185.

251 Deng X, den Bakker HC, Hendriksen RS. Genomic epidemiology: whole-genome-sequencing–powered surveillance and outbreak investigation of foodborne bacterial pathogens. *Annu Rev Food Sci Technol* 2016; **7**: 353–74.

252 Comas I. Genomic epidemiology of tuberculosis. *Adv Exp Med Biol* 2017; **1019**: 79–93.

253 Lakhundi S, Zhang K. Methicillin-resistant *Staphylococcus aureus*: Molecular characterization, evolution, and epidemiology. *Clin Microbiol Rev* 2018; **31**: e00020-18.

254 De Silva D, Peters J, Cole K, *et al.* Whole-genome sequencing to determine transmission of *Neisseria gonorrhoeae*: an observational study. *Lancet Infect Dis* 2016; **16**: 1295–303.

255 Demczuk W, Lynch T, Martin I, *et al.* Whole-genome phylogenomic heterogeneity of *Neisseria gonorrhoeae* isolates with decreased cephalosporin susceptibility collected in Canada between 1989 and 2013. *J Clin Microbiol* 2015; **53**: 191–200.

256 Ezewudo MN, Joseph SJ, Castillo-Ramirez S, *et al.* Population structure of *Neisseria gonorrhoeae* based on whole genome data and its relationship with antibiotic resistance. *PeerJ* 2015; **2015**: 1–23.

257 Al Suwayyid BA, Coombs GW, Speers DJ, Pearson J, Wise MJ, Kahler CM. Genomic epidemiology and population structure of *Neisseria gonorrhoeae* from remote highly endemic Western Australian populations. *BMC Genomics* 2018; **19**: 165.

258 Lee RS, Seemann T, Heffernan H, *et al.* Genomic epidemiology and antimicrobial resistance of *Neisseria gonorrhoeae* in New Zealand. *J Antimicrob Chemother* 2018; **73**: 353–64.

259 Boiko I, Golparian D, Jacobsson S, *et al.* Genomic epidemiology and antimicrobial resistance determinants of *Neisseria gonorrhoeae* isolates from Ukraine, 2013–2018. *APMIS* 2020; **128**: 465–75.

260 Golparian D, Bazzo ML, Golfetto L, *et al.* Genomic epidemiology of *Neisseria gonorrhoeae* elucidating the gonococcal antimicrobial resistance and lineages/sublineages across Brazil, 2015–16. *J Antimicrob Chemother* 2020; **75**: 3163–72.

261 Alfsnes K, Eldholm V, Olsen AO, *et al.* Genomic epidemiology and population structure of *Neisseria gonorrhoeae* in Norway, 2016–2017. *Microb Genom* 2020; **6**: e000359.

262 Peirano G, Matsumura Y, Adams MD, *et al.* Genomic Epidemiology of global carbapenemase-producing *Enterobacter* spp., 2008–2014. *Emerg Infect Dis* 2018; **24**: 1010–9.

263 Meumann EM, Anstey NM, Currie BJ, *et al.* Genomic epidemiology of severe community-onset *Acinetobacter baumannii* infection. *Microb Genom* 2019; **5**: e000258.

264 Cabot G, López-Causapé C, Ocampo-Sosa AA, *et al.* Deciphering the resistome of the widespread *P. aeruginosa* ST175 international high-risk clone through whole genome sequencing. *Antimicrob Agents Chemother* 2016; **60**: 7415-23.

265 Cobo F, Cabezas-Fernández MT, Cabeza-Barrera MI. Antimicrobial susceptibility and typing of *Neisseria gonorrhoeae* strains from Southern Spain, 2012–2014. *Enferm Infecc Microbiol Clin* 2016; **34**: 3–7.

266 Ibargoyen García U, Nieto Toboso MC, Azpeitia EM, *et al.* Epidemiological surveillance study of gonococcal infection in Northern Spain. *Enferm Infecc Microbiol Clin* 2020; **38**: 59–64.

267 Harris SR, Cole MJ, Spiteri G, *et al.* Public health surveillance of multidrug-resistant clones of *Neisseria gonorrhoeae* in Europe: a genomic survey. *Lancet Infect Dis* 2018; **18**: 758–68.

268 Kwong JC, Chow EPF, Stevens K, *et al.* Whole-genome sequencing reveals transmission of gonococcal antibiotic resistance among men who have sex with men: an observational study. *Sex Transm Infect* 2018; **94**: 151–7.

269 Maduna LD, Kock MM, van der Veer BMJW, *et al.* Antimicrobial

resistance of *Neisseria gonorrhoeae* isolates from high-risk men in Johannesburg, South Africa. *Antimicrob Agents Chemother* 2020; **64**: e00906-20.

270 Xu X, Chow EPF, Ong JJ, *et al.* Modelling the contribution that different sexual practices involving the oropharynx and saliva have on *Neisseria gonorrhoeae* infections at multiple anatomical sites in men who have sex with men. *Sex Transm Infect* 2020; DOI: 10.1136/sextrans-2020-054565.

271 Painset A, Day M, Doumith M, *et al.* Comparison of phenotypic and WGS-derived antimicrobial resistance profiles of *Campylobacter jejuni* and *Campylobacter coli* isolated from cases of diarrhoeal disease in England and Wales, 2015–16. *J Antimicrob Chemother* 2020; **75**: 883–9.

272 Davies TJ, Stoesser N, Sheppard AE, *et al.* Reconciling the potentially irreconcilable? Genotypic and phenotypic amoxicillin-clavulanate resistance in *Escherichia coli*. *Antimicrob Agents Chemother* 2020; **64**: e02026-19.

273 Urmi UL, Nahar S, Rana M, *et al.* Genotypic to phenotypic resistance discrepancies identified involving β-lactamase genes, bla_{KPC}, bla_{IMP}, bla_{NDM-1}, and bla_{VIM} in uropathogenic *Klebsiella pneumoniae*. *Infect Drug Resist* 2020; **13**: 2863–75.

274 Ou C-Y, Ciesielski CA, Myers G, *et al.* Molecular epidemiology of HIV transmission in a dental practice. *Science* 1992; **256**: 1165–71.

275 Birch CJ, McCaw RF, Bulach DM, *et al.* Molecular analysis of human immunodeficiency virus strains associated with a case of criminal transmission of the virus. *J Infect Dis* 2000; **182**: 941–4.

276 Machuca R, Jørgensen LB, Theilade P, Nielsen C. Molecular investigation of transmission of human immunodeficiency virus Type 1 in a criminal case. *Clin Diagnostic Lab Immunol* 2001; **8**: 884–90.

277 Metzker ML, Mindell DP, Liu X-M, Ptak RG, Gibbs RA, Hillis DM. Molecular evidence of HIV-1 transmission in a criminal case. *Proc Natl Acad Sci* 2002; **99**: 14292–7.

278 Scaduto DI, Brown JM, Haaland WC, Zwickl DJ, Hillis DM, Metzker ML. Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences. *Proc Natl Acad Sci* 2010; **107**: 21242–7.

279 González-Candelas F, Bracho MA, Wróbel B, Moya A. Molecular evolution in court: analysis of a large hepatitis C virus outbreak from an evolving source. *BMC Biol* 2013; **11**: 76.

280 Jernigan DB, Raghunathan PL, Bell BP, *et al.* Investigation of bioterrorism-related anthrax, United States, 2001: epidemiologic findings. *Emerg Infect Dis* 2002; **8**: 1019–28.

281 Hoffmaster AR, Fitzgerald CC, Ribot E, Mayer LW, Popovic T. Molecular subtyping of *Bacillus anthracis* and the 2001 bioterrorism-associated anthrax outbreak, United States. *Emerg Infect Dis* 2002; **8**: 1111–6.

282 Rasko DA, Worsham PL, Abshire TG, *et al. Bacillus anthracis* comparative genome analysis in support of the Amerithrax investigation. *Proc Natl Acad Sci* 2011; **108**: 5027–32.

283 Budowle B, Murch R, Chakraborty R. Microbial forensics: the next forensic challenge. *Int J Legal Med* 2005; **119**: 317–30.

284 Budowle B, Connell ND, Bielecka-Oder A, *et al.* Validation of high throughput sequencing and microbial forensics applications. *Investig Genet* 2014; **5**: 9.

285 Vicente D, Esnal O, Marimon JM, Gastesi C, Pérez-Trallero E. *Neisseria meningitidis* W-135 in the Basque Country, northern Spain. *Clin Microbiol Infect* 2006; **12**: 812–5.

286 Unemo M, Golparian D, Sánchez-Busó L, *et al.* The novel 2016 WHO *Neisseria gonorrhoeae* reference strains for global quality assurance of laboratory investigations: phenotypic, genetic and reference genome characterization. *J Antimicrob Chemother* 2016; **71**: 3096–108.

287 Maiden MCJ. Multilocus Sequence Typing of Bacteria. *Annu Rev Microbiol* 2006; **60**: 561-88.

288 Unemo M, Dillon J-AR. Review and international recommendation of

methods for typing *Neisseria gonorrhoeae* isolates and their implications for improved knowledge of gonococcal epidemiology, treatment, and biology. *Clin Microbiol Rev* 2011; **24**: 447–58.

289 Sathirareuangchai S, Phuangphung P, Leelaporn A, Boon-yasidhi V. The usefulness of *Neisseria gonorrhoeae* strain typing by Pulse-Field Gel Electrophoresis (PFGE) and DNA detection as the forensic evidence in child sexual abuse cases: a case series. *Int J Legal Med* 2015; **129**: 153–7.

290 Leopold SR, Goering R V., Witten A, Harmsen D, Mellmann A. Bacterial whole-genome sequencing revisited: portable, scalable, and standardized analysis for typing and detection of virulence and antibiotic resistance genes. *J Clin Microbiol* 2014; **52**: 2365–70.

291 Kovanen SM, Kivisto RI, Rossi M, *et al.* Multilocus Sequence Typing (MLST) and whole-genome MLST of *Campylobacter jejuni* isolates from human infections in three districts during a seasonal peak in Finland. *J Clin Microbiol* 2014; **52**: 4147–54.

292 Lüth S, Kleta S, Al Dahouk S. Whole genome sequencing as a typing tool for foodborne pathogens like *Listeria monocytogenes* – The way towards global harmonisation and data exchange. *Trends Food Sci Technol* 2018; **73**: 67–75.

293 Quainoo S, Coolen JPM, van Hijum SAFT, *et al.* Whole-genome sequencing of bacterial pathogens: the future of nosocomial outbreak analysis. *Clin Microbiol Rev* 2017; **30**: 1015–63.

294 Budowle B, Johnson MD, Fraser CM, Leighton TJ, Murch RS, Chakraborty R. Genetic analysis and attribution of microbial forensics evidence. *Crit Rev Microbiol* 2005; **31**: 233–54.

295 Schmedes SE, Sajantila A, Budowle B. Expansion of microbial forensics. *J Clin Microbiol* 2016; **54**: 1964–74.

296 Hammerschlag MR, Guille□n CD. Medical and legal implications of testing for sexually transmitted infections in children. *Clin Microbiol Rev* 2010; **23**: 493–506.

297 Kellogg N, the Committee on Child Abuse and Neglect. The evaluation

of sexual abuse in children. *Pediatrics* 2005; **116**: 506–12.

298 Klemm E, Dougan G. Advances in understanding bacterial pathogenesis gained from whole-genome sequencing and phylogenetics. *Cell Host Microbe* 2016; **19**: 599–610.

299 Duchêne S, Holt KE, Weill F-X, *et al.* Genome-scale rates of evolutionary change in bacteria. *Microb Genom* 2016; **2**: e000094.

300 Maltezou HC, Tryfinopoulou K, Katerelos P, *et al.* Consecutive *Serratia marcescens* multiclone outbreaks in a neonatal intensive care unit. *Am J Infect Control* 2012; **40**: 637–42.

301 Geyter D De, De Geyter D, Blommaert L, *et al.* The sink as a potential source of transmission of carbapenemase-producing *Enterobacteriaceae* in the intensive care unit. *Antimicrob Resist Infect Control* 2017; **6**: 24.

302 Regev-Yochay G, Smollan G, Tal I, *et al.* Sink traps as the source of transmission of OXA-48–producing *Serratia marcescens* in an intensive care unit. *Infect Control Hosp Epidemiol* 2018; **39**: 1307–15.

303 Rohit A, Suresh Kumar D, Dhinakaran I, *et al.* Whole-genome-based analysis reveals multiclone *Serratia marcescens* outbreaks in a non-neonatal intensive care unit setting in a tertiary care hospital in India. *J Med Microbiol* 2019; **68**: 616–21.

304 Saralegui C, Ponce-Alonso M, Pérez-Viso B, *et al.* Genomics of *Serratia marcescens* isolates causing outbreaks in the same pediatric unit 47 years apart: Position in an updated phylogeny of the species. *Front Microbiol* 2020; **11**: 451.

