UNIVERSITAT DE BARCELONA

Epidemiology, population structure, and pathogenesis of opportunistic respiratory tract colonising bacteria

Anna Carreras Salinas

“The essential is invisible to the eye...”

The Little Prince, Antoine de Saint-Exupéry

Epidemiology, population structure, and pathogenesis of opportunistic respiratory tract colonising bacteria

Doctoral Thesis

Anna Carrera Salinas

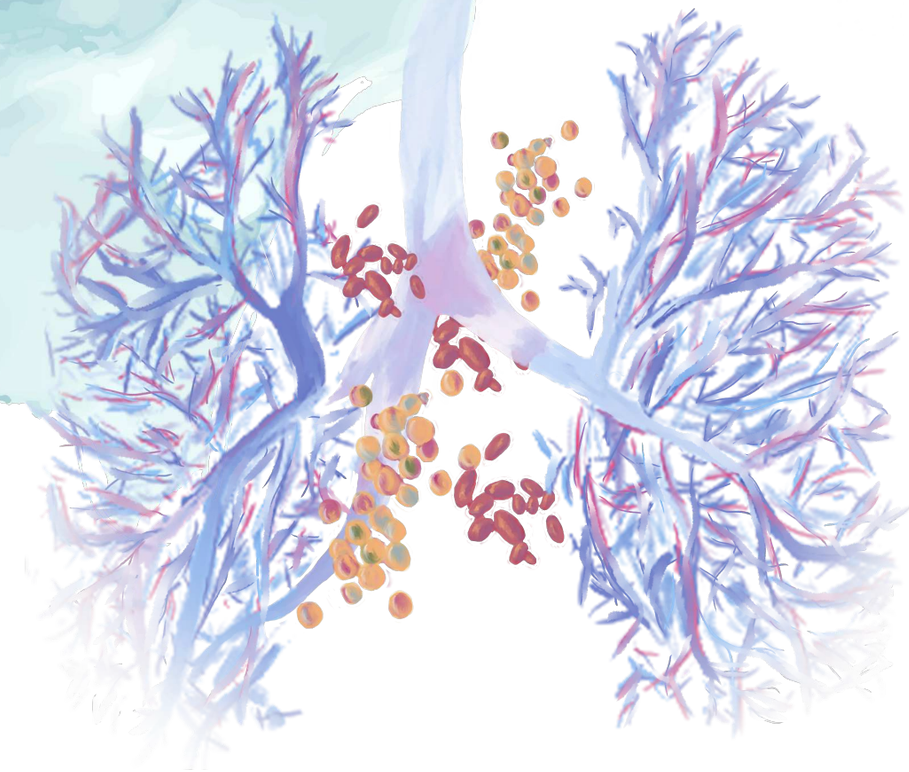
2022

Doctoral Thesis

Epidemiology, population structure, and pathogenesis of opportunistic respiratory tract colonising bacteria

Anna Carrera Salinas

September 2022



IDIBELL Institut d'Investigació Biomèdica de Bellvitge

Universitat de Barcelona

Bellvitge Hospital Universitari

UNIVERSITAT DE BARCELONA

Doctoral Programme in Biomedicine

Research area: Metabolism, metabolic signalling, and associated pathologies

Faculty of Medicine, Universitat de Barcelona

Department of Pathology and Experimental Therapy

Epidemiology, population structure, and pathogenesis of opportunistic respiratory tract colonising bacteria

Dissertation presented to obtain the Degree of Doctor from the University of Barcelona

Author	Co-supervisor	Co-supervisor
Anna Carrera Salinas	Mª Ángeles Domínguez Luzón, MD, PhD	Sara Martí Martí, PhD

Barcelona, 2022

Als meus pares, Jaume i Cristina
A la padrina de Gerb, Tresa

Acknowledgements

This thesis was supported by:

- Centro de Investigaciones Biomédicas en Red de Enfermedades Respiratorias (CIBERES). CB06/06/0037.
- Centro de Investigaciones Biomédicas en Red de Enfermedades Infecciosas (CIBERINFEC). CB21/13/00009.
- Fondo de Investigaciones Sanitarias, Instituto de Salud Carlos III:
    - PI16/00977: Análisis global del tratamiento prolongado con azitromicina en el microbioma y el metaboloma de pacientes EPOC grave y exacerbaciones frecuentes. Identificación de biomarcadores sistémicos.
    - PI16/01382: Epidemiología y patogenia de las infecciones por *Staphylococcus aureus* resistente a meticilina: Estudio de la dinámica poblacional y de la persistencia bacteriana.
- Sociedad Española de Neumología y Cirugía Torácica (SEPAR):
    - SEPAR 418/2017: Estudio de la dinámica de poblaciones microbianas en pacientes con Enfermedad Pulmonar Obstructiva Crónica grave y tratamiento prolongado con azitromicina.
    - SEPAR 1116/2020: Impacto del tratamiento continuo y prolongado con azitromicina en la composición del microbioma y la diversidad de sus productos metabólicos.

The PhD candidate was supported by:

- Formación de Profesorado Universitario, Ministerio de Educación, Cultura y Deporte (FPU16/02202).

The international stay at the Institute of Medical Microbiology in Münster (Germany) was financially supported by:

- Sociedad Española de Neumología y Cirugía Torácica (SEPAR):
    - SEPAR 418/2017.

L'experiència de fer un doctorat m'ha fet créixer com a professional, però també com a persona. Quan miro enrere i veig tot el que ha passat des de que vaig començar el doctorat, veig que he fet i he aprés moltes coses que mai m'hauria imaginat. Però tot això no hauria estat possible sense la col·laboració i l'ajuda de moltes persones, a les quals m'agradaria donar les gràcies a continuació.

A les meves dos directores de tesi, la Mª Ángeles Domínguez i la Sara Martí, moltes gràcies per tot l'esforç, per haver-me ajudat i orientat sempre que ho he necessitat, per haver confiat en mi i per haver-me acompanyat en tot aquest procés d'aprenentatge.

Als meus companys de Recerca: la Meri, el Dani, l'Aida, la Sara M, el Yanik, la Irene, la Lucía i la Sara C. Heu estat fonamentals tot aquest temps. He aprés molt de tots i treballar amb vosaltres ha estat un plaer. Però sobretot, moltes gràcies per la vostra paciència i per haver-me suportat tot aquest temps. També vull donar les gràcies als estudiants de pràctiques que heu anat passant pel nostre grup, per la vostra feina i per haver-me fet gaudir ensenyant-vos tot el que sé.

A la Carmen Ardanuy, perquè aquelles pràctiques d'estiu que vaig fer quan encara era una estudiant de grau ho van canviar tot. Sense l'oportunitat que em vas donar, avui moltes coses serien diferents.

A la Dàmaris, una de les primeres persones que vaig conèixer al laboratori. Moltes gràcies per la teva ajuda, però sobretot, per les teves abraçades!

A tots els facultatius del Servei de Microbiologia per ajudar-me sempre que ho he necessitat, tan dins com fora del laboratori, per resoldre tots els dubtes que he tingut i per donar-me noves idees i perspectives que m'han ajudat a créixer professionalment.

A tots els residents, gràcies per la vostra disposició a ajudar-me i animar-me, i per la vostra energia, entusiasme i alegria.

A tots els tècnics del Servei de Microbiologia us vull agrair que, potser sense ser-ne conscients, m'heu fet sentir una més de l'equip. El simple fet d'intercanviar un somriure,

una salutació, o quatre paraules pel passadís, m'ha fet feliç. Especialment, gràcies a la Laura Arauz, el Javi Garcia i el Fernando Rey, per la vostra amabilitat infinita i per haver-me ajudat sempre que ho he necessitat.

A la Yuliana Pascual, l'Ester Cuevas, l'Alicia Marín, la Conchita Montón i la Salud Santos, ja que a través dels diferents projectes que hem compartit, he tingut l'oportunitat d'aprendre moltíssim i adquirir un punt de vista diferent que m'ha resultat molt enriquidor.

To Silke Niemann from the Institute of Medical Microbiology of Münster, for receiving me and for your guidance in *S. aureus* and cell cultures. Also, thanks to the whole team for your help, especially Michaela, Gitte, Beate and Carmen. And to Maria and Ileana, for the good times we shared and for all your support. I learned a lot from you, and I really had a great time.

To Josh Chang Mell and Rachel Ehrlich from the Microbiology and Immunology Department of Drexel University, for your guidance on genome analysis.

A la Mariona Vinaixa i la Laura Marcos-Zambrano, perquè gràcies a la vostra ajuda he pogut aprendre un munt de coses que m'han motivat moltíssim i que pensava que mai seria capaç de fer.

A més a més, he de donar les gràcies a totes aquelles persones que heu format part d'aquest camí des d'una altra perspectiva i que m'heu donant forces per seguir endavant. M'he sentit molt estimada i estic molt orgullosa d'estar envoltada de persones tan valuoses com vosaltres. A l'Anna Requena, perquè sempre he pogut contar amb tu, perquè sempre m'has fet costat, i espero seguir veient com anem creixent, superant els obstacles i aconseguint tots aquells reptes que ens proposem. A la Marta Nogué, per les bones estones que passem juntes. Al Mikel, per entendre'm i divertir-me amb les teves històries. A la Raquel i el Carlos, per tots els sushis que m'encanta compartir amb vosaltres. Al Marc, l'Alba, l'Oriol i tots els sardanistes que m'han fet gaudir ballant sardanes.

A la Gaby, el Juan i el Víctor, and also to Hannes, Juli and Monique, per acollir-me tan bé des del primer moment i per fer-me sentir com a casa i una més de la família.

A la Cinta Fornons, el Sergi Palou, la Mari Florensa, i especialment el Joaquim Sabaté, per ser una inspiració i una referència. Heu estat una gran motivació i m'heu ensenyat a superar-me cada dia una mica més.

A tota la meva família, per donar-me suport, cuidar-me i estimar-me tant. Als padrins m'agradaria dir-vos que us estimo molt i que us trobo molt a faltar. Espero que esteu tan orgullosos de jo, com jo ho estic de vosaltres.

A la Ori, per fer-me tan feliç.

Al Yanik, pels teus consells, la teva paciència, el teu suport en els moments difícils, i per ser el millor Q del món.

I finalment, al Papa i a la Mama, per creure en mi, per haver-me animat a seguir endavant i a gaudir de tot el que faig, i per estimar-me tant. Gràcies per tot el que heu fet perquè jo pogués arribar fins aquí. Us estimo molt!

Abstract

Bacterial colonisation of the respiratory tract is critical for health by establishing a symbiotic relationship between the bacterial community and the host. ***Haemophilus influenzae*** and ***Staphylococcus aureus*** are two opportunistic microorganisms that usually colonise the respiratory tract, and while their interaction is usually asymptomatic, this colonisation can be the first step in the development of **severe infection**. This thesis focuses on the **epidemiology** (Study 1), **population structure** (Studies 1 and 2) and the **pathogenesis** (Studies 3 and 4, and Additional Results) of these microorganisms, assuming that changes in some of these factors may influence the development of serious diseases.

Epidemiology of *H. influenzae* invasive disease

Since its introduction, the conjugate vaccine against *H. influenzae* serotype b has caused a change in the epidemiology of the invasive disease caused by this microorganism. Furthermore, the rise in β-lactam resistance has led to the inclusion of *H. influenzae* in the World Health Organisation priority list, requiring epidemiological surveillance to track both the evolution of non-vaccine preventable clones and the rise in antimicrobial resistance.

Study 1 provides an update on the status of *H. influenzae* disease at Hospital Universitari de Bellvitge in 2014-2019 (n = 68) and compares its evolution with that in 2008-2013 (n = 82). The overall incidence of *H. influenzae* disease (2.07 cases per 100,000 population) remained constant between the two periods (2.12 and 2.02 cases per 100,000, respectively); however, it was lower in young adults (≤ 64 years; 0.89 cases per 100,000) than in older adults (≥ 65 years; 6.90 cases per 100,000). Comorbidity was present in 82.0% of the patients, with immunosuppression (40.0%) and solid organ malignancy (36.0%) most common. The 30-day mortality reduced from the 2008-2013 period (20.7%) to the 2014-2019 period (14.7%).

During the 2014-2019 period, the rate of ampicillin resistance increased from 10% to 17.6%, mainly due to β-lactamase TEM-1 production. Moreover, 4.4% of strains were

resistant to fluoroquinolones, implying the emergence of resistance because no fluoroquinolone-resistant strains were detected in the 2008-2013 period.

Non-typeable *H. influenzae* (NTHi) strains were the leading cause of invasive disease in both study periods (84.3% and 85.3%, respectively). Two of the most common sequence types (STs), ST103 and ST160, were associated with β-lactamase TEM-1 production, and the prevalence of both increased from the 2008-2013 period. By contrast, capsulated strains were uncommon in both periods, with serotype f the most prevalent (12.9% and 8.8%, respectively), followed by serotypes e (1.4% and 4.4%, respectively) and b (1.4% and 1.5%, respectively).

Population structure of *H. influenzae*

NTHi strains have high genetic heterogeneity and are associated with various STs, complicating the study of their population structure. Using multi-locus sequence typing (MLST), details of the genetic diversity of the NTHi population can be missed, resulting in **Study 1** proposing classification into six clades (I-VI) based on the presence and absence of 17 accessory genes. Although NTHi strains isolated from invasive disease in 2014-2019 (n = 58) belonged to 47 STs and were genetically diverse, most were clustered in clade V (n = 30). Strains from this clade shared the same phylogenetic origin and grouped the most frequent NTHi STs (ST3, ST103, ST160), including 9 of the 12 strains producing β-lactamase. Clade VI was the second most abundant (31.0%), but this was highly heterogeneous compared to clade V.

Contrasting with these results, capsulated strains were highly clonal, with each serotype associated with a limited number of lineages. However, serotypes f and a are rising among cases of invasive disease in Europe and North America, respectively. We must therefore characterise these extensively to develop, if necessary, strategies that can effectively control their possible emergence.

Study 2, therefore, characterises the genomic diversity and the pangenome of capsulated strains, including the genomes available in the National Center for

Biotechnology and Information and European Nucleotide Archive databases, focusing on serotype f by using strains isolated in different countries (Portugal, the Netherlands and Spain). MLST and phylogenetic analysis of the capsulated genomes revealed a few clonal complexes (CCs) associated with each serotype. These were highly clonal. Serotype a clones were classified into three groups: CC23 and CC1755 (clade I), CC62 and CC372 (clade II), and CC4 (clade III). Serotype b clones were mostly assigned to CC6 (n = 151), while CC50 (n = 13) and CC464 (n = 1) were clustered to a different phylogenetic group. Serotypes c and d were associated with CC7 and CC10, respectively. These showed less genetic variability than the other serotypes, probably due to the small number of genomes. Serotype e was exclusively associated with CC18. Serotype f clones were phylogenetically distributed into two groups: CC124 (n = 222) and CC16 (n = 12).

The pangenome analysis revealed that serotype b genomes had a larger gene pool (n = 3,517 genes) than the other serotypes, which ranged from 1,896 to 2,969 genes. It also had more accessory genes per genome (mean = 46.6%), followed by serotypes f (mean = 25.3%), a (mean = 25.1%), and e (mean = 20.7%). Single nucleotide polymorphism (SNP) analysis showed that the genomes of each serotype could be classified into CCs by the number of SNPs, which revealed that the total number of SNPs in serotype f was significantly lower than in serotypes a, b, and e (p-value < 0.0001), indicating low variability within the CCs of serotype f. These findings support the high genetic stability of this serotype in strains isolated from different countries, suggesting that no relationship exists between phylogeny and geographical origin.

Adaptive evolution of *H. influenzae* during persistent colonisation

Characterising bacterial adaptation during colonisation can improve our understanding of disease progression associated with *H. influenzae*, such as chronic obstructive pulmonary disease (COPD). Study 3 examines the genetic changes in this microorganism during colonisation based on the respiratory isolates from 15 patients with COPD who received prolonged azithromycin therapy (250 mg, three doses/week).

Furthermore, *H. parainfluenzae* isolates from these patients were included to investigate their role in the horizontal transmission of resistance genes.

Among the 15 patients, 4 (P02, P08, P11, and P13) had persistent *H. influenzae* colonisation, and 2 (P04 and P11) had persistent *H. parainfluenzae* colonisation. All persistent lineages isolated before treatment were susceptible to azithromycin, but antibiotic pressure led to resistance in the first few months, except for P02 who discontinued the treatment due to adverse effects. Different resistant genetic determinants were identified for each lineage. Concerning *H. influenzae*, P08 ST107 isolates acquired successive mutations in 23S rRNA, while P11 ST2480 and P13 ST165 had mutations in the genes encoding ribosomal proteins L4 and L22. Concerning *H. parainfluenzae*, persistent isolates in P04 acquired changes in the *rlmC* gene that encodes a 23S rRNA methyltransferase, whereas in P11 isolates, the genes encoding MefE and MsrD efflux pumps were found within the *tet*(M)-MEGA element, carried in an integrative conjugative element. The identification of this element in the persistent ST147 *H. influenzae* lineage isolated from the same patient supported the role of *H. parainfluenzae* as a reservoir of resistance genes.

Other genetic variation mainly occurred in genes associated with cell wall and inorganic ion transport and metabolism. Persistent *H. influenzae* strains showed phase-variable changes in *licA* and *hgpB* genes linked to lipooligosaccharide synthesis and heme uptake, respectively. Other genes (*lex1*, *lic3A*, *hgpC*, and *fadL*) had genetic variation in multiple lineages; and some regions of the genome accumulated many genetic changes or large insertions and deletions that corresponded to prophage-associated regions.

Biofilm formation in *S. aureus*

Biofilm formation allows bacteria to colonise different surfaces for extended periods in both the host and the hospital environment. Biofilms serve as reservoirs for bacteria, facilitating the transmission and spread of serious infection. Study 4 compares biofilm formation for endemic (CC5, CC8, and CC22) and sporadic (CC1, CC30, CC45, CC72,

and CC398) methicillin-resistant *S. aureus* (MRSA) clones in our healthcare setting. The endemic clones produced more biofilms than the sporadic clones, although the CC22 isolates had biofilm formation comparable to that of sporadic clones, producing less biofilm than the CC5 and CC8 isolates.

A correlation existed between biofilm and the *sasG* gene: strains carrying the gene produced more biofilm than those without the gene. Allelic variant analysis revealed that CC5 and CC8 strains carried variant 1, while CC1, CC22, CC72, and CC88, carried variant 2. Strains with variant 1 formed more biofilm than those with variant 2 or those without the gene. Furthermore, variants 1 and 2 had different gene expressions. Thus, *sasG* variant 1 may offer an adaptive advantage over strains lacking the gene or carrying variant 2, favouring bacterial survival and dissemination.

Intracellular invasion in *S. aureus*

The interior of human cells provides an optimal environment, free from the action of the immune system and some antimicrobial agents, for intracellular survival and persistent bacterial colonisation. The **Additional Results** examine the genetic changes associated with intracellular invasion in three patients with respiratory colonisation and subsequent bacteraemia caused by the same *S. aureus* lineage.

The bacteraemic isolates showed a significantly lower ability to invade A549 lung epithelial cells than the respiratory isolates. Genetic analysis revealed differences between the respiratory and bacteraemic isolates, with the latter showing changes in genes encoding proteins involved in the interaction of epithelial cells with the bacterial surface. The genetic alterations were found in three genes: *ebh*, which encodes extracellular matrix binding protein and is associated with fibronectin-binding; *clfB*, which encodes clumping factor B associated with fibrinogen- and cytokeratin-binding; and *fnbA*, which encodes fibronectin-binding protein A.

Resum

La **colonització bacteriana del tracte respiratori** és essencial per la salut per establir una relació simbiòtica entre la comunitat bacteriana i l'hostatger. ***Haemophilus influenzae*** i ***Staphylococcus aureus*** són dos dels microorganismes oportunistes que habitualment colonitzen el tracte respiratori, i mentre la seva interacció normalment és asimptomàtica, aquesta colonització pot ser el pas previ en el desenvolupament d'**infeccions greus**. Aquesta tesi es centra en l'**epidemiologia** (Estudi 1), l'**estructura poblacional** (Estudis 1 i 2) i la **patogènesi** (Estudis 3 i 4, i Resultats Addicionals) d'aquests microorganismes, assumint que canvis en alguns d'aquests factors poden influenciar el desenvolupament de malalties greus.

Epidemiologia de la malaltia invasiva causada per *H. influenzae*

Des de la seva implementació, la vacuna conjugada contra el serotip b de *H. influenzae* ha produït un canvi en l'epidemiologia de la malaltia invasiva causada per aquest microorganisme. A més a més, l'augment de la resistència a β-lactàmics ha provocat la inclusió d'aquest microorganisme a la llista de vigilància de la Organització Mundial de la Salut, requerint la vigilància epidemiològica per monitoritzar tant l'evolució dels clons no evitables per la vacuna com l'augment de la resistència antimicrobiana.

L'**Estudi 1** proporciona una actualització en l'estat de la malaltia invasiva causada per *H. influenzae* en l'Hospital Universitari de Bellvitge entre 2014-2019 (n = 68) i compara la seva evolució amb el període 2008-2013 (n = 82). La incidència global de la malaltia invasiva causada per *H. influenzae* (2,07 casos per 100.000 habitants) es va mantenir constant entre els dos períodes (2,12 i 2,02 casos per 100.000 habitants, respectivament); no obstant, es va mostrar més baixa en adults joves (≤ 64 anys; 0,89 casos per 100.000 habitants) que en adults més grans (≥ 65 anys; 6,90 casos per 100.000). El 82,0% dels pacients presentaven alguna comorbiditat, sent la immunosupressió (40,0%) i la malignitat d'òrgans sòlids (36,0%) les més freqüents. La mortalitat als 30 dies es va reduir entre el període 2008-2013 (20,7%) i el període 2014-2019 (14,7%).

Durant el període 2014-2019, la taxa de resistència a ampicil·lina va augmentar del 10% al 17,6%, principalment a causa de la producció de β-lactamases TEM-1. A més a més, el 4,4% de les soques eren resistents a fluoroquinolones, cosa que suposava l'emergència de la resistència ja que en el període 2008-2013 no es van detectar soques resistents a fluoroquinolones.

Les soques no capsulades de *H. influenzae* (NTHi) eren la principal causa de malaltia invasiva en els dos períodes d'estudi (84,3% i 85,3%, respectivament). Dos dels seqüenciotips (STs) més comuns, el ST103 i el ST160, s'associaven a la producció de la β-lactamasa TEM-1, i ambdós van presentar un augment de la prevalença respecte el període 2008-2013. D'altra banda, les soques capsulades eren poc freqüents en els dos períodes, amb el serotip f sent el més prevalent (12,9% i 8,8%, respectivament), seguit pels serotips e (1,4% i 4,4%, respectivament) i b (1,4% i 1,5%, respectivament).

Estructura poblacional de *H. influenzae*

Les soques NTHi tenen una elevada heterogeneïtat genètica i s'associaven a diversos STs, complicant l'estudi de la seva estructura poblacional. Utilitzant el *multi-locus sequence typing* (MLST), els detalls de la diversitat genètica de la població de NTHi poden passar desapercebuts, de manera que en l'Estudi 1 es proposa una classificació en sis clades (I-VI) basada en la presència i absència de 17 gens accessoris. Tot i que les soques NTHi aïllades en malaltia invasiva entre 2014 i 2019 (n = 58) pertanyien a 47 STs i eren genèticament diversos, la majoria s'agrupaven en el clade V (n = 30). Les soques d'aquest clade compartien el mateix origen filogenètic i agrupaven els STs de NTHi més freqüents (ST3, ST103, ST160), incloent 9 de les 12 soques productores de β-lactamases. El clade VI era el segon més abundant (31,0%), però era altament heterogeni comparat amb el clade V.

En comparació amb aquests resultats, les soques capsulades eren altament clonals, amb cada serotip associat a un nombre limitat de llinatges. No obstant, els serotips f i a estan en auge entre els casos de malaltia invasiva a Europa i a Nord d'Amèrica,

respectivament. Així doncs, és necessari caracteritzar-los extensivament per tal de poder desenvolupar, en cas necessari, estratègies efectives que puguin controlar efectivament la seva possible emergència.

Així doncs, l'**Estudi 2** caracteritza la diversitat genòmica i el pangenoma de les soques capsulades, incloent els genomes disponibles en les bases de dades del *National Center for Biotechnology Information* i del *European Nucleotide Archive*, aprofundint en el serotip f utilitzant soques aïllades en diferents països (Portugal, Països Baixos i Espanya). El MLST i l'anàlisi filogenètic dels genomes capsulats van revelar un nombre reduït de complexes clonals (CCs) associats a cada serotip. Aquests eren altament clonals. Els clons del serotip a es van classificar en tres grups: CC23 i CC1755 (clade I), CC62 i CC372 (clade II), i CC4 (clade III). Els clons del serotip b principalment pertanyien al CC6 (n = 151), mentre que els clons del CC50 (n = 13) i el CC464 (n = 1) s'agrupaven en un grup filogenètic diferent. Els serotips c i d es van associar amb els CC7 i CC10, respectivament. Aquests presentaven menys variabilitat genètica que els altres serotips, probablement a causa del baix nombre de genomes. El serotip e estava exclusivament associat amb el CC18. Els clons del serotip f es van agrupar filogenèticament en dos grups: CC124 (n = 222) i CC16 (n = 12).

L'anàlisi del pangenoma va mostrar que els genomes del serotip b tenien un patrimoni gènic més gran (n = 3.517 gens) que els altres serotips, els quals oscil·laven entre 1.896 i 2.969 gens. Aquest serotip també tenia més gens accessoris per genoma (mitjana = 46,6%), seguit pels serotips f (mitjana = 25,3%), a (mitjana = 25,1%), i e (20,7%). L'anàlisi dels polimorfismes de nucleòtids simples (SNPs) va mostrar que els genomes de cada serotip es podien classificar en CCs segons el nombre de SNPs, els quals van mostrar que el nombre total de SNPs en el serotip f era significativament menor que en els serotips a, b, i e (p-valor < 0,0001), cosa que indicava una baixa variabilitat dins dels CCs del serotip f. Aquests resultats donen suport a l'elevada estabilitat genètica d'aquest serotip en soques aïllades en diferents països, suggerint que no existeix una relació entre la filogènia i l'origen geogràfic.

Evolució adaptativa de *H. influenzae* durant la colonització persistent

La caracterització de l'adaptació bacteriana produïda durant la colonització pot millorar la nostra comprensió de la progressió de les malalties associades amb *H. influenzae*, com la malaltia pulmonar obstructiva crònica (MPOC). L'Estudi 3 examina els canvis genètics en aquest microorganisme durant la colonització utilitzant els aïllats respiratoris de 15 pacients amb MPOC que rebien un tractament prolongat amb azitromicina (250 mg, 3 dosis/setmana). A més a més, els aïllats de *H. parainfluenzae* d'aquests pacients es van incloure per investigar el seu paper en la transmissió horitzontal de gens de resistència.

Dels 15 pacients, 4 (P02, P08, P11, i P13) tenien una colonització persistent per *H. influenzae*, i 2 (P04 i P11) tenien una colonització persistent per *H. parainfluenzae*. Tots els llinatges persistents aïllats abans del tractament eren sensibles a azitromicina, però la pressió antibiòtica va provocar l'aparició de resistència en els primers mesos, excepte en el P02, en el que es va interrompre el tractament a causa d'efectes adversos. Es van identificar diferents determinants genètics de resistència en cada llinatge. En *H. influenzae*, els aïllats del P08 ST107 havien acumulat mutacions successives en el 23S rRNA, mentre que els P11 ST2480 i P13 ST165 tenien mutacions en els gens codificants per les proteïnes ribosomals L4 i L22. En *H. parainfluenzae*, els aïllats persistents del P04 van adquirir canvis en el gen *rlmC* que codifica la metiltransferasa del 23S rRNA, mentre que en els aïllats del P11, els gens codificants de les bombes d'expulsió MefE i MsrD es van trobar dins de l'element *tet*(M)-MEGA, present en un element de conjugació i integració. La identificació d'aquest element en el llinatge persistent de *H. influenzae* ST147 aïllat en el mateix pacient donaria suport al paper de *H. parainfluenzae* com a reservori de gens de resistència.

Altres variacions genètiques principalment tenien lloc en gens associats amb la paret cel·lular i el transport i metabolisme de ions inorgànics. Les soques persistents de *H. influenzae* mostraven canvis de variació de fase en els gens *licA* i *hgpB* associats a la síntesi de lipooligosacàrid i a la captació d'hemo, respectivament. Altres gens (*lex1*,

lic3A, *hgpC*, i *fadL*) presentaven variacions genètiques en diversos llinatges; i algunes regions del genoma acumulaven molts canvis genètics o grans insercions i delecions que corresponien a regions associades a profags.

Formació de biofilm en *S. aureus*

La formació de biofilm permet als bacteris colonitzar diferents superfícies durant llargs períodes, tant en l'hostatger com en l'entorn hospitalari. Els biofilms actuen com a reservoris bacterians, facilitant la transmissió i la disseminació d'infeccions greus. L'**Estudi 4** compara la formació de biofilm dels clons de *S. aureus* resistents a meticil·lina (MRSA) endèmics (CC5, CC8, i CC22) i esporàdics (CC1, CC30, CC45, CC72, i CC398) en el nostre entorn hospitalari. Els clons endèmics produïen més biofilm que els clons esporàdics, tot i que els aïllats del CC22, presentaven una formació de biofilm comparable a aquella dels clons esporàdics, produint menys biofilm que els aïllats del CC5 i del CC8.

Existia una correlació entre el biofilm i el gen *sasG*: les soques que tenien el gen produïen més biofilm que aquelles que no el tenien. L'anàlisi de les variants al·lèliques va mostrar que les soques dels CC5 i CC8 tenien la variant 1, mentre les dels CC1, CC22, CC72, i CC88, tenien la variant 2. Les soques amb la variant 1 formaven més biofilm que aquelles amb la variant 2 o les que no tenien el gen. A més a més, les variants 1 i 2 tenien diferents expressions gèniques. Així doncs, la variant 1 de *sasG* podria oferir un avantatge adaptatiu respecte les soques sense el gen o amb la variant 2, afavorint la supervivència i la disseminació bacteriana.

Invasió intracel·lular en *S. aureus*

L'interior de les cèl·lules humanes proporciona un entorn òptim, lliure de l'acció del sistema immunitari i d'alguns agents antimicrobians, per la supervivència intracel·lular i la colonització bacteriana persistent. Els **Resultats Addicionals** examinen els canvis genètics associats a la invasió intracel·lular en tres pacients amb una colonització respiratòria i una posterior bacterièmia causada pel mateix llinatge de *S. aureus*.

Els aïllats bacterièmics presentaven una capacitat significativament menor d'envair les cèl·lules epitelials pulmonars A549 que els aïllats respiratoris. L'anàlisi genètic va revelar diferències entre els aïllats respiratoris i bacterièmics, i aquests últims mostraven canvis en gens que codifiquen proteïnes implicades en la interacció de les cèl·lules epitelials amb la superfície bacteriana. Les alteracions genètiques es van identificar en tres gens: *ebh*, que codifica la proteïna d'unió a matriu extracel·lular i s'associa amb la unió a fibronectina; *clfB*, que codifica el factor d'aglutinació B associat amb la unió a fibrinogen i citoqueratina; i *fnbA*, que codifica la proteïna d'unió a fibronectina A.

Scientific production

La Dra. Mª Ángeles Domínguez Luzón, Professora Agregada del Departament de Patologia i Terapèutica Experimental de la Universitat de Barcelona - cap del Servei de Microbiologia, Hospital Universitari de Bellvitge (Barcelona) i la Dra. Sara Martí Martí, Professora Associada del Departament de Medicina de la Universitat de Barcelona - investigadora Postdoctoral sènior en el *Centro de Investigación Biomédica en Red de Enfermedades Respiratorias* (CIBERes),

FAN CONSTAR que les revistes científiques on es recullen els articles derivats de la tesi doctoral "**Epidemiology, population structure and pathogenesis of opportunistic respiratory tract colonising bacteria"**, tenen els següents **factors d'impacte** (Journal Citation Reports 2021):

Estudi 1

Anna Carrera-Salinas, Aida González-Díaz, Laura Calatayud, Julieta Mercado-Maza, Carmen Puig, Dàmaris Berbel, Jordi Càmara, Fe Tubau, Imma Grau, M Ángeles Domínguez, Carmen Ardanuy, Sara Martí. Epidemiology and population structure of *Haemophilus influenzae* causing invasive disease. *Microbial genomics* 2021; 7(12):000723. doi: 10.1099/mgen.0.000723.

Factor d'impacte: 4.868

Participació: Treball experimental, anàlisi, redacció del manuscrit

Estudi 2

Aida González-Díaz, **Anna Carrera-Salinas**, Miguel Pinto, Meritxell Cubero, Arie van der Ende, Jeroen D Langereis, M Ángeles Domínguez, Carmen Ardanuy, Paula Bajanca-Lavado, Sara Martí. Comparative pangenome analysis of capsulated *Haemophilus influenzae* serotype f highlights their high genomic stability. *Scientific reports* 2022; 12(1):3189. doi: 10.1038/s41598-022-07185-5.

Factor d'impacte: 4.996

Participació compartida amb Aida González-Díaz, qui va utilitzar una versió preliminar d'aquest treball en la seva tesi doctoral defensada el 2020. Aquest treball va ser finalitzat amb la col·laboració d'Anna Carrera-Salinas, tal i com es defineix a continuació:

Aida González-Díaz: Seqüenciació (Illumina); anàlisi bioinformàtic i filogènia (Linux); redacció del manuscrit; preparació de la figura 1 i les figures suplementàries S1-S4.

Anna Carrera-Salinas: Anàlisi bioinformàtic i filogènia (R-studio); *SNP typing*; anàlisi pangenòmic (*gene pool analysis*); redacció del manuscrit; preparació de les figures 2-3, i la figura suplementària S5.

Estudi 4

Anna Carrera-Salinas, Aida González-Díaz, Daniel Antonio Vázquez-Sánchez, Mariana Camoez, Jordi Niubó, Jordi Càmara, Carmen Ardanuy, Sara Martí, M Ángeles Domínguez. *Staphylococcus aureus* surface protein G (*sasG*) allelic variants: correlation between biofilm formation and their prevalence in methicillin-resistant *S. aureus* (MRSA) clones. *Research in Microbiology* 2022; 173(3):103921. doi: 10.1016/j.resmic.2022.103921

Factor d'impacte: 3.946

Participació: Treball experimental, anàlisi, redacció del manuscrit

I per deixar constància, firmen el present document.

L'Hospitalet de Llobregat, 19 de Setembre de 2022

Mª Ángeles Domínguez Luzón

Sara Martí Martí

PUBLICATIONS

Publications in international peer-reviewed journals

1. Anna Carrera-Salinas, Aida González-Díaz, Laura Calatayud, Julieta Mercado-Maza, Carmen Puig, Dàmaris Berbel, Jordi Càmara, Fe Tubau, Imma Grau, M Ángeles Domínguez, Carmen Ardanuy, Sara Martí. **Epidemiology and population structure of *Haemophilus influenzae* causing invasive disease**. Microbial genomics. 2021. 7(12):000723. Impact factor in 2021 is 4.868 (1st quartile).
2. Aida González-Díaz, Anna Carrera-Salinas, Miguel Pinto, Meritxell Cubero, Arie van der Ende, Jeroen D Langereis, M Ángeles Domínguez, Carmen Ardanuy, Paula Bajanca-Lavado, Sara Martí. **Comparative pangenome analysis of capsulated *Haemophilus influenzae* serotype f highlights their high genomic stability**. Scientific reports. 2022.12(1):3189. Impact factor in 2021 is 4.996 (1st tercile).
3. Anna Carrera-Salinas, Aida González-Díaz, Daniel Antonio Vázquez-Sánchez, Mariana Camoez, Jordi Niubó, Jordi Càmara, Carmen Ardanuy, Sara Martí, M Ángeles Domínguez. ***Staphylococcus aureus* surface protein G (*sasG*) allelic variants: correlation between biofilm formation and their prevalence in methicillin-resistant *S. aureus* (MRSA) clones**. Research in Microbiology. 2022. 173(3):103921. Impact factor in 2021 is 3.946 (3rd quartile).

Collaborations

1. Yanik Sierra, Fe Tubau, Aida González-Díaz, Anna Carrera-Salinas, Javier Moleres, Paula Bajanca-Lavado, Junkal Garmendia, M Ángeles Domínguez, Carmen Ardanuy, Sara Martí. **Assessment of trimethoprim-sulfamethoxazole susceptibility testing methods for fastidious *Haemophilus* spp**. Clinical Microbiology and Infection. 2020. 26(7):944.e1-944.e7. Impact factor in 2020 is 8.067 (1st decile).
2. Yanik Sierra, Aida González-Díaz, Anna Carrera-Salinas, Dàmaris Berbel, Daniel Antonio Vázquez-Sánchez, Fe Tubau, Meritxell Cubero, Junkal Garmendia, Jordi Càmara, Josefina Ayats, Carmen Ardanuy, Sara Martí. **Genome-wide analysis of urogenital and respiratory multidrug-resistant *Haemophilus parainfluenzae***. The Journal of Antimicrobial Chemotherapy. 2021. 76(7):1741-1751. Impact factor in 2021 is 5.758 (1st quartile).

3. Aida González-Díaz, Dàmaris Berbel, María Ercibengoa, Emilia Cercenado, Nieves Larrosa, M Dolores Quesada, Antonio Casabella, Meritxell Cubero, Jose Maria Marimón, M Ángeles Domínguez, Anna Carrera-Salinas, Jordi Càmara, Antonio J Martín-Galiano, Jose Yuste, Sara Martí, Carmen Ardanuy. **Genomic features of predominant non-PCV13 serotypes responsible for adult invasive pneumococcal disease in Spain**. The Journal of Antimicrobial Chemotherapy. 2022. 77(9):2389-2398. Impact factor in 2021 is 5.758 (1st quartile).

COMMUNICATIONS

Oral presentations in international meetings

1. Anna Carrera-Salinas, Aida González-Díaz, Ileana P Salto, María Mrakovcic, Meritxell Cubero, M Ángeles Domínguez, Silke Niemann, Sara Martí. **Host adaptative changes of *Staphylococcus aureus* through respiratory colonisation and bloodstream infection.** Accepted as oral communication in the 30th ECCMID. Paris, France – Published in the abstract book.

Oral presentations in national meetings

1. Anna Carrera-Salinas, Aida González-Díaz, Irene Cadenas-Jiménez, Lucía Saiz-Escobedo, Daniel Antonio Vázquez-Sánchez, Sara Calvo-Silveria, Fe Tubau, Ester Cuevas, Daniel Huertas, Xavier Pomares, M Ángeles Domínguez, Carmen Ardanuy, Alicia Marín, Conchita Montón, Salud Santos, Sara Martí. **Adquisición de resistencia a los macrólidos y adaptación genética de *Haemophilus influenzae* y *Haemophilus parainfluenzae* durante el tratamiento prolongado con azitromicina en pacientes con EPOC.** 25th SEIMC Congress. Granada, Spain – 02-04/06/2022.

2. Anna Carrera-Salinas, Aida González-Díaz, Yanik Sierra, Ileana P Salto, Maria Mrakovcic, M Ángeles Domínguez, Silke Niemann, Sara Martí. **Transition from carriage to invasive *Staphylococcus aureus* disease: genetic modifications during the intracellular colonization stage.** 8th Barcelona Boston Lung Conference. Online – 28/01/2021.

Poster presentations in international meetings

1. Anna Carrera-Salinas, Mariana Camoez, Sara Martí, Oriol Gasch, Miquel Pujol, Emilia Cercenado, Elena Loza, M Nieves Larrosa, Enrique Ruiz, Fernando Chaves, Luis Martínez-Martínez, M Ángeles Domínguez. **Detection of genes related with adhesion and biofilm formation among bacteraemic methicillin-resistant *Staphylococcus aureus* (MRSA) clones.** 28th ECCMID. Madrid, Spain – 21-24/10/2018.

2. Carla Ferrero, Aida González-Díaz, Anna Carrera-Salinas, Daniel Huertas Marti, Fe Tubau, Jordi Càmara, Xavier Pomares, Junkal Garmendia, Carmen Ardanuy, Marian García-Núñez, Alicia Marín, Concepción Montón, Salud Santos, Sara Martí. **Dynamics of persistent colonization by *Haemophilus influenzae* in Chronic Obstructive Pulmonary Disease (COPD) patients under long-term treatment with azithromycin.** 29th ECCMID. Amsterdam, The Netherlands – 13-16/04/2019.

3. Anna Carrera-Salinas, Aida González-Díaz, Mariana Camoez, Jordi Niubó, Sara Martí, M Ángeles Domínguez. **The role of *Staphylococcus aureus* surface protein G (*sasG*) and its allelic variants in biofilm formation .** 30th ECCMID . Paris , France – Published in the abstract book.

4. Anna Carrera-Salinas, Aida González-Díaz, Laura Calatayud, Julieta Mercado-Maza, Irene Cadenas-Jiménez, Lucía Saiz-Escobedo, Daniel Antonio Vázquez-Sánchez, Carmen Puig, Dàmaris Berbel, Jordi Càmara, Fe Tubau, M Ángeles Domínguez, Sara Martí. ***Haemophilus influenzae* population structure in invasive disease based on clade-related classification.** 31st ECCMID. Online – 09-12/07/2021.

5. Anna Carrera-Salinas, Aida González-Díaz, Irene Cadenas-Jiménez, Daniel Antonio Vázquez-Sánchez, Lucía Saiz-Escobedo, Sara Calvo-Silveria, Fe Tubau, Ester Cuevas, Daniel Huertas, Xavier Pomares, M Ángeles Domínguez, Carmen Ardanuy, Alicia Marín, Conchita Montón, Salud Santos, Sara Martí. ***Haemophilus influenzae* genetic adaptation during long-term azithromycin therapy in patients with chronic obstructive pulmonary disease (COPD).** 32nd ECCMID. Lisboa, Portugal – 23-26/04/2022.

Poster presentations in national meetings

1. Anna Carrera-Salinas, Aida González-Díaz, Ester Cuevas, Daniel Huertas, Yanik Sierra, Fe Tubau, Junkal Garmendia, Xavier Pomares, Carmen Ardanuy, Marian García-Núñez, Alicia Marín, Concepción Montón, Salud Santos, Sara Martí. **Estudio de la dinámica de las poblaciones microbianas en pacientes con EPOC grave y tratamiento prolongado con azitromicina.** 53th SEPAR Congress. Sevilla, Spain – 12-15/11/2020.

2. Anna Carrera-Salinas, Aida González-Díaz, Ester Cuevas, Daniel Huertas, Yanik Sierra, Fe Tubau, Junkal Garmendia, Xavier Pomares, Carmen Ardanuy, M García-Núñez, Alicia Marín, C Montón, Salud Santos, Sara Martí. **Estudio de la dinámica de las poblaciones microbianas en pacientes con EPOC grave y tratamiento prolongado con azitromicina.** 53th SEPAR Congress. Online – 12-14/11/2020.

3. Anna Carrera-Salinas, Aida González-Díaz, Laura Calatayud, Julieta Mercado-Maza, Carmen Puig, Dàmaris Berbel, Jordi Càmara, Fe Tubau, M Ángeles Domínguez, Carmen Ardanuy, Sara Martí. **Epidemiología de la enfermedad invasiva por *Haemophilus influenzae* en adultos (Hospital Universitario de Bellvitge, 2008-2019).** 25th SEIMC. Online – 06-11/06/2021.

4. Anna Carrera-Salinas, Aida González-Díaz, Yanik Sierra, Ileana P Salto, Maria Mrakovcic, M Ángeles Domínguez, Silke Niemann, Sara Martí. **De la colonización respiratoria a la enfermedad invasiva: cambios genéticos adaptativos en *Staphylococcus aureus*.** 54th SEPAR Congress. Sevilla – 04-06/11/2021.

1. Introduction

1. Microbial pathophysiology of the respiratory tract

The microbiota of the respiratory tract is essential for maintaining a healthy respiratory system. The respiratory tract of a healthy person has specific physiological properties that vary along the airways and provide different growth conditions, influencing the density and composition of microbial communities[1] (Figure 1). In healthy individuals, the upper respiratory tract has high bacterial densities, dominated by *Staphylococcus*, *Propionibacterium*, *Corynebacterium*, *Moraxella*, *Streptococcus*, and *Haemophilus* spp. Despite often being thought sterile, the lower respiratory tract can be colonised by microorganisms through the inhalation of upper respiratory tract secretions . Therefore , the lower airways are frequently colonised by low bacterial densities of *Prevotella*, *Veillonella*, *Streptococcus*, and *Rothia* spp[2].

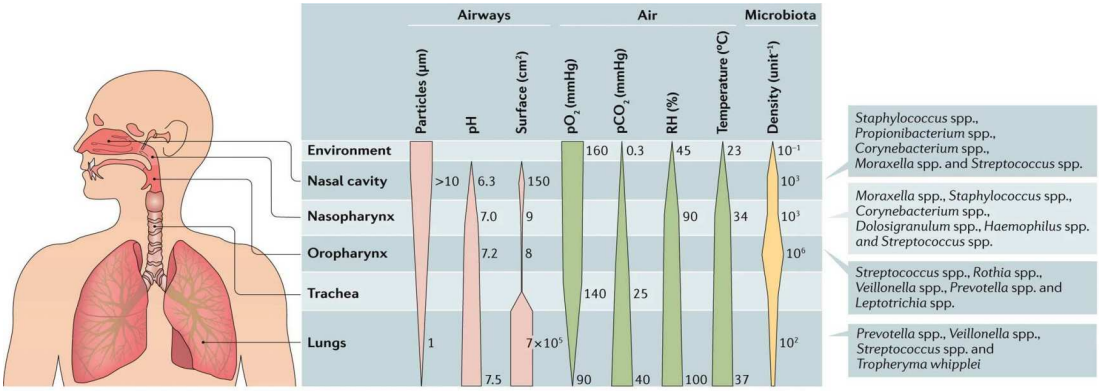


Figure 1. Physiological and microbial gradients along the healthy respiratory tract. Image adapted from Man WH, *et al*.[1].

A healthy microbiota is diverse and maintains microbial **homeostasis** by secreting antimicrobial peptides that prevent pathogenic species from overgrowing. The balance between microbial growth and elimination is also mediated by host defences, resulting in a **symbiotic relationship** between the microbial community and host[2,3]. However, various factors disrupt this balance in patients with chronic respiratory diseases and result in **dysbiosis**. Factors include decreased bacterial clearance, decreased oxygen availability, and excessive mucus production. Antimicrobial treatment and

hospitalisations further affect the bacterial community. Together, these contribute to the potential for overgrowth by pathogenic bacteria (Figure 2), such as ***Haemophilus influenzae*** and ***Staphylococcus aureus***, which can cause chronic inflammation, progression of the underlying disease, and chronic infection. As respiratory disease progresses, the microbial communities become dysbiotic and less diverse[2,4].

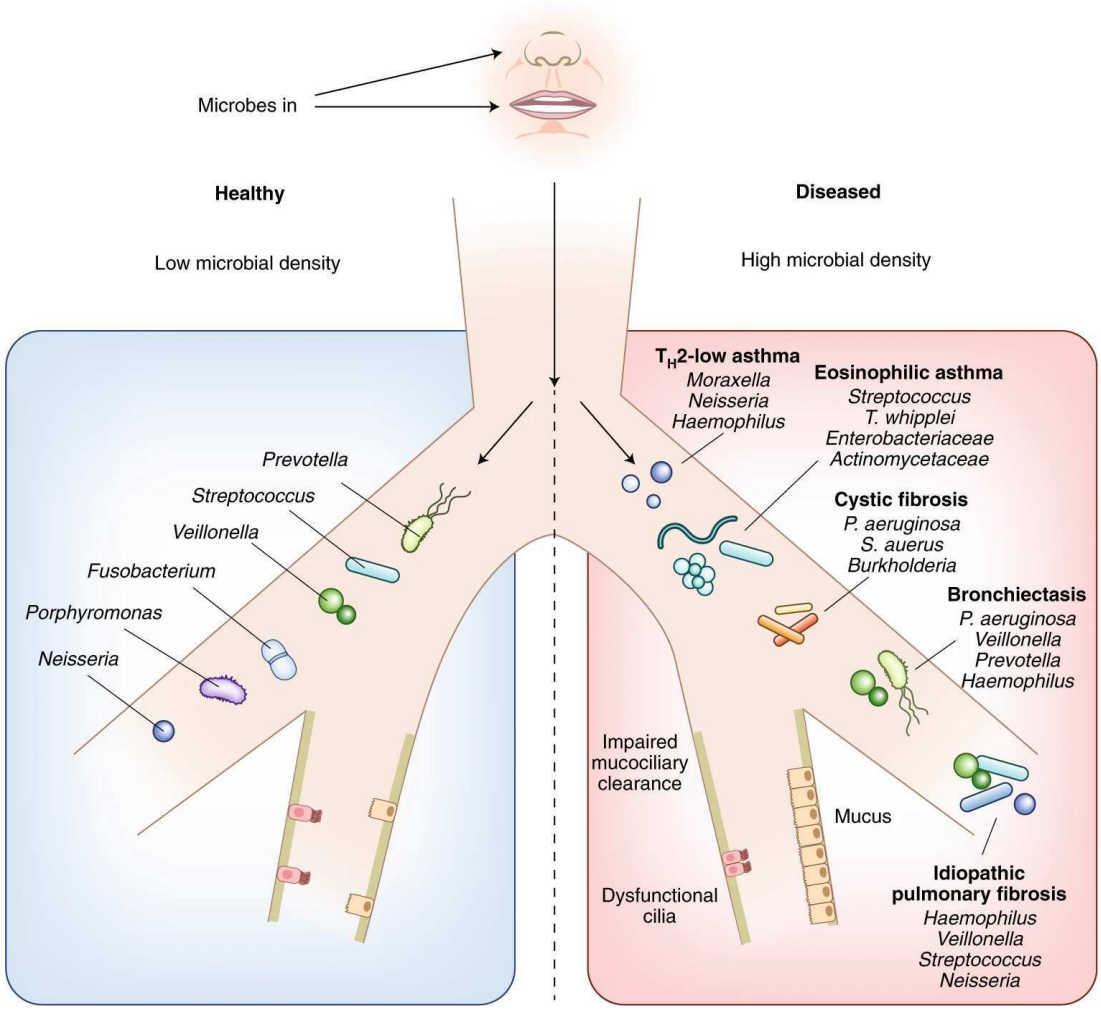


Figure 2. Comparison of the lung microbiome in health and disease. Mechanisms of dysbiosis and the predominant bacteria in certain chronic respiratory diseases are displayed (red, right). Image adapted from Wypych TP, *et al*.[3].

1.1. Bacterial adaptation in the respiratory tract

For persistent colonisation and survival in the respiratory tract, bacteria must adapt to the selective pressures imposed by the immune system, antimicrobial agents, other species in their ecological niche, and environmental stressors such as hypoxia, oxidative stress, nutrient deficiency, toxic concentrations of metals, or pH changes[5,6].

The microbial genome evolves through point mutations, insertions, deletions, and rearrangements via intragenomic recombination. These **genetic changes** may result in novel or altered antigens, with natural selection favouring these microorganisms due to their ability to adapt to and overcome host immune responses or antimicrobials[7,8]. Furthermore, phase variation is associated with reversible mutations in sequence repeats located in promoter regions or in open reading frames, so that changes in the number of repeats affect transcription or cause a shift in the reading frame of the gene. This results in a gene expression regulation mechanism that allows bacteria to switch between two phenotypes or phases, with one differing from the other by expressing an additional protein[9,10]. Horizontal gene transfer, by transformation and conjugation, facilitates the transfer and uptake of exogenous DNA from the environment and other pathogens, leading to the acquisition of antimicrobial resistance genes and virulence factors[2,6].

Some gram-negative bacteria can produce **outer membrane vesicles** (OMVs), which are spherical elements surrounded by a membrane that carry nucleic acids, virulence factors, and immunomodulatory molecules that can be passed to other bacteria or host cells. This establishes cooperative and competitive relationships with other bacteria, such as the delivery of resistance genes or peptidoglycan hydrolases that provide antimicrobial resistance or cause the death of gram-positive bacteria, respectively. By contrast, OMVs can affect host interaction, either increasing bacterial survival within host cells or activating inflammatory or death-signalling pathways[2,11].

Biofilm formation is a strategy used by many bacteria to colonise, survive, and spread in the host. A biofilm is a microbial community that attaches to a surface (e.g.,

living tissue or indwelling medical devices) and is surrounded by an extracellular polymeric substance (EPS) matrix[12]. Biofilm formation comprises of four stages (Figure 3): initial attachment, when microorganisms adhere to a surface; early biofilm formation, when microorganisms divide and produce EPS; biofilm maturation, when the EPS matrix envelops and protects the bacteria, creating a microenvironment for different bacteria to interact; and dispersal, when the biofilm disaggregates and the bacteria are released, returning to a planktonic state to colonise or invade other regions[12]. When compared to planktonic cells, bacteria in biofilm exhibit changes in gene expression that produce metabolic changes and slower growth rates. Furthermore, the proximity of bacteria in biofilms provides an ideal environment for genetic exchange and communication via quorum sensing. These structures also limit the diffusion of extracellular products, including immune cells and antimicrobials[2,13].

Another mechanism of bacterial adaptation in response to environmental stressors is the development of a **small colony variant** (SCV) phenotype. SCVs are a bacterial subpopulation that exhibits distinctive characteristics, including slow growth, atypical colony morphology, such as shortness and non-pigmentation, and altered biochemical processes, such as hemin, thiamine, menadione, thymidine, and unsaturated fatty acids auxotrophies. Slow growth could be caused by a reduction in electron transport, which reduces ATP production and affects cell wall formation. Impaired electron transport also compromises the formation of carotenoid pigments, explaining why SCVs are often colourless. Finally, the SCV phenotype provides better adaptation than the wild type, probably due to greater resistance to host immune defences and antibiotics, as well as a greater ability to invade and adhere to respiratory epithelial cells[14,15].

Aside from strictly intracellular microorganisms, other bacterial species can adhere to, invade, and survive within host cells. Moreover, bacteria have developed protective mechanisms to avoid being degraded by host cells and to maintain colonisation. The intracellular environment protects against immune system defence and antimicrobial attack, serving as a reservoir for recurrence and infection[16].

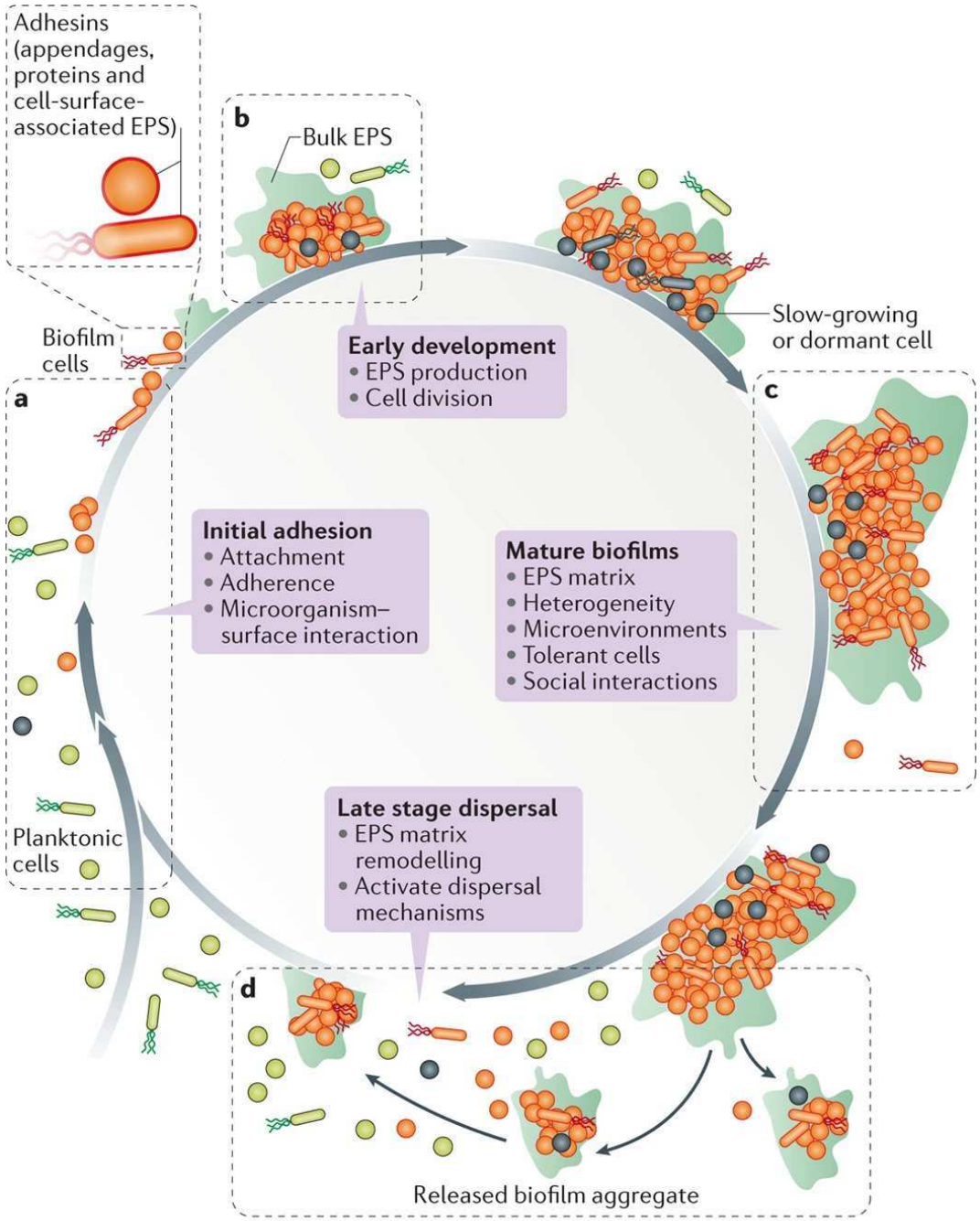


Figure 3. Biofilm formation cycle. **(a)** Initial adhesion. **(b)** Early biofilm formation. **(c)** Biofilm maturation. **(d)** Biofilm dispersal. EPS, extracellular polymeric substance. Image adapted from Koo H, *et al*.[12].

2. *Haemophilus influenzae*

2.1. History, taxonomy, and microbiological characteristics

In 1892, the physician and bacteriologist Richard Pfeiffer identified the aetiological agent responsible for the influenza pandemic, which was initially known as Pfeiffer's bacillus or *Bacillus influenzae*. Its poor growth on conventional media, which improved with the addition of haemoglobin, explained the failure to isolate the bacterium in all influenza patients and led to its subsequent renaming as *H. influenzae*, derived from the Greek *haima* (blood) and *philos* (lover) to reflect this growth requirement, but retained the Italian *influenza* (influenza) due to its association with the disease. However, it was finally proven in the 1930s that influenza was caused by a virus, and that the bacteria *H. influenzae*, *Streptococcus pneumoniae*, *S. aureus*, and other bacteria caused secondary infections associated with high mortality[17,18].

The genus *Haemophilus* is one of the 30 genera identified in the ***Pasteurellaceae*** family, which also includes the genera *Actinobacillus*, *Aggregatibacter*, *Avibacterium*, and *Pasteurella*. *Pasteurellaceae*, which belongs to the order **Pasteurellales**, the class **Gammaproteobacteria**, and the phylum **Proteobacteria**, forms a compact phylogenetic group distinct from other nearby families based on 16S rRNA sequences[19]. Taxonomic definition at the genus level is more complex and is constantly changing due to genetic diversity among its members. In 2021, the genus *Haemophilus* included 14 *Haemophilus* species (Figure 4), of which 11 are hosted exclusively by humans (*H. influenzae*, *Haemophilus aegyptius*, *Haemophilus parainfluenzae*, *Haemophilus haemolyticus*, *Haemophilus parahaemolyticus*, *Haemophilus ducreyi*, *Haemophilus paraphrohaemolyticus*, *Haemophilus pittmaniae*, *Haemophilus sputorum*, *Haemophilus seminalis*, and *Haemophilus massiliensis*) and the remainder have higher specificity for animal reservoirs (*Haemophilus felis*, *Haemophilus haemoglobinophilus*, and *Haemophilus cuniculus*)[18].