305 Pearce ME, Alikhan N-F, Dallman TJ, Zhou Z, Grant K, Maiden MCJ. Comparative analysis of core genome MLST and SNP typing within a European *Salmonella* serovar Enteritidis outbreak. *Int J Food Microbiol* 2018; **274**: 1–11.

306 Martineau C, Li X, Lalancette C, *et al. Serratia marcescens* outbreak in a neonatal intensive care unit: new insights from next-generation sequencing applications. *J Clin Microbiol* 2018; **56**: 235-18.

307 Lee RS, Behr MA. Does choice matter? Reference-based alignment for molecular epidemiology of tuberculosis. *J Clin Microbiol* 2016; **54**: 1891–5.

308 Ferrario C, Ricci G, Milani C, *et al. Lactococcus garvieae*: Where is it from? A first approach to explore the evolutionary history of this emerging pathogen. *PLoS One* 2013; **8**: e84796.

309 Shahi N, Mallik SK. Emerging bacterial fish pathogen *Lactococcus garvieae* RTCLI04, isolated from rainbow trout (*Oncorhynchus mykiss*): Genomic features and comparative genomics. *Microb Pathog* 2020; **147**: 104368.

310 Gyles C, Boerlin P. Horizontally transferred genetic elements and their role in pathogenesis of bacterial disease. *Vet Pathol* 2014; **51**: 328–40.

311 Awadalla P. The evolutionary genomics of pathogen recombination. *Nat Rev Genet* 2003; **4**: 50–60.

312 Lefébure T, Stanhope MJ. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biol* 2007; **8**: R71.

313 Petersen L, Bollback JP, Dimmic M, Hubisz M, Nielsen R. Genes under positive selection in *Escherichia coli*. *Genome Res* 2007; **17**: 1336–43.

314 Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci* 2009; **106**: 19126–31.

315 Snitkin ES, Zelazny AM, Thomas PJ, *et al.* Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. *Sci Transl Med* 2012; **4**: 148ra116-148ra116.

316 Struelens MJ, Brisse S. From molecular to genomic epidemiology: transforming surveillance and control of infectious diseases. *Eurosurveillance* 2013; **18**. DOI:10.2807/ese.18.04.20386-en.

317 Grad YH, Harris SR, Kirkcaldy RD, *et al.* Genomic epidemiology of gonococcal resistance to extended-spectrum cephalosporins, macrolides, and fluoroquinolones in the United States, 2000-2013. *J Infect Dis* 2016; **214**: 1579–87.

318 Valiente-Mullor C, Beamud B, Ansari I, *et al.* One is not enough: on the effects of reference genome for the mapping and subsequent analyses of short-reads. *bioRxiv* 2020; : 2020.04.14.041004.

319 Linz B, Ivanov Y V., Preston A, *et al.* Acquisition and loss of virulence-associated factors during genome evolution and speciation in three clades of *Bordetella* species. *BMC Genomics* 2016; **17**: 767.

320 Nakano K, Shiroma A, Shimoji M, *et al.* Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. *Hum Cell* 2017; **30**: 149–61.

321 Berbers B, Saltykova A, Garcia-Graells C, *et al.* Combining short and long read sequencing to characterize antimicrobial resistance genes on plasmids applied to an unauthorized genetically modified *Bacillus*. *Sci Rep* 2020; **10**: 4310.

APPENDICES

I. VERSIONS OF THE COMPUTER PROGRAMS USED

Program[a]	Chapter 1	Chapter 2	Chapter 3	Chapter 4
APE	5.2	4.1	5.2	
ARIBA	3.6.9			
BCFtools	1.4	1.4	1.4	
BLAST	2.9.0+	2.8.0+	2.9.0+	2.9.0+
BWA-MEM	0.7.5a-r405	0.7.5a-r405	0.7.5a-r405	
CD-HIT-EST	4.8.1			
Circlator				1.5.0
CLUMPP	1.1.2			
dendextend	1.13.4			
EMBOSS tools	6.6.0.0			
FastQC	0.11.5	0.11.5	0.11.5	
GATK	3.5	3.5	3.5	
GBlocks	0.91b	0.91b	0.91b	
Gubbins	1.4.10			
IDBA-UD	1.1.1	1.1.1	1.1.1	
IQ-TREE	1.6.1	1.5.5	1.6.1	1.6.1
Kraken			1.0	
LSD2	1.9.7			
MACSE				1.2
Mash			2.0	
MultiQC	1.7	1.7	1.7	
MUMMER	3.23	3.1	3.23	
Picard tools	1.141	1.141	1.141	
poppr	2.8.5			

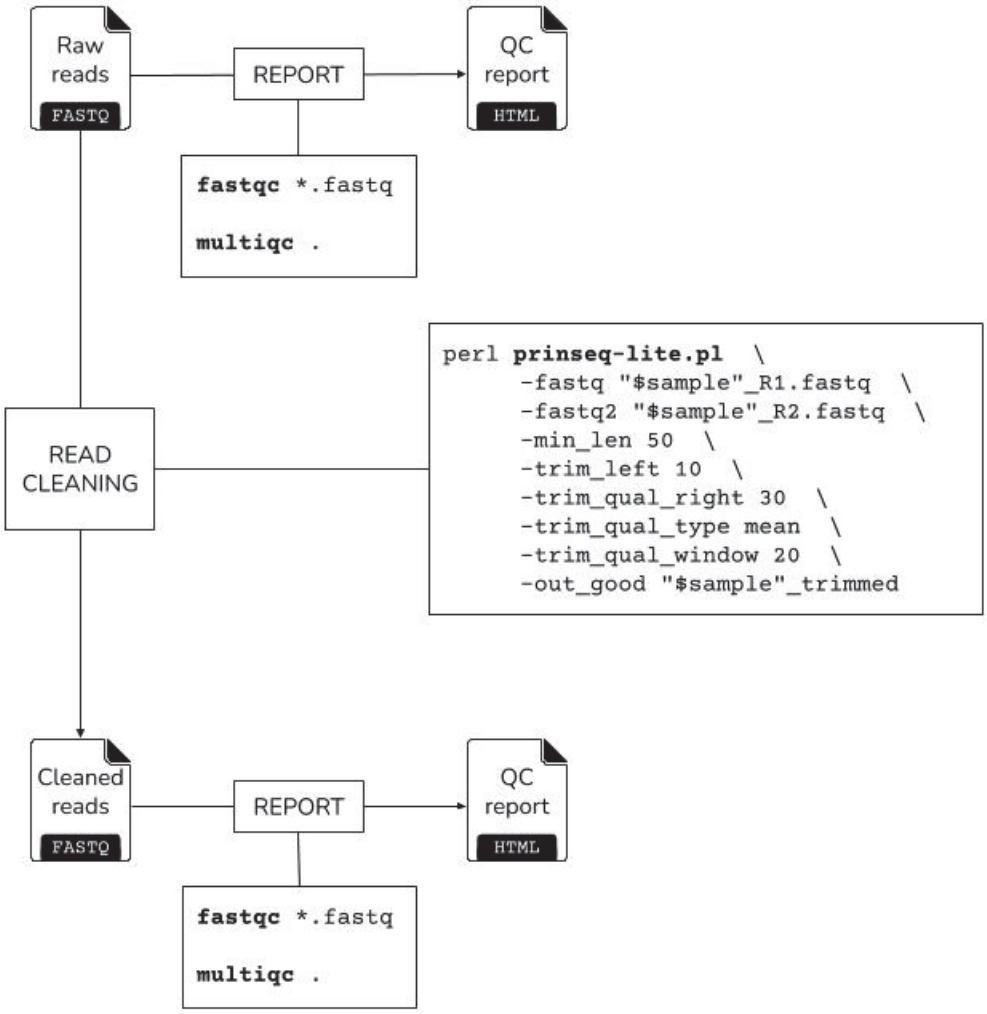
Continued on next page.

Program[a]	Chapter 1	Chapter 2	Chapter 3	Chapter 4
PRINSEQ –lite	0.20.4	0.20.4	0.20.4	
Prokka		1.13		1.11
Proteinortho		5.16b		5.16b
pyani				0.2.10
QUAST	5.0.2	4.3	5.0.2	
R	3.6.3	3.4.1	3.4.4	3.6.3
SAMtools	1.4	1.4	1.4	
seqtk	1.2-r101	1.2-r101	1.2-r101	
SMRT				2.3.0
SPAdes	3.13.0	3.9.0	3.13.0	
SRST2	0.2.0	0.2.0		
STRUCTURE	2.3.4			
STRUCTURE Harvester	0.6.94			
TempEst	1.5.3			
WGS-assembler				7.0

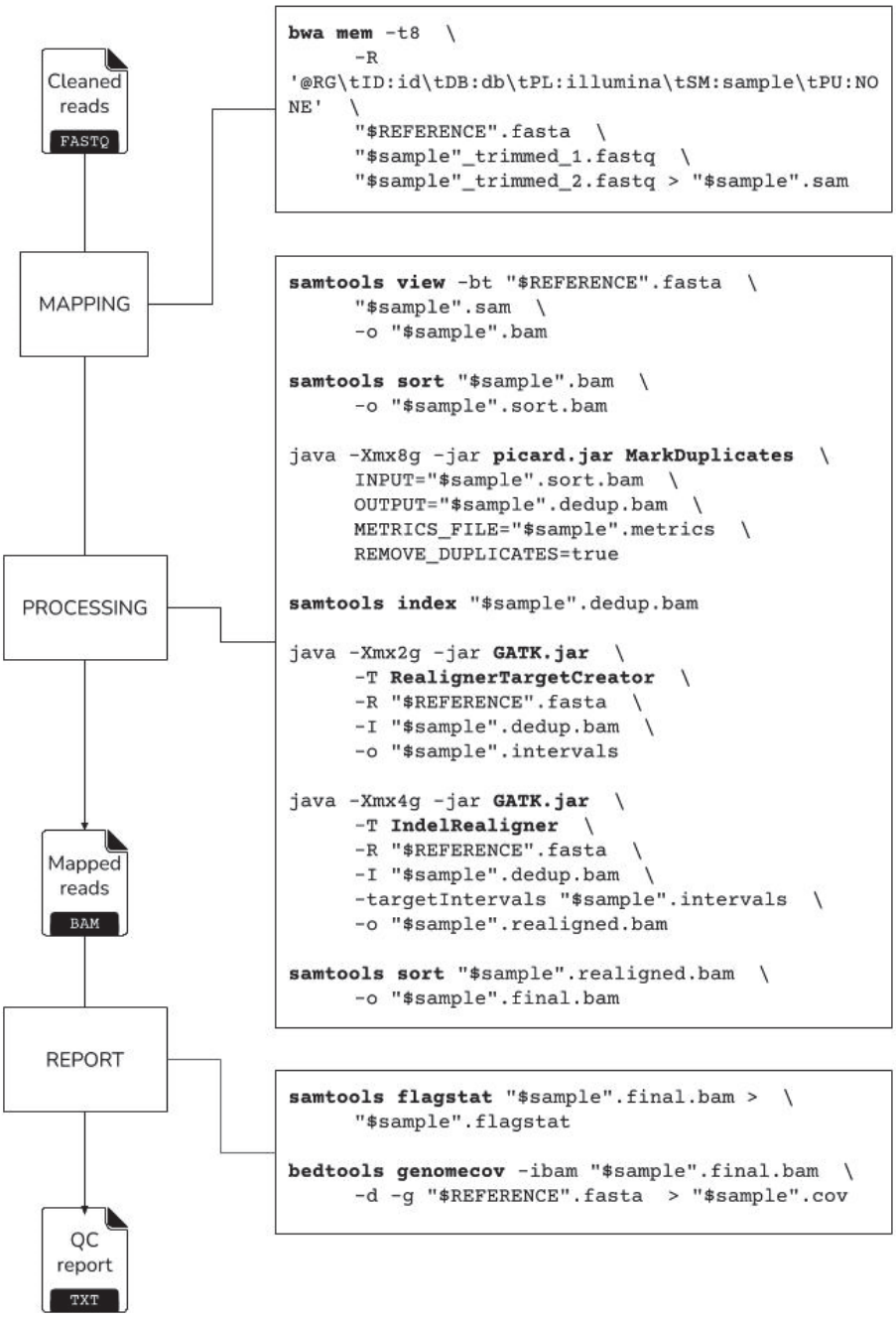
a. Online programs were excluded.