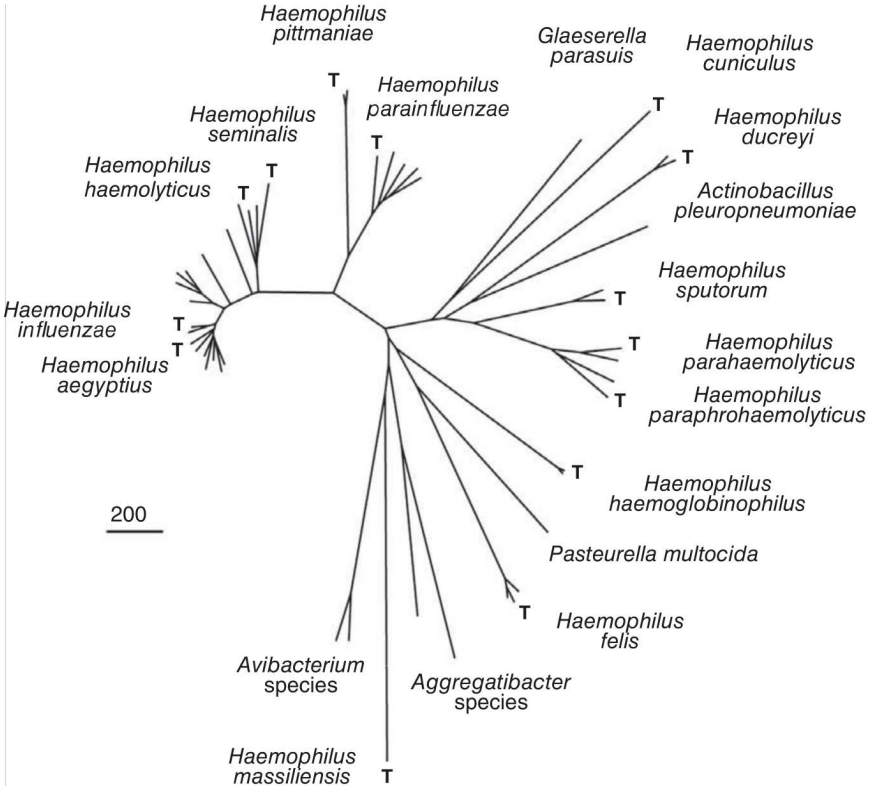


Figure 4. Neighbour-joining dendrogram of 27 core genes from species currently classified in the genus *Haemophilus*. T, type strains. Image adapted from Nørskov-Lauritsen N.[18].

H. influenzae is a non-motile **gram-negative coccobacillus** with pleomorphic ability (width, 0.3-0.5 μm; length, 0.5-3 μm). It forms non-haemolytic, smooth, translucent colonies with grey pigmentation on chocolate agar plates, and its optimal growth conditions are 35ºC-37ºC in the presence of 5%-7% CO_2[18]. Dependence on **factor V** (nicotinamide adenine dinucleotide [NAD]) and **factor X** (heme) differentiate *H. influenzae* from other species in the genus (Table 1). Initially, growth factor dependence was used to define and distinguish the genus *Haemophilus* from other genera in the *Pasteurellaceae* family, but this classification has become obsolete because other genera in the family also exhibit factor dependence. Members obtain NAD from the environment, and although some can convert nicotinamide to NAD via nicotinamide phosphoribosyltransferase (NadV), its absence results in V factor dependence in some species. Alternatively, protoporphyrin, is synthesized by δ-aminolevulinic acid and

transformed to haemin through the action of ferrochelatase on ferrous iron. The lack of biosynthetic enzymes in this pathway results in factor X dependence[18,20].

Table 1. Phenotypic characteristics of genus *Haemophilus*. +, positive; -, negative; v, variable (11-89% of strains positive); NA, no data. Adapted from Nørskov-Lauritzen N, Inzana TJ, *et al*., Zheng ML, *et al*., and Public Health of England[18,21–23].

	H. influenzae	*H. aegyptius*	*H. haemolyticus*	*H. ducreyi*	*H. felis*	*H. haemoglobinophilus*	*H. massiliensis*	*H. cuniculus*	*H. parahaemolyticus*	*H. parainfluenzae*	*H. paraphrohaemolyticus*	*H. pittmaniae*	*H. seminalis*	*H. sputorum*
Growth factor dependence														
Factor V	+	+	+	-	+	-	NA	+	+	+	+	+	+	+
Factor X	+	+	+	+	+	+	NA	+	-	-	-	-	-	-
β-haemolysis	-	-	+	v	v	-	NA	-	+	-	+	+	-	+
Sugar fermentation														
D-ribose	+	+	+	-	-	v	NA	-	-	-	-	NA	NA	-
Xylose	+	-	v	-	-	+	+	-	-	-	-	-	NA	-
Biotyping														
Tryptophanase	v	-	v	-	-	+	-	+	-	v	-	-	+	-
Urease	v	+	+	-	-	-	+	+	+	v	+	-	NA	+
Ornithine decarboxylase	v	-	-	-	-	-	NA	+	-	v	-	-	-	-

H. influenzae is a **facultative anaerobe**, and its carbohydrate metabolism is based on a fructose-specific phosphotransferase system. When fructose levels are low, the transport and metabolism of other monosaccharides (e.g., fucose, xylose, and ribose) is upregulated. Heterogeneity in the production of tryptophanase, urease, and ornithine decarboxylase (Table 1) can be used to distinguish eight biotypes (Table 2)[24].

Table 2. Biotypes of *H. influenzae*. +, positive; -, negative. Adapted from Ledeboer NA, *et al*.[24].

	I	II	III	IV	V	VI	VII	VIII
Tryptophanase	+	+	-	-	+	-	+	-
Urease	+	+	+	+	-	-	-	-
Ornithine decarboxylase	+	-	-	+	+	+	-	-

Margaret Pittman described the presence of a **polysaccharide capsule** in some *H. influenzae* strains in 1931. The capsule is composed of disaccharide repeats, with six different disaccharide combinations defined that result in six capsule types (a to f). This feature has been used for typing, allowing differentiation between capsulated and non-capsulated strains, also known as non-typeable (NTHi)[25].

2.2. Genomic profile and population structure

In 1995, *H. influenzae* became the first bacterial genome to be completely sequenced. Its genome has an average length of **1.85 Mb**, an average GC content of 38%, and encodes an average of 1,722 proteins[26].

The **pangenome** refers to all genes in a species, including core (shared by almost all strains), accessory (present in some, but not all strains), and unique genes (to individual strains). The *H. influenzae* pangenome comprises approximately 1,300 core and >6,000 accessory genes, mostly associated with hypothetical and phage-associated proteins[27,28]. This large pangenome aids bacterial diversity and survival under different selective pressures, with *H. influenzae* adapting its virulence by modifying gene expression and content through phase variation, homologous recombination, transformation, and phage or transposon acquisition[29]. However, some studies have found a higher recombination ratio in NTHi than in capsulated strains. The absence of a capsule in NTHi strains, combined with the presence of competence-related genes (e.g., type IV pilus or type II secretion genes) likely explains the observed differences in recombination ratios[20,28,30]. Capsulated strains are highly clonal, whereas NTHi strains show high genetic heterogeneity (Figure 5).

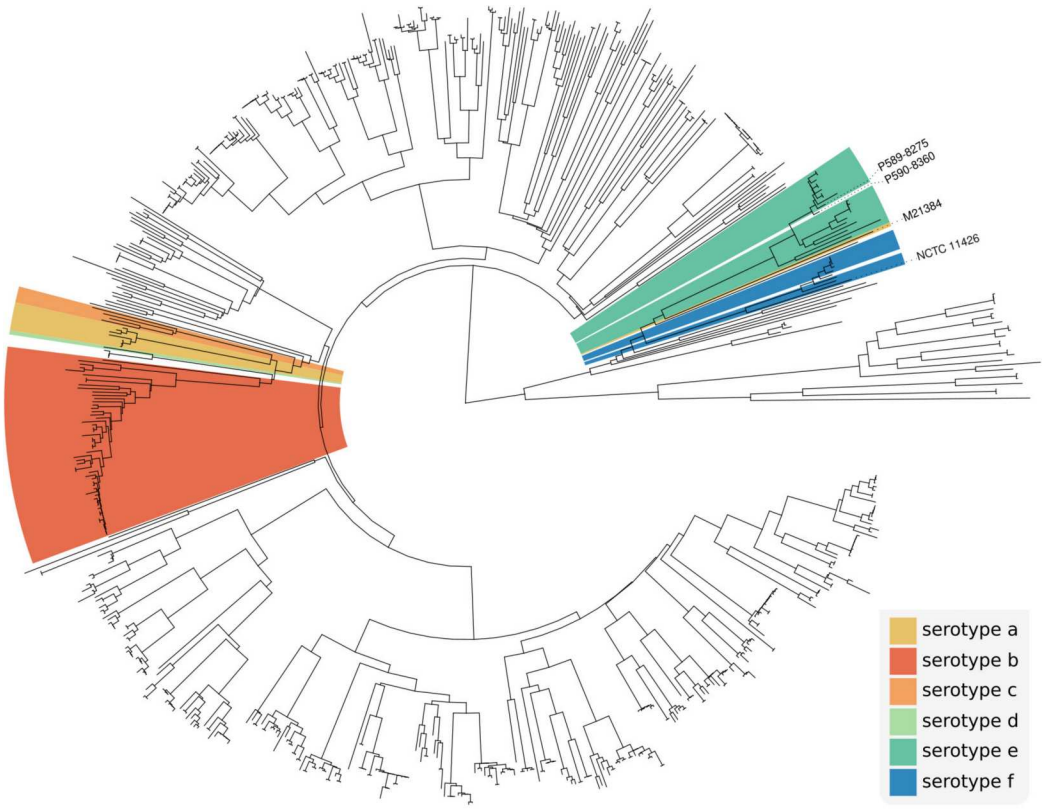


Figure 5. Whole genome phylogeny of *H. influenzae*. Image adapted from Watts SC, *et al*.[31].

The **capsulated strains** are grouped into two superclades within the *H. influenzae* phylogeny: one containing serotypes a, b, c, and d and the other containing serotypes e and f. Each serotype has its own phylogenetic cluster and is associated with only a few sequence types (STs), indicating a high level of genetic conservation and evolution from a common ancestor. Although uncommon (0.1%-0.3%), non-capsulated strains can be found within serotype-specific clusters if partial or complete deletion of the *cap* locus occurs. Furthermore, the presence of serotype a strains within the serotype e cluster could be explained by recombination between strains of these serotypes[31,32].

By contrast, **NTHi** isolates exhibit greater genetic diversity, with diverse phylogenetic origins and diverse STs. Studying the NTHi population by multi-locus sequence typing (MLST) is complex because most STs are not sufficiently related to

form clonal complexes (CCs). Thus, some studies support a clade-related classification (I-VI) based on the presence or absence of 17 accessory genes, which reveals that most of NTHi belong to clades V and VI[28,30,33].

2.3. Virulence factors and regulation

Numerous virulence factors are involved in the pathogenesis of *H. influenzae*, with most related to adherence, immune system evasion, biofilm formation, and host cell invasion. This high genetic variability, particularly in NTHi, combined with virulence factor regulation, contribute to excellent bacterial adaptation that ensures the successful survival and infection of this species.

2.3.1. Cell wall surface

The cell wall of *H. influenzae* is about 20 nm thick and has the typical structure of a gram-negative bacterium, including (from the outside to the inside) the capsule polysaccharide, which is present in some strains; the outer membrane, which is composed of lipooligosaccharide (LOS); a thin layer of peptidoglycan; and the inner cell membrane[18].

The **capsule polysaccharide** is a virulence factor that protects against complement-mediated killing and opsonisation, thereby preventing phagocytosis[34]. The ***cap* locus**, comprising regions I to III, contains the genes required for capsule synthesis: region I contains *bex* operon required for capsule polysaccharide translocation; region II codes for genes involved in polysaccharide biosynthesis; and region III contains the *hcs* operon, required for post-translational processing. Six capsule types (a to f) have been described (Figure 6), which share regions I and III but differ in region II based on the polysaccharide they synthesise[31]. The *cap* locus is susceptible to duplication, deletion, and interruption, which can affect virulence. Duplication of this locus results in a thicker capsule that increases bacterial protection. Most serotype b, and some serotype a, strains have duplicated *cap* loci, which may explain why they are clinically more significant than other serotypes[34–36].

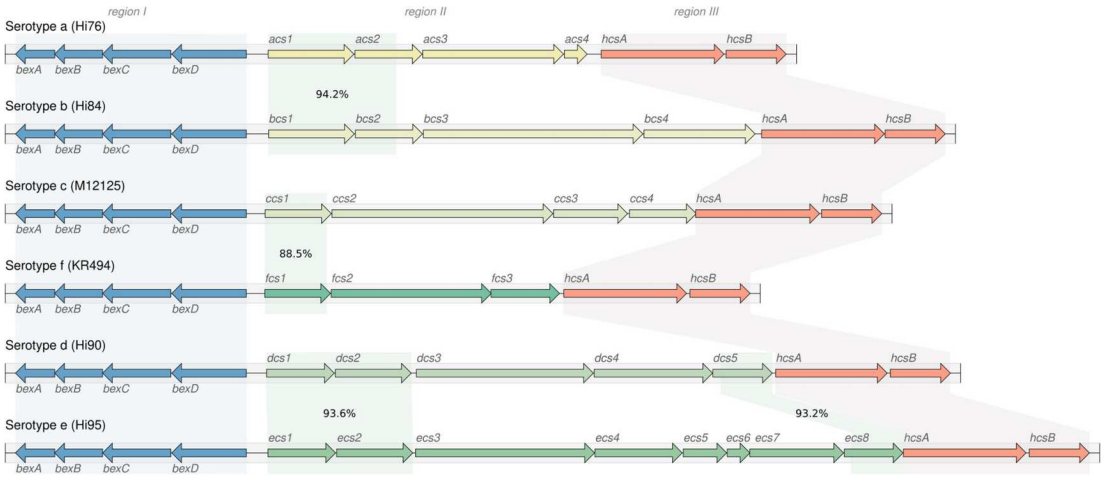


Figure 6. Representation of the six *cap* loci described in *H. influenzae*. The shading represents the homology of the different zones (87% and 90%, in region I and III, respectively). Adapted from Watts CS, *et al*.[31].

LOS is a glycolipid formed from lipid A. Its core, unlike that of lipopolysaccharide (LPS), lacks the O-antigen formed from oligosaccharide polymers of varying length. Lipid A is anchored to the outer membrane and attached to the inner core via 2-keti-3-deoxyctulosonic acid (Kdo). The inner core comprises three heptoses that bind to an outer core with different combinations of glucose, galactose, N-acetylglucosamine, phosphorylcholine (PCho), and sialic acid[37,38] (Figure 7). Biosynthesis of the outer glycolipid core varies between strains due to phase variation of some LOS biosynthetic genes, which contain four base pair repeats within their 5′ coding region that affect protein production. PCho is essential for LOS-mediated pathogenicity. It activates the complement system by binding to C-reactive protein and mimics platelet-activating factor (PAF) by binding to PAF-receptors on epithelial cells. Colonising strains appear to contain more PCho, which promotes adhesion. By contrast, invasive strains have less PCho due to phase variation in *lic1A*, encoding choline kinase and responsible for the addition of PCho to the LOS. Furthermore, the uptake and addition of sialic acid to LOS, mediated by sialyltransferases (Lic3A, Lic3B, SiaA, and LsgB) and influenced by phase variation, also affects complement system activation[34,39].

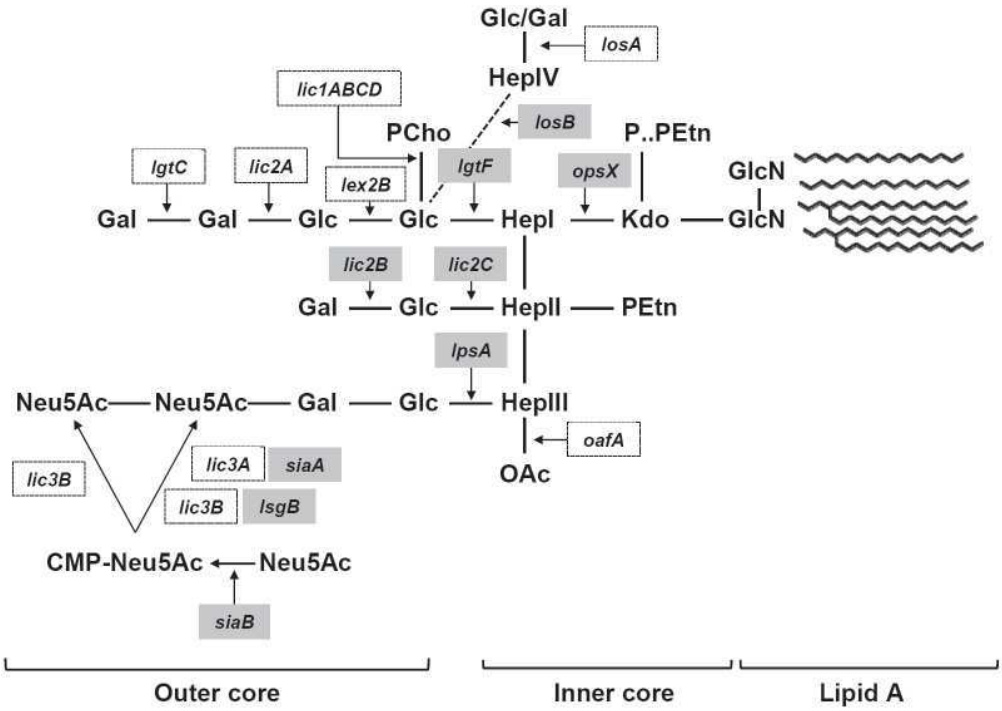


Figure 7. Lipooligosaccharide (LOS) structure. Genes encoding enzymes involved in the LOS biosynthesis are labelled white (phase-variable genes) or grey (non-phase-variable genes). GlcN, glucosamine; Kdo, 2-keto-3-deoxyoctulosonic; PEtn, phosphoethanolamine; Hep, heptose; Glc, glucose; Gal, galactose; Neu5Ac, sialic acid; PCho, phosphorylcholine; OAc, O-acetyl group. Image adapted from Garmendia J, *et al*.[38].

The **outer membrane proteins** (OMPs) serve as adhesins and immune system evasins (Figure 8). The most abundant OMP, P2, interacts with mucin, to facilitate infection in conditions of altered mucus clearance. Although this protein is highly immunogenic, some regions exhibit a high degree of variability in both size and amino acid sequence, allowing bacteria to evade the immune system and develop into chronic infections. The P5 protein, which also varies antigenically, acts as an adhesin that can bind to mucins, ICAM-1, and CEACAM1, and facilitates escape from the complement system by binding factor H. The P4 protein is involved in the uptake of NAD and the binding of laminin, fibronectin, and vitronectin, whereas the P6 protein is associated with peptidoglycan and laminin binding[39–42]. Protein D binds to monocytes and epithelial cells

and is related to intracellular invasion and persistence. Although the mechanism is unknown, it also impairs ciliary activity and damages ciliated cells[39,41]. Protein E is a 16 kDa lipoprotein found in most NTHi strains and is involved in adhesion and intracellular invasion. Its interaction with laminin improves adhesion, whereas its interactions with plasminogen and vitronectin affect the complement system by degrading C3 and inhibiting membrane attack complex (MAC) formation, respectively[42–44]. Protein F binds to laminin and vitronectin, whereas protein H, found primarily in serotype b and f strains, binds to factor H and vitronectin[40]. FadL (also called OmpP1) binds CEACAM1 to favour bacterial interaction with host cells and epithelial infection. It also binds to the free long-chain fatty acids released by damaged epithelial cells and that have bactericidal effects. Loss of FadL has been described as an adaptive mechanism used by *H. influenzae* to persist, and although its ability to infect the epithelium is reduced, survival against the bactericidal effects of long-chain fatty acids is increased[45].

Other cell wall surface proteins associated with adhesion to human epithelial cells are classified as **piliated adhesins**. These include haemagglutinating pili, type IV pilus, and **non-piliated adhesins**, such as high molecular weight adhesin 1 and 2 (HMW1/2), *H. influenzae* adhesin (Hia), *Haemophilus* surface fibril (Hsf), and *Haemophilus* adhesion and penetration protein (Hap) (Figure 8).

The **haemagglutinating pili** are approximately 450 nm in length and are distributed peritrichously. They adhere to epithelial cells and agglutinate with AnWj-expressing erythrocytes. The haemagglutinating pili is encoded by the *hif* operon, which comprises five genes (*hifA-hifE*): *hifA* encodes the major structural pilin subunit, *hifB* encodes a chaperone, *hifC* encodes an assembly platform, *hifD* encodes a protein that is most likely involved in the regulation of pili polymerisation, and *hifE* encodes a protein found at the tip of the pili that is thought to contain the epithelial cell binding site. The promoter regions of *hifA* and *hifB* have a variable numbers of TA repeats, and the expression of the operon can be regulated by phase variation, increasing during colonisation and decreasing during infection[39,46].

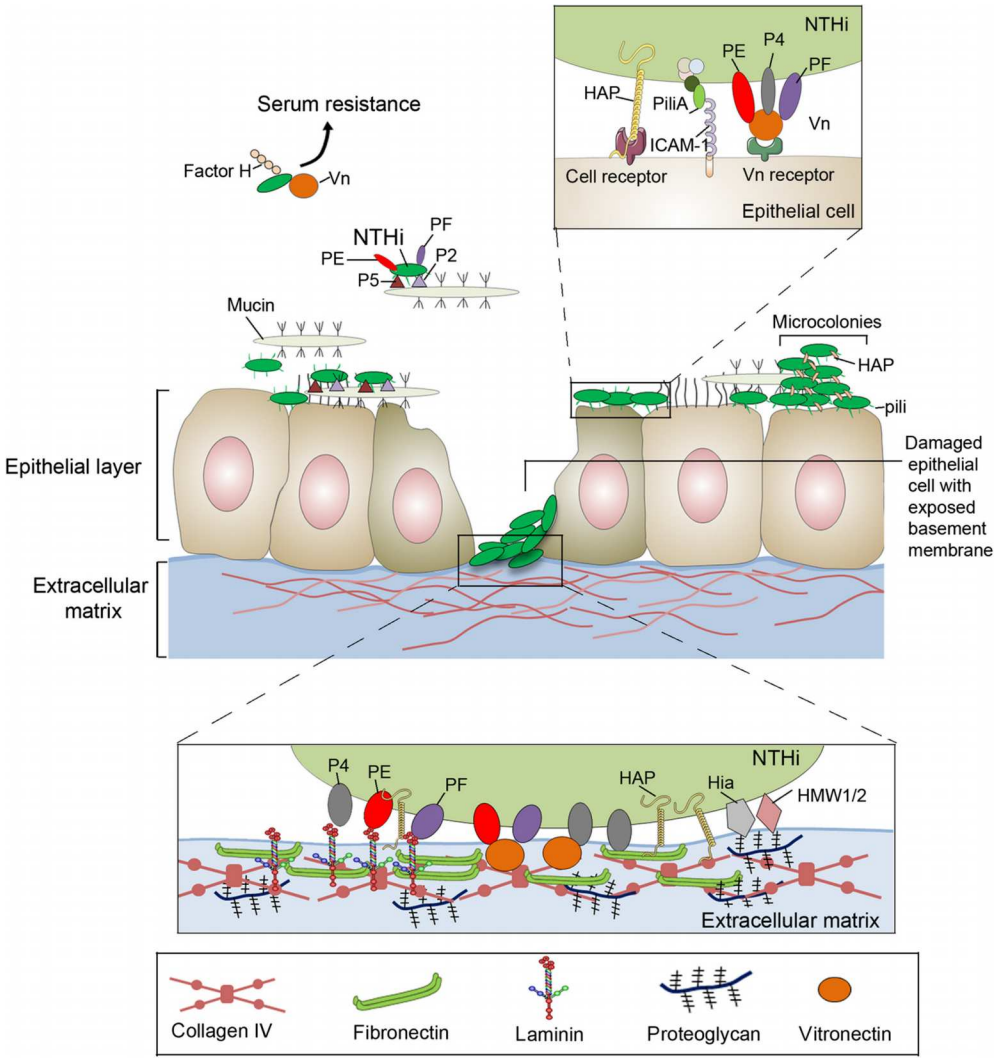


Figure 8. Interaction between cell wall virulence factors and host. Vn, vitronectin; PE, protein E; PF, protein F. Adapted from Duell BL, *et al*.[43].

The **type IV pilus** (Tfp), encoded by *pil* and *com* operons, is implicated in competence, nasopharyngeal colonisation, motility, and biofilm formation. Furthermore, it is capable of binding to ICAM1 and mucin. The *pil* operon encodes for PilA (the major pilin subunit), PilB (the assembly ATPase), PilC (a predicted cytoplasmatic membrane protein), and PilD (a type IV prepilin peptidase), whereas the *com* operon encodes for ComA, ComB, ComC, and ComD, which stabilise the ComE secretin[43,47].

HMW1 and HMW2, which are found in 80% of NTHi isolates, are the most common non-piliated adhesins. They are encoded by two distinct chromosome loci, each containing three genes: *hmw1A* or *hmw2A*, encoding adhesins; *hmw1B* or *hmw2B*, encoding outer membrane translocators; and *hmw1C* or *hmw2C*, encoding cytoplasmatic proteins that stabilise the adhesins prior to translocation. Furthermore, *hmw1B-hmw2B* and *hmw1C-hmw2C* are functionally interchangeable (99% and 97% identical, respectively). Although both proteins are involved in the adhesion to respiratory epithelial cells, their binding specificities differ, resulting in interactions with different host cell receptors. HMW1 recognises a sialylated glycoprotein, whereas the receptor for HMW2 is unknown[39,48].

Despite the absence of HMW1/2 adhesins in 20% to 30% of NTHi strains, most can adhere independent of pili, implying the presence of non-piliated adhesins. Hia, Hsf, Hap, and IgA1 protease are members of the autotransporter family, being distinguished by an N-terminal domain containing a secretion signal peptide, an extracellular surface-exposed passenger domain with binding or protease functions, and a C-terminal β-barrel domain anchored to the outer membrane[49] (Figure 9).

Hia and Hsf are 72% identical, with their N- and C-terminal regions showing the highest homology. However, Hia is expressed by most NTHi strains without HMW1/2 and Hsf is ubiquitous among capsulated strains. They share adhesive specificities for epithelial cells, although Hia has two binding domains and Hsf has three. Moreover, Hsf binds to vitronectin, preventing MAC formation and complement-mediated killing[34,37,49].

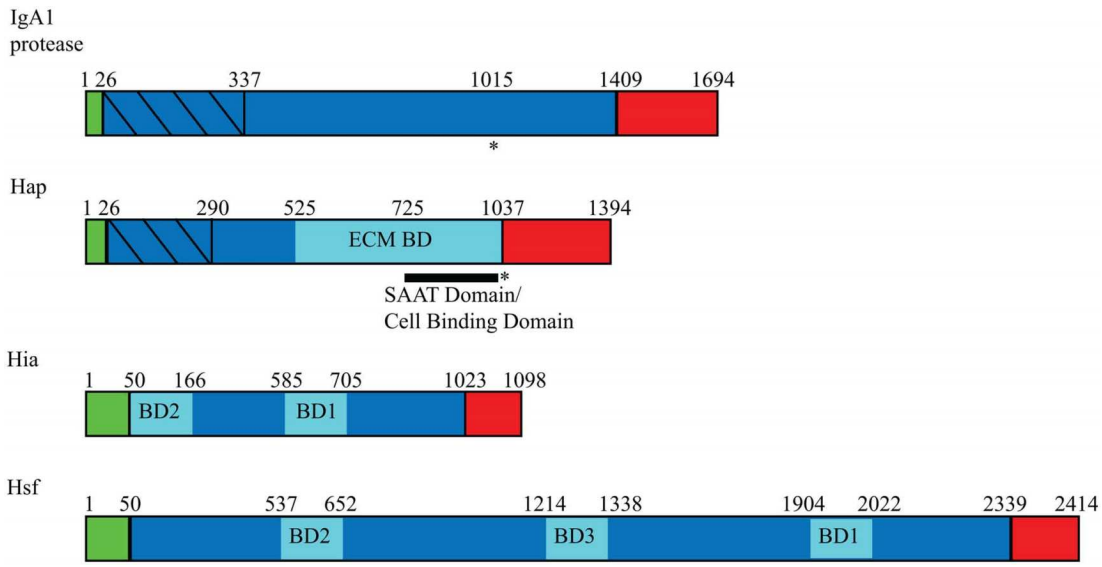


Figure 9. Autotransporter family. Representation of signal peptide in green, passenger domain in blue, and β-barrel domain in red. Binding domains (BD) are represented in light blue, and protease domains with diagonal lines. Asterisks (*) indicate autoproteolysis regions. SAAT, self-associating autotransporter. Adapted from Spahich NA, *et al*.[49].

Hap is found in most *H. influenzae* isolates. It promotes human epithelial adhesion and invasion by binding extracellular matrix components such as laminin, fibronectin, and collagen IV. It also has a self-associating autotransporters (SAAT) domain that promotes Hap-Hap interactions, resulting in the formation of microcolonies, aggregation, and biofilms[49]. It also demonstrates autoproteolysis mediated by serine protease activity, with a catalytic site composed of Lys-98, Asp-140, and Ser-243 that causes the extracellular release of the passenger domain. Because of the autoproteolytic activity, some bacteria from the biofilm may be released to colonise elsewhere. However, the secretory leucocyte protease inhibitor, a component of respiratory secretions, is upregulated in inflammatory conditions and can inhibit Hap autoproteolysis, resulting in Hap accumulation on the bacterial surface and increased adherence to epithelial cells and extracellular matrix components[37,43].

Despite being a secreted protein, IgA1 protease has the characteristic structure of the autotransporter family and possesses autoproteolysis activity that results in

passenger domain release from the membrane-anchored β-barrel domain by cleaving the hinge region of human IgA1. This is the predominant IgA in the respiratory mucosa, which binds bacteria to prevent their attachment to epithelial cells and mediates antibody-dependent cytotoxicity. The proteolysis of the IgA1 hinge region releases the crystallisable region (Fc) and the antigen-binding (Fab) fragments, leading to evasion from mucosal immunity and enhanced bacterial colonisation. Contrasting with capsulated isolates, NTHi isolates present highly variable IgA1 proteases that help them avoid immune detection and that result in different protease activities[49]. Furthermore, two IgA1 proteases have been described, one coded by the *igaA* gene, present in all NTHi strains, and the other coded by *igaB*, present in 40% of NTHi strains. The IgA B protease also favours intracellular persistence by disrupting lysosomal acidification through cleavage of the hinge region of lysosome-associated membrane protein 1 (LAMP1), which has microbicidal activity[41].

Other cell wall-associated proteins participate in the uptake of exogenous porphyrin or iron to compensate for the inability of *H. influenzae* to synthesise protoporphyrin IX, the heme precursor. Passive diffusion allows the transport of small molecules through porins, but energetic transport is required for molecules containing heme or iron, which are larger and in lower concentrations. *H. influenzae* has a diverse range of mechanisms for haem and iron uptake. **TonB-dependent receptors**, such as haemoglobin-haptoglobin binding proteins (Hgp), haem-haemopexin acquisition proteins (Hxu), and haem-utilisation protein (Hup), are OMPs that require the proton motive force of the cytoplasmic membrane transduced by the TonB complex (TonB, ExbB, and ExbD) to transport of exogenous haem and iron. Genetic alterations in TonB-dependent receptors reduce the virulence of *H. influenzae* infections. By contrast, **TonB-independent receptors**, such as the periplasmic haem-binding protein (HbpA), and the sensitivity to antimicrobial peptides protein A (SapA), bind haem via ABC transporters[50,51].

2.3.2. Regulation

The signalling pathways that control the virulence factors of *H. influenzae* are mostly unknown. However, the regulation of gene expression by **phase variation** is common in this species, and it has been described in many genes involved in the synthesis of LOS, OMPs, haemagglutinating pili, and some adhesins[40,43]. In addition, two **transcriptional regulatory systems** have been described, ArcAB and FNR, that can modulate the expression of the genes required for adaptation to changes in oxygen availability[52]. Moreover, *H. influenzae* has a **quorum sensing** mechanism, mediated by autoinducer 2 (AI-2), which bacteria secrete to sense cell density. LuxS is required for AI-2 synthesis and is highly expressed during the infection process, modulating biofilm formation, intracellular invasion, and pathogenicity[53].

2.4. Molecular epidemiology

Serotype b was previously a main cause of invasive *H. influenzae* and the leading cause of bacterial meningitis in children under 5 years of age, being associated with severe sequelae and high mortality[54,55]. The incidence of meningitis in infants varied by geography: 8-27 cases per 100,000 in Europe, 32-60 in North America, 17-25 in South America, 25-60 in Africa, 3-163 in Asia, and 25-27 in Australia and New Zealand. However, some subpopulations had even higher incidences with rates of 152-282 cases per 100,000 among the Navajo, Apache Indians, and Native Alaskans[56].

The introduction of the **serotype b vaccine** led to changes in the epidemiology of *H. influenzae* disease. Initially, a polysaccharide vaccine was developed, but it failed to induce antibody production in infants under the age of 2 years (the main risk group) due to their immature B-cell function. Consequently, in the early 1990s, a serotype b polyribosyl-ribitol phosphate polysaccharide was combined with a carrier protein (e.g., the tetanus toxoid) to produce a conjugate vaccine that could induce antibody production in infants younger than 2 months old[57,58]. Currently, the Spanish vaccination schedule uses a hexavalent vaccine against polio, diphtheria, tetanus, pertussis, and

hepatitis b, which is administered at 2, 4, and 11 months of age[59]. Adding the vaccine has caused a significant reduction in invasive *H. influenzae* serotype b disease in children under the age of 5 years, from 8.13 million in 2000 to 340,000 in 2015. Although serotype b vaccination is available worldwide, its inclusion has been slower in low- and middle-income countries, while vaccination is still voluntary in other countries (e.g., China). Therefore, invasive serotype b disease remains a matter of concern[60].

In the United States between 1989 and 2009-2015, the overall incidence of invasive *H. influenzae* disease due to capsulated or NTHi strains, decreased from 4.39 to 1.7 cases per 100,000 population, respectively[61,62]. In Europe, the overall incidence of invasive *H. influenzae* disease in 2018 was 0.8 cases per 100,000 population[63] (Figure 10).

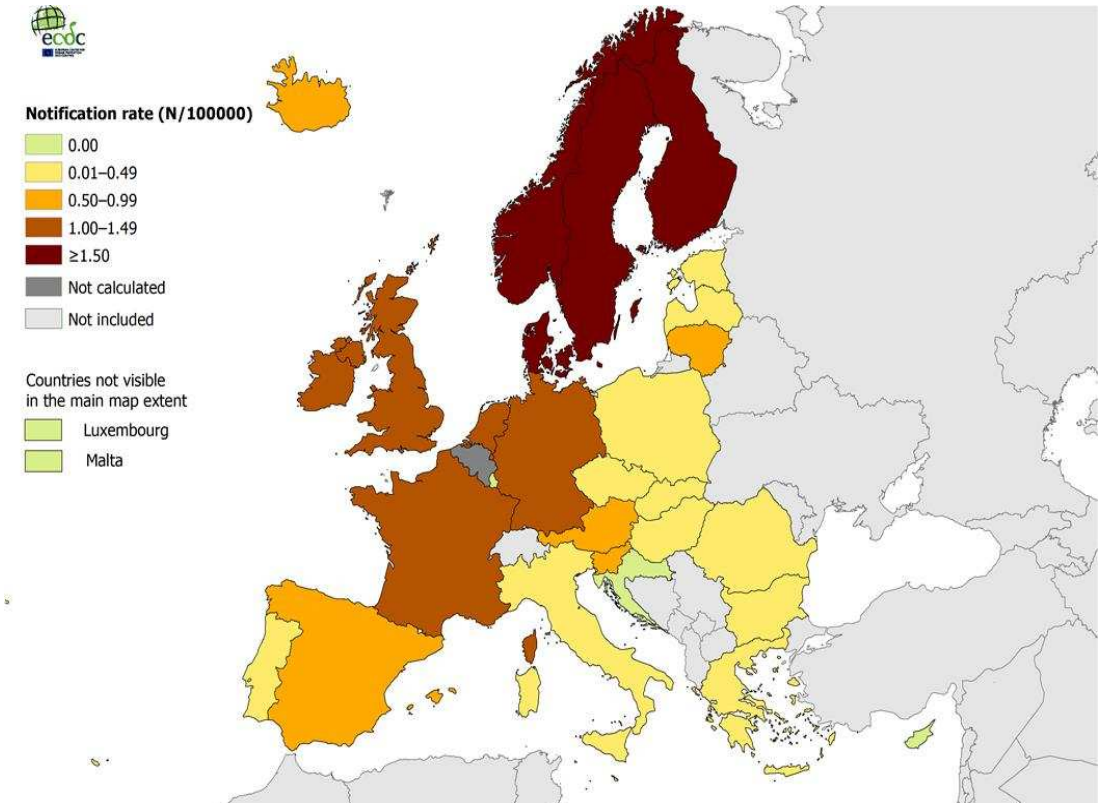


Figure 10. Confirmed *H. influenzae* disease cases per 100,000 population by country in Europe (2018). Image adapted from European Centre for Disease Prevention and Control[63].

NTHi is currently the leading cause of invasive *H. influenzae* disease worldwide, especially in neonates, the elderly, and those with underlying conditions. It accounts for 54.6% to 82% of cases and has shown a 3% annual increase[64–67]. However, the frequency of capsular serotypes varies by geographical region, with **serotypes f** and **e** most common in Europe (74.1% and 21.4%, respectively), and **serotype a** isolates increasing at annual rate of 13% in North America between 2002 and 2015. Severe cases due to serotype a, notably ST23 and related strains, are being observed in indigenous communities in Canada and the United States, where they represent 23.1% of invasive *H. influenzae* disease cases. Although the reasons are unknown, population genetics, environmental factors, and socioeconomic factors may contribute to increased colonisation and higher transmission of serotype a in certain populations. Continuous surveillance of the epidemiology of *H. influenzae* disease is necessary to control the evolution of the bacterial population and develop effective prevention strategies[55,68].

2.5. Colonisation

The **upper respiratory tract** is the main reservoir of *H. influenzae*, which allows person-to-person transmission through salivary droplets. Carriage rates vary by age and location (nasopharynx or oropharynx). For unknown reasons, children are more likely to be colonised than adults, with up to 60% having nasopharyngeal carriage. Most cases since the introduction of the conjugate vaccine are associated with NTHi strains, though 1% still carry serotype b[69,70]. Other risk factors for colonisation include primary immunodeficiencies, sickle cell disease, viral respiratory infection, and large family size[39,71]. The **lower respiratory tract**, which was previously thought to be sterile, can also be colonised by *H. influenzae*. However, this is uncommon in healthy individuals (only 4%), while patients with cystic fibrosis or chronic obstructive pulmonary disease (COPD) show NTHi carriage in up to 30% and 50% of clinically stable cases, respectively. Persistent lung colonisation in these diseases causes tissue damage and a gradual decline in the lung function, which lead to clinical manifestations[6,72,73].

NTHi preferentially adheres to **mucus**, **non-ciliated cells**, and **damaged epithelium**, which is enhanced by the presence of peritrichous pilus, the autotransporter family of proteins, and some OMPs, such as P2 and P5[37,46]. In addition, *H. influenzae* can interact with macrophages and epithelial cells via **intracellular invasion** and **paracytosis** (Figure 11). These mechanisms allow immune system evasion and persistent colonisation[37,46].

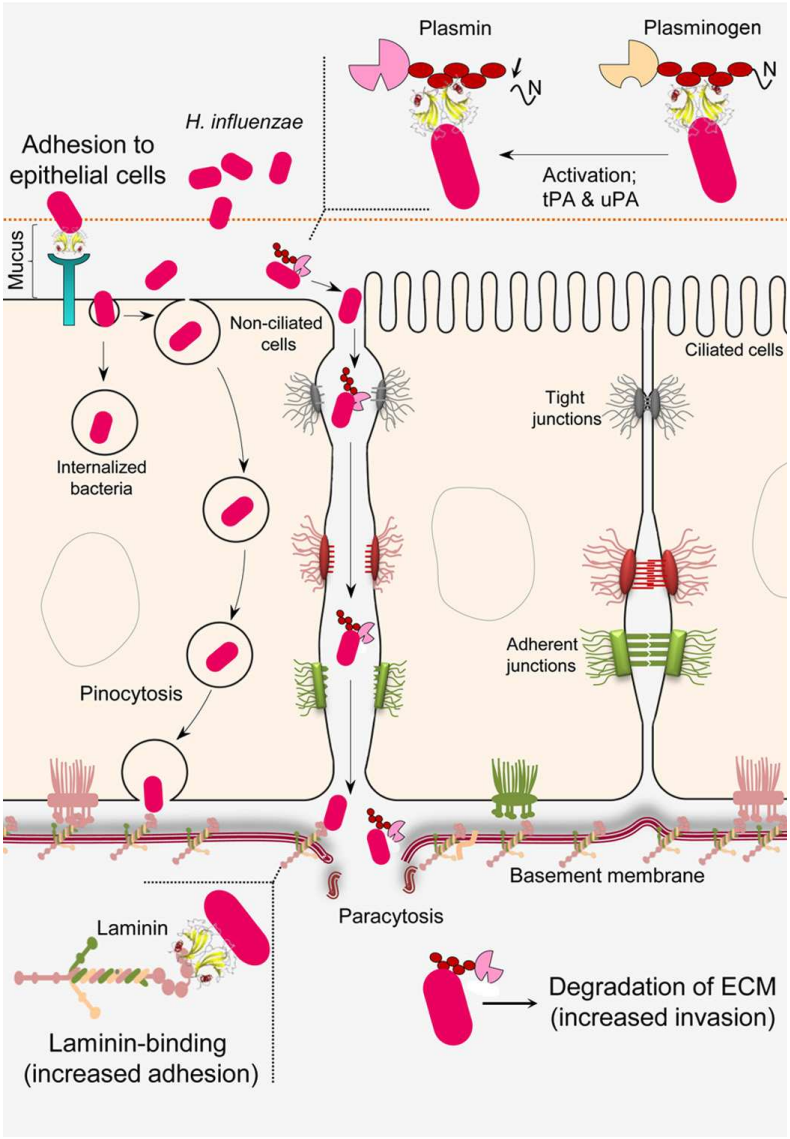


Figure 11. Interaction of *H. influenzae* with host cells. ECM, extracellular matrix. Image adapted from Singh B, *et al*.[44].

To facilitate colonisation, *H. influenzae* blocks the action of antimicrobial peptides like β-defensin and cathelicidin LL-37 via the SapA transporter and LOS, respectively. Some strains also show reduced susceptibility to the complement system due to the activity of LOS glycosyltransferase Lic2B, whereas others inhibit the action of IgA, via Ig1A protease. Phase variation, along with changes in amino acid sequences, cause a change in immunogenic surface proteins[72,74]. These changes allow continued **immune system evasion**. In addition, **metabolic adaptation** facilitates successful colonisation in environments with limited nutrient availability, low pH, and high levels of reactive oxygen species. Thus, *H. influenzae* can produce iron-metabolising proteins, urease, and antioxidants involved in its survival in the respiratory tract[72].

The high genetic heterogeneity and plasticity of NTHi allow different isolates to colonise at the same time, ensuring survival of the species by interacting with the host in different ways. Adaptation to the host improves adhesion to host cells, immune evasion, dissemination to adjacent respiratory tract structures, and competition with other colonisers, thereby facilitating infection[38].

2.6. Infections caused by *H. influenzae*

Before the conjugate vaccine, serotype b was the leading cause of invasive *H. influenzae* disease. However, because the vaccine has no effect on NTHi strains, these have emerged as a leading cause of both non-invasive and invasive *H. influenzae* infections (Figure 12). The risk of contracting *H. influenzae* is higher for unvaccinated children younger than 5 years, people older than 65 years, immunocompromised individuals, and some indigenous populations. Underlying conditions, such as human immunodeficiency virus (HIV), asplenia, or sickle cell disease, also increase risk[42,75].

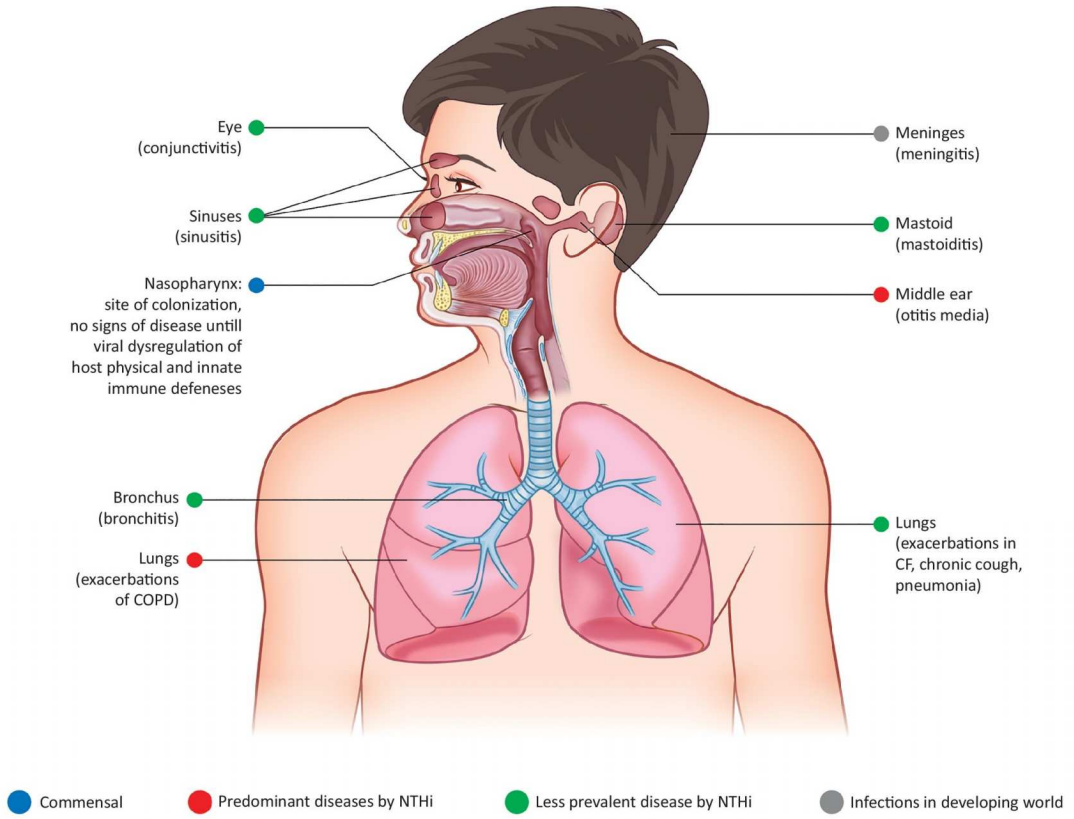


Figure 12. Diseases caused by *H. influenzae*. COPD, chronic obstructive pulmonary disease; CF, cystic fibrosis. Image adapted from Bakaletz LO, *et al.*[76].

H. influenzae colonises the upper respiratory tract, where it is linked to **mucosal infections**. Otitis media can occur at any age but is most common in infants aged 6-24 months. It is the second most common paediatric diagnosis after respiratory infections, with 23% of children younger than age of 1 year likely to have had at least one episode and NTHi accounting for nearly half[42,77]. NTHi is also present in 41% of children with sinusitis and in 50% and 25% of children and adults with conjunctivitis, respectively. Of note, 20%-70% of children with conjunctivitis subsequently develop acute otitis media caused by the same NTHi strain[73]. Viral infections, such as rhinovirus or respiratory syncytial virus, can disrupt the epithelial barrier and promote the overgrowth of colonising *H. influenzae*, which will lead to secondary respiratory mucosal infections[42].

H. influenzae is also a leading cause of **lower respiratory tract infections**, causing 80% of persistent bronchitis (characterised by a chronic cough) in children younger than 5 years. By contrast, acute bronchitis occurs when a viral infection damages the respiratory epithelium, and it is also common in patients with COPD and asthma. In patients with cystic fibrosis and COPD, respiratory colonisation by NTHi increases the inflammation and tissue damage, triggering *Pseudomonas aeruginosa* colonisation and acute exacerbations, respectively. In fact, NTHi is the most common aetiological agent in acute exacerbations of COPD, accounting for more than 50% of all cases, followed by *S. pneumoniae*, *Moraxella catarrhalis*, and *P. aeruginosa*. Exacerbations are characterised by fever, worsening cough, and increased dyspnoea and sputum production. They not only reduce quality of life but also increase the risk of hospitalisation and death[42,73,75].

Another disease caused by *H. influenzae* is pneumonia, which is an inflammation of the lung parenchyma. Community-acquired pneumonia affects people not exposed to healthcare settings, whereas hospital-acquired pneumonia refers to patients currently or recently exposed to healthcare settings, including hospitals, day care centres, and nursing homes. Community-acquired pneumonia is caused by the aspiration of upper respiratory tract secretions and is associated with viral infections, particularly in elderly patients or those with underlying diseases. *H. influenzae* is the second most common cause of community-acquired pneumonia after *S. pneumoniae*, accounting for 7%-16% of cases and manifesting as fever, chills, productive cough, dyspnoea, and chest pain, particularly in patients who have completed their pneumococcal vaccination schedule. Hospital-acquired pneumonia, which has a worse prognosis, is also caused by upper respiratory tract microbiota, but is more often associated with mechanical ventilation and intubation[75,78,79].

In terms of **invasive disease**, *H. influenzae* is no longer the most common cause of meningitis since the vaccine. It now causes only sporadic cases of meningitis in the elderly, people with underlying conditions, and unvaccinated children. Colonising

bacteria from the upper respiratory tract can invade the bloodstream, cross the blood-brain barrier, and infiltrate the subarachnoid space, where they cause an intense inflammatory response that can spread to the brain parenchyma and cause cerebral infarctions and oedema. These patients exhibit fever, headache, photophobia, nausea, altered consciousness, and nuchal rigidity. Meningitis is associated with fatality rates of 5%-28% and with long-term complications in 26% of survivors, including hearing loss, intellectual impairment, and behavioural disorders. Apart from meningitis, *H. influenzae* can cause other severe invasive diseases, such as bacteraemia, epiglottitis, septic arthritis, osteomyelitis, pericarditis, and cellulitis[60,75,78].

2.7. Treatment and antibiotic resistance

Infections caused by *H. influenzae* are usually treated with **β-lactams**, such as aminopenicillins or cephalosporins, with **fluoroquinolones** and **macrolides** as alternative therapeutic options, particularly in patients with β-lactam-resistant strains[55,78,80].

Haemophilus species are generally susceptible to β-lactams, fluoroquinolones, tetracyclines, co-trimoxazole, macrolides, and aminoglycosides. However, **β-lactamase** production (e.g., TEM-1 and ROB-1) has led to an increase in the number of ampicillin-resistant strains (approximately 15% of all strains). TEM-1 is present in 90%-95% of β-lactamase-producing strains. ROB-1, which is present in a pB1000 plasmid, is less common because these plasmids affect bacterial fitness[18,55,80]. Although the presence of a β-lactamase is the primary resistance mechanism, some ampicillin-resistant strains lack this enzyme and are known as **β-lactamase-negative ampicillin-resistant** (BLNAR) strains. Dabernat *et al*. classified BLNAR strains into three groups (I-III) according to the amino acid substitutions in penicillin-binding protein 3 (PBP3), encoded by *ftsI*, which reduce the affinity of β-lactams for their target (Table 3). Groups I and II are found worldwide and are associated with low-level aminopenicillin resistance, whereas group III strains are found mostly in Asia and are associated with high-level aminopenicillin resistance and cephalosporin resistance[55,80,81].

Table 3. Classification of amino acid substitutions in PBP3. -, no substitution; *, additional substitutions not found in all the strains in the group. Adapted from Dabernat H, *et al*.[81] and García-Cobos S, *et al*.[82].

Group	Asp-350	Ser-357	Ala-368	Met-377	Ser-385	Leu-389	Ala-437	Ile-449	Gly-490	Ala-502	Arg-517	Asn-526
I	Asn*	Asn*	-	Ile*	-	-	-	-	Glu*	Thr* Val*	His	-
IIa	Asn*	-	-	-	-	-	-	-	Glu*	-	-	Lys
IIb	Asn*	-	-	Ile*	-	-	Ser*	-	Glu*	Val	-	Lys
IIc	Asn*	-	Thr*	-	-	-	Ser*	-	-	Thr	-	Lys
IId	-	-	-	-	-	-	-	Val	-	Thr*	-	Lys
III	Asn	Asn	-	Ile	Thr	Phe	-	-	-	-	-	Lys
III-like	Asn	Asn	-	Ile	Thr	Phe*	-	-	-	-	His	-

Fluoroquinolones are frequently used to treat respiratory tract infections in adults. Although fluoroquinolone-resistant strains are uncommon (0.4% in Spain[83] and 2.7% in China[84]), they are emerging globally and should be monitored. Resistance to fluoroquinolones is mainly associated with amino acid substitutions in Ser-84 and Asp-88 of **DNA gyrase** (GyrA) and Gly-82, Ser-84, and Glu-88 of **DNA topoisomerase IV** (ParC). In general, GyrA mutations appear first and decrease the susceptibility to quinolones, with mutations in these genes increasing resistance in a stepwise manner[55].

Macrolides are used to treat respiratory infections, but they must be used with caution to prevent the selection of resistant strains. The main mechanisms of macrolide resistance in *H. influenzae* are the acquisition of **efflux pumps** encoded by *mef* and *msr* genes, the acquisition of the **23S rRNA methylase** encoded by *erm* genes, the overexpression of the **AcrAB efflux pump**, mutations near the macrolide-binding region (A2058) in the V domain of **23S rRNA**, and modifications in **ribosomal proteins L4 and L22** encoded by *rplD* and *rplV*, respectively[55,80,85].

3. *Staphylococcus aureus*

3.1. History, taxonomy, and microbiological characteristics

The first descriptions of the genus *Staphylococcus* date back to the late 19th century. In 1880, Alexander Ogston identified round microorganisms in grape-shaped clusters as a cause of purulent surgical wound infection, naming them *Staphylococcus*, from the Greek *staphylē* (bunch of grapes) and *kókkos* (grain). In 1884, Friedrich Julius Rosenbach isolated and differentiated the first two species based on the colour of their colonies, naming them *S. aureus* and *Staphylococcus albus* (later renamed *Staphylococcus epidermidis*), from the Latin *aurum* (gold) and *albus* (white), respectively[86].

Taxonomic classification of the genus has evolved with improvements in phylogenetic analysis techniques. Staphylococci and micrococci were initially classified in the *Micrococcaceae* family due to their microscopic resemblance, but differences in GC content, oxidase activity, and genetic differences between species has led to a change in classification[87]. Nowadays, *Staphylococcus* is classified based on 16S rRNA sequences, together with the *Aliicoccus*, *Auricoccus-Abyssicoccus*, *Corticicoccus*, *Jeotgalicoccus*, *Macrococcus*, *Nosocomiicoccus*, and *Salinicoccus* genera, in the ***Staphylococcaceae*** family (Figure 13) of the order **Bacillales**, the class **Bacilli**, and the phylum **Firmicutes**[88,89]. Advances in molecular techniques have led to the identification of over 50 staphylococcal species with a 16S rRNA sequence similarity of at least 96.5%, differing from the other genera in this family that share a similarity of 93.4%-95.3%[90]. Staphylococci are usually found as skin and mucous membrane colonisers in humans and animals, but they can also be isolated in the environment or food. Species such as *S. epidermidis*, *Staphylococcus hominis*, *Staphylococcus lugdunensis*, *Staphylococcus haemolyticus*, and *Staphylococcus capitis*, show low pathogenicity, while others are opportunistic pathogens that can cause soft tissue and skin infection, foreign body infection, bloodstream infection, and food poisoning. However, due to the

severity of the infections and antibiotic resistance, *S. aureus* is the most pathogenic species in this genus[91–93].

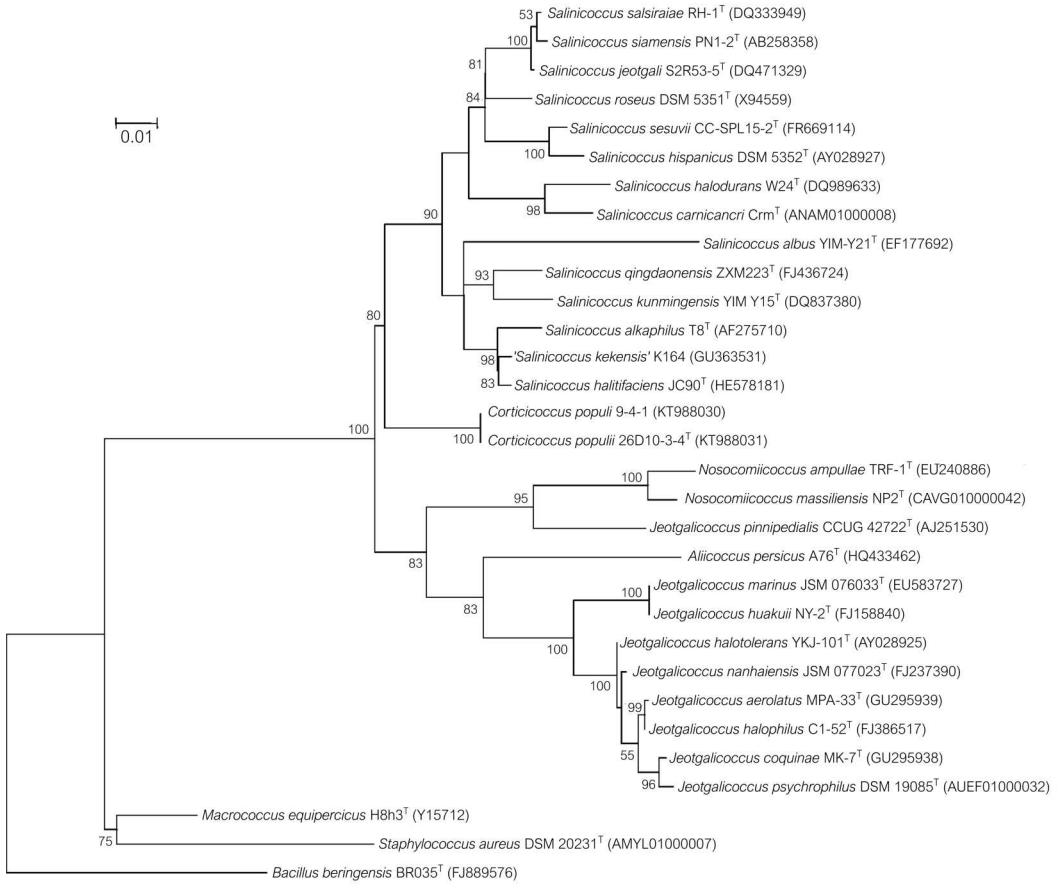


Figure 13. Phylogenetic tree based on the 16S rRNA sequences of *Staphylococcaceae* members. *Bacillus beringensis* (BR035^T) used as an outgroup. Bootstrap values are represented at the nodes. Scale bar = substitutions / nucleotide site. Adapted from Li Y, *et al*.[94].

S. aureus is a non-motile **gram-positive coccus** with a diameter of 0.5-1 μm that groups in pairs, tetrads, and grape-shaped clusters based on its division into different planes. It forms smooth, raised, and shiny colonies with entire margins on blood agar plates and grey to yellow pigmentation, due to the production of triterpenoid carotenoid pigments and their derivates[90]. It can grow in media containing 10% NaCl and at temperatures of 10ºC-45°C, although the optimal temperature range is 30ºC-37°C for

growth. It is a **facultative anaerobe** capable of both respiratory and fermentative metabolisms. Under aerobic conditions, it primarily metabolises glucose via glycolysis and, to a lesser extent, via the pentose phosphate pathway, while under anaerobic conditions it can switch to lactic acid fermentation. Furthermore, its ability to regulate amino acid biosynthesises can result in amino acid requirement if mutations occur in biosynthetic genes[95].