II. PIPELINES OF THE ANALYSES

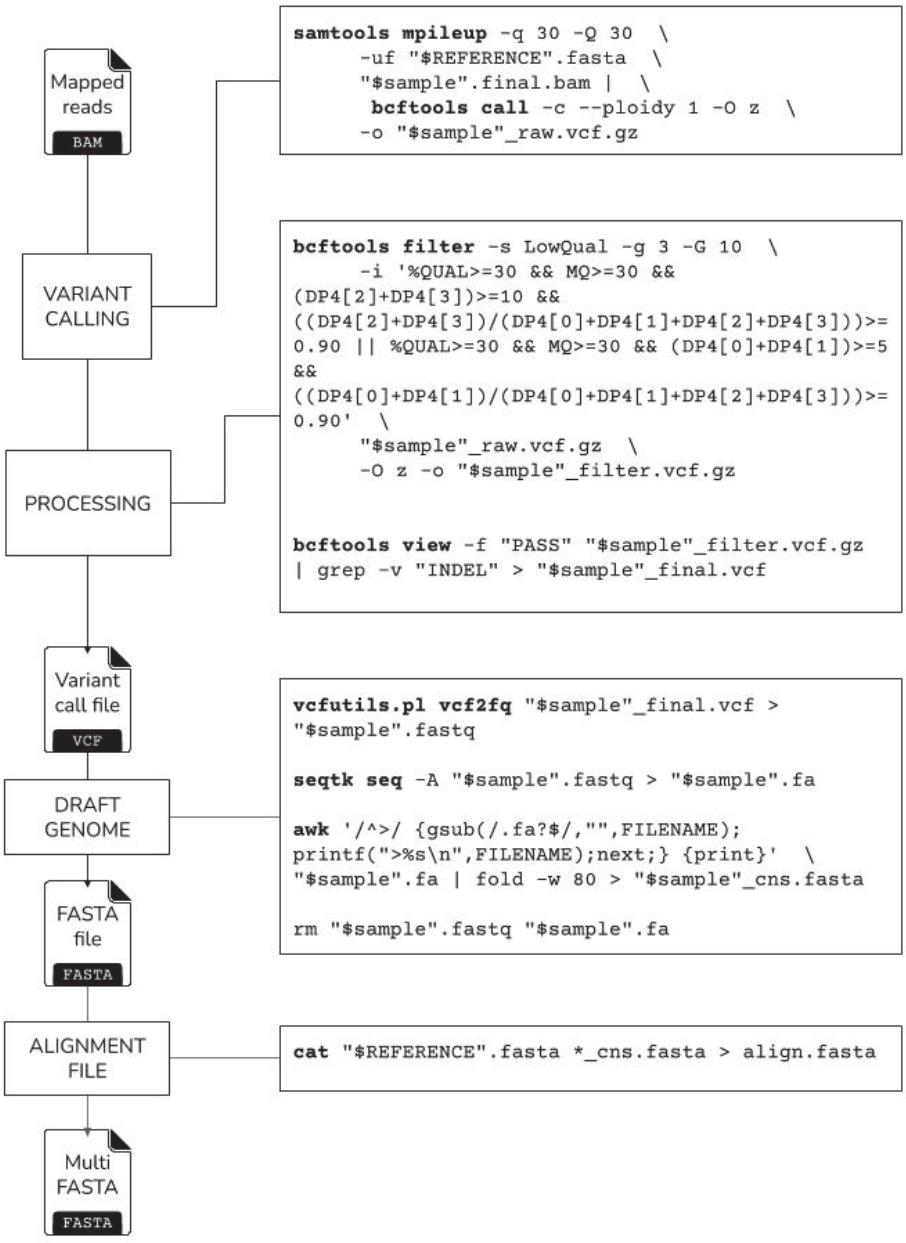
1. Primary analysis



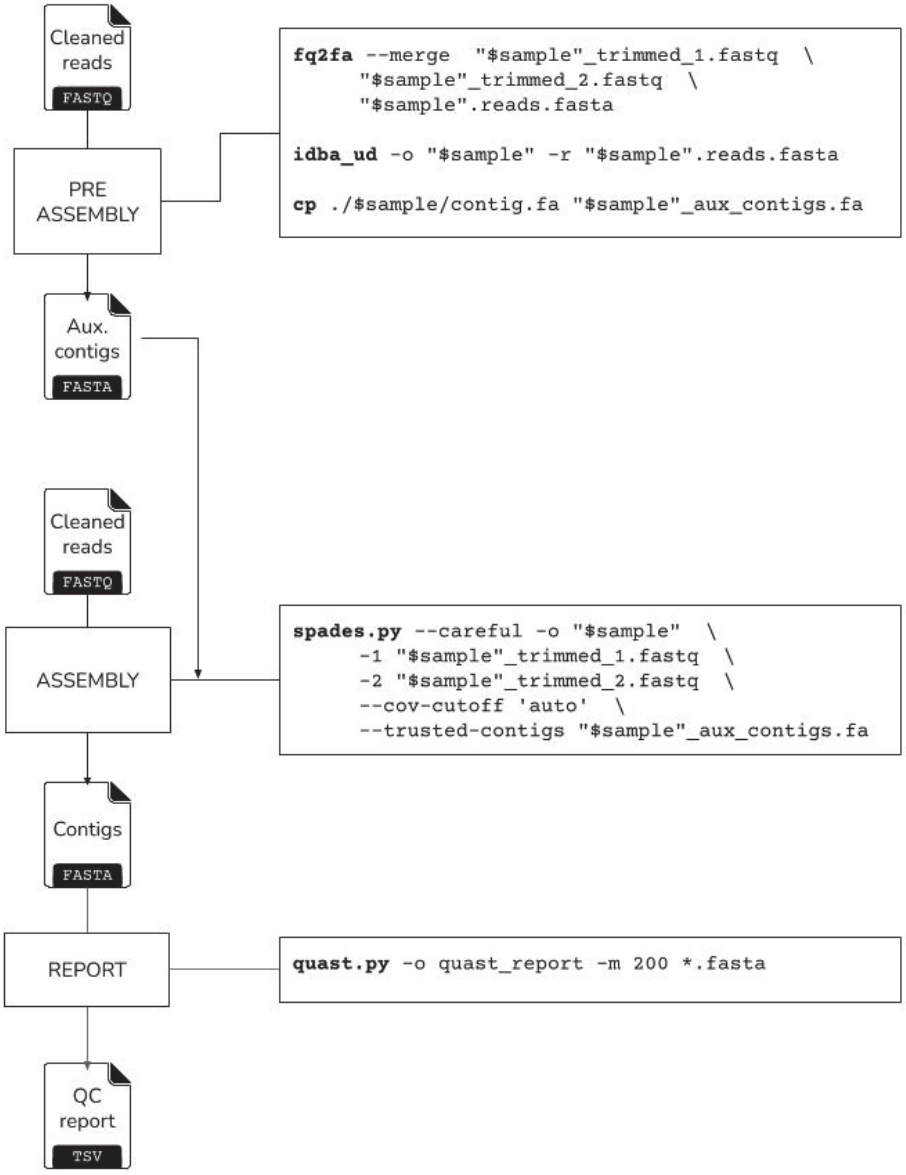
2. Secondary analysis: Mapping



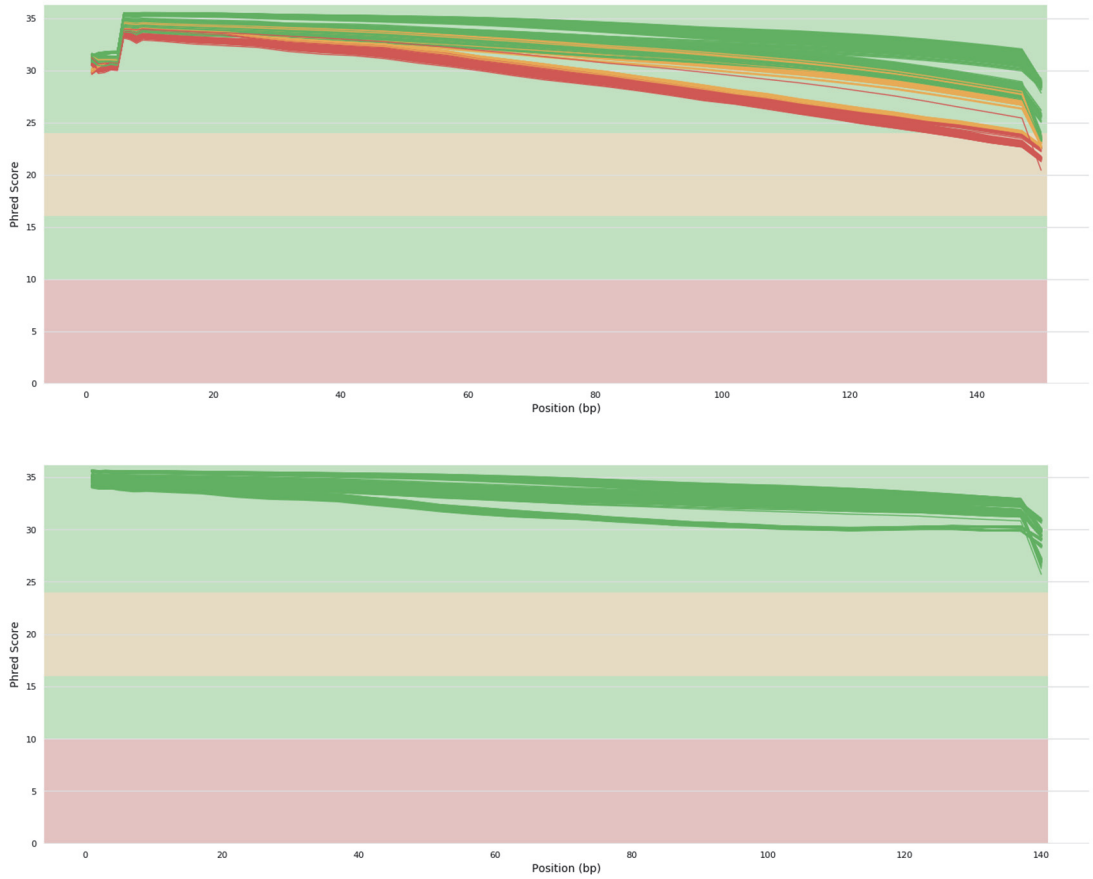
3. Secondary analysis: Variant Calling



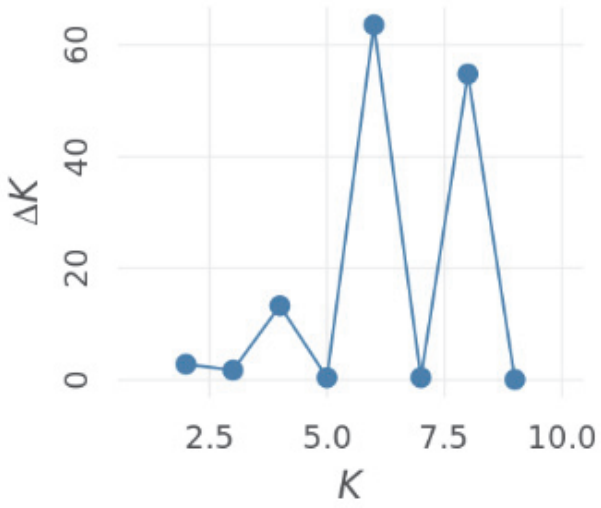
4. Secondary analysis: Assembly



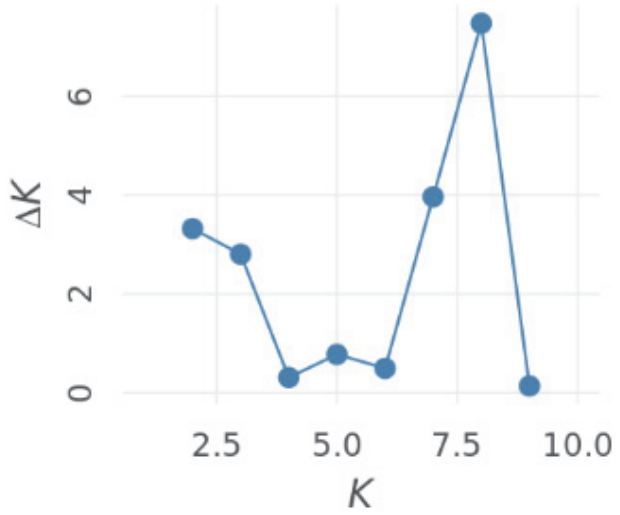
III. SUPPLEMENTARY MATERIAL FOR CHAPTER 1



Supplementary Figure 1 | Quality of the raw reads by nucleotide of *Neisseria gonorrhoeae* isolates (top) and after quality filters applying (bottom).



Supplementary Figure 2 | Plot of the Evanno method to estimate the number of *K* populations of the gonococcal dataset before recombinant genes removal. The peak in *K*=6 indicates that the dataset has 6 populations.



Supplementary Figure 3 | Plot of the Evanno method to estimate the number of *K* populations of the gonococcal dataset after recombinant genes removal. The peak in *K*=8 indicates that the dataset has 8 populations.