Most *S. aureus* strains produce **catalase**, **coagulase**, and **DNase**, but are oxidase deficient. The clustering observed on Gram staining, along with the production of catalase, are the main features that distinguish most staphylococci from streptococci, which group in chains and cannot produce catalase. The production of coagulase allows *S. aureus* to be distinguished from coagulase-negative staphylococci, which include most staphylococci except for some strains of *Staphylococcus hyicus*, *Staphylococcus intermedius*, and *Staphylococcus schleiferi* subspecies *coagulans*[95,96].

3.2. Genomic profile and population structure

The genome of *S. aureus* has an average length of **2.84 Mb**, an average GC content of 32.7%, and it encodes an average of 2,679 proteins[97]. The pangenome contains 7,457 genes, of which 1,441 are core, 2,871 are accessory, and 3,145 are unique[98]. Most core genes are involved in metabolism, transcription, and translation. Prophages, plasmids, pathogenicity islands, transposons, and cassette chromosomes account for approximately 15%-20% of the *S. aureus* genome, and a significant proportion of accessory and unique genes are associated with these mobile genetic elements[98,99].

On average, an *S. aureus* genome contains about 2,800 genes, implying that core genes account for more than half of the genome. Furthermore, although natural recombination events may occur, they are uncommon, with a relative substitution rate <1 due to recombination vs mutation (r/m)[100,101]. These factors explain the genetic stability and clonal structure of the *S. aureus* population, which is organised in CCs

defined by MLST. Strains that share at least five of the seven MLST loci are part of the same CC. Phylogenetic analyses show high clonality within each CC, but a greater degree of diversity between them[98,102] (Figure 14). Moreover, the level of homoplasy (i.e., independent acquisition or loss of a trait in two different organisms over the course of evolution) is very low within the same CC (0.62%), and higher among different CCs (37.8%). Accordingly, the proportion of accessory genes shared by clones from the same CC is higher than the proportion shared by clones from different CCs[103].

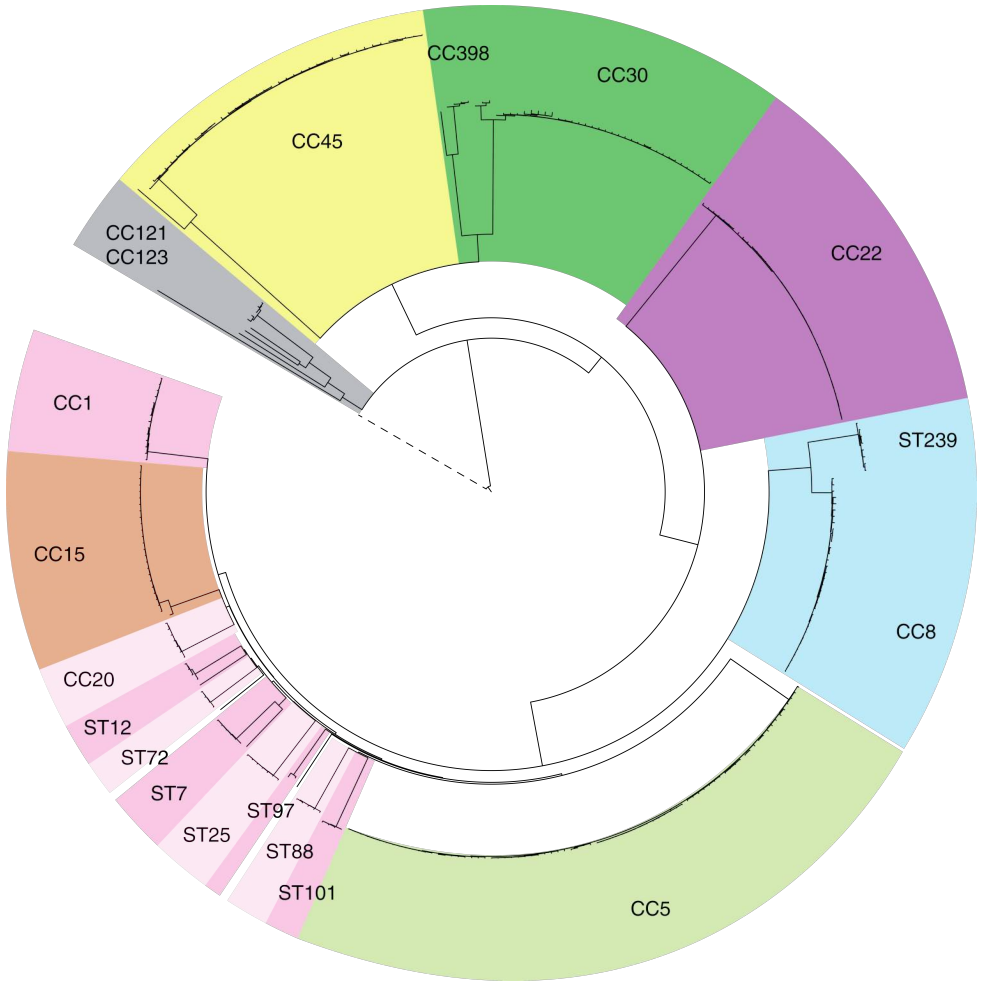


Figure 14. Core SNP phylogenetic tree of *S. aureus* population. Lineages are highlighted according to their corresponding clonal complex (CC) or sequence type (ST). Adapted from Aanensen, DM, *et al.*[103].

3.3. Virulence factors and regulation

S. aureus has a wide range of virulence factors, as evidenced by the variety of associated colonisation sites and infection types. The expression of genes coding for proteins involved in adhesion and immune system defence is upregulated early, whereas after establishing successful colonisation, the expression of toxins increases to facilitate the spread of infection[104,105]. The main virulence factors and regulatory mechanisms of *S. aureus* are described below.

3.3.1. Cell wall surface

The cell surface contains capsular polysaccharides and cell wall anchored (CWA) proteins that act as virulence factors and immune system modulators.

Capsular polysaccharides help to avoid opsonisation and neutrophil-mediated phagocytosis, with the resulting capsulated strains being more virulent than capsule-defective strains. Although 11 capsular serotypes have been identified, most strains belong to serotypes 5 and 8[106]. By contrast, **CWA proteins** have different functions, including host cell adhesion and invasion, immune system evasion, and biofilm formation. Most have the same structure: a cell wall anchor domain in the C-terminus, followed by a hydrophobic transmembrane region, ending with a secretion signal in the N-terminus that carries the protein to the bacterial surface. Their extracellular portions contain a ligand binding domain that allows the adhesin to act[107]. CWA proteins are divided into four groups according to their structure: microbial surface components recognising adhesive matrix molecules (MSCRAMMs), near iron transporter (NEAT) motif, three-helical bundle, and G5-E repeat families[108] (Figure 15; Table 4).

Most CWA proteins are members of the **MSCRAMM family**, which are characterised by their ability to bind to molecules of the host extracellular matrix, such as fibronectin, fibrinogen, and collagen, and therefore promote host cell adhesion and invasion. This protein family includes serine-aspartate repeat group (Sdr), clumping factor A and B (ClfA and ClfB), sialo-binding protein (Bbp), fibronectin-binding proteins (FnBPs), and collagen adhesin (Cna) (Figure 15; Table 4)[104,108,109].

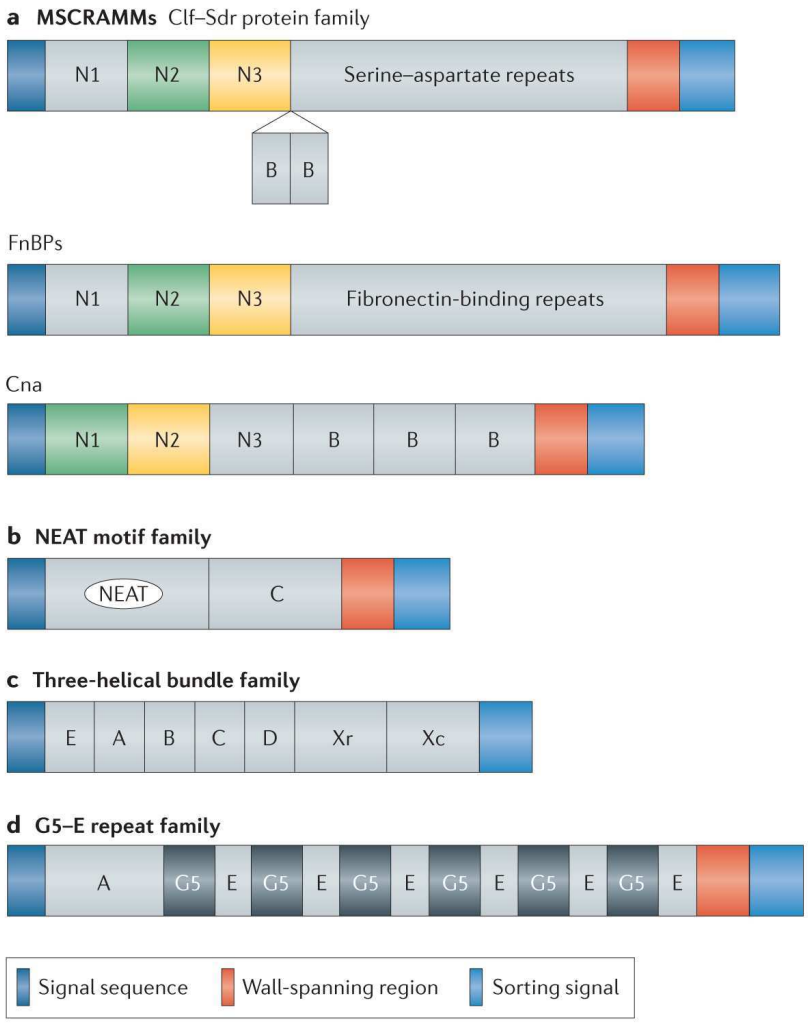


Figure 15. Classification of cell wall-anchored proteins. (a) Microbial surface component recognising adhesive matrix molecules (MSCRAMMs) include clumping factor and serine-aspartate repeat group (Clf-Sdr), fibronectin-binding proteins (FnBPs) and collagen adhesin (Cna). The N-terminal A domain contains three subdomains (N1, N2 and N3), followed by a B domain associated with a variable number of repeats and ligand binding. **(b)** Near iron transporter (NEAT) motif with haemoglobin binding (NEAT) and hydrophilic (C) domains. **(c)** Three-helical bundle family with an N-terminal IgG-binding domain (EABCD), followed by repeat (Xr) and non-repeat (Xc) regions. **(d)** G5-E repeat family with an A domain (A) followed by five glycine residues (G5) repeats, separated by 50-residue sequences (E). Image adapted from Foster TJ. *et al*.[108].

Table 4. Major cell-wall anchored proteins: ligands and functions. Adapted from Foster TJ, *et al*.[108,110].

CWA group	Protein	Ligand	Function
MSCRAMMs	Serine-aspartate repeat group SdrC, SdrD, SdrE	β-neurexin Complement H Desmoglein-1 SdrC homophilic interactions	Adhesion Biofilm formation Colonisation Immune evasion
	Clumping factor group ClfA, ClfB	Complement factor I Cytokeratin 8 and 10 Fibrinogen Loricrin	Adhesion Biofilm formation Colonisation Immune evasion
	Bbp	Fibrinogen Bone sialoprotein	Adhesion
	FnBPs FnBPA, FnBPB	Elastin Fibrinogen Fibronectin FnBPA/B homophilic interactions Plasminogen	Adhesion Biofilm formation Invasion of host cells
	Cna	Collagen type I Complement C1q Laminin	Adhesion Colonisation Immune evasion
NEAT motif family	Iron-regulated surface proteins IsdA, IsdB, IsdH	β3 integrins Cytokeratin 10 Fibrinogen Fibronectin Haemoglobin Loricrin	Adhesion Iron acquisition Invasion of host cells Resistance to antimicrobial molecules Immune evasion
Three-helical bundle family	Spa (Protein A)	IgG IgM TNFR1 von Willebrand factor	Immune evasion Endovascular infection
G5-E repeat family	SasG	SasG homophilic interactions[111] Unknown	Adhesion Biofilm formation

NEAT motif proteins uptake and transport haem from haemoglobin in the extracellular space to the bacterial cytoplasm, where iron is released, allowing bacterial survival in iron-restricted environments. This family includes iron-regulated surface determinant (Isd) proteins that are characterised by the presence of one or more NEAT motifs responsible for haemoglobin binding (Figure 15; Table 4)[108,109].

Protein A (Spa), the only member of the **three-helical bundle family**, is distinguished by an IgG-binding domain (EABCD) in the N-terminal region that forms helical bundles. Ig-G binding inhibits opsonisation and phagocytosis, while Spa disrupts adaptive immune responses and immunological memory[112].

G5-E repeat family has variable repeats of five conserved glycine residues (G5) separated by 50-residue sequences (E) (Figure 15). Proteolytic removal of the A domain and some G5-E repeats exposes the remaining G5-E region on the cell wall surface, promoting cell aggregation and biofilm formation[108]. The main member of this family is *S. aureus* surface protein G (SasG). The cleavage of G5-E repeats causes the release of the SasG N-terminal region and promotes interaction of the remaining G5-E repeats on bacterial surface with other exposed G5-E repeats, favouring cell-to-cell interaction and biofilm formation[111]. However, two SasG variants distinguished by a variable region in the A domain have been described, with some studies suggesting that these variants could have different functions[113].

3.3.2. Secreted proteins

Virulence factors are expressed later in the colonisation process and play important roles in immune system, host cell, and tissue disruption, promoting both dissemination and infection. Secreted proteins can be classified in superantigens, cytotoxins, exoenzymes, and other proteins.

Superantigens cause uncontrolled T-cell activation at the systemic level due to the overproduction of cytokines. The production of one or more superantigens, such as toxic shock syndrome toxin 1 (TSST-1), exfoliative toxins (ETs), staphylococcal enterotoxins, and staphylococcal enterotoxin-like toxins, is associated with clinically relevant strains and severe clinical conditions[104]. Moreover, ETs are serine proteases

that disrupt keratinocyte junctions by desmoglein 1 hydrolysis, causing bulla formation and skin desquamation[114].

Cytotoxins generate pores in the membranes of epithelial, endothelial, and immune cells, causing their destruction by osmotic lysis. The cytotoxin repertoire includes cytolysins (α-, β-, γ- and δ-toxins), the leukocidin family, such as Panton-Valentine leukocidin (PVL), and phenol-soluble modulins (PSMs)[104,115].

S. aureus also secretes **exoenzymes** that destroy host tissues and inactivate host antimicrobial mechanisms. Staphylokinase (Sak) produces plasmin, which initiates the fibrinolytic process by degrading fibrin, C3, and Ig on the bacterial surface, promoting biofilm degradation and bacterial dissemination, and neutralising some defensins. The metalloprotease aureolysin can also degrade antimicrobial peptides, while the antioxidant enzymes catalase, superoxide dismutase, and staphyloxanthin neutralise reactive oxygen and nitrogen species. Finally, *S. aureus* secretes nuclease, which degrades neutrophil extracellular traps and produces deoxyadenosine, leading to caspase-3-mediated immune cell death[109,116].

S. aureus promotes fibrin clot formation and coagulation through von Willebrand factor-binding protein and coagulase. Each binds prothrombin to form staphylotrombin, which in turn, cleaves fibrinogen to fibrin. Other secreted proteins, such as chemotaxis inhibitory protein of *S. aureus*, formyl peptide receptor-like inhibitory proteins, staphylococcal superantigen-like proteins, extracellular adherence protein, and staphylococcal inhibitor of complement, are responsible for immune system modulation, blocking neutrophil chemotaxis, C3 convertases, or Ig binding[115,116].

3.3.3. Regulation

Regulatory genes are essential for managing and changing virulence factor expression, especially under stress due to antibiotic exposure or the immune response, and allow bacteria to adapt, survive, and spread. This explains the high success rate of *S. aureus* infections.

One of the most important regulatory mechanisms is the **accessory gene regulatory** (Agr) quorum sensing system (Figure 16). *S. aureus* produces basal levels of the auto-inducing peptide (AIP), and as bacterial density increase, AIP accumulates, inducing the expression of the *agr* operon. Briefly, AIP binds to the transmembrane histidine kinase AgrC, creating a conformational change that causes its own phosphorylation and both the phosphorylation and activation of the DNA-binding response regulator AgrA. In turn, AgrA binds to the *agr* intergenic region (*agr*-IR) and activates the operon promoters P2 and P3, allowing AgrBDCA and RNAIII transcription, respectively. The transcription of AgrBDCA genes results in the production of AgrC and AgrA, which are required for continued activation of this system. AgrD results in the production of an immature AIP that is processed and exported by the transmembrane AgrB. RNAIII expression causes an upregulation of secreted virulence factors (e.g., α- and δ-haemolysins and PVL), and a downregulation of cell surface associated proteins (e.g., Spa and FnBPs). The *agr* operon has four different allelic variants, each producing different AIPs that act as activators of inhibitors of other AgrC variants[117,118].

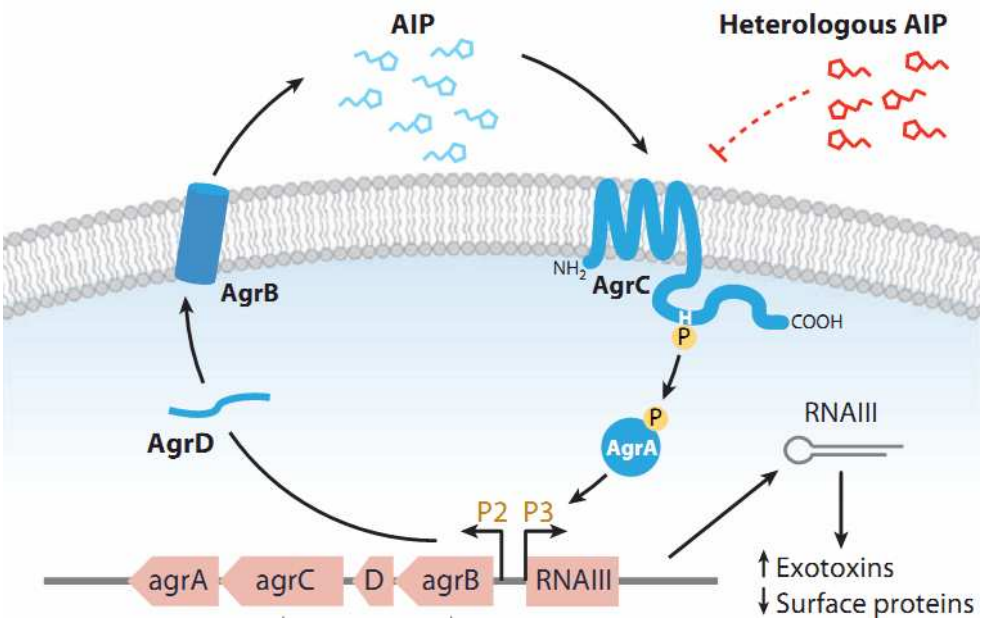


Figure 16. The Agr quorum sensing accessory gene regulator system. Image adapted from Novick RP, *et al*.[119].

Other regulatory systems, such as repressor of toxins (Rot), autolysis regulated locus (ArlRS), sigma factor B (SigB), SarA family, or *S. aureus* exoprotein expression (SaeRS), also affect the expression of different virulence factors. The Rot system is involved in the onset of infection by upregulating the expression of adhesins and immune evasion proteins. However, this system is inhibited by the action of the Agr system in later stages of infection[118]. ArlRS and SigB not only inhibit Agr activity but also promote the clumping and control of approximately 200 genes, respectively. By contrast, the SarA family upregulates the Agr system, promoting AgrBDCA and RNAIII transcription by binding to the P2 and P3 promoters of the *agr* operon. Finally, the SaeRS system modulates the transcription of different virulence factors, including coagulase, nuclease, secreted toxins, and adhesins, by sensing external stimuli, such as changes in pH and sodium chloride concentrations, or the presence of antimicrobial peptides[120].

3.4. Molecular epidemiology

S. aureus is an antibiotic-susceptible species, but the increased use of antibiotics in both humans and animals has resulted in the emergence of antibiotic-resistant clones since the introduction of penicillin in the mid-1940s. Indeed, the first methicillin-resistant strains were reported within a year of introducing methicillin in 1960[121]. The importance of methicillin resistance has even led to the reclassification of this species into **methicillin-susceptible *S. aureus*** (MSSA) and **methicillin-resistant *S. aureus*** (MRSA). The emergence of MRSA clones did not represent a clonal replacement of, but rather an addition to, MSSA[105,122]. A surveillance study across 26 European countries in 2016 revealed that MSSA and MRSA accounted for 60% and 40% of the isolates, respectively[103].

The MSSA population shows great diversity of lineages, including CC7, CC9, CC12, CC15, CC25, CC51, and CC101. At some point, MSSA lineages CC1, CC5, CC8, CC22, CC30, and CC45, acquired the **SCC*mec* cassette**, resulting in the emergence of MRSA (Figure 17). There is no evidence that MRSA clones evolved from predominant

MSSA clones, suggesting that not all MSSA clones can acquire and maintain the SCC*mec* cassette[122,124].

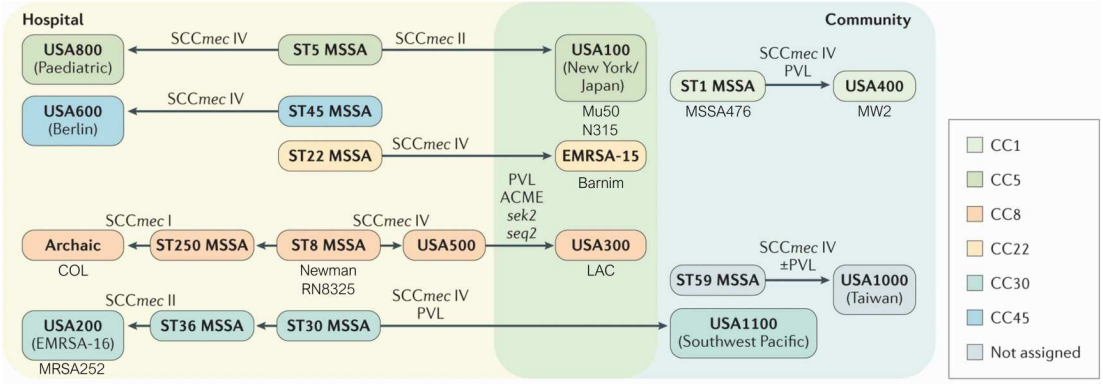


Figure 17. Evolution from MSSA to dominant MRSA clones. The most studied *S. aureus* strains are shown below the corresponding clones[123]. ACME, arginine catabolic mobile element; EMRSA, epidemic MRSA; PVL, Panton–Valentine leukocidin; *sek2* and *seq2*, staphylococcal genes encoding enterotoxins. Image adapted from Lee, AS, *et al*.[105].

MRSA clones are widely distributed across different geographical regions (Figure 18). To better understand the epidemiology of MRSA infections, MRSA clones are classified into healthcare-associated (HA-MRSA), community-associated (CA-MRSA), and livestock-associated (LA-MRSA), based on where they are acquired[122].

When first discovered, most cases were **HA-MRSA** and caused by the ST250 or archaic clone (CC8), with the ST147 or Iberian clone (CC8) the predominant lineage in Spain. However, the epidemiology of HA-MRSA infection evolved, and these clones were gradually replaced by new ones[122,125]. Today, epidemic ST5 (CC5), ST8 or USA300 (CC8), and ST22 or EMRSA-15 (CC22) clones predominate in most European countries. Although CC5 is the most common cause of invasive infection in Europe, some sequence types are endemic in specific geographical areas (e.g., ST125 in Spain, ST225 in Germany and the Czech Republic, and ST339 in Croatia). It is also important to note the EMRSA-15 clone, which originated in the United Kingdom and spread quickly, becoming the dominant HA-MRSA clone in Europe[103,122].

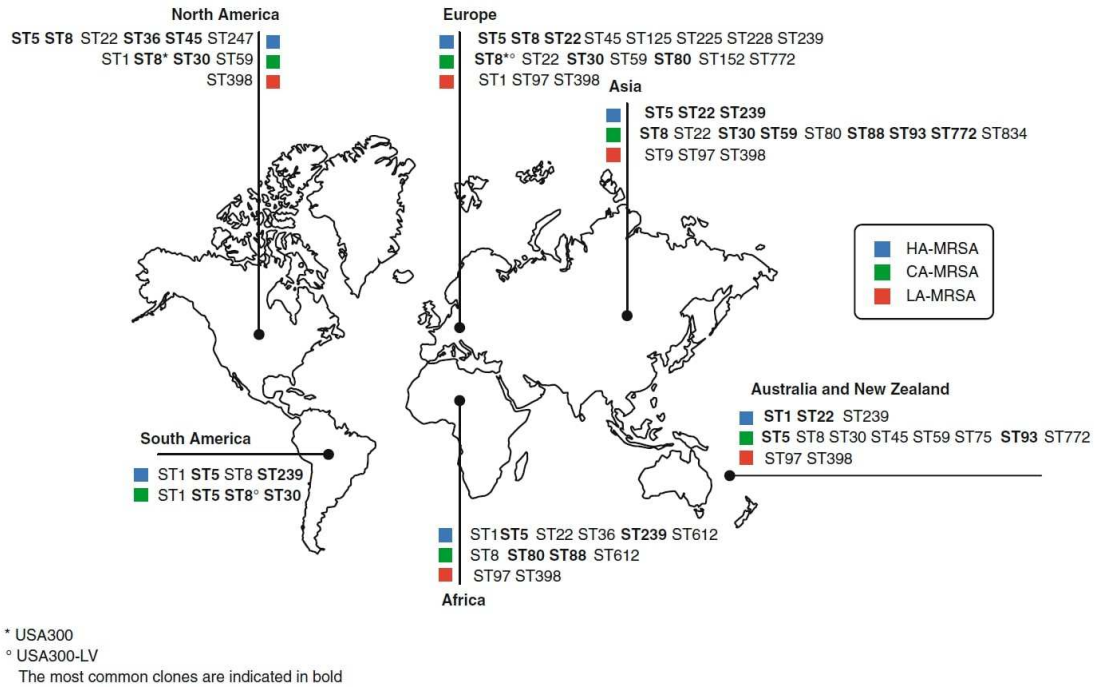


Figure 18. Global distribution of prevalent MRSA clones. Classification: healthcare-associated (HA-), community-associated (CA-), and livestock-associated (LA-) MRSA. Image adapted from Monaco, M, *et al*.[122].

CA-MRSA infections typically affect the skin and soft tissue in healthy and young people who have little contact with the healthcare system. However, they are also emerging in nosocomial settings for patients with risk factors. These clones are generally more susceptible to non-β-lactam antimicrobials than HA-MRSA clones and they are associated with SCC*mec* IV and PVL genes. First detected in remote areas of Australia in the 1980s, when they were associated with ST8 strains, CA-MRSA infections later emerged in the United States in the 1990s due to ST1 or USA400 (CC1) among children without risk factors. This clone was eventually replaced by ST8 or USA300, which has caused large outbreaks of severe invasive disease, including necrotising pneumonia, sepsis, and osteomyelitis[112,122,126]. In Europe, the first CA-MRSA outbreaks were also reported in the early 1990s due to the ST80 clone[127].

Animals, particularly pigs and cattle, serve as a reservoir of MRSA, and have been described as the focus of infection in livestock workers. **LA-MRSA** clones are associated with resistance to tetracycline and trimethoprim-sulfamethoxazole, both of which are commonly used in livestock. The first cases of LA-MRSA infection were identified in Europe in 2005 due to ST398 (CC398) clones. Since then, ST1 (CC1), ST9 (CC9), and ST97 (CC97) clones have been isolated in pigs and from bovine mastitis worldwide[105,122]. These MRSA clones are now widespread, sporadically colonising and infecting both individuals in contact with livestock and patients with underlying disease[128].

3.5. Colonisation

S. aureus colonises the **upper respiratory tract** and **skin**, particularly the hands, perineum, and axillae, even in healthy individuals. Nasal colonisation promotes extra-nasal colonisation: the higher the bacterial load, the more likely it is that other areas of the body will be colonised[129,130] (Figure 19). Furthermore, epithelial, endothelial, and inflammatory cells may serve as habitats for **intracellular** *S. aureus*, protecting it from host defence mechanisms and antimicrobial agents[131]. *S. aureus* has also been isolated from a wide variety of **fomites** in both home and hospital settings, including mobile phones and clothing, where it can survive for months. Hands are the primary vector of transmission from contaminated surfaces or the nasal cavity to different body regions, although this can also occur through the air, contaminating new reservoirs, or through close contact with colonised individuals. This explains the high risk of dissemination among household members, athletes, livestock keepers, and individuals with frequent healthcare contact[126].

Individuals are classified into three groups based on nasal colonisation patterns: 20% are **persistent carriers**; 60% are **intermittent carriers**; and 20% are **non-carriers**. Persistent carriers are typically colonised by a single *S. aureus* strain for long periods, whereas intermittent carriers may carry different strains at different times[126]. The rate of persistent colonisation is 45% during the first weeks of life, but it drops to 21% by 6

months. During adolescence, there is a transition from persistent carriage to intermittent or non-carrying states, reaching an average **nasal carriage rate of about 30%** in the adult population[129,130]. It should be noted that persistent carriers have higher bacterial loads, and subsequently, a higher risk of infections. Some studies demonstrate that most bacteraemia-causing strains in nasal carriers are endogenous to the patient, emphasising the importance of controlling nasal colonisation[132,133]. Decolonisation with mupirocin or chlorhexidine may help prevent subsequent infections in specific situations, such as recurrent skin and soft tissue infection or before cardiac surgery[134].

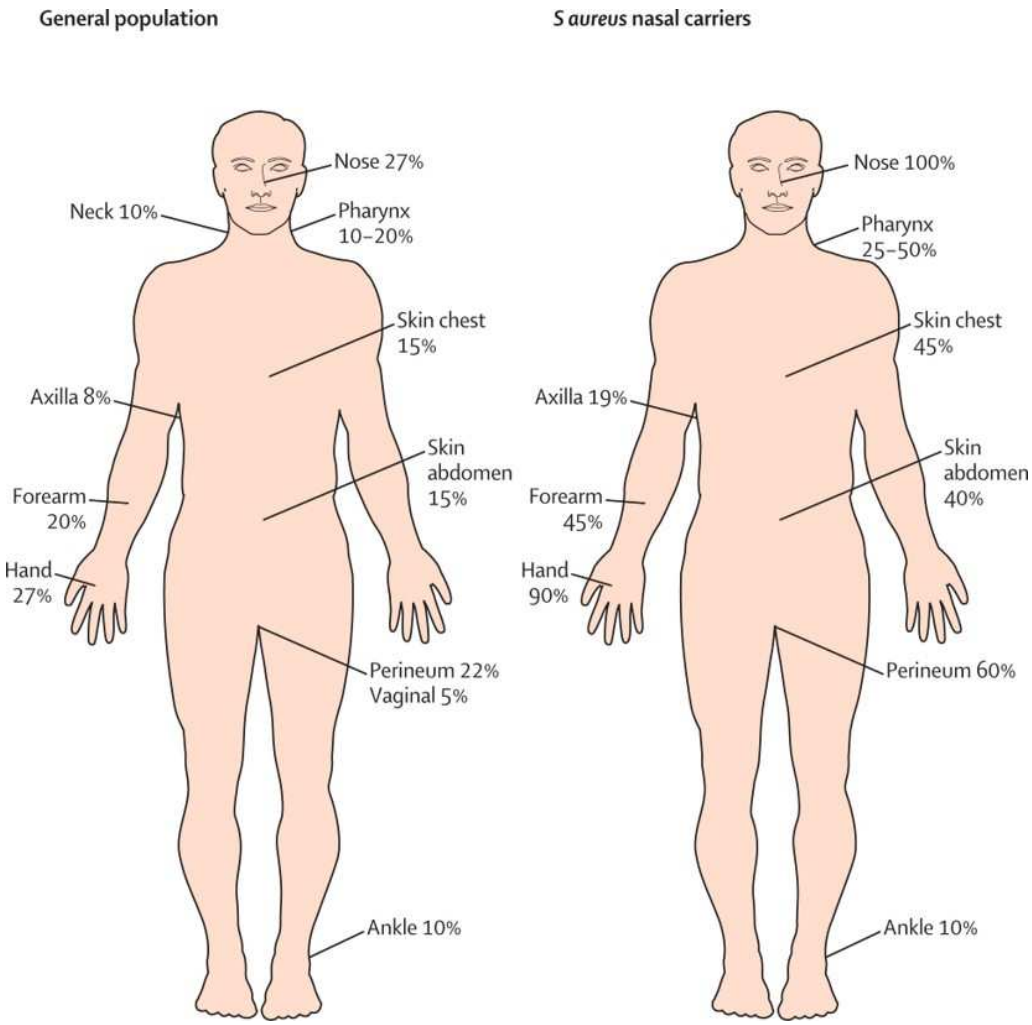


Figure 19. *S. aureus* carriage rates per body site in the general population and nasal carriers. Image adapted from Wertheim, HFL, *et al*.[130].

Nasal colonisation occurs through four key steps: *S. aureus* must enter the nasal cavity, attach to nasal receptors, overcome host defences, and multiply and spread throughout the nasal niche. These are dependent on the interaction between multiple host and microbiological factors (Figure 20).

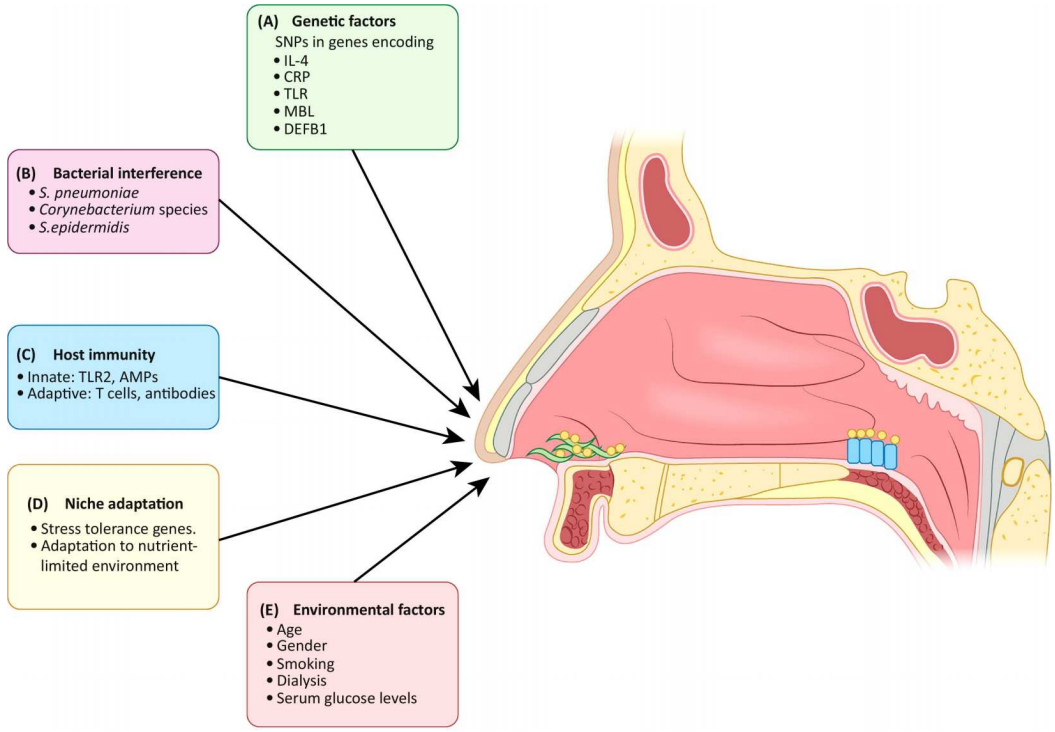


Figure 20. Factors involved in the successful *S. aureus* nasal colonisation. *S. aureus* (yellow circles) colonise squamous epithelial cells in the anterior nares (green) and columnar ciliated epithelium in the posterior cavity (blue). IL-4, interleukin-4; CRP, C-reactive protein; TLR, Toll-like receptor; MBL, mannose-binding lectin; DEFB1, β-defensin 1; AMPs, antimicrobial peptides. Image adapted from Mulcahy ME, *et al*.[135].

Some **risk factors** increase the risk of carriage: patient factors include male sex, increasing age, diabetes mellitus, skin disease, HIV, obesity, and dialysis; healthcare-related factors include antibiotic intake or surgery; and occupational factors include farming and veterinary work[136]. The role of smoking as a protective or risk factor is

controversial, but the degree of tobacco exposure and the genetic background of each strain may explain the differences observed between studies[137].

Host-genetic factors that influence colonisation include single nucleotide polymorphisms (SNPs) in genes that encode for immunity factors, such as Toll-like receptor 2 (TLR2), glucocorticoid receptor, mannose-binding lectin, interleukin-4, β-defensin 1, and C-reactive protein[135]. It is also regulated by **innate and adaptive immune responses**. However, colonising strains can delay TLR2 expression by downregulating antimicrobial peptide production and promoting colonisation. T-cell-mediated immunity enhances nasal clearance, but impaired immune responses, such as those observed in patients with HIV or atopic dermatitis, can change carriage status[135].

Interference with other bacterial species in the nasal cavities may affect *S. aureus* colonisation. Some coagulase-negative staphylococci inhibit *S. aureus* colonisation: *S. epidermidis* secretes extracellular serine protease (Esp), which degrades the adhesins and ligands required for *S. aureus* epithelial binding; and *S. lugdunensis* produces lugdunin, an antimicrobial that destroys *S. aureus*. In addition, *S. pneumoniae* releases hydrogen peroxide (H_2O_2), resulting in staphylococcal lysis, while *Propionibacterium* spp. and *Corynebacterium* spp. modulate the adhesion and aggregation of *S. aureus* in nasal colonisation[129,138]. Respiratory viruses predispose to *S. aureus* carriage by immune system impairment, tight junction disruption (promoting intracellular invasion), and receptors alteration (those involved in bacterial adhesion and internalisation)[139].

The anterior and posterior nasal cavities are composed of keratinised stratified squamous and columnar ciliated epithelial cells respectively, necessitating different attachment mechanisms. Bacterial adhesion begins in the **posterior nasal cavity** when wall teichoic acid polymers bind to the surface of columnar ciliated cells through the scavenger receptor class F member 1 (SREC1). *S. aureus* colonises the squamous stratified cells of the **anterior nasal nares** later. It binds using several microbial surface components: ClfB, binds to loricrin and cytokeratin 10; IsdA, expressed under iron-

limited conditions, binds to loricrin, cytokeratin 10, involucrin, fibronectin, and fibrinogen; and SdrD binds to desmoglein 1[109,129,138].

3.6. Infections caused by *S. aureus*

Despite usually causing no symptoms in humans, *S. aureus* can cause a wide range of diseases that range in severity from mild to life-threating. Contact with colonised people or contaminated objects, skin and nasal carriage, and the presence of wounds or the medical implants can all contribute to the spread of these infections. Risk factors for staphylococcal infection include diabetes, cancer, vascular or pulmonary disease, immunosuppression, surgical procedures, intubation, and drug injection[115,140].

S. aureus is the main cause of **skin and soft tissue infections** (SSTI). Disruptions of the skin barrier by trauma, injury, or surgery facilitate *S. aureus* entry into the subcutaneous tissues, resulting in infection that manifests as purulent exudates. Folliculitis is caused by a hair follicle infection and can progress to deeper abscesses, furuncles, or even carbuncles. Moreover, a localised skin infection can spread, resulting in erysipelas, cellulitis, and fasciitis. Other soft tissue infections include mastitis (mammary gland infection) and hidradenitis suppurative (apocrine sweat gland infection). All SSTIs are important foci for bacteraemia[104,115,141].

Toxins produced by *S. aureus* can cause a variety of **toxin-related diseases**. TSST-1, which is produced by up to 10% of the *S. aureus* population, causes toxic shock syndrome, a severe illness characterised by high fever, hypotension, vomiting, altered consciousness, and renal and hepatic impairment. This disease was first observed in women using highly absorbent menstrual tampons in the early 1980s, and in patients with influenza or wounds colonised by TSST-1-producing strains[142]. Staphylococcal-scalded skin syndrome is caused by exfoliative toxins (ETA and ETB) and mainly affects the epidermidis of neonates and infants. ETs cause bullous impetigo and skin desquamation, which can lead to dehydration, hypothermia, and secondary infections

like cellulitis, sepsis, or pneumonia[114]. Finally, staphylococcal enterotoxis production has been associated with food poisoning. After consuming contaminated food, patients develop abdominal pain, diarrhoea, nausea, vomiting, and signs of dehydration occur after an incubation period of 2-6 hours[104].

S. aureus also causes **bone and joint infections** via open fractures, orthopaedic surgery, diabetic foot or pressure ulcers, and the bloodstream, resulting in inflammation and abscess formation. Osteomyelitis causes bone destruction, necrosis, and formation, usually in the epiphyses of long bones in children and adolescents, and in the vertebrae and intervertebral discs of adults. By contrast, septic arthritis affects joints, causing inflammation, tension, and movement difficultly. *S. aureus* and coagulase-negative staphylococci are associated with 50%-60% of joint prosthesis infections. These organisms can contaminate either the prosthesis or periprosthetic tissue through direct contact or aerosols, but can also spread from an inadequately healed surgical wound or from nearby traumatised or surgically treated tissues. Risk factors for prosthetic joint infection include prosthetic joint surgery, obesity, rheumatoid arthritis, and immunosuppression. Treatment usually requires surgery and antimicrobial therapy, which can result in bacteraemia if done incorrectly[104,143,144].

Respiratory tract infection can be acquired in the community or in hospital. Community-acquired pneumonia, which typically results from commensal *S. aureus* aspiration from the upper respiratory tract after an influenza-like illness, is more common in healthy children and young adults. The rapid progression of necrotising pneumonia in these cases appears to result from epithelial damage, deficient ciliary function, and impaired bacterial clearance caused by PVL production. Conversely, hospital-acquired pneumonia is more common in the elderly and people on prolonged mechanical ventilation. In these cases, *S. aureus* infects the lower respiratory tract through haematogenous spread or endotracheal tube colonisation in ventilated patients, which can cause lung abscesses or empyema[95,104].

S. aureus is the most common cause of **bacteraemia**, showing an annual incidence of 4-40 cases per 100,000 population. The incidence in the elderly population exceeds 100 cases per 100,000 population and is attributable to increased healthcare exposure and underlying comorbidities. The presence of a prosthetic device (e.g., catheter, orthopaedic prosthesis, or implant) is a major risk factor, although nasal colonisation also appears relevant because clonally identical strains appear in 80% of patients with nasal colonisation who developed bacteraemia. Intravenous drug use, haemodialysis, and underlying conditions (e.g., diabetes, HIV infection, malignancy, and immunosuppressive therapy) also increase the risk of bacteraemia. *S. aureus* bacteraemia represents a major clinical burden associated with various clinical manifestations, ranging from hypotension and febrile illness to severe shock and multi-organic failure, causing respiratory distress, coagulopathy, and 30-day mortality rates of 20%-30%[115,143,145].

Patients with bacteraemia are more likely to develop endocarditis, particularly if they have valvular heart disease or prosthetic heart valves. Prosthetic valve **endocarditis** can be fulminant and is characterised by the formation of myocardial abscesses and progression to valve failure. Otherwise, endocarditis caused by intravenous drug injection mainly affects young people with no history of valvular disease. Mortality is lower for these cases when compared to prosthetic valve endocarditis, although the prognosis may be poor if patients have advanced HIV infection[104,107].

3.7. Treatment and antibiotic resistance

S. aureus adapts and develops resistance mechanisms to the different available antimicrobials, complicating the treatment of infection. Although **β-lactams** are usually effective, the emergence of MRSA has limited their use[146]. Figure 21 details alternative antimicrobials that are effective against MRSA infections.

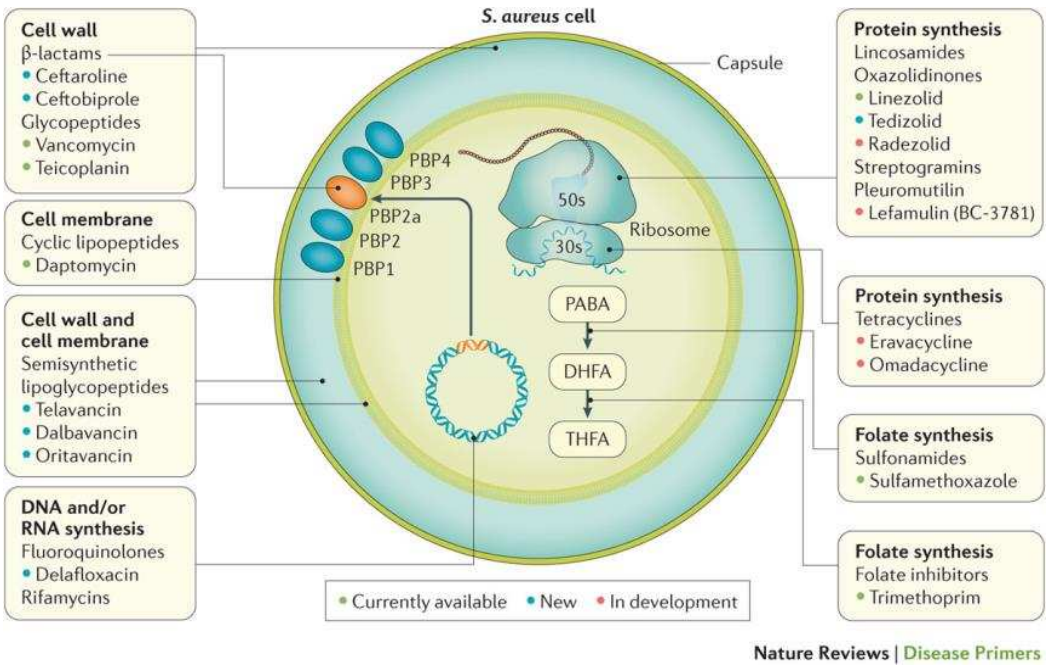


Figure 21. Bacterial targets of antibiotics active against MRSA. PBP, penicillin-binding protein; PABA, para-aminobenzoic acid; DHFA, dihydrofolic acid; THFA, tetrahydrofolic acid. Image adapted from Lee AS, *et al*.[105].

The impact of MRSA infections varies by geographical region. Northern Europe (Norway, Sweden, Denmark, the Netherlands) has a very low rate of MRSA among invasive *S. aureus* infections (<5%), whereas Southern Europe (Portugal, Spain, Italy, Greece), America, and Asia have much higher rates (25%-50%). These disparities could be attributed to differences in infection control and antimicrobial stewardship[105,112].

Methicillin resistance occurs through the acquisition of a **staphylococcal cassette chromosome *mec*** (SCC*mec* I-XIII). This transposable element contains the following: *orfX* for cassette integration into the chromosome; a recombinase (*ccrA*, *ccrB* or *ccrC*) for the excision and integration of the cassette; transposons (*tnp* and *Tn554*) required for transposition; flanking insertion-sequence elements (*IS431* and *IS1272*); the *mecA* gene and its regulatory genes (*mecR* and *mecI*), which are associated with β-lactam resistance; and other genes associated with heavy metal and antimicrobial resistance (e.g., macrolides or aminoglycosides)[99,147]. β-lactams bind covalently to penicillin-

binding proteins (PBPs), inhibiting their ability to synthesise the bacterial cell wall. The *mecA* gene, which codes for PBP2a, has a low affinity for most β-lactams, preventing their antimicrobial action while allowing cell wall synthesis[148]. Despite *mecA* being the most common cause of methicillin resistance, *mecB* and *mecC* homologues have been identified on a plasmid and SCC*mec* type XI, respectively. These appear to be associated with methicillin resistance[149].

In the early 1990s, antimicrobial therapy with vancomycin was effective for treating severe MRSA infections. However, the first *S. aureus* with **reduced susceptibility to vancomycin** appeared after only a few years, and was soon followed by the emergence of **vancomycin-resistant strains** (VRSA). The main mechanisms associated with reduced susceptibility relate to increased cell wall thickness, changes in peptidoglycan synthesis and growth, disruption of the surface protein profile, and dysfunction of the Agr system. Although uncommon, vancomycin resistance has also been associated with acquiring the ***vanA* operon** in a conjugative plasmid. This modifies peptidoglycan precursors, preventing vancomycin binding and allowing bacterial cell wall synthesis[150].

The emergence of MRSA strains with reduced susceptibility to vancomycin limits the available therapeutic options. **Daptomycin** is a lipopeptide antibiotic that causes functional disruption of the cytoplasmic membrane, leading to membrane depolarisation and cell death. It is the main alternative for the treatment of SSTIs, endocarditis, and bacteraemia, especially when caused by MRSA strains with reduced or absent susceptibility to vancomycin. Interaction with pulmonary surfactant inhibits daptomycin activity and makes it ineffective for the treatment of respiratory infections[151]. Although daptomycin resistance is rare, some cases have been reported, mainly due to mutations in *mprF*. However, these alterations have a high fitness cost, which could explain the low dissemination of daptomycin-resistant strains[152]. **Linezolid**, of the oxazolidinone family, binds to 23S rRNA to inhibit bacterial protein synthesis. It is commonly used to treat both SSTIs and pneumonia caused by *S. aureus* and has a low resistance rate

(<1%). Some linezolid-resistant strains have appeared carrying the *cfr* gene or mutations in the 23S rRNA gene[153,154].

Promising antimicrobial agents have been introduced for the treatment of MRSA infections over recent years. **Dalbavancin**, a semi-synthetic lipoglycopeptide antibiotic approved by the European Medicines Agency (EMA) in 2015, inhibits cell wall synthesis and is very active against *S. aureus*. **Ceftaroline** and **ceftobiprole** are new generation cephalosporines approved by the EMA in 2012 and 2013, respectively. These interfere with cell wall synthesis by binding to PBPs, including PBP2a, and in contrast to the other cephalosporines, show bactericidal activity against MRSA[155].

2. Justification of the Study and Objectives

Haemophilus influenzae and *Staphylococcus aureus* are opportunistic pathogens that commonly colonise the human respiratory tract. Bacterial adaptation to the selective pressures imposed by the immune system and antimicrobial agents is critical to survival in the adverse respiratory tract environment. Given that colonisation may precede severe infections (e.g., pneumonia, meningitis, and bacteraemia) improved epidemiological surveillance of invasive clones and a better understanding of the mechanisms involved in colonisation and pathogenesis may help to control and prevent severe infection, ultimately improving quality of life for patients. Therefore, this thesis focuses on the epidemiological characterisation, population structure, and pathophysiological mechanisms associated with *H. influenzae* and *S. aureus* colonisation. This is achieved through two objectives:

OBJECTIVE 1: To determine the epidemiology and population structure of *H. influenzae*.

1.1: Current epidemiology of invasive *H. influenzae* disease at Bellvitge University Hospital.

1.2: Characterisation of the *H. influenzae* population structure.

1.3: Analysis of the genomic diversity and pangenome of *H. influenzae* capsulated clones.

OBJECTIVE 2: To characterise the mechanisms of adaptive evolution and persistence in *H. influenzae* and *S. aureus*.

2.1: Characterisation of the *H. influenzae* population colonising the respiratory tract in patients with chronic obstructive pulmonary disease (COPD) and prolonged azithromycin treatment.

2.2: Analysis of the genetic determinants in clinical *S. aureus* isolates associated with biofilm formation and intracellular host cell invasion.

3. Results

OBJECTIVE 1: To determine the epidemiology and population structure of *H. influenzae*.

STUDY 1 – Epidemiology and population structure of *Haemophilus influenzae* causing invasive disease.

Objective 1.1: Current epidemiology of invasive *H. influenzae* disease at Bellvitge University Hospital.

Objective 1.2: Characterisation of the *H. influenzae* population structure.

STUDY 2 – Comparative pangenome analysis of capsulated *Haemophilus influenzae* serotype f highlights their high genomic stability.

Objective 1.2: Characterisation of the *H. influenzae* population structure.

Objective 1.3: Analysis of the genomic diversity and pangenome of *H. influenzae* capsulated clones.

OBJECTIVE 2: To characterise the mechanisms of adaptive evolution and persistence in *H. influenzae* and *S. aureus*.

STUDY 3 – Acquisition of macrolide resistance and genetic adaptation of *Haemophilus* spp during persistent respiratory tract colonisation in chronic obstructive pulmonary disease (COPD) patients receiving long-term azithromycin treatment.

Objective 2.1: Characterisation of the *H. influenzae* population colonising the respiratory tract in patients with chronic obstructive pulmonary disease (COPD) and prolonged azithromycin treatment.

STUDY 4 – *Staphylococcus aureus* surface protein G (*sasG*) allelic variants: correlation between biofilm formation and their prevalence in methicillin-resistant *S. aureus* (MRSA) clones.

Objective 2.2: Analysis of the genetic determinants in clinical *S. aureus* isolates associated with biofilm formation and intracellular host cell invasion.

ADDITIONAL RESULTS – Host adaptive changes of *Staphylococcus aureus* through respiratory colonisation and bloodstream infection.

Objective 2.2: Analysis of the genetic determinants in clinical *S. aureus* isolates associated with biofilm formation and intracellular host cell invasion.

Study 1

Epidemiology and population structure of *Haemophilus influenzae* causing invasive disease

Anna Carrera-Salinas, Aida González-Díaz, Laura Calatayud, Julieta Mercado-Maza, Carmen Puig, Dàmaris Berbel, Jordi Càmara, Fe Tubau, Imma Grau, M. Ángeles Domínguez, Carmen Ardanuy, Sara Martí

Dr. Mª Ángeles Domínguez Luzón, Associate Professor of the Department of Pathology and Experimental Therapeutics of the Universitat de Barcelona and Head of the Microbiology Department, Hospital Universitari de Bellvitge (Barcelona), and Dr. Sara Martí Martí, Assistant Professor of the Department of Medicine of the Universitat de Barcelona and Senior Postdoctoral Researcher at the Centre for Biomedical Research Network on Respiratory Diseases (CIBERes),

DECLARE that the original article entitled "**Epidemiology and population structure of *Haemophilus influenzae* causing invasive disease**" has been published in Microbial Genomics (Impact factor in 2021: 4.868). The contribution of Anna Carrera-Salinas is detailed below:

- Experimental work.
- Data analysis.
- Manuscript redaction.

Signatures of thesis supervisors

MªÁngeles Domínguez Luzón **Sara Martí Martí**

MICROBIAL GENOMICS

RESEARCH ARTICLE

Carrera-Salinas *et al.*, *Microbial Genomics* 2021;7:000723

DOI 10.1099/mgen.0.000723

Epidemiology and population structure of *Haemophilus influenzae* causing invasive disease

Anna Carrera-Salinas[1], Aida González-Díaz[1,2], Laura Calatayud[1,2], Julieta Mercado-Maza[1], Carmen Puig[1], Dàmaris Berbel[1,2], Jordi Càmara[1,2], Fe Tubau[1,2], Imma Grau[2,3], M. Ángeles Domínguez[1,4,5], Carmen Ardanuy[1,2,5] and Sara Martí[1,2,6,*]

Abstract

This study provides an update on invasive *Haemophilus influenzae* disease in Bellvitge University Hospital (2014–2019), reporting its evolution from a previous period (2008–2013) and analysing the non-typeable *H. influenzae* (NTHi) population structure using a clade-related classification. Clinical data, antimicrobial susceptibility and serotyping were studied and compared with those of the previous period. Population structure was assessed by multilocus sequence typing (MLST), SNP-based phylogenetic analysis and clade-related classification. The incidence of invasive *H. influenzae* disease remained constant between the two periods (average 2.07 cases per 100000 population), while the 30 day mortality rate decreased (20.7–14.7%, respectively). Immunosuppressive therapy (40%) and malignancy (36%) were the most frequent comorbidities. Ampicillin and fluoroquinolone resistance rates had increased between the two periods (10–17.6% and 0–4.4%, respectively). NTHi was the main cause of invasive disease in both periods (84.3 and 85.3%), followed by serotype f (12.9 and 8.8%). NTHi displayed high genetic diversity. However, two clusters of 13 (*n*=20) and 5 sequence types (STs) (*n*=10) associated with clade V included NTHi strains of the most prevalent STs (ST3 and ST103), many of which showed increased frequency over time. Moreover, ST103 and ST160 from clade V were associated with β-lactam resistance. Invasive *H. influenzae* disease is uncommon, but can be severe, especially in the elderly with comorbidities. NTHi remains the main cause of invasive disease, with ST103 and ST160 (clade V) responsible for increasing β-lactam resistance over time.

DATA SUMMARY

Sequence reads were deposited in the European Nucleotide Archive (ENA) under the project accession number PRJEB42971 (Dataset S1, available in the online version of the article).

INTRODUCTION

Haemophilus influenzae is a Gram-negative coccobacillus that colonizes the human nasopharynx and throat in up to 80% of children and 20–30% of adults [1, 2]. Capsulated strains produce different capsular polysaccharides (identified as serotypes a to f), a distinctive feature that distinguishes them from the non-capsulated strains that are commonly known as non-typeable *H. influenzae* (NTHi) [3].

Despite the usual asymptomatic interaction with humans, *H. influenzae* is responsible for respiratory infections ranging from otitis media, sinusitis and conjunctivitis to deeper infections such as pneumonia and exacerbations of chronic obstructive pulmonary

Received 19 April 2021; Accepted 20 October 2021; Published 13 December 2021

Author affiliations: [1]Microbiology Department, Bellvitge University Hospital, IDIBELL-UB, Barcelona, Spain; [2]Research Network for Respiratory Diseases (CIBERES), ISCIII, Madrid, Spain; [3]Infectious Diseases Department, Bellvitge University Hospital, IDIBELL-UB, Barcelona, Spain; [4]Spanish Network for Research in Infectious Diseases (REIPI), ISCIII, Madrid, Spain; [5]Department of Pathology and Experimental Therapeutics, School of Medicine, University of Barcelona, Barcelona, Spain; [6]Department of Medicine, School of Medicine, University of Barcelona, Barcelona, Spain.

***Correspondence:** Sara Martí, smartinm@bellvitgehospital.cat

Keywords: epidemiology; *Haemophilus influenzae*; invasive disease; structure population.

Abbreviations: CCI, Charlson comorbidity index; CLSI, Clinical and Laboratory Standards Institute; COPD, chronic obstructive pulmonary disease; ENA, European Nucleotide Archive; MALDI-TOF, matrix-assisted laser desorption/ionization time-of-flight; MIC, minimum inhibitory concentration; MLST, multilocus sequence typing; MST, minimum spanning tree; NTHi, non-typeable H. influenzae; PBP, Penicillin-binding protein; %R, resistant; r/m, Relative impact of recombination and mutation; S, Susceptible; SLVs, single-locus variants; SNP, Single nucleotide polymorphism; ST, Sequence type; WGS, whole-genome sequencing.

Sequence reads were deposited in the European Nucleotide Archive (ENA) under the project accession number PRJEB42971.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Two supplementary tables, three supplementary figures and supplementary dataset are available with the online version of this article.

disease (COPD) and even invasive disease [4–7]. Before the widespread introduction of the *H. influenzae* type b conjugate vaccine, this serotype was the most common cause of invasive disease. Vaccine implementation has produced an epidemiological shift, with most invasive infections currently occurring in elderly patients with underlying conditions and caused mainly by NTHi followed by non-type-b serotypes [3, 4, 8]. As the epidemiology of invasive *H. influenzae* disease evolves, concerns are now focused on non-vaccine-preventable isolates and the growing proportion of β-lactam resistance by the acquisition of β-lactamases and alterations in penicillin-binding proteins (PBPs) [9, 10].

The structure of the bacterial population provides information on the evolution of a bacterial species, its genetic diversity and the dynamics of antimicrobial resistance determinants and pathogenic mechanisms [11, 12]. Classification based on multilocus sequence typing (MLST) is useful for clustering capsulated *H. influenzae* isolates, as they are highly clonal and associated with few sequence types (STs). By contrast, NTHi strains show high genetic heterogeneity [13]. The classification of NTHi into clades according to the presence or absence of some surface-associated proteins and virulence determinants, as previously proposed [12, 14], could be a valuable tool to address this issue.

Epidemiological surveillance of invasive *H. influenzae* clones, along with their antibiotic resistance profile and genetic variability, is essential to design useful prevention strategies. In this study, we provide an update on the epidemiology of invasive *H. influenzae* disease in Bellvitge University Hospital (2014–2019), monitoring changes from a previous period (2008–2013) [15].