Supplementary Table 1 | Metadata for gonococci dataset.

Available at:

https://docs.google.com/spreadsheets/d/13KhYsO2YRV5NwnJkohF7D5CIPbS641nhCURCBuVHLG8/edit?usp=sharing

Supplementary Table 2 | Quality control and mapping statistics.

Available at:

https://docs.google.com/spreadsheets/d/1X3A3bmketpZaqk07qDiQvv97yN6lOuYSb1iPy6OIBkE/edit?usp=sharing

Supplementary Table 3 | Recombinant genes of *Neisseria gonorrhoeae*.

Available at:

https://docs.google.com/spreadsheets/d/15HkYuFqcH7Sft1QoFhANhAuaA924A40D2upm0peXQF8/edit?usp=sharing

Supplementary Table 4 | NG-MAST genogroups that include multiple STs.

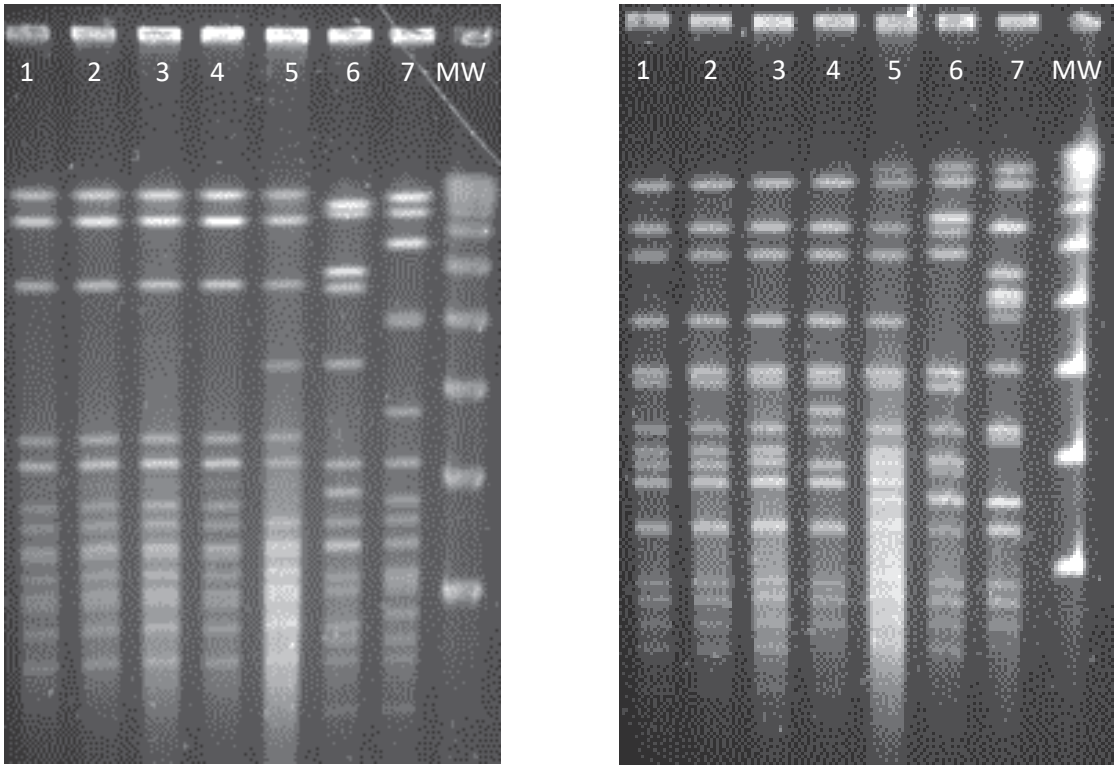
Genogroup (n)	**STs (n)**
G3378 (34)	1407 (5), 2212 (1), 3128 (1), 3149 (1), 3378 (12), 4120 (6), 8993 (2), 5619 (1), 5622 (1), 8921 (1), 11990 (1), 13123 (1)
G2400 (29)	2400 (16), 4943 (1), 6360 (7), 9184 (4), 14305 (1)
G2992 (28)	2992 (23), 5119 (3), 7636 (2)
G5441 (16)	5441 (14), 13489 (1), 18362 (1)
G11461 (11)	11461 (9), 14764 (2)
G21 (14)	5 (2), 21 (7), 1034 (2), 5445 (1), 8329 (2)
G437 (10)	225 (3), 437 (5), 289 (1), 880 (1)
G11547 (7)	7072 (1), 11547 (4), 13070 (2)
G4186 (6)	4186 (5), 15708 (1)
G5526 (5)	5526 (2), 6902 (1), 10688 (1), 12790 (1)
G51 (3)	51 (2), 881 (1)
G3935 (2)	3935 (1), 6765 (1)
G26 (2)	26 (1), 5364 (1)

Supplementary Table 5 | Comparison between ARIBA and the BLAST-based method results.

Available at:

https://docs.google.com/spreadsheets/d/1lawlHgoorqYi-0uiksrA6L-l8fQqOexPX3Y6rgmlvwE/edit?usp=sharing

IV. SUPPLEMENTARY MATERIAL FOR CHAPTER 2



Supplementary Figure 4 | PFGE profiles of *Neisseria gonorrhoeae* isolates after *Nhe*I (left) and *Spe*I (right) restriction. 1: victim isolate; 2: suspect isolate; 3: local control isolate 3 (LC3); 4: LC1; 5: LC2; 6: LC4; 7: LC5; MW: DNA molecular weight control (50kb ladder). The isolates from 1 to 5 were ST9363. Isolates 6 and 7 were ST7827 and ST7363, respectively and they were excluded from the study. PFGE patterns of isolates 1 to 4 after *Nhe*I restriction (left) were indistinguishable. PFGE patterns of isolates 1 to 3 were indistinguishable after *Spe*I (right).

Supplementary Table 6 | Number of reads before and after cleaning, and mapping statistics for the 31 *N. gonorrhoeae* samples analyzed.

			N. reads		Mapped reads		
ID	**Origin**	**MLST**	**Raw**	**Cleaned**	**N.**	**%**	**Coverage**
V	Don	9363	935,938	856,274	741,791	86.63	46.15
S	Don	9363	1,691,820	1,554,926	1,428,694	91.88	89.01
LC 1	Don	9363	2,860,862	2,613,098	2,102,826	80.47	131.02
LC 2	Don	9363	1,648,870	1,523,371	1,266,352	83.13	79.10
LC 3	Don	9363	1,614,184	1,510,483	1,322,620	87.56	82.57
AC 1	CV	9363	1,782,102	1,528,596	1,358,616	88.88	80.14
AC 2	CV	9363	1,312,538	1,143,083	1,088,076	95.19	64.64
AC 3	CV	9363	1,823,716	1,613,134	1,419,806	88.02	84.52
AC 4	CV	9363	2,122,428	1,841,661	1,308,302	71.04	77.34
AC 5	CV	9363	1,349,496	1,184,541	1,039,879	87.79	61.84
AC 6	CV	9363	1,645,828	1,432,996	1,305,494	91.10	77.43
AC 7	CV	9363	1,249,748	1,081,253	998,968	92.39	59.05
AC 8	CV	9363	1,761,924	1,536,652	1,459,616	94.99	86.83
AC 9	CV	9363	1,206,800	1,059,289	1,007,631	95.12	59.94
AC 10	CV	9363	1,438,902	1,209,761	1,181,816	97.69	68.80
AC 11	CV	9363	1,661,282	1,451,672	1,390,344	95.78	82.67

Continued on next page.

			N. reads		Mapped reads		
ID	**Origin**	**MLST**	**Raw**	**Cleaned**	**N.**	**%**	**Coverage**
AC 12	CV	9363	1,199,568	1,045,723	1,002,558	95.87	59.46
AC 13	CV	9363	1,535,454	1,338,361	1,284,042	95.94	76.19
AC 14	CV	9363	1,288,880	1,135,164	1,091,028	96.11	64.87
AC 15	CV	9363	1,293,178	1,130,548	1,086,659	96.12	64.45
AC 16	CV	9363	1,014,218	880,028	791,393	89.93	46.65
AC 17	Cat	9363	1,401,892	1,251,724	1,051,715	84.02	61.78
AC 18	Cat	9363	819,974	640,054	472,935	73.89	25.30
AC 19	Cat	9363	1,026,874	982,405	905,365	92.16	48.38
AC 20	Cat	9363	921,218	867,286	770,621	88.85	40.62
AC 21	Cat	9363	499,996	469,055	399,334	85.14	21.94
AC 22	Cat	9363	866,460	814,282	706,440	86.76	38.81
AC 23	Cat	9363	718,610	733,549	705,572	96.19	34.96
AC 24	Cat	9363	665,070	653,857	584,235	89.35	30.42
AC 25	Cat	9363	901,550	837,001	765,942	91.51	42.00
AC 26	Cat	9363	1,298,624	1,242,078	1,121,979	90.33	59.87

V: victim; S: suspect; LC: local control; AC: additional control; Don: Donostia; CV: Comunidad Valenciana; Cat: Catalonia

Supplementary Table 7 | Pairwise distance matrix between the isolates.

Available at:

https://docs.google.com/spreadsheets/d/1GjToXnfVD5DFPfNZ7fOXaROSakytQyQ2xFyYLD0iVg4/edit?usp=sharing

Supplementary Table 8 | Quality control of the *de novo* assemblies for the accessory genome of victim/suspect and the two closest local control isolates.

	VICTIM	**SUSPECT**	**LC 3**	**LC 1**
Total contigs	17	24	25	23
Total length	70,043	68,173	71,410	71,421
N. contigs ≥ 250 bp	10	12	13	12
Total length ≥ 250 bp	69,029	65,952	69,812	65,642
Largest contig (bp)	56,062	55,868	56,062	55,868
N_{50}	56,062	55,868	56,062	55,868
N's per 100 Kb	0.00	0.00	0.00	0.00

Supplementary Table 9 | Cluster of orthologous genes of the accessory genome (excluding plasmid) of the case isolates and the two closest controls.

Available at:

https://drive.google.com/file/d/1bfAfOx9pWZU_q3D0my_z5CzpwblwXGlF/view?usp=sharing

Supplementary Table 10 | Best BLAST results for the identification of plasmids with the unmapped reads of the suspect, victim and the two most similar controls (LC 1 and LC 3).

Available at:

https://docs.google.com/spreadsheets/d/1tO2A_nZivVisa86nyRuayhduj1aDYS5Likgrc2Esgqg/edit?usp=sharing

V. SUPPLEMENTARY MATERIAL FOR CHAPTER 3

Supplementary information | Antimicrobial susceptibility profile of the isolates of the outbreak B (same profile for all isolates).

- **Penicillins:** ampicillin (R), amoxicillin/clavulanic acid (R), piperacillin/ tazobactam (S).
- **Carbapenems:** imipenem (S), ertapenem (S).
- **ESCs:** ceftazidime (S), cefotaxime (S), cefuroxime (R), cefepime (S), cefoxitin (I), cefuroxime axetil (R).
- **Fluoroquinolones:** ciprofloxacin (S), nalidixic acid (S).
- **Aminoglycosides:** amikacin (S), gentamicin (S).
- **Sulfonamide:** co-trimoxazole (trimethoprim/sulfamethozaxole) (S).
- **Tetracycline:** tigecycline (S).

Supplementary Table 11 | Demographic and clinical data of outbreak A isolates.

Isolate ID	Strain name[a]	Patient	Sex[b]	Specimen type[c]	Sampling date	Discharge date[b]
P1a	SM-Elx-1	1	M	Conjunctival	09/06/17	11/13/17
P1b	SM-Elx-2	1		Blood culture	09/11/17	
P1c	SM-Elx-3	1		Incubator tubing	09/19/17	
P2	SM-Elx-4	2	M	Perineal smear	09/14/17	10/26/17
P3a	SM-Elx-5	3	M	Conjunctival	09/05/17	09/27/17
P3b	SM-Elx-6	3		Mother's milk	09/15/17	
P4a	SM-Elx-7	4	M	Perineal smear	09/13/17	10/02/17
P4b	SM-Elx-8	4		Incubator hood	09/15/17	
P4c	SM-Elx-9	4		Mother's milk	09/25/17	
P5	SM-Elx-10	5	F	Perineal smear	09/13/17	11/14/17
P6a	SM-Elx-11	6	F	Pharyngeal	09/13/17	10/02/17
P6b	SM-Elx-12	6		Mother's milk	09/14/17	
P7a	SM-Elx-13	7	M	Conjunctival	09/08/17	10/10/17
P7b	SM-Elx-14	7		Mother's milk	09/24/17	
P7c	SM-Elx-15	7		Mother's milk	09/21/17	
Env1	SM-Elx-16	Env	-	Milk waste container	09/15/17	

Continued on next page

Isolate ID	Strain name[a]	Patient	Sex[b]	Specimen type[c]	Sampling date	Discharge date[b]
Env2	SM-Elx-17	Env	-	NICU's sink	09/11/17	
Env3	SM-Elx-18	Env	-	NICU's sink	09/11/17	
-	SM-Elx-19	Control	-	Worker hands	09/15/17	
C	SM-Elx-20	Control	-	Mother's milk	10/18/17	
-	SM-Elx-21	Control	-	Mother's milk	10/18/17	

a. Strains SM-Elx-19 and SM-Elx-21 were identified as *S. liquefaciens* and they were removed from the study.
b. Sex and date of discharge was showed once by patient to avoid redundancy. M: male; F: female; Env: environmental.
c. Mother's milk specimens correspond to breast pumps contamination or by the mothers' manipulation during milk extraction, and not to mastitis cases.

Supplementary Table 12 | Demographic and clinical data of outbreak B isolates.

Isolate ID	Strain name	Patient	Sex	Specimen type[a]	Sampling date	Discharge date
P1	SM-Cs-6	1	F	Perianal	04/24/18	07/04/18
P2	SM-Cs-4	2	F	Peripheral catheter	04/23/18	05/14/18
P3	SM-Cs-7	3	F	Perianal	04/23/18	05/22/18
P4	SM-Cs-1	4	M	Perianal	05/07/18	05/30/18
P5	SM-Cs-5	5	F	Perineal smear	04/26/18	05/11/18
P6	SM-Cs-2	6	M	Endotracheal tube	04/17/18	07/20/18
C	SM-Cs-3	Control	-	Blood culture	10/14/17	11/07/17

a. The control isolate in this case was an unrelated case of *S. marcescens* bacteremia.

Supplementary Table 13 | List of NCBI reference genomes of *S. marcescens* used in this study.

Strain	NCBI accession	NCBI project	Collection date	Country	Source[a]
AR_0027	NZ_CP026702.1	PRJNA292901	unknown	USA	C
B3R3	NZ_CP013046.2	PRJNA299742	2011	China	E (plant)
CAV1492	NZ_CP011642.1	PRJNA246471	12/2011	USA	C (respiratory)
Db11	NZ_HG326223.1	PRJEB4201	1980	Sweden	E (insect)
FDAARGOS_65	NZ_CP026050.1	PRJNA231221	19/10/2013	USA	C (respiratory)
RSC-14	NZ_CP012639.1	PRJNA294721	2013	South Korea	E (plant)
SM39	NZ_AP013063.1	PRJDB1121	1999	Japan	C (bacteremia)
SMB2099	NZ_HG738868.1	PRJEB4597	2012	Germany	C
SmUNAM836	NZ_CP012685.1	PRJNA284857	2005	Mexico	C (respiratory)
U36365	NZ_CP016032.1	PRJNA317568	30/12/2015	India	C (urine)
UMH1	NZ_CP018915.1	PRJNA357595	11/2013	USA	C (bacteremia)
UMH2	NZ_CP018924.1	PRJNA357595	01/2014	USA	C (bacteremia)
UMH3	NZ_CP018925.1	PRJNA357595	03/2014	USA	C (bacteremia)
UMH5	NZ_CP018917.1	PRJNA357595	04/2014	USA	C (bacteremia)
UMH6	NZ_CP018926.1	PRJNA357595	07/2013	USA	C (bacteremia)
UMH7	NZ_CP018919.1	PRJNA357595	09/2013	USA	C (bacteremia)

Continued on next page.

Strain	NCBI accession	NCBI project	Collection date	Country	Source[a]
UMH8	NZ_CP018927.1	PRJNA357595	08/2013	USA	C (bacteremia)
UMH9	NZ_CP018923.1	PRJNA357595	08/2014	USA	C (bacteremia)
UMH10	NZ_CP018928.1	PRJNA357595	05/2014	USA	C (bacteremia)
UMH11	NZ_CP018929.1	PRJNA357595	05/2014	USA	C (bacteremia)
UMH12	NZ_CP018930.1	PRJNA357595	08/2014	USA	C (bacteremia)
WW4	CP003959.1	PRJNA88659	2013	Taiwan	E (paper machine)

a. C: clinical, E: environmental

Supplementary Table 14 | List of NCBI reference plasmids of *S. marcescens* used in this study.

Plasmid name	NCBI accession	*S. marcescens* strain	Length (bp)	GC content (%)
R478	NC_005211.1	-	274,762	45.50
pRK10	NC_010796.1	-	4,241	52.72
pRIO-5	NC_019267.1	-	12,957	53.35
R830b	NC_019344.1	-	81,793	53.20
unnamed1	NC_CP020506.1	95	219,979	51.58
unnamed2	NC_CP020505.1	95	47,999	45.59
unnamed3	NC_CP020504.1	95	13,878	52.95
pNDM_7209	NZ_CM008885.1	7209	87,593	56.07
p7209-15	NZ_CM008886.1	7209	11,781	50.86
pNDM_9580	NZ_CM008884.1	9580	105,614	55.15
pNDM_12TM	NZ_CM008895.1	12TM	119,046	54.15
p12TM-94	NZ_CM008896.1	12TM	80,890	51.79
p14ES-6400	NZ_CM008883.1	14ES	3,221	56.32
pNDM_4TM	NZ_CM008879.1	4TM	119,047	54.15
p4TM-92	NZ_CM008880.1	4TM	80,889	51.78
unitig_1_pilon	NZ_CP026703.1	AR_0027	22,569	65.15
pSERAS01	NZ_AP019010.1	AS-1	104,121	54.77

Continued on next page.

Plasmid name	NCBI accession	*S. marcescens* strain	Length (bp)	GC content (%)
pATCC	NZ_CP041234.1	ATCC_13880	43,151	58.42
pSM22	NC_015972.2	B-6493	43,190	58.45
unnamed1	NZ_CP013047.2	B3R3	123,171	51.67
unnamed	NZ_CP020508.1	BWH-35	204,208	51.46
pCAV1492-3223	NZ_CP011637.1	CAV1492	3,223	56.34
pCAV1492-6393	NZ_CP011638.1	CAV1492	6,393	52.70
pKPC_CAV1492	NZ_CP011639.1	CAV1492	69,158	49.21
pCAV1492-73	NZ_CP011640.1	CAV1492	73,100	53.36
pCAV1492-199	NZ_CP011641.1	CAV1492	199,444	51.13
pCAV1761-3223	NZ_CP029444.1	CAV1761	3,223	56.34
pCAV1761-6393	NZ_CP029445.1	CAV1761	6,393	52.70
pKPC_CAV1761	NZ_CP029446.1	CAV1761	69,158	49.20
pCAV1761-73	NZ_CP029447.1	CAV1761	73,100	53.36
pCAV1761-205	NZ_CP029448.1	CAV1761	204,825	51.50
pE28_001	NZ_CP042513.1	E28	186,249	51.98
pE28_002	NZ_CP042514.1	E28	87,731	52.79
pE28_003	NZ_CP042515.1	E28	67,074	56.34
pE28_004	NZ_CP042516.1	E28	2,694	46.81
unnamed	NZ_CP027797.1	EL1	38,997	34.44

Continued on next page.

Plasmid name	NCBI accession	*S. marcescens* strain	Length (bp)	GC content (%)
unnamed	NZ_CP027799.1	KS10	38,978	34.42
pSmN45	NZ_CP031315.1	N4-5	11,089	43.50
unnamed_1	NZ_CP027301.1	SGAir0764	76,484	53.96
pSMC1	NZ_AP013064.1	SM39	41,517	61.46
pSMC2	NZ_AP013065.1	SM39	58,929	51.90
pSmUNAM836	NZ_CP012686.1	SmUNAM836	26,346	43.52
unnamed1	NZ_CP018916.1	UMH1	73,532	54.90
unnamed2	NZ_CP018918.1	UMH5	100,699	52.34
unnamed3	NZ_CP018920.1	UMH7	111,810	53.09
unnamed4	NZ_CP018921.1	UMH7	47,264	46.42
unnamed5	NZ_CP018922.1	UMH7	21,738	49.05
pWVU-005-1	NZ_CP041127.1	WVU-005	91,252	51.64
pWVU-005-2	NZ_CP041128.1	WVU-005	63,265	56.28
pWVU-009	NZ_CP041133.1	WVU-009	77,267	55.19
pSmWW4	NC_020212.1	WW4	3,248	47.81

Supplementary Table 15 | Number of raw and clean reads from both outbreaks isolates.

Outbreak	Isolate ID	Raw reads	Clean reads
Outbreak A	P1a	1,115,720	830,798
	P1b	1,116,196	850,350
	P1c	1,432,380	1,066,224
	P2	900,026	669,528
	P3a	767,334	559,120
	P3b	1,118,770	809,188
	P4a	1,709,528	1,285,638
	P4b	1,463,528	1,093,300
	P4c	2,179,754	1,681,124
	P5	1,544,626	1,128,012
	P6a	1,663,634	1,252,752
	P6b	2,643,420	1,988,044
	P7a	1,499,434	1,106,532
	P7b	1,224,248	901,536
	P7c	1,599,964	1,182,196
	Env1	1,540,400	1,122,428
	Env2	1,184,466	886,596
	Env3	795,646	557,366
	C	791,326	567,684

Continued on next page.

Outbreak	Isolate ID	Raw reads	Clean reads
Outbreak B	P1	2,244,634	1,952,528
	P2	2,874,302	2,517,050
	P3	2,468,760	2,168,034
	P4	3,410,598	2,990,070
	P5	1,716,034	1,483,848
	P6	2,417,898	2,121,092
	C	2,049,380	1,779,348

Supplementary Table 16 | Mapping statistics of outbreak A against the UMH9 reference strain.

		Mapped reads		Coverage		Unmapped reads	
Isolate ID	**Total reads[a]**	**N**	**%**	**Depth (X)**	**Breadth (%)**	**N**	**%**
P1a	816,978	816,474	99.94	15.53	93.88	504	0.06
P1b	835,660	835,108	99.93	16.01	94.77	552	0.07
P1c	1,047,630	1,047,028	99.94	19.89	97.13	602	0.06
P2	659,213	658,760	99.93	12.51	88.27	453	0.07
P3a	551,130	550,571	99.90	10.42	81.26	559	0.10
P3b	796,352	795,801	99.93	14.96	93.02	551	0.07
P4a	1,261,723	1,260,979	99.94	24.12	98.42	744	0.06
P4b	1,072,956	1,072,395	99.95	20.28	97.55	561	0.05
P4c	1,643,636	1,642,675	99.94	31.83	98.90	961	0.06
P5	1,108,764	1,107,922	99.92	20.80	97.54	842	0.08
P6a	1,229,292	1,228,508	99.94	23.49	98.21	784	0.06
P6b	1,938,909	1,938,090	99.96	36.63	99.07	819	0.04
P7a	1,088,354	1,087,462	99.92	20.40	97.23	892	0.08
P7b	888,058	887,503	99.94	16.62	94.97	555	0.06
P7c	1,161,516	1,160,673	99.93	21.73	97.52	843	0.07
Env1	1,104,528	1,103,859	99.94	20.55	97.13	669	0.06
Env2	870,915	870,478	99.95	16.49	95.13	437	0.05
Env3	549,979	549,481	99.91	10.15	78.76	498	0.09
C	559,018	476,050	85.16	8.65	66.01	82,968	14.84

a. Note that the number of reads is smaller than that of clean reads specified in Supplementary Table V. This is because secondary alignments occur sometimes in the mapping step, and the reads involved are marked as supplementary not being included in the mapping statistics. For more information, see the SAM format specification (https://samtools.github.io/hts-specs/SAMv1.pdf).

Supplementary Table 17 | Quality report of the assembled accessory genome of outbreak A against UMH9 reference strain.

Isolate ID	Number of contigs	Total length	Largest contig size	N_{50}	N's per 100 Kb
C	321	655,961	12,303	2,661	0.00

Supplementary Table 18 | Mapping statistics of outbreak A against the Db11 reference strain.