Impact Statement

Haemophilus influenzae is an opportunistic pathogen in the human respiratory tract that may cause community-acquired invasive infections with high mortality rates. Moreover, the rise in β-lactam resistance due to β-lactamase production is cause for concern. We provide an update on the current epidemiology of invasive *H. influenzae* disease, as well as an analysis of the heterogeneous non-typable *H. influenzae* (NTHi) population structure using a clade-related classification based on accessory genome analysis. Multilocus sequence typing (MLST) is useful for clustering capsulated *H. influenzae* isolates because they are clonal and have few sequence types (STs), but it is not the best method for studying the heterogeneous NTHi population. The use of a clade-related classification based on the presence or absence of some surface-associated proteins and virulence determinants is a valuable tool to address this issue. Interestingly, we have observed that despite the genetic diversity, more than half of the invasive NTHi isolates shared a common phylogenetic origin, clustering them in the same clade. Moreover, these isolates showed an increased frequency over time and were linked to β-lactamase production. This should raise our awareness of the possibility of clonal dissemination, and we should consider monitoring the evolution of these clones in invasive disease.

METHODS

Study design and bacterial strains

A laboratory-based surveillance study on the epidemiology of invasive *H. influenzae* disease between 2008 and 2019 (first period: 2008–2013 [15]; second period: 2014–2019) was undertaken. This study was conducted at Bellvitge University Hospital, a tertiary care centre for adult patients that provides health assistance to the southern area of Barcelona, Spain. Cases of invasive disease were defined as the isolation of *H. influenzae* from the blood, cerebrospinal fluid or pleural fluid (Table S1), the same sample types considered in the previous study period (2008–2013) [15] to facilitate comparison of the results. All clinical *H. influenzae* isolates were identified by mass spectrometry using MALDI-TOF (MALDI Biotyper, Bruker).

The medical records of patients with invasive *H. influenzae* disease were reviewed to obtain demographic and clinical data, including information on gender, age, source of infection, underlying conditions and 30 day mortality. The incidence of invasive *H. influenzae* disease was expressed as the number of cases per 100000 population and was estimated using the adult population that received health coverage in Bellvitge University Hospital as the denominator, according to the Statistical Institute of Catalonia (https://www.idescat.cat).

Antimicrobial susceptibility testing

Antimicrobial susceptibility was assessed by the microdilution method, using commercial panels (STRHAE2, Sensititre) and following the recommendations for clinical breakpoints of the Clinical and Laboratory Standards Institute (CLSI) [16]. β-lactamase activity was measured using the chromogenic cephalosporin method (Cefinase disc, BD).

Whole genome sequencing

All 68 invasive *H. influenzae* isolates identified during the second period (2014–2019) were subjected to whole-genome sequencing (WGS) for serotyping, MLST classification and the study of resistance-related mutations. Genomic DNA was extracted using the QIAamp DNA Mini Kit (Qiagen) and quantified by the QuantiFluor dsDNA System (Promega). Nextera XT was used to prepare the libraries, followed by paired-end sequencing on a MiSeq platform (Illumina). Sequences were assembled with the INNUca v4.2 pipeline (github.com/BUMMI/INNUca) using default parameters. MLST was performed using the mlst v2.4 software (Seemann T, mlst Github, github.com/tseemann/mlst). New allele and ST numbers were registered in PubMLST (pubmlst.org). Full MST algorithm with PHYLOViZ [17] was used to analyse the relationship between the STs. *In silico* serotyping was conducted using hicap (github.com/scwatts/hicap) [18].

The mechanisms of acquired resistance to antibiotics were screened using ResFinder [19]. The screening of mutations in genes involved in antibiotic resistance was performed using Geneious R9 (Biomatters), with the closed genome of *H. influenzae* Rd KW20 (NC_000907) used as the reference. For phylogenetic analysis, core-SNPs were extracted with Snippy's core module (snippy-core). Subsequently, a novel core-SNP phylogenetic tree, using RAxML-NG [20], was built using a discrete GAMMA model of rate heterogeneity, and 100 bootstrap replicates to determine the phylogeny of the 68 strains included in this study and the strains of clade V, separately. Relative bootstrap values below 0.75 indicated poor support [21]. The Hi375 strain (CP009610) was used as a reference. The phylogenetic trees were visualized using Microreact (microreact.org). The core-genome SNP alignment obtained with Snippy's core module was subjected to the prediction of recombinant regions using the Gubbins v2.3.1 software to assess the proportion of sites affected by recombination, which was expressed as the relative impact of recombination and mutation (r/m) [22].

NTHi clade analysis

NTHi isolates were grouped into clades, as previously proposed [12, 14]. A total of 213 NTHi genomes were included in this analysis: 98 genomes from De Chiara *et al.* [12], 57 genomes from Pinto *et al.* [14], and 58 genomes from this study. Six NTHi clades (I to VI) were defined using patho_typing (github.com/B-UMMI/patho_typing) based on the presence or absence of 17 gene sequences [14]. Phylogenetic analysis was performed by constructing an assembly-based core-SNP phylogenetic tree, using Parsnp from the Harvest suite [23] with default parameters, apart from parameter -C, which was adjusted to 5000 to maximize the reference coverage. The Hi375 strain (CP009610) was used as a reference. Tree visualization was performed using the ggtree R package [24] (available at github.com/microbellvitge/Phylogenetic-analysis-invasive-HINF).

Statistical analysis

Statistical analyses were carried out using the GraphPad Prism 5 software, applying Fisher's exact test or unpaired *t*-test when appropriate. *P*-values<0.05 were considered statistically significant.

RESULTS

Clinical characteristics

A total of 8191 *H. influenzae* strains were isolated from adult patients admitted to Bellvitge University Hospital (2008–2019). Of these, 150 (1.8%) were responsible for invasive disease, most of which had been isolated from blood (88.0%) and less frequently from cerebrospinal and pleural fluids (8.7 and 3.3%, respectively). In the first period (2008–2013), 82 of the 3433 (2.4%) *H. influenzae* isolates caused invasive disease, compared to 68 of the 4758 isolates (1.4%) in the second period (2014–2019).

The overall incidence of invasive *H. influenzae* disease in the adult population was 2.07 cases per 100000 population, with no significant differences between the two periods (2.12 and 2.02 episodes per 100000 population, respectively). By age groups, the overall incidence was lower in young adults (≤64 years) than among those aged 65 or older (0.89 and 6.90 cases per 100000, respectively) (*P*-value<0.0001), as observed in each study period separately (1.10 and 6.80 cases per 100000 in the first period, and 0.68 and 6.97 per 100000 in the second period). The incidence in young adults decreased in the second period (*P*-value=0.03), while the incidence in older adults remained stable.

Table 1 shows the demographics, clinical characteristics and underlying conditions of the patients. The most common sources of infection were respiratory tract infections (67.4%), followed by meningitis and biliary tract infections (9.3%, each) (Fig. 1a). Although respiratory infections were common in the entire population, they were significantly more frequent in the elderly than in young adults (76.3 and 50.9%, respectively) (Fig. 1b, c). No statistically significant differences in the focus of infection were observed between the two periods.

The most common comorbidities were immunosuppressive therapy (40.0%) and solid organ malignancies (36.0%), which were strongly associated with 30 day mortality. Diabetes, heart disease and COPD were more common in the elderly, while HIV was more frequent in younger adults. The only difference observed between the two periods was an increase in the number of patients with heart disease. The overall 30 day mortality was 18.0% (*n*=27), which decreased between the periods (20.7 and 14.7%, respectively) and did not differ between the age groups (18.9 and 17.5%, respectively) (Table 1). The Charlson comorbidity index (CCI) was significantly higher in the elderly adults (5.5±1.8) compared to young adults (3.4±3.0) (*P*-value<0.0001). Nevertheless, 30 day mortality in young adults was associated with a higher CCI (5.2±3.8) compared to young adults who survived (3.0±2.7) (*P*-value=0.04), while no differences were observed in the elderly (5.9±1.6 and 5.4±1.8, respectively) (Table S2).

Antimicrobial susceptibility and resistance determinants

Susceptibility rates for all the antibiotics tested were similar in the two periods (Table 2). All the strains were susceptible to amoxicillin/clavulanic acid, cefotaxime, ceftriaxone, imipenem, chloramphenicol and tetracycline. On the other hand, 37 strains (24.7%) were resistant to cotrimoxazole, 19 (12.7%) to ampicillin, 3 (2.0%) to ciprofloxacin, and 1 (0.7%) to cefuroxime, cefepime and azithromycin.

During the second period (2014–2019), there was a slight increase in β-lactam resistance and an emergence of resistance to ciprofloxacin (Table 2). Twelve strains (17.6%) were resistant to ampicillin due to β-lactamase production: eleven NTHi strains had a $bla_{\text{TEM-1}}$ gene and one serotype f strain carried a $bla_{\text{ROB-1}}$ gene in a pB1000 plasmid (GU080063). The latter was also resistant to cefuroxime and cefepime. Amino acid substitutions in

Table 1. Demographic data, clinical characteristics and underlying conditions of patients with invasive *H. influenzae* disease

	Overall study period (2008–2019) *n*=150				Comparison between periods		
		Young adults (<65 years) *n*=53	Elderly adults (≥65 years) *n*=97	*P*-value	First period (2008–2013) *n*=82	Second period (2014–2019) *n*=68	*P*-value
Characteristics [no. (%)]							
Age (mean±SD; range)	67.5±14.9; 21–96	51.6±11.4; 21–64	76.1±7.7; 65–96	**<0.0001**	64.3±16.1; 21–96	71.2±12.3; 37–93	**0.0043**
Male sex	88 (58.7)	30 (56.6)	58 (59.8)	0.7311	52 (63.4)	36 (52.9)	0.2439
Underlying conditions							
Immunosuppressive therapy	60 (40.0)	22 (41.5)	38 (39.2)	0.86	28 (34.1)	32 (47.1)	0.13
Solid organ malignancy	54 (36.0)	17 (32.1)	37 (38.1)	0.48	26 (31.7)	28 (41.2)	0.24
Diabetes	39 (26.0)	8 (15.1)	31 (32.0)	**0.03**	18 (22.0)	21 (30.9)	0.26
Heart disease	40 (26.7)	7 (13.2)	33 (34.0)	**0.01**	15 (18.3)	25 (36.8)	**0.02**
COPD	30 (20.0)	4 (7.5)	26 (26.8)	**0.01**	18 (22.0)	12 (17.6)	0.54
Chronic liver disease	18 (12.0)	10 (18.9)	8 (8.2)	0.07	11 (13.4)	7 (10.3)	0.62
Hematologic malignancy*	16 (10.7)	7 (13.2)	9 (9.3)	0.58	9 (11.0)	7 (10.3)	1.00
Cerebrovascular disease	7 (4.7)	0 (0.0)	7 (7.2)	0.05	5 (6.1)	2 (2.9)	0.46
Organ transplant†	10 (6.7)	6 (11.3)	4 (4.1)	0.17	4 (4.9)	6 (8.8)	0.51
HIV	3 (2.0)	3 (5.7)	0 (0.0)	**0.04**	2 (2.4)	1 (1.5)	1.00
No underlying conditions	27 (18.0)	15 (28.3)	12 (12.4)	**0.02**	18 (22.0)	9 (13.2)	0.20
Acquisition							
Community acquired	130 (86.7)	43 (81.1)	87 (89.7)	0.21	71 (86.6)	59 (86.8)	1.00
Nosocomial	20 (13.3)	10 (18.9)	10 (10.3)	0.21	11 (13.4)	9 (13.2)	1.00
Source of infection							
Respiratory tract infection	101 (67.3)	27 (50.9)	74 (76.3)	**0.01**	49 (59.8)	52 (76.5)	**0.04**
Meningitis	14 (9.3)	5 (9.4)	9 (9.3)	1.00	9 (11.0)	5 (7.4)	0.58
Biliary tract infection	14 (9.3)	7 (13.2)	7 (7.2)	0.25	9 (11.0)	5 (7.4)	0.58
Primary bacteremia	9 (6.0)	6 (11.3)	3 (3.1)	0.07	7 (8.5)	2 (2.9)	0.18
Peritonitis	3 (2.0)	2 (3.8)	1 (1.0)	0.28	2 (2.4)	1 (1.5)	1.00
Other‡	9 (6.0)	6 (11.3)	3(3.1)	0.07	6 (7.3)	3 (4.4)	0.51
Charlson comorbidity index	4.8±2.5	3.4±3.0	5.5±1.8	**<0.0001**	4.3±2.4	5.3±2.5	**0.01**
30 day mortality	27 (18.0)	10 (18.9)	17 (17.5)	0.83	17 (20.7)	10 (14.7)	0.40

*Myeloma (*n*=8, 5.3%), leukaemia and lymphoma (*n*=4, 2.7% each).
†Liver (*n*=4, 2.7%), bone marrow (*n*=3, 2.0%), kidney (*n*=2, 1.3%) and heart (*n*=1, 0.7%).
‡Epiglotittis, liver abscess (*n*=2, 1.3% each), endometritis, facial cellulites, septic shock of an aortic valve prosthetic origin, surgical wound infection, and urinary tract infection (*n*=1, 0.7% each).

PBP3 were observed in 26 strains (38.2%) and were classified according to Dabernat *et al.* [25] (Table 3). Eleven strains were classified in group II, characterized by the K526N substitution, and showed reduced susceptibility to ampicillin. Fluoroquinolone resistance was observed in three strains (4.4%), which, despite being low, indicated the emergence of resistance to fluoroquinolones since no ciprofloxacin-resistant strains were identified in the previous period. Resistance was associated with changes in GyrA (S84L and D88G in two of the strains,

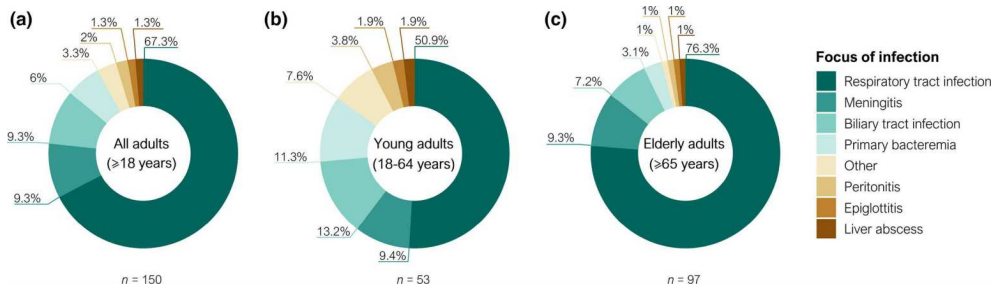


Fig. 1. Frequencies of invasive *H. influenzae* infection by focus (2008–2019). (a) Focus of infection in the overall adult population. (b) Focus of infection in the young adult population (18–64 years). (c) Focus of infection in the elderly adult population (≥65 years). Infections caused by only one isolate were grouped as other: facial cellulitis, endometritis, urinary tract infection, and surgical wound infection (corresponding to young adults) and septic shock of an aortic valve prosthetic origin (corresponding to elderly adults).

and S84L and D88N in one strain) and ParC (S84I in all the ciprofloxacin-resistant strains) (Dataset S1).

Serotype distribution and molecular epidemiology

NTHi was the main cause of invasive disease in both periods in contrast to the overall low percentage of capsulated strains, mainly serotype f and serotype e, both of which had disappeared at the end of the study (Fig. 2a). Serotype b maintained a low frequency, with only one case detected in each period. Overall, 117 cases were caused by NTHi (n=59 in the first period and n=58 in the second period), 15 were caused by serotype f (n=9 and n=6, respectively), four were due to serotype e (n=1 and n=3, respectively), and two were due to serotype b strains (n=1 and n=1, respectively) (Fig. 2b). The remaining 12 strains from the first period were not available for microbiological analysis. There were no differences in serotype distribution by age group nor by study period.

Table 2. Comparison of antimicrobial resistance rates of invasive *H. influenzae* during the two study periods

Antimicrobial compound	MIC breakpoints (CLSI)		First period*			Second period			*P*-value
			2008–2013 (n=70)			2014–2019 (n=68)			
	S ≤	R ≥	MIC_{50} (mg l^{-1})	MIC_{90} (mg l^{-1})	%R	MIC_{50} (mg l^{-1})	MIC_{90} (mg l^{-1})	%R	
Ampicillin	1	4	0.25	2	10	0.25	>4	17.6	0.22
Amoxicillin/clavulanic acid†	4	8	≤0.5	2	0	≤0.5	1	0	1
Cefuroxime	4	16	1	2	0	≤0.5	1	1.5	0.49
Cefepime	2	4	≤0.25	≤0.25	0	≤0.25	≤0.25	1.5	0.49
Cefotaxime	2	4	≤0.06	≤0.06	0	≤0.06	≤0.06	0	1
Ceftriaxone	2	4	≤0.12	≤0.12	0	≤0.12	≤0.12	0	1
Imipenem	4	8	0.5	1	0	≤0.12	0.25	0	1
Chloramphenicol	2	8	≤1	≤1	0	≤1	≤1	0	1
Tetracycline	2	8	≤1	2	0	≤1	≤1	0	1
Ciprofloxacin	1	2	≤0.03	≤0.03	0	≤0.03	≤0.03	4.4	0.12
Co-trimoxazole‡	0.5	4	≤0.5	>2	26	≤0.5	>2	27.9	0.85
Azithromycin	4	8	1	2	1.4	2	4	0	0.49

*Only 70 of the 82 strains of the first period were viable.
†The ratio of amoxicillin/clavulanic acid was 2:1.
‡The ratio of co-trimoxazole was 1:19.
MIC, Minimum inhibitory concentration; R, resistant; %R, resistance rate; S, susceptible.

Table 3. Mechanisms of β-lactam antimicrobial resistance among invasive *H. influenzae* strains (2014–2019). Amino acid substitutions in PBP3 and classification according to Dabernat *et al.* [25], and β-lactamase production

	No. isolates	MIC (mg l^{-1})		Amino acid substitutions										β-lactamase*	
		AMP	AMC†	Ile	Asp	Ala	Met	Ala	Ile	Gly	Ala	Asn	Ala	*bla*$_{\text{TEM-1}}$	*bla*$_{\text{ROB-1}}$
				348	350	368	377	437	449	490	502	526	530		
No changes	32	≤0.12–0.25	≤0.5											–	–
	10	4 ->4	≤0.5–2											+	–
Group IIa	1	0.5	1									Lys	Ser	–	–
	1	1	1		Asn					Glu		Lys	Ser	–	–
	1	1	1		Asn							Lys		–	–
Group IIb	1	1	1							Glu	Val	Lys		–	–
	2	0.5–1	1		Asn		Ile				Val	Lys		–	–
Group IIc	2	1	1								Thr	Lys		–	–
	2	1	1		Asn						Thr	Lys		–	–
Group IId	1	1	1						Val			Lys		–	–
Miscellaneous	10	0.25–0.5	≤0.5–1		Asn									–	–
	1	≤0.12	≤0.5	Val										–	–
	1	>4	2		Asn									–	+
	1	>4	≤0.5		Asn									+	–
	1	≤0.12	≤0.5		Asn			Ser						–	–
	1	0.25	≤0.5			Thr								–	–

*β-lactamase production: positive (+), negative (–).
†The ratio of amoxicillin/clavulanic acid was 2:1.
AMC, amoxicillin/clavulanic acid; AMP, ampicillin; MIC, Minimum inhibitory concentration.

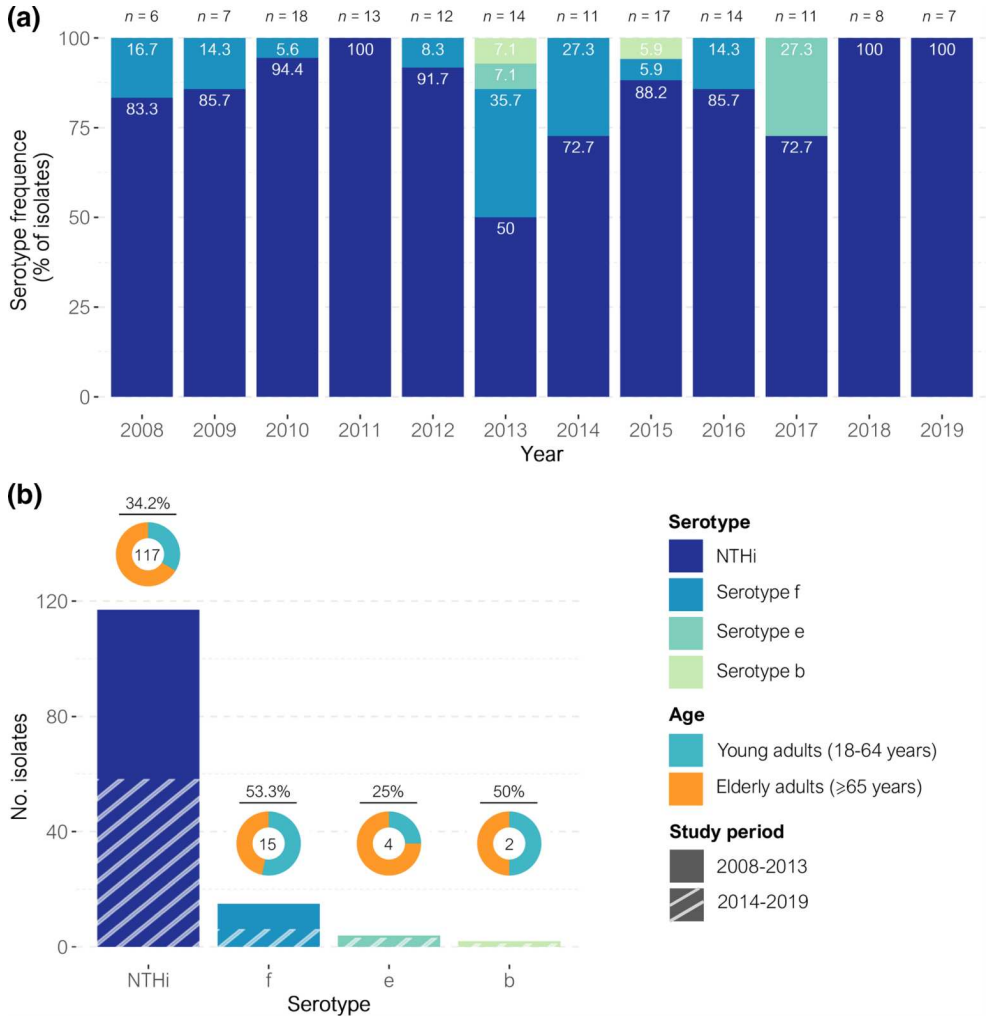


Fig. 2. Distribution of *H. influenzae* serotypes in invasive disease (2008–2019). (a) Frequencies of *H. influenzae* serotypes causing invasive disease per year. Frequencies are displayed at the upper part of each bar. (b) Distribution of *H. influenzae* serotypes causing invasive disease between 2008 and 2019. Donut charts, at the top of each bar, show the age group frequency; the frequency (%) in young adults is shown above. The number of isolates presenting each serotype is shown inside the donut charts.

MLST and PHYLOViZ analysis showed high genetic heterogeneity, with 84 different STs identified among the 117 NTHi strains. The most common in the first period was ST57 (*n*=3), while ST103 was the most prevalent in the second period (*n*=4), followed by ST3 and ST160 (*n*=3, each). Two of the most common STs, ST103 and ST160, produced a TEM-1 *β*-lactamase, and both of them showed increased prevalence from the previous period: ST103 from one to four isolates, while ST160, which was not detected in the first period, caused three cases of invasive disease in the second period. Despite the great variability among NTHi, PHYLOViZ revealed two clusters, which each consisted of clones that were linked to one another by sharing at least five MLST loci with another clone in the group (Fig. S1). One of the clusters was related to ST3 and included 13 STs (*n*=20), while the other was associated with ST103 and included five STs (*n*=10). Moreover, there was an increase in the number of isolates belonging to these clusters, with 13 in the first period (ten associated with the ST3 cluster and three with the ST103 cluster) and 17 in the second period (ten and seven, respectively). On the other hand, capsulated strains showed high clonality, with a few related STs (single-locus variants, SLVs) identified for each serotype in both periods: ST124 (*n*=12) and ST2354, ST2361, and ST2366 (*n*=1, each) for serotype f; ST18 and ST386 (*n*=2,

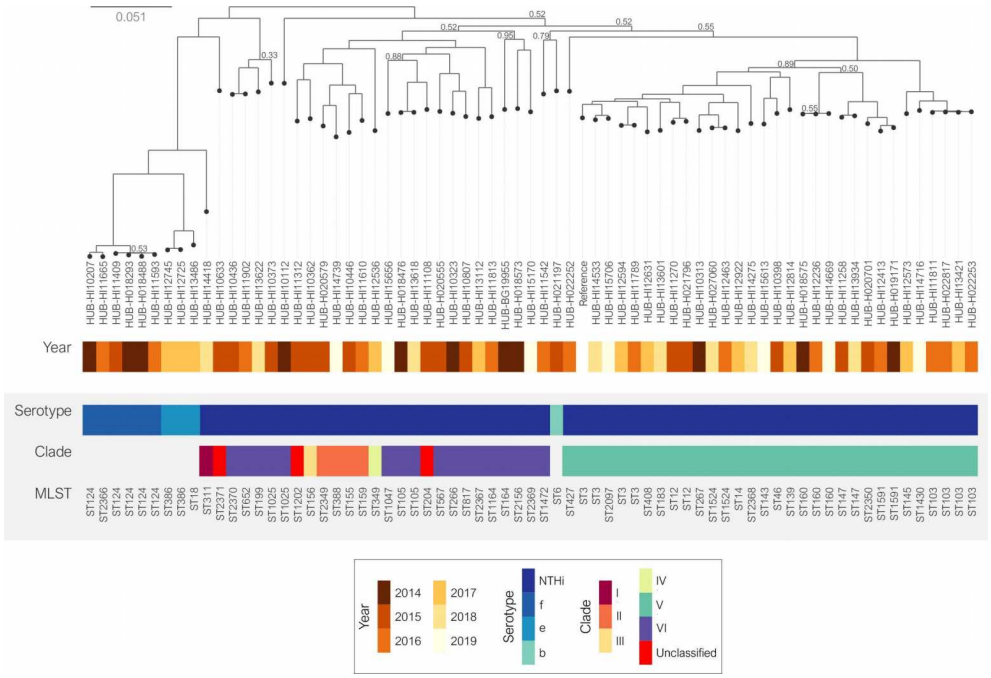


Fig. 3. Core-SNP phylogenetic tree of invasive *H. influenzae* isolates (2014–2019). NTHi was classified into clades (I–VI) as defined by De Chiara *et al.* [12]. Relative bootstrap values other than one are given at branch nodes. Relative bootstrap values below 0.75 indicate that the branches are poorly supported. Strain Hi375 (CP009610) was used as the reference.

each) for serotype e; and ST6 and ST2364 (n=1, each) for serotype b. No differences in ST distribution were observed in the capsulated strains between the periods.

Phylogeny and the NTHi population structure

In the first study period [15], only the seven MLST genes were sequenced. However, in the second period, due to improved methodological techniques, all invasive *H. influenzae* genomes were sequenced through WGS.

The MLST results were confirmed by the phylogenetic analysis (Fig. 3): 28 NTHi strains showed high genetic diversity, while the remaining 30 strains, most of which were clustered by MLST, shared the same monophyletic origin. The invasive NTHi isolates were classified into clades based on the presence or absence of 17 accessory genes as previously described [12, 14]. To provide a context for the invasive NTHi population, we combined our genome collection with the 155 NTHi assemblies selected for clade definition [12, 14] (Fig. 4) . This classification showed that most of the NTHi strains belonged to clades V and VI (32.5 and 37.7%, respectively). Three strains belonging to ST204, ST1202 and ST2371 could not be categorized into any of the clades. According to their monophyletic origins, the ST204 and ST1202 strains were related to clade VI and III, respectively, while ST2371 had a phylogenetic origin unrelated to the other clades (Fig. 3).

The SNP-based phylogenetic analysis revealed high genetic similarity between the clade V isolates, which included the 17 strains belonging to the large MLST clusters formed by the most prevalent STs and the strains belonging to ST160 (n=3), ST12, ST147 and ST1524 (n=2, each), and ST183, ST836, ST1591 and ST2350 (n=1, each). Moreover, nine of the 12 β-lactamase producing strains were included in this clade (Figs 3 and 5).

In addition, each clade constituted a different monophyletic lineage, except for clades II, III and VI, in which the strains showed more diverse phylogenetic origins (Fig. 4). The recombination analyses revealed that a large proportion of sites were affected by recombination, especially in NTHi strains (r/m=7.85, SD=12.42) compared to capsulated isolates (r/m=1.14, SD=1.38) (Figs S2 and S3).