		Mapped reads		Coverage		Unmapped reads	
Isolate ID	**Total reads**	**N**	**%**	**Depth (X)**	**Breadth (%)**	**N**	**%**
P1a	816,944	725,128	88.76	13.20	78.43	91,816	11.24
P1b	835,841	742,324	88.81	13.63	79.36	93,517	11.19
P1c	1,047,478	931,765	88.95	16.96	82.26	115,713	11.05
P2	659,234	584,279	88.63	10.62	72.76	74,955	11.37
P3a	551,091	488,480	88.64	8.85	66.25	62,611	11.36
P3b	796,268	701,539	88.10	12.61	77.41	94,729	11.90
P4a	1,261,098	1,123,461	89.09	20.59	84.04	137,637	10.91
P4b	1,072,603	959,687	89.47	17.39	82.74	112,916	10.53
P4c	1,644,013	1,447,162	88.03	26.85	84.96	196,851	11.97
P5	1,108,465	977,802	88.21	17.57	82.68	130,663	11.79
P6a	1,229,104	1,090,106	88.69	19.96	83.74	138,998	11.31
P6b	1,939,399	1,712,957	88.32	31.01	85.35	226,442	11.68
P7a	1,087,833	970,443	89.21	17.44	82.44	117,390	10.79
P7b	887,773	784,479	88.36	14.06	79.63	103,294	11.64
P7c	1,161,125	1,035,335	89.17	18.58	82.84	125,790	10.83
Env1	1,103,870	984,854	89.22	17.57	82.45	119,016	10.78
Env2	870,884	775,004	88.99	14.07	79.82	95,880	11.20
Env3	549,861	488,125	88.77	8.64	64.27	61,736	11.23
C	559,238	520,991	93.16	9.50	74.79	38,247	6.84

Supplementary Table 19 | Quality report of the assembled accessory genome of outbreak A against the Db11 reference strain.

Isolate ID	Number of contigs	Total length	Largest contig size	N_{50}	N's per 100 Kb
P1a	194	531,236	20,861	4,610	0.00
P1b	191	530,385	18,583	4,611	0.00
P1c	185	533,386	20,897	4,745	0.00
P2	215	521,357	20,077	3,846	0.00
P3a	257	510,372	13,924	2,761	0.00
P3b	198	528,060	20,879	4,072	0.00
P4a	184	541,777	20,810	5,467	0.00
P4b	181	538,421	20,005	5,367	0.00
P4c	180	543,984	20,854	5,621	0.00
P5	183	538,693	20,929	5,255	0.00
P6a	186	541,659	20,834	5,158	0.00
P6b	174	545,252	20,852	5,527	0.00
P7a	186	534,264	20,937	5,524	0.00
P7b	202	531,122	20,862	4,198	0.00
P7c	193	536,857	20,902	4,289	0.00
Env1	193	534,135	20,853	4,687	0.00
Env2	203	535,314	20,855	3,960	0.00
Env3	278	501,243	10,461	2,669	0.00
C	124	318,456	12,270	3,555	0.00

Supplementary Table 20 | Mapping statistics of outbreak B against the UMH9 reference strain.

		Mapped reads		Coverage		Unmapped reads	
Isolate ID	**Total reads**	**N**	**%**	**Depth (X)**	**Breadth (%)**	**N**	**%**
P1	1,890,489	1,678,202	88.77	41.42	83.03	212,287	11.23
P2	2,427,524	2,147,808	88.48	53.06	83.63	279,716	11.52
P3	2,099,176	1,861,041	88.66	45.91	83.39	238,135	11.34
P4	2,880,415	2,549,250	88.50	63.01	83.82	331,165	11.50
P5	1,443,503	1,283,135	88.89	31.54	81.03	160,368	11.11
P6	2,049,984	1,816,265	88.60	44.90	83.24	233,719	11.40
C	1,732,597	1,456,290	84.05	35.98	83.90	276,307	15.95

Supplementary Table 21 | Quality report of the assembled accessory genome of outbreak B against the UMH9 reference strain.

Isolate ID	Number of contigs	Total length	Largest contig size	N50	N's per 100 Kb
P1	184	530,869	24,466	4,795	0.00
P2	188	529,834	24,469	4,620	0.00
P3	183	528,691	24,464	4,928	0.00
P4	186	533,575	24,466	5,604	0.00
P5	178	521,013	24,466	4,881	0.00
P6	183	526,175	24,468	4,803	0.00
C	205	720,630	58,394	8,061	0.00

Supplementary Table 22 | Mapping statistics of outbreak B against the Db11 reference strain.

		Mapped reads		Coverage		Unmapped reads	
Isolate ID	**Total reads**	**N**	**%**	**Depth (X)**	**Breadth (%)**	**N**	**%**
P1	1,891,403	1,711,754	90.50	41.69	84.90	179,649	9.50
P2	2,429,139	2,190,910	90.19	53.41	85.48	238,229	9.81
P3	2,100,416	1,899,465	90.43	46.26	85.24	200,951	9.57
P4	2,882,611	2,598,957	90.16	63.39	85.63	283,654	9.84
P5	1,444,195	1,308,972	90.64	31.74	82.96	135,223	9.36
P6	2,051,110	1,852,531	90.32	45.19	85.10	198,579	9.68
C	1,733,278	1,437,349	82.93	34.74	80.63	295,929	17.07

Supplementary Table 23 | Quality report of the assembled accessory genome of outbreak B against the Db11 reference strain.

Isolate ID	Number of contigs	Total length	Largest contig size	N_{50}	N's per 100 Kb
P1	120	416,301	34,150	6,418	0.00
P2	124	416,978	34,150	6,016	0.00
P3	126	415,863	34,145	6,365	0.00
P4	122	416,911	34,353	6,268	0.00
P5	123	413,923	34,149	6,405	0.00
P6	121	414,515	34,145	6,405	0.00
C	194	774,488	58,370	10,225	0.00

VI. SUPPLEMENTARY MATERIAL FOR CHAPTER 4

Supplementary information | Antimicrobial susceptibility profile of the Lg-Granada strain.

- Penicillin (I).
- Cefotaxime (S).
- Erythromycin (S).
- Vancomycin (S).
- Daptomycin (S).
- Levofloxacin (S).
- Clindamycin (R).

Supplementary Table 24 | Unique genes of Lg-Granada strain.

Location	Locus	Product	GO Term		
			Cellular component	Molecular function	Biological process
Chromosome	_00042	Hypothetical protein			
Chromosome	_00224	Hypothetical protein			
Chromosome	_00280	Mevalonate kinase	Cytoplasm	Catalytic activity/Binding	Biosynthetic process
Chromosome	_00410	Hypothetical protein			
Chromosome	_00417	Hypothetical protein			
Chromosome	_00430	Hypothetical protein			
Chromosome	_00434	Hypothetical protein			
Chromosome	_00438	Hypothetical protein			
Chromosome	_00455	Hypothetical protein			
Chromosome	_00668	Hypothetical protein			
Chromosome	_00669	Hypothetical protein			
Chromosome	_00670	Hypothetical protein			
Chromosome	_00682	Hypothetical protein			
Chromosome	_00685	Hypothetical protein			
Chromosome	_00687	Phage terminase, small subunit			
Chromosome	_00688	Phage terminase			
Chromosome	_00689	Phage portal protein			
Chromosome	_00691	Phage capsid family protein			
Chromosome	_00693	Phage gp6-like head-tail connector protein			

Continued on next page.

Location	Locus	Product	GO Term		
			Cellular component	Molecular function	Biological process
Chromosome	_00694	Hypothetical protein			
Chromosome	_00695	Hypothetical protein			
Chromosome	_00696	Hypothetical protein			
Chromosome	_00697	Hypothetical protein			
Chromosome	_00698	Hypothetical protein			
Chromosome	_00699	Chromosome partition protein Smc			
Chromosome	_00700	Phage tail protein			
Chromosome	_00703	Hypothetical protein			
Chromosome	_00709	Hypothetical protein			
Chromosome	_00814	Hypothetical protein			
Chromosome	_00967	Host cell surface-exposed lipoprotein			
Chromosome	_00978	PD-(D/E)XK nuclease superfamily protein			
Chromosome	_00979	Hypothetical protein			
Chromosome	_00986	Hypothetical protein			
Chromosome	_00988	Hypothetical protein			
Chromosome	_00990	Hypothetical protein			
Chromosome	_00997	Phage terminase large subunit			
Chromosome	_01001	Hypothetical protein			
Chromosome	_01003	Hypothetical protein			
Chromosome	_01029	Hypothetical protein			
Chromosome	_01242	Hypothetical protein			

Continued on next page.

Location	Locus	Product	GO Term		
			Cellular component	Molecular function	Biological process
Chromosome	_01249	Hypothetical protein			
Chromosome	_01299	Hypothetical protein			
Chromosome	_01313	Hypothetical protein			
Chromosome	_01324	Hypothetical protein			
Chromosome	_01325	ERF superfamily protein			
Chromosome	_01390	Hypothetical protein			
Chromosome	_01427	Hypothetical protein			
Chromosome	_01560	4-deoxy-L-threo-5-hexosulose-uronate ketol-isomerase 1		Catalytic activity/Binding	Catabolic process
Chromosome	_01692	Potassium/sodium uptake protein NtpJ	Membrane	Transporter activity	Transport
Chromosome	_01977	Serine/threonine exchanger SteT	Membrane	Transporter activity	
Chromosome	_02072	Hypothetical protein			
Chromosome	_02085	Hypothetical protein		Binding	
Chromosome	_02167	Hypothetical protein			
Plasmid	_02184	Hypothetical protein	Membrane		Transport
Plasmid	_02190	Hypothetical protein			
Plasmid	_02191	Hypothetical protein			
Plasmid	_02198	Hypothetical protein			
Plasmid	_02200	Hypothetical protein			
Plasmid	_02201	Hypothetical protein			
Plasmid	_02202	Hypothetical protein			

Continued on next page.

Location	Locus	Product	GO Term		
			Cellular component	Molecular function	Biological process
Plasmid	_02209	Hypothetical protein			
Plasmid	_02210	Hypothetical protein			
Plasmid	_02215	Zinc-transporting ATPase	Membrane	Catalytic activity/Binding	Transport
Plasmid	_02244	Hypothetical protein			
Plasmid	_02245	Hypothetical protein			
Plasmid	_02246	Hypothetical protein			
Plasmid	_02247	Integrase core domain protein			

Supplementary Table 25 | Genes of Lg-Granada strain shared exclusively with other *Lactococcus garvieae* strains grouped by their isolation source.

Location	Locus	Isolation source of the other strains	Strains	Product	GO Term		
					Cellular component	Molecular function	Biological process
Chromosome	_01304	Human	21881	Hypothetical protein			
Chromosome	_00041	Animal	DSM 20684; M79	Hypothetical protein			
Chromosome	_00221	Animal	8831; PAQ102015-99	Hypothetical protein			
Chromosome	_00222	Animal	8831; DSM 20684; M79; NBRC 100934; PAQ102015-99; TRF1	Lipoteichoic acid synthase	Membrane	Catalytic activity/Binding	Cell wall organization
Chromosome	_00437	Animal	DCC43	Hypothetical protein			
Chromosome	_00443	Animal	LG9; M79; NBRC 100934; UNIUD074	HNH endonuclease			
Chromosome	_00444	Animal	DSM 20684; LG9; M79; NBRC 100934; UNIUD074	Phage terminase, small subunit			
Chromosome	_00445	Animal	DSM 20684; LG9; M79; NBRC 100934; UNIUD074	Phage Terminase			

Continued on next page.

Location	Locus	Isolation source of the other strains	Strains	Product	GO Term		
					Cellular component	Molecular function	Biological process
Chromosome	_00446	Animal	DSM 20684; LG9; M79; NBRC 100934; UNIUD074	Phage portal protein			
Chromosome	_00447	Animal	DCC43; DSM 20684; LG9; M79; NBRC 100934; UNIUD074	ATP-dependent Clp protease proteolytic subunit 1	Cytoplasm	Catalytic activity	
Chromosome	_00448	Animal	DSM 20684; LG9; NBRC 100934; UNIUD074	Phage capsid family protein			
Chromosome	_00449	Animal	DSM 20684; LG9; M79; NBRC 100934; UNIUD074	Phage gp6-like head-tail connector protein			
Chromosome	_00450	Animal	LG9; M79; NBRC 100934; UNIUD074	Phage head-tail joining protein			
Chromosome	_00451	Animal	DSM 20684; LG9; NBRC 100934; UNIUD074	Hypothetical protein			
Chromosome	_00452	Animal	DSM 20684; LG9; NBRC 100934; UNIUD074	Hypothetical protein			
Chromosome	_00448	Animal	DSM 20684; LG9; NBRC 100934; UNIUD074	Phage capsid family protein			

Continued on next page

Location	Locus	Isolation source of the other strains	Strains	Product	GO Term		
					Cellular component	Molecular function	Biological process
Chromosome	_00449	Animal	DSM 20684; LG9; M79; NBRC 100934; UNIUD074	Phage gp6-like head-tail connector protein			
Chromosome	_00450	Animal	LG9; M79; NBRC 100934; UNIUD074	Phage head-tail joining protein			
Chromosome	_00451	Animal	DSM 20684; LG9; NBRC 100934; UNIUD074	Hypothetical protein			
Chromosome	_00453	Animal	DSM 20684; LG9; M79; NBRC 100934; UNIUD074	Phage major tail protein			
Chromosome	_00454	Animal	LG9; UNIUD074	Hypothetical protein			
Chromosome	_00456	Animal	LG9; UNIUD074	Chromosome partition protein Smc			
Chromosome	_00666	Animal	DCC43	Hypothetical protein			
Chromosome	_00667	Animal	DCC43; NBRC 100934	Hypothetical protein			
Chromosome	_00686	Animal	TRF1	HNH endonuclease			
Chromosome	_00991	Animal	DCC43	Hypothetical protein			
Chromosome	_01019	Animal	DCC43	Hypothetical protein			
Chromosome	_01074	Animal	NBRC 100934; UNIUD074	Hypothetical protein			
Chromosome	_01278	Animal	UNIUD074	Hypothetical protein			

Continued on next page

Location	Locus	Isolation source of the other strains	Strains	Product	GO Term		
					Cellular component	Molecular function	Biological process
Chromosome	_01295	Animal	ATCC 49156; Lg2	Hypothetical protein			
Chromosome	_01298	Animal	LG9; UNIUD074	Hypothetical protein			
Chromosome	_01307	Animal	LG9; UNIUD074	Hypothetical protein			
Chromosome	_01314	Animal	LG9; UNIUD074	Hypothetical protein			
Chromosome	_01333	Animal	LG9; M79; UNIUD074	Hypothetical protein			
Chromosome	_01391	Animal	DSM 20684	Hypothetical protein			
Chromosome	_01604	Animal	122061; ATCC 49156; Lg2	Tagatose-6-phosphate kinase		Catalytic activity/Binding	Catabolic process
Chromosome	_02074	Animal	M79	Putative BsuMI modification methylase subunit YdiO		Catalytic activity	Response to stimulus
Plasmid	_02192	Animal	TRF1	Hypothetical protein			
Plasmid	_02217	Animal	TRF1	Hypothetical protein			
Chromosome	_00164	Food	I113	L-lactate dehydrogenase 2	Cytoplasm	Catalytic activity	Metabolic process
Chromosome	_00416	Food	KS1546	Hypothetical protein			
Chromosome	_00921	Food	Tac2	Hypothetical protein			
Chromosome	_01002	Food	M14	Hypothetical protein			

Continued on next page.

Location	Locus	Isolation source of the other strains	Strains	Product	GO Term		
					Cellular component	Molecular function	Biological process
Chromosome	_01005	Food	M14	Hypothetical protein			
Chromosome	_01273	Food	IPLA 31405	Abi-like protein			
Chromosome	_01320	Food	I113	Hypothetical protein			
Chromosome	_01336	Food	KS1546	tRNA_anti-like protein			
Chromosome	_01688	Food	IPLA 31405; M14; Tac2	Putative quinone oxidoreductase YhfP	Cytoplasm	Catalytic activity/Binding	
Chromosome	_02062	Food	Tac2	Hypothetical protein			
Chromosome	_02073	Food	Tac2	Hypothetical protein			
Chromosome	_02075	Food	Tac2	Hypothetical protein			
Chromosome	_02076	Food	Tac2	Hypothetical protein			
Plasmid	_02193	Food	M14	Extracellular cysteine protease	Cell wall	Catalytic activity	Pathogenesis
Plasmid	_02194	Food	M14	Hypothetical protein			
Plasmid	_02195	Food	IPLA 31405; KS1546; Tac2	Extracellular cysteine protease	Cell wall	Catalytic activity	Pathogenesis
Plasmid	_02196	Food	KS1546; M14; Tac2	Hypothetical protein			
Plasmid	_02235	Food	KS1546; Tac2	Hypothetical protein			
Plasmid	_02236	Food	KS1546; Tac2	Hypothetical protein			

Continued on next page.

Location	Locus	Isolation source of the other strains	Strains	Product	GO Term		
					Cellular component	Molecular function	Biological process
Plasmid	_02237	Food	Tac2	Serine/threonine-protein kinase StkP	Membrane	Catalytic activity/Binding	Pathogenesis
Plasmid	_02241	Food	IPLA 31405; M14	Lactococcin-like family protein	Membrane		Defense response to bacterium
Chromosome	_00673	Soil	A1	Hypothetical protein			
Plasmid	_02218	Soil	A1	Hypothetical protein			

Supplementary Table 26 | *L. garvieae* strains used together with Lg-Granada for the species-level core genome inference (all genomes available at NCBI as of October 2018).

Strain	Length (Mb)	GC%	CDS [a]	NCBI accession [b]	Isolation source	Country
21881	2.16	37.9	2,082	NZ_AFCC00000000.1	Human (septicemia)	Spain
UNIUD074	2.17	38.7	2,017	NZ_AFHF00000000.1	Animal (rainbow trout)	Italy
8831	2.09	38.0	1,944	NZ_AFCD00000000.1	Animal (rainbow trout)	Spain
ATCC 49156	1.95	38.8	1,863	**NC_015930.1**	Animal (yellowtail)	Japan
Lg2	1.96	38.8	1,875	**NC_017490.1**	Animal (yellowtail)	Japan
TB25	2.01	38.1	1,915	NZ_AGQX00000000.1	Food (cheese)	Italy
LG9	2.08	38.5	2,014	NZ_AGQY00000000.1	Animal (rainbow trout)	Italy
IPLA 31405	2.05	38.5	1,950	NZ_AKFO00000000.1	Food (cheese)	Spain
DCC43	2.24	37.8	2,145	NZ_AMQS00000000.1	Animal (mallard duck)	Norway
I113	2.18	37.9	2,051	NZ_AMFD00000000.1	Food (meat)	Italy
Tac2	2.24	38.2	2,101	NZ_AMFE00000000.1	Food (meat)	Italy
TRF1	2.20	38.5	1,781	NZ_AVFE00000000.1	Animal (timber rattlesnake)	USA
NBRC 100934	2.03	38.5	1,923	NZ_BBJW00000000.1	Animal (bovine mastitis)	Japan

Continued on next page.

Strain	Length (Mb)	GC%	CDS [a]	NCBI accession [b]	Isolation source	Country
M14	2.25	37.7	2,125	NZ_CCXC00000000.1	Food (fermented milk)	Algeria
Lg-ilsanpaik-gs201105	1.96	38.1	1,841	NZ_JPUJ00000000.1	Human (cholecystitis)	South Korea
PAQ102015-99	2.07	38.0	1,934	NZ_LXWL00000000.1	Animal (rainbow trout)	USA
122061	2.00	38.2	1,785	**NZ_AP017373.1**	Animal (yellowtail)	Japan
M79	2.16	38.6	2,032	NZ_FOTJ00000000.1	Animal (camel rumen)	Australia
A1	2.04	38.0	1,932	NZ_NBBK00000000.1	Environment (soil)	Turkey
UBA5784	2.03	38.6	2,048	DIEC00000000.1	NA	NA
DSM 20684	2.02	38.5	1,920	NZ_JXJV00000000.1	Animal (bovine mastitis)	NA
KS1546	2.18	37.8	2,072	NZ_LTDA00000000.1	Food (milk)	Norway
UBA11300	1.95	38.6	1,892	DQHM00000000.1	NA	NA

a. Coding sequence (CDS) files available from NCBI, except for strain UBA5784 which was annotated with Prokka.

b. Closed genomes accession numbers are marked with bold.

Supplementary Table 27 | ANI values for all *L. garvieae* strains used in this study.

Available at:

https://drive.google.com/file/d/1yFc3N7Hfvo6GzpU7QoAxDx3IjB-D34Wx/view?usp=sharing

Supplementary Table 28 | Recombinant genes detected in *L. garvieae* core genome.

Available at:

https://drive.google.com/file/d/1gyjZSnIiKZ-M9a_ZlcopAujYjLRYVQZo/view?usp=sharing

Supplementary Table 29 | Recombination events involving 2 or more genes at the species level.

Lg-Granada chromosome coordinates		Length	N. of genes	Lg-Granada loci
Start	**End**			
2709	9588	6880	2	_00003, _00004
16254	20222	3969	3	_00012, _00013, _00014
24674	26889	2216	2	_00026, _00027
62818	66222	3405	2	_00062, _00063
78201	82476	4276	3	_00080, _00081, _00082
84685	86632	1948	2	_00087, _00088
129573	132674	3102	2	_00129, _00130
175158	177096	1939	2	_00187, _00188
177770	180189	2420	2	_00190, _00191
219882	225832	5951	4	_00225, _00226, _00227, _00228
242604	250318	7715	5	_00246, _00247, _00248, _00249, _00250
278266	279982	1717	2	_00281, _00282
304687	307686	3000	3	_00303, _00304, _00305
340310	342017	1708	2	_00345, _00346
374214	375529	1316	2	_00372, _00373
411082	413186	2105	2	_00401, _00402
457739	460636	2898	2	_00464, _00465
532812	534785	1974	2	_00538, _00539
549205	551471	2267	2	_00557, _00558
552895	560287	7393	4	_00560, _00561, _00562, _00563
581786	585742	3957	3	_00588, _00589, _00590
600393	602675	2283	2	_00606, _00607
608383	611348	2966	2	_00614, _00615
620122	621194	1073	2	_00625, _00626

Continued on next page.

Lg-Granada chromosome coordinates		Length	N. of genes	Lg-Granada loci
Start	End			
633523	635177	1655	2	_00637, _00638
641726	644622	2897	2	_00645, _00646
648292	651666	3375	2	_00653, _00654
745189	747608	2420	2	_00770, _00771
772344	775335	2992	2	_00793, _00794
870803	874597	3795	3	_00896, _00897, _00898
941283	944095	2813	2	_00955, _00956
945366	946772	1407	2	_00958, _00959
951462	953345	1884	2	_00964, _00965
1001928	1003196	1269	2	_01039, _01040
1013927	1016467	2541	3	_01053, _01054, _01055
1040914	1044546	3633	2	_01085, _01086
1060210	1064131	3922	2	_01101, _01102
1077709	1081198	3490	4	_01119, _01120, _01121, _01122
1083756	1088634	4879	3	_01126, _01127, _01128
1090986	1093629	2644	2	_01131, _01132
1124729	1125680	952	2	_01161, _01162
1148893	1152000	3108	2	_01185, _01186
1174746	1176756	2011	2	_01211, _01212
1185208	1188335	3128	4	_01221, _01222, _01223, _01224
1196664	1201230	4567	4	_01235, _01236, _01237, _01238
1202158	1203308	1151	2	_01240, _01241
1275760	1277393	1634	2	_01339, _01340
1300809	1302942	2134	2	_01367, _01368
1311247	1317799	6553	4	_01378, _01379, _01380, _01381

Continued on next page.

Lg-Granada chromosome coordinates		Length	N. of genes	Lg-Granada loci
Start	End			
1345889	1351015	5127	2	_01415, _01416
1364573	1366305	1733	2	_01431, _01432
1368310	1370327	2018	2	_01437, _01438
1371717	1374285	2569	2	_01440, _01441
1375691	1381160	5470	4	_01443, _01444, _01445, _01446
1382947	1388274	5328	4	_01448, _01449, _01450, _01451
1392697	1394595	1899	2	_01457, _01458
1400515	1407025	6511	5	_01464, _01465, _01466, _01467, _01468
1421028	1425222	4195	3	_01480, _01481, _01482
1439637	1443857	4221	3	_01496, _01497, _01498
1444514	1446209	1696	2	_01501, _01502
1450242	1455737	5496	4	_01507, _01508, _01509, _01510
1494005	1496468	2464	2	_01549, _01550
1497346	1500808	3463	4	_01552, _01553, _01554, _01555
1536483	1541079	4597	3	_01589, _01590, _01591
1552831	1554594	1764	2	_01600, _01601
1568106	1573786	5681	4	_01619, _01620, _01621, _01622
1575821	1577507	1687	2	_01626, _01627
1578694	1589304	10611	10	_01629, _01630, _01631, _01632, _01633, _01634, _01635, _01636, _01637, _01638
1596231	1597841	1611	2	_01648, _01649
1602014	1604037	2024	2	_01657, _01658
1605924	1608493	2570	3	_01662, _01663, _01664
1654803	1655868	1066	2	_01720, _01721
1707824	1710766	2943	2	_01787, _01788
1726194	1731949	5756	5	_01805, _01806, _01807, _01808, _01809

Continued on next page.

Lg-Granada chromosome coordinates		Length	N. of genes	Lg-Granada loci
Start	**End**			
1738444	1740509	2066	2	_01817, _01818
1741764	1746008	4245	4	_01820, _01821, _01822, _01823
1755023	1758654	3632	3	_01833, _01834, _01835
1789350	1792948	3599	2	_01861, _01862
1818085	1820323	2239	2	_01886, _01887
1822836	1825020	2185	2	_01890, _01891
1871820	1874783	2964	3	_01942, _01943, _01944
1885459	1887472	2014	2	_01956, _01957
1901322	1905234	3913	2	_01973, _01974
1907503	1909714	2212	2	_01979, _01980
2036949	2040184	3236	2	_02116, _02117
2040624	2043133	2510	2	_02119, _02120

Supplementary Table 30 | *Lactococcus* strains used together with Lg-Granada for the genus-level core genome inference.

Species	Strain	Length (Mb)	GC%	CDS	NCBI accession	Isolation source [a]	Country
L. lactis subsp. *lactis*	IL1403	2.37	35.3	2,219	NC_002662.1	Dairy fermentation	NA
L. lactis subsp. *lactis*	275	2.75	35.4	2,579	NZ_CP015897.1	Dairy fermentation	NA
L. lactis subsp. *cremoris*	MG1363	2.53	35.7	2,319	NC_009004.1	NA	UK
L. lactis subsp. *cremoris*	JM1	2.75	35.8	2,415	NZ_CP015899.1	Dairy fermentation	Ireland
L. raffinolactis	WiKim0068	2.29	39.7	2,141	NZ_CP023392.1	Fermented vegetables	South Korea
L. piscium	MKFS47	2.50	38.6	2,375	NZ_LN774769.1	MAP broiler filet strips	Finland

a. MAP: modified-atmosphere packaging.

Supplementary Table 31 | ANI values for all *Lactococcus* species used in this study.

Available at:

https://drive.google.com/file/d/1smfn3sF_2sWD7cYBMGbGgdTll_clVNip/view?usp=sharing

Supplementary Table 32 | Recombinant genes detected in the *Lactococcus* core genome.

Available at:

https://drive.google.com/file/d/1VWyIvcykahztYJH-Txu3Ukw1TqnSnZ_y/view?usp=sharing

Supplementary Table 33 | Recombination events involving 2 or more genes at the genus level.

Lg-Granada chromosome coordinates		Length	N. of genes	Lg-Granada loci
Start	**End**			
441906	443807	1902	4	_00450, _00451, _00452, _00453 (From Lg-Granada to *L. lactis* subsp. *cremoris* MG1363)
532812	534785	1947	2	_00538, _00539 (From Lg-Granada to *L. lactis* subsp. *lactis* 275)
1843389	1845894	2506	2	_01916, _01917 (From *L. lactis* to Lg-Granada)

Supplementary Table 34 | Other members of *Bacilli* class used together with Lg-Granada for the class-level core genome inference.

Species	Strain	Length (Mb)	GC%	CDS	NCBI accession	Isolation source	Country
Oceanobacillus iheyensis	HTE831	3.63	35.7	3,460	NC_004193.1	Deep-sea sediment	Japan
Listeria monocytogenes	EGD-e	2.94	38.0	2,867	NC_003210.1	Murine macrophage	France
Listeria ivanovii	WSLC 30151	3.05	37.1	2,860	NZ_CP009576.1	NA	NA
Listeria welshimeri	SLCC5334	2.81	36.4	2,721	NC_008555.1	Plants	USA
Aerococcus viridans	CCUG4311	2.20	39.4	1,887	NZ_CP014164.1	Air sample	UK
Carnobacterium maltaromaticum	LMA28	3.65	34.5	3,274	NC_019425.2	Cheese	France
Enterococcus faecium	DO	3.05	37.9	2,765	NC_017960.1	Human	USA
Enterococcus faecalis	V583	3.36	37.35	3,172	NC_004668.1	Human	USA
Tetragenococcus halophilus	NBRC 12172	2.56	36.0	2,377	NC_016052.1	Soy sauce	Japan
Vagococcus penaei	CD276	2.37	35.1	2,123	NZ_CP019609.1	Spoiled cooked shrimp	Pacific Ocean
Lactobacillus plantarum	WCFS1	3.35	44.5	3,023	NC_004567.2	Human saliva	The Netherlands

Continued on next page.

Species	Strain	Length (Mb)	GC%	CDS	NCBI accession	Isolation source	Country
Lactobacillus delbrueckii	ATCC 11842	1.87	49.7	1,568	NC_008054.1	Yogurt	Bulgaria
Lactobacillus reuteri	DSM 20016	2.00	38.9	1,885	NC_009513.1	Human gut	USA
Lactobacillus salivarius	UCC118	2.13	33.0	1,973	NC_007929.1	Human gut	Ireland
Lactobacillus brevis	ATCC 367	2.34	46.0	2,171	NC_008497.1	NA	NA
Leuconostoc mesenteroides	ATCC 8293	2.08	37.7	1,975	NC_008531.1	NA	NA
Streptococcus pneumoniae	R6	2.04	39.7	1,862	NC_003098.1	Human	USA
Streptococcus pyogenes	SF370	1.85	38.5	1,722	NC_002737.2	Human	USA
Streptococcus mutans	UA159	2.03	36.8	1,858	NC_004350.2	Human	USA

Supplementary Table 35 | ANI values for all species of *Bacilli* class used in this study.

Available at:

https://drive.google.com/file/d/1C2sMmrhS1g15cVjXJ7wDh9I6a-qItvSl/view?usp=sharing

Supplementary Table 36 | Recombinant genes detected in the *Bacilli* core genome.

Available at:

https://drive.google.com/file/d/1djEG8j-r3I4uD5_2kYO4Akh-GrWE5Y9S/view?usp=sharing

Supplementary Table 37 | Genes transferred to *L. garvieae* strain Lg-Granada from other species.

Lg-Granada locus	**Donor**
_00006	*Aerococcus viridans*
_00051	*Streptococcus pneumoniae*
_00065	*Tetragenococcus halophilus*
_00091	*Lactococcus lactis* subsp. *cremoris*
_00113	*Lactococcus lactis* subsp. *lactis* strain 275
_00114	*Lactococcus piscium*
_00121	*Lactococcus lactis* subsp. *lactis* strain 275
_00153	*Carnobacterium maltaromaticum*
_00160	*Lactobacillus salivarius*
_00165	*Leuconostoc mesenteroides*
_00198	*Streptococcus pneumoniae*
_00200	External
_00205	*Vagococcus penaei*
_00214	*Streptococcus mutans*
_00240	*Streptococcus* spp.
_00248	*Leuconostoc mesenteroides*
_00330	*Streptococcus* spp.
_00334	*Vagococcus penaei*

Continued on next page.

Lg-Granada locus	Donor
_00342	*Enterococcus* spp.
_00356	External
_00366	*Enterococcus faecium*
_00381	*Enterococcus faecalis*
_00384	*Streptococcus pyogenes*
_00406	*Streptococcus mutans*
_00489	*Listeria* spp.
_00500	*Vagococcus penaei*
_00523	*Streptococcus pneumoniae*
_00526	*Vagococcus penaei*
_00531	*Enterococcus* spp.
_00532	*Aerococcus viridans*
_00537	*Tetragenococcus halophilus*
_00549	*Streptococcus pneumoniae*
_00559	*Lactococcus* spp.
_00562	*Listeria welshimeri*
_00581	*Enterococcus faecium*
_00585	*Lactococcus lactis* subsp. *cremoris* strain JM1
_00588	*Lactobacillus delbrueckii*
_00589	*Streptococcus pneumoniae*

Continued on next page.

Lg-Granada locus	Donor
_00606	*Aerococcus viridans*
_00613	*Lactococcus piscium*
_00618	*Lactococcus* spp.
_00635	*Enterococcus faecium*
_00653	*Aerococcus viridans*
_00720	*Lactococcus piscium*
_00771	*Oceanobacillus iheyensis*
_00773	*Aerococcus viridans*
_00779	*Leuconostoc mesenteroides*
_00791	*Vagococcus penaei*
_00794	*Streptococcus pneumoniae*
_00795	*Carnobacterium maltaromaticum*
_00897	*Aerococcus viridans*
_00902	*Lactococcus lactis*
_00912	*Vagococcus penaei*
_00996	*Lactococcus* spp.
_01040	*Lactococcus raffinolactis*
_01041	*Lactococcus* spp.
_01070	*Streptococcus* spp.
_01089	*Lactococcus* spp.

Continued on next page.

Lg-Granada locus	Donor
_01117	*Streptococcus mutans*
_01121	*Tetragenococcus halophilus*
_01133	*Listeria* spp.
_01141	*Enterococcus faecium*
_01149	External
_01167	*Leuconostoc mesenteroides*
_01189	*Streptococcus pneumoniae*
_01196	*Lactobacillus brevis*
_01217	*Aerococcus viridans*
_01243	*Lactococcus lactis* subsp. *lactis*
_01334	*Lactococcus lactis* subsp. *cremoris* strain MG1363
_01389	*Lactobacillus reuteri*
_01447	*Leuconostoc mesenteroides*
_01472	*Aerococcus viridans*
_01482	*Enterococcus faecium*
_01507	*Vagococcus penaei*
_01509	*Leuconostoc mesenteroides*
_01526	*Aerococcus viridans*
_01600	External
_01628	*Carnobacterium maltaromaticum*

Continued on next page

Lg-Granada locus	**Donor**
_01633	*Enterococcus faecalis*
_01639	*Tetragenococcus halophilus*
_01662	*Enterococcus* spp.
_01676	*Streptococcus* spp.
_01702	*Lactobacillus* spp.
_01724	*Lactococcus lactis* subsp. *lactis* strain Il1403
_01755	*Enterococcus faecalis*
_01799	*Aerococcus viridans*
_01808	*Vagococcus penaei*
_01836	*Streptococcus mutans*
_01842	*Streptococcus mutans*
_01872	*Lactobacillus* spp.
_01884	*Vagococcus penaei*
_01916	*Lactococcus lactis*
_01917	*Lactococcus lactis*
_01924	*Streptococcus pneumoniae*
_01935	*Leuconostoc mesenteroides*
_02009	External
_02028	*Aerococcus viridans*

Continued on next page.

Lg-Granada locus	**Donor**
_02033	*Leuconostoc mesenteroides*
_02040	*Aerococcus viridans*
_02079	*Aerococcus viridans*
_02086	*Streptococcus mutans*
_02087	*Aerococcus viridans*
_02144	*Streptococcus mutans*
_02165	*Enterococcus* spp.

Supplementary Table 38 | Genes transferred from *L. garvieae* strain Lg-Granada to other species.

Lg-Granada locus	Recipient
_00031	*Leuconostoc mesenteroides*
_00051	*Lactobacillus delbrueckii*
_00198	*Lactobacillus salivarius* & *Lactobacillus reuteri*
_00214	*Leuconostoc mesenteroides*
_00244	*Lactobacillus* spp.
_00248	*Lactobacillus plantarum* & *Enterococcus* spp.
_00253	*Tetragenococcus halophilus*
_00295	*Lactobacillus reuteri* & *Lactobacillus brevis*
_00314	*Lactococcus lactis* subsp. *lactis* strain Il1403
_00331	*Lactococcus lactis* subsp. *cremoris* strain JM1
_00366	*Lactococcus piscium* & *Lactococcus raffinolactis*
_00529	*Streptococcus pneumoniae*
_00538	*Lactococcus lactis*
_00539	*Lactococcus lactis*
_00549	*Aerococcus viridans; Vagococcus penaei* & *Lactobacillus brevis*
_00558	*Lactococcus lactis* subsp. *cremoris* strain MG1363 & *Lactococcus raffinolactis*
_00579	*Lactococcus piscium* & *Lactococcus raffinolactis*

Continued on next page.

Lg-Granada locus	**Recipient**
_00581	*Lactococcus lactis* subsp. *cremoris* strain JM1
_00585	*Lactococcus piscium* & *Lactococcus raffinolactis*
_00864	*Oceanobacillus iheyensis*
_00897	Leuconostoc *mesenteroides*
_00902	*Lactococcus lactis* subsp. *lactis*
_00930	*Lactobacillus delbrueckii*
_00949	*Tetragenococcus halophilus*
_01141	*Aerococcus viridans* & *Lactobacillus plantarum*
_01196	*Oceanobacillus iheyensis*
_01359	*Leuconostoc mesenteroides*; *Oceanobacillus iheyensis* & *Vagococcus penaei*
_01483	Aerococcus *viridans*
_01509	*Enterococcus* spp.
_01524	*Tetragenococcus halophilus*
_01562	*Lactococcus lactis* subsp. *cremoris* strain JM1
_01722	*Lactococcus lactis* subsp. *cremoris* strain MG1363
_01724	*Lactococcus piscium* & *Lactococcus raffinolactis*
_01741	*Vagococcus penaei*
_01836	*Enterococcus faecalis*
_01884	*Leuconostoc mesenteroides*

Continued on next page.

Lg-Granada locus	**Recipient**
_01916	*Lactococcus lactis* subsp. *cremoris* strain MG1363 & *Lactococcus lactis* subsp. *lactis* strain Il1403
_01917	*Lactococcus lactis* subsp. *cremoris* strain MG1363 & *Lactococcus lactis* subsp. *lactis* strain Il1403
_02078	*Aerococcus viridans*
_02081	*Leuconostoc mesenteroides*
_02127	*Vagococcus penaei*