DISCUSSION

Invasive *H. influenzae* disease is uncommon, but can be severe, especially in adults with comorbidities. *H. influenzae* is a matter of concern due to its community-acquired burden and the high prevalence of ampicillin resistance it presents [26]. For these reasons, surveillance has become crucial for the control of *H. influenzae*. In this study, we provide an update on the epidemiology of invasive *H. influenzae* disease

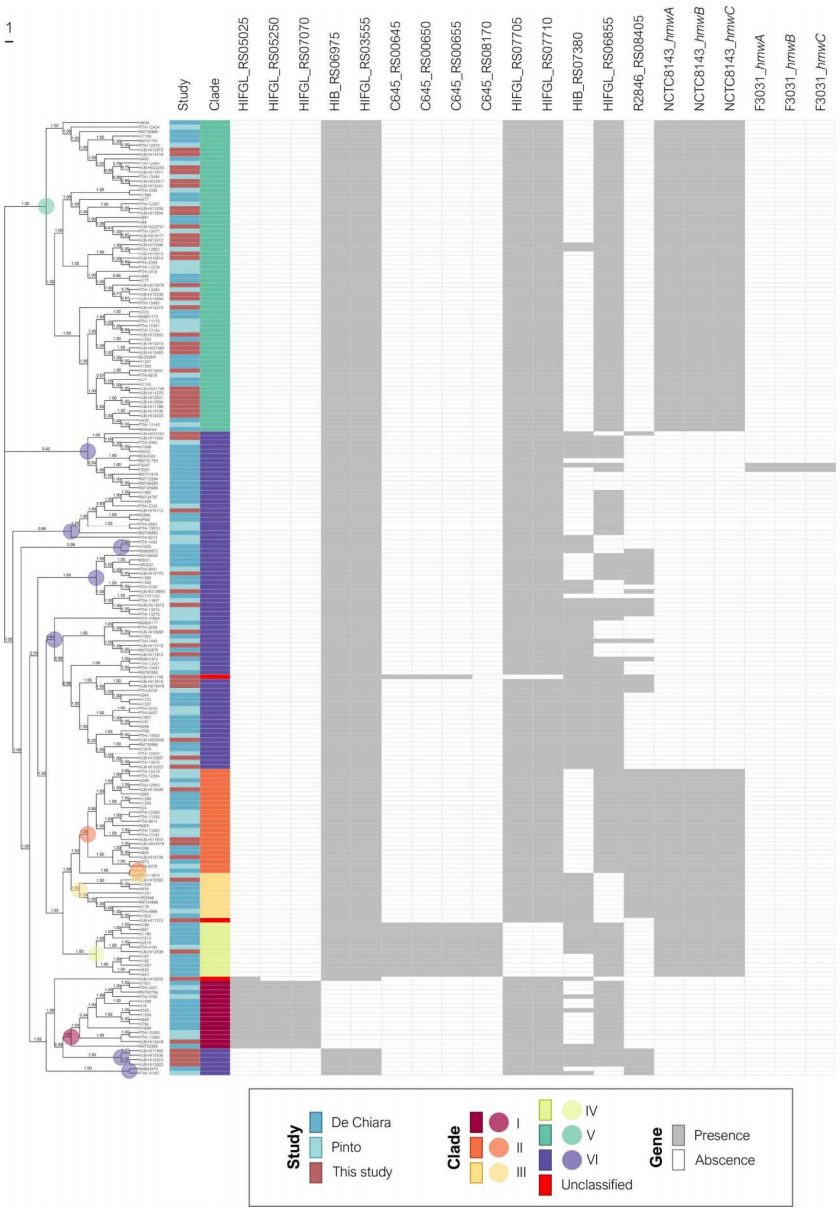


Fig. 4. Assembly-based core-SNP phylogenetic tree and clade distribution of nontypeable *H. influenzae*. The tree includes 213 NTHi genomes from three studies: De Chiara *et al.* [12], Pinto *et al.* [14] and Carrera-Salinas *et al.* (this study). Genomes were distributed in different clades (I–VI), as previously proposed according to the presence or absence of 17 accessory genes [12, 14]: HIFGL_RS05025: *dinB* (DNA polymerase IV); HIFGL_RS05250: 'YbhB/YbcL family Raf kinase inhibitor-like protein'; HIFGL_RS07070: 'ABC transporter substrate-binding protein'; HIB_RS06975: '7-carboxy-7-deazaguanine synthase QueE'; HIFGL_RS03555: 'DEAD/DEAH box helicase family protein'; C645_RS00645: 'Hypothetical protein'; C645_RS00650: 'pirin family protein'; C645_RS00655: 'DUF1016 family protein'; C645_RS08170: 'Hypothetical protein'; HIFGL_RS07705: 'nucleotidyltransferase domain-containing protein'; HIFGL_RS07710: 'nucleotidyltransferase substrate-binding subunit'; HIB_RS07380: 'ABC transporter ATP-binding protein'; HIFGL_RS06855: '5-oxoprolinase/urea amidolyase family protein'; R2846_RS08405: 'TonB-dependent receptor'; NCTC8143_*hmwA/B/C* and F3031_*hmwA/B/C: hmwA/B/C* (high molecular weight protein A, B or C). The coloured dots show the most recent common ancestor (MRCA) found in each clade. Relative bootstrap values are given at branch nodes (values below 0.75 are poorly supported).

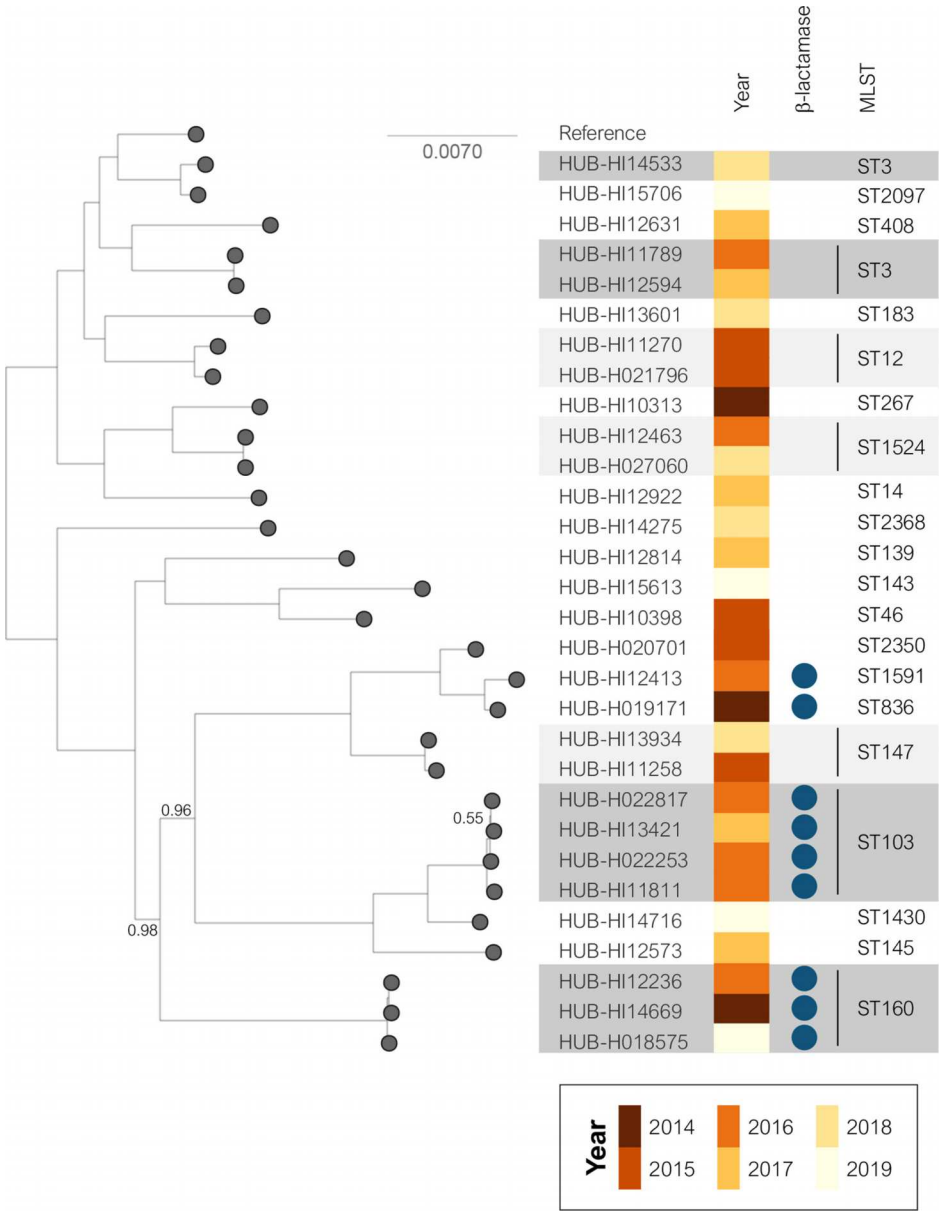


Fig. 5. Phylogenetic tree of invasive NTHi isolates from clade V (2014–2019). Blue circles indicate the presence of a TEM-1 β-lactamase. Shadowed areas highlight the most common STs (dark grey for ≥3 isolates and light grey for two isolates included in the ST). Relative bootstrap values other than one are given at branch nodes. Relative bootstrap values below 0.75 indicate that the branches are poorly supported. Strain Hi375 was used as the reference.

over a 12 year period and analysed the population structure of NTHi using a clade-related classification.

Between 2008 and 2019, the overall incidence of invasive disease in Bellvitge University Hospital was 2.07/100000, which is consistent with the incidence rate found in Europe and the USA [27, 28]. Invasive disease had a higher incidence in individuals over 65 years of age, in line with the underlying conditions that are more common in older adults such as COPD, heart disease and diabetes [8, 29, 30]. By contrast, young adults presented more cases of invasive disease due

to unusual origins, such as peritonitis, liver abscess, endometritis, facial cellulitis or urinary tract infection, which, although less common, typically affect immunocompetent young adults [31–33]. The overall 30 day mortality was 18.0%, within the range of 10–20% found in other studies [34–36]. It should be noted that the 30 day mortality decreased between the two periods (20.7 and 14.7%, respectively), probably due to a better diagnosis of bloodstream infections and a more efficient clinical management of these patients. Moreover, severe underlying conditions were the main cause of mortality, especially in young adults.

Invasive *H. influenzae* infections caused by NTHi and non-type-b serotypes have gradually increased in Europe between 2000 and 2014 [8]. Our study showed a prevalence of 84.8% for NTHi, 10.9% for serotype f, 2.9% for serotype e and 1.4% for serotype b. Likewise, other studies have reported that NTHi strains are the most frequent cause of invasive disease, ranging from 54.6–82% of the cases [27, 36–38], while the frequency of the capsulated serotypes varies according to the geographical region. The results reported to the ECDC [39] indicated that serotype b isolates caused 7% of invasive infections, showing a marked reduction since the introduction of the serotype b vaccination, while prevalence of serotype f, which caused 9% of invasive infections, was increasing annually. However, in some countries such as Italy and Portugal [9, 38], serotype b caused more cases of invasive disease than serotype f despite high vaccine coverage, highlighting the importance of epidemiological surveillance. Although rare in the pre-vaccine era, serotype a is currently emerging as a cause of serious morbidity and mortality in Canada, USA, Brazil and Australia [40], causing up to 23.1% of invasive infections in Canada [37]. Langereis *et al.* [34] observed that most of the infections caused by this serotype occurred in children, probably explaining why it was not detected in our study since it only included the adult population.

Antimicrobial resistance rates remained constant between the two periods of our study, with only a slight increase in β-lactam and fluoroquinolone resistance. TEM-1 β-lactamase was the most common cause of resistance to ampicillin, while one isolate carried a plasmid with the ROB-1 enzyme. San Millan *et al.* [41] demonstrated that the presence of the pB1000 plasmid carrying the ROB-1 β-lactamase reduced the fitness of *H. influenzae*, which could explain the low prevalence of this β-lactamase type. Despite being low, fluoroquinolone resistance emerged in the second period. Fluoroquinolones are commonly used to treat respiratory infections, which may lead to resistance acquisition. Although the fluoroquinolone resistance rate is low worldwide [42, 43], vigilance is required to monitor its progression.

NTHi had higher genetic variability and a larger proportion of sites affected by recombination compared to capsulated isolates or other bacterial species [44]. Most NTHi are composed of individual STs, even if some strains may be grouped occasionally into clonal complexes [9, 38]. SNP-based phylogenetic analysis together with clade-based classification [12, 14] could be useful in identifying NTHi subpopulations with specific pathogenesis and epidemiology. More than half of our invasive NTHi strains were grouped into the same cluster by PHYLOViZ, SNP-based phylogenetic analysis or clade-based classification. All these NTHi isolates were clustered into clade V and belonged to the most prevalent STs (ST3, ST14, ST103, and ST139). They shared a common phylogenetic origin with the clones identified in the first period, and some of them showed increased frequency over time. In addition, in clade V, we found 9 of the 12 invasive isolates with β-lactamase activity, which were associated mainly with ST103, ST160, ST836 and ST1591. Some studies have already reported that ST103 and ST160 are among the most frequent invasive isolates [9, 38]. It should therefore be considered whether the increase in β-lactam resistance is due to the rise of these clones or whether β-lactamase production gives an advantage over the other less frequent clones. In any case, the evolution of the clade V clones should be monitored.

The second most common NTHi clade was VI followed by clade II, while clades I, III and IV were rarely observed. Isolates belonging to clades II and III showed high genetic variability, which could be solved by classification into subclades defined by their different monophyletic origins. On the other hand, clade VI isolates had different monophyletic origins, despite sharing the pattern of accessory genes described by Pinto *et al.* [14]. The inclusion of more samples and the consideration of other accessory genome genes in the analysis may improve this clade classification. Based on their clinical characteristics, it was difficult to differentiate NTHi isolates from among each clade since they were clinically diverse, which is consistent with other studies showing no correlation between the phylogenetic classification and clinical or geographical origins [12, 45]. However, Kc *et al.* [46] classified NTHi isolates into eight clades in which, despite having COPD strains distributed among all the clades, they could differentiate between COPD and non-COPD strains according to the composition of the accessory genome. This suggests that further studies are needed to improve the classification of NTHi strains and to find an association between the clades, the composition of the accessory genome, clinical data and pathogenesis.

Although MLST is useful for classifying capsulated isolates and it did cluster some of the NTHi isolates, it is not the most appropriate technique to study the heterogeneous NTHi population. Details of the heterogeneity of the *H. influenzae* population can be missed if only MLST is performed, particularly if the sample size is small and the analysed clones do not belong to the most prevalent STs. For this reason, we suggest that the combination of phylogenetic analysis and clade-related classification based on the accessory genome is a good approach to monitor the evolution of invasive *H. influenzae* clones and to control the rise of clones associated with β-lactamase production. Further studies will elucidate clades with high phylogenetic variability.

In conclusion, invasive *H. influenzae* disease is mainly associated with NTHi and its incidence increases with age, especially in patients with underlying disease. Serotype f remains the most common among the capsulated strains, followed by

serotype e, although both disappeared by the end of the study. Continued surveillance of invasive *H. influenzae* disease is necessary to control the emergence of NTHi and non-vaccine serotype isolates, as well as the rising resistance to β-lactams and fluoroquinolones. We propose that the NTHi population structure should be studied through phylogenetic analysis and clade-related classification, as it clusters the main STs associated with invasive disease. This is an alternative to overcome the limitations of MLST and to comprehend the genetic diversity and dynamics of NTHi.

Funding information
This study was funded by Instituto de Salud Carlos III (ISCIII) through the Projects from the Fondo de Investigaciones Sanitarias 'PI16/00977' to SM, and CIBER de Enfermedades Respiratorias (CIBERES–CB06/06/0037), co-funded by the European Regional Development Fund/European Social Fund (ERDF/ESF, 'Investing in your future'). Bioinformatic analysis was supported by an Amazon Web Services (AWS) research grant to SM. AC was supported by FPU grant 'FPU16/02202' (Formación de Profesorado Universitario, Ministerio de Educación, Spain), and SM was supported by Miguel Servet contract 'CP19/00096' (ISCIII).

Acknowledgements
We would like to thank the staff of the Microbiology Laboratory of Bellvitge University Hospital who contributed daily to this project.

Author contributions
M.A.D., C.A. and S.M. supervised the study. A.C., A.G., J.M. and C.P. performed laboratory assays. A.C., A.G., L.C., D.B., J.C., F.T., I.G., M.A.D., C.A. and S.M. analysed and interpreted the data. A.C. wrote the manuscript with the assistance of A.G. and S.M. All authors read and approved the final manuscript.

Conflicts of interest
The authors declare that there are no conflicts of interest.

Ethical statement
This study was in accordance with the Declaration of Helsinki from the World Medical Association. It was approved by the Clinical Research Ethics Committee of Bellvitge University Hospital (PR126/21). Written informed consent was not required as this was a retrospective and observational study with isolates obtained as part of routine microbiological tests. Patient confidentiality was always protected and all personal data were anonymized following the current legal normative in Spain (LOPD 15/1999 and RD 1720/2007). Moreover, this project followed Law 14/2007 of Biomedical Investigation for the management of biological samples in clinical research.

References

1. LaCross NC, Marrs CF, Gilsdorf JR. Population structure in nontypeable *Haemophilus influenzae*. *Infect Genet Evol* 2013;14:125–136.
2. Slack MPE. A review of the role of *Haemophilus influenzae* in community-acquired pneumonia. *Pneumonia (Nathan)* 2015;6:26–43.
3. Nørskov-Lauritsen N. Classification, identification, and clinical significance of *Haemophilus* and *Aggregatibacter* species with host specificity for humans. *Clin Microbiol Rev* 2014;27:214–240.
4. Agrawal A, Murphy TF. *Haemophilus influenzae* infections in the *H. influenzae* type b conjugate vaccine era. *J Clin Microbiol* 2011;49:3728–3732.
5. Sethi S, Murphy TF. Infection in the pathogenesis and course of chronic obstructive pulmonary disease. *N Engl J Med* 2008;359:2355–2365.
6. Cardines R, Giufrè M, Pompilio A, Fiscarelli E, Ricciotti G, *et al*. *Haemophilus influenzae* in children with cystic fibrosis: Antimicrobial susceptibility, molecular epidemiology, distribution of adhesins and biofilm formation. *Int J Med Microbiol* 2012;302:45–52.
7. van de Beek D, Brouwer M, Hasbun R, Koedel U, Whitney CG, *et al*. Community-acquired bacterial meningitis. *Nat Rev Dis Primers* 2016;2:16074.
8. Wang S, Tafalla M, Hanssens L, Dolhain J. A review of *Haemophilus influenzae* disease in Europe from 2000-2014: challenges, successes and the contribution of hexavalent combination vaccines. *Expert Rev Vaccines* 2017;16:1095–1105.
9. Heliodoro CIM, Bettencourt CR, Bajanca-Lavado MP, Portuguese Group for the Study of Haemophilus influenzae invasive infection. Molecular epidemiology of invasive *Haemophilus influenzae* disease in Portugal: an update of the post-vaccine period, 2011-2018. *Eur J Clin Microbiol Infect Dis* 2020;39:1471–1480.
10. Wen S, Feng D, Chen D, Yang L, Xu Z. Molecular epidemiology and evolution of *Haemophilus influenzae*. *Infect Genet Evol* 2020;80:104205.
11. Robinson DA, Thomas JC, Hanage WP. Population structure of pathogenic bacteria. In: *Genetics and Evolution of Infectious Diseases*. Elsevier Inc, 2011. pp. 43–57.
12. De Chiara M, Hood D, Muzzi A, Pickard DJ, Perkins T, *et al*. Genome sequencing of disease and carriage isolates of nontypeable *Haemophilus influenzae* identifies discrete population structure. *Proc Natl Acad Sci U S A* 2014;111:5439–5444.
13. Staples M, Graham RMA, Jennison AV. Characterisation of invasive clinical *Haemophilus influenzae* isolates in Queensland, Australia using whole-genome sequencing. *Epidemiol Infect* 2017;145:1727–1736.
14. Pinto M, González-Díaz A, Machado MP, Duarte S, Vieira L, *et al*. Insights into the population structure and pan-genome of *Haemophilus influenzae*. *Infect Genet Evol* 2019;67:126–135.
15. Puig C, Grau I, Marti S, Tubau F, Calatayud L, *et al*. Clinical and molecular epidemiology of *Haemophilus influenzae* causing invasive disease in adult patients. *PLoS One* 2014;9:e112711.
16. Clinical Laboratory Standards Institute. Performance standards for antimicrobial susceptibility testing: 29th Edition. In: *CLSI Supplement Document M100*. Wayne, PA, 2019.
17. Francisco AP, Vaz C, Monteiro PT, Melo-Cristino J, Ramirez M, *et al*. PHYLOViZ: Phylogenetic inference and data visualization for sequence based typing methods. *BMC Bioinformatics* 2012;13:87.
18. Watts SC, Holt KE. HICAP: In silico serotyping of the Haemophilus influenzae capsule locus. *J Clin Microbiol* 2019;57:e00190-19.
19. Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, *et al*. ResFinder 4.0 for predictions of phenotypes from genotypes. *J Antimicrob Chemother* 2020;75:3491–3500.
20. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 2019;35:4453–4455.
21. Pattengale ND, Alipour M, Bininda-Emonds ORP, Moret BME, Stamatakis A. How many bootstrap replicates are necessary? *J Comput Biol* 2010;17:337–354.
22. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, *et al*. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* 2015;43:e15.
23. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol* 2014;15:524.
24. Yu G, Smith DK, Zhu H, Guan Y, Lam TT, *et al*. ggtree : an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* 2016;8:28–36.
25. Dabernat H, Delmas C, Seguy M, Pelissier R, Faucon G, *et al*. Diversity of beta-lactam resistance-conferring amino acid substitutions in penicillin-binding protein 3 of *Haemophilus influenzae*. *Antimicrob Agents Chemother* 2002;46:2208–2218.
26. Tacconelli E, Carrara E, Savoldi A, Harbarth S, Mendelson M, *et al*. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect Dis* 2018;18:318–327.

27. Whittaker R, Economopoulou A, Dias JG, Bancroft E, Ramliden M, *et al*. Epidemiology of Invasive *Haemophilus influenzae* Disease, Europe, 2007-2014. *Emerg Infect Dis* 2017;23:396–404.
28. Soeters HM, Blain A, Pondo T, Doman B, Farley MM, *et al*. Current epidemiology and trends in invasive *Haemophilus influenzae* Disease-United States, 2009-2015. *Clin Infect Dis* 2018;67:881–889.
29. GBD Chronic Respiratory Disease Collaborators. Prevalence and attributable health burden of chronic respiratory diseases, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Respir Med* 2020;8:585–596.
30. Haro JM, Tyrovolas S, Garin N, Diaz-Torne C, Carmona L, *et al*. The burden of disease in Spain: results from the global burden of disease study 2010. *BMC Med* 2014;12:236.
31. Fujii M, Gomi H, Ishioka H, Takamura N. Bacteremic renal stone-associated urinary tract infection caused by nontypable *Haemophilus influenzae*: A rare invasive disease in an immunocompetent patient. *IDCases* 2017;7:11–13.
32. Stærk M, Tolouee SA, Christensen JJ. Nontypable *Haemophilus influenzae* septicemia and urinary tract infection associated with renal stone disease. *Open Microbiol J* 2018;12:243–247.
33. Martin D, Dbouk RH, Deleon-Carnes M, del Rio C, Guarner J. *Haemophilus influenzae* acute endometritis with bacteremia: Case report and literature review. *Diagn Microbiol Infect Dis* 2013;76:235–236.
34. Langereis JD, de Jonge MI. Invasive disease caused by nontypeable *Haemophilus influenzae*. *Emerg Infect Dis* 2015;21:1711–1718.
35. Chiappini E, Inturrisi F, Orlandini E, de Martino M, de Waure C. Hospitalization rates and outcome of invasive bacterial vaccine-preventable diseases in Tuscany: A historical cohort study of the 2000-2016 period. *BMC Infect Dis* 2018;18:396.
36. Blain A, MacNeil J, Wang X, Bennett N, Farley MM, *et al*. Invasive *Haemophilus influenzae* Disease in Adults ≥65 Years, United States, 2011. *Open Forum Infect Dis* 2014;1:ofu044.
37. Tsang RSW, Shuel M, Whyte K, Hoang L, Tyrrell G, *et al*. Antibiotic susceptibility and molecular analysis of invasive *Haemophilus influenzae* in Canada, 2007 to 2014. *J Antimicrob Chemother* 2017;72:1314–1319.
38. Giufrè M, Fabiani M, Cardines R, Riccardo F, Caporali MG, *et al*. Increasing trend in invasive non-typeable *Haemophilus influenzae* disease and molecular characterization of the isolates, Italy, 2012-2016. *Vaccine* 2018;36:6615–6622.
39. European Centre for Disease Prevention and Control (ECDC). *Haemophilus influenzae annual epidemiological report for 2018*. Stockholm, 2020.
40. Ulanova M, Tsang RSW. *Haemophilus influenzae* serotype a as a cause of serious invasive infections. *Lancet Infect Dis* 2014;14:70–82.
41. San Millan A, Garcia-Cobos S, Escudero JA, Hidalgo L, Gutierrez B, *et al*. *Haemophilus influenzae* clinical isolates with plasmid pB1000 bearing blaROB-1: Fitness cost and interspecies dissemination. *Antimicrob Agents Chemother* 2010;54:1506–1511.
42. Shoji H, Shirakura T, Fukuchi K, Takuma T, Hanaki H, *et al*. A molecular analysis of quinolone-resistant *Haemophilus influenzae*: Validation of the mutations in quinolone resistance-determining regions. *J Infect Chemother* 2014;20:250–255.
43. Puig C, Tirado-Vélez JM, Calatayud L, Tubau F, Garmendia J, *et al*. Molecular characterization of fluoroquinolone resistance in nontypeable *Haemophilus influenzae* clinical isolates. *Antimicrob Agents Chemother* 2015;59:461–466.
44. Vos M, Didelot X. A comparison of homologous recombination rates in bacteria and archaea. *ISME J* 2009;3:199–208.
45. Pettigrew MM, Ahearn CP, Gent JF, Kong Y, Gallo MC, *et al*. *Haemophilus influenzae* genome evolution during persistence in the human airways in chronic obstructive pulmonary disease. *Proc Natl Acad Sci U S A* 2018;115:E3256–E3265.
46. Kc R, Leong KWC, Harkness NM, Lachowicz J, Gautam SS, *et al*. Whole-genome analyses reveal gene content differences between nontypeable *Haemophilus influenzae* isolates from chronic obstructive pulmonary disease compared to other clinical phenotypes. *Microbial Genomics* 2020;6:1–12.

SUPPLEMENTARY MATERIAL

Supplementary Dataset S1 is available with the online version of this article.

Supplementary Table S1. Types of samples from which *H. influenzae* was isolated over the study period.

Sample type	Number of isolates		
	First period 2008-2013	Second period 2014-2019	All study period 2008-2019
Adenopathy	2	0	2
Ascitic fluid	2	0	2
Bile	8	5	13
Blood*	70	62	132
Cerebrospinal fluid*	8	5	13
Joint tissue	1	1	2
Lower respiratory tract[a]	3184	4560	7744
Ocular[b]	38	8	46
Pleural fluid*	4	1	5
Rectal swab	1	2	3
Skin	2	1	3
Ulcer or abscess	34	37	71
Upper respiratory tract[c]	15	16	31
Urogenital tract[d]	55	52	107
Wound	9	8	17
Total	3433	4758	8191

[a]Lower respiratory tract samples: sputum, bronchial or tracheal aspiration, and bronchoalveolar lavage.
[b]Ocular samples: conjunctival or corneal exudate, ocular prothesis, and contact lens.
[c]Upper respiratory tract samples: nasal or pharyngeal exudate and nasal or pharyngeal swab.
[d]Urogenital tract samples: endometrial, vaginal or urethral swab, semen, and genital ulcer.
*Samples included in the study.

Supplementary Table S2. Demographic data, clinical characteristics, and underlying conditions of patients with invasive *H. influenzae* disease associated to 30-day mortality.

	30-day mortality in the overall study period (2008-2019) n = 150			30-day mortality in young adults (<65 years) n = 53			30-day mortality in elderly adults (≥65 years) n = 97			30-day mortality comparison between periods		
	30-day mortality n = 27	30-day survival n = 123	p-value	30-day mortality n = 10	30-day survival n = 43	p-value	30-day mortality n = 17	30-day survival n = 80	p-value	First period (2008-2013) n = 17	Second period (2014-2019) n = 10	p-value
Characteristics [no. (%)]												
Age (mean ± SD; range)	66.6 ± 14.1; 21-96	67.6 ± 15.1; 23-93	0.74	54.0 ± 12.8; 21-63	51.1 ± 11.1; 23-64	0.47	74.0 ± 8.4; 65-96	76.3 ± 7.5; 65-93	0.21	67.0 ± 16.7; 21-96	65.9 ± 8.6; 55-86	0.85
Male sex	15 (55.6)	73 (59.3)	0.83	7 (70.0)	23 (53.5)	0.48	8 (47.1)	50 (62.5)	0.28	11 (64.7)	4 (40.0)	0.26
Underlying conditions												
Immunosuppressive therapy	17 (63.0)	43 (35.0)	**0.01**	6 (60.0)	16 (37.2)	0.29	11 (64.7)	27 (33.8)	**0.03**	9 (52.9)	8 (80.0)	0.23
Solid organ malignancy	16 (59.3)	38 (30.9)	**0.01**	6 (60.0)	11 (25.6)	0.06	10 (58.8)	27 (33.8)	0.06	8 (47.1)	8 (80.0)	0.12
Diabetes	7 (25.9)	32 (26.0)	1	2 (20.0)	6 (14.0)	0.64	5 (29.4)	26 (32.5)	1	5 (29.4)	2 (20.0)	0.68
Heart disease	8 (29.6)	32 (26.0)	0.81	3 (30.0)	4 (9.3)	0.11	5 (29.4)	28 (35.0)	0.78	5 (29.4)	3 (30.0)	1
COPD	6 (22.2)	24 (19.5)	0.79	0 (0)	4 (9.3)	1	6 (35.3)	20 (25.0)	0.38	4 (23.5)	2 (20.0)	1
Chronic liver disease	4 (14.8)	14 (11.4)	0.74	1 (10.0)	9 (20.9)	0.67	3 (17.6)	5 (6.3)	0.14	2 (11.8)	2 (20.0)	0.61
Hematologic malignancy[a]	4 (14.8)	12 (9.8)	0.49	2 (20.0)	5 (11.6)	0.6	2 (11.8)	7 (8.8)	0.66	4 (23.5)	0 (0.0)	0.26
Cerebrovascular disease	2 (7.4)	5 (4.1)	0.61	0 (0)	0 (0)	1	2 (11.8)	5 (6.3)	0.6	2 (11.8)	0 (0.0)	0.56
Organ transplant[b]	0 (0.0)	10 (8.1)	0.21	0 (0)	6 (14.0)	0.58	0 (0.0)	4 (5.0)	1	0 (0.0)	0 (0.0)	1
HIV	2 (7.4)	1 (0.8)	0.08	2 (20.0)	1 (2.3)	0.09	0 (0.0)	0 (0.0)	1	1 (5.9)	1 (10.0)	1
No underlying conditions	0 (0.0)	27 (22.0)	**0.004**	0 (0)	15 (34.8)	**0.05**	0 (0.0)	12 (15.0)	0.12	0 (0.0)	0 (0.0)	1
Acquisition												
Community acquired	21 (77.8)	109 (88.6)	0.21	8 (80.0)	35 (81.4)	1	15 (88.2)	71 (88.8)	1	13 (76.5)	8 (80.0)	1
Nosocomial	6 (22.2)	14 (11.4)	0.21	2 (20.0)	8 (18.6)	1	2 (11.8)	9 (11.3)	1	4 (23.5)	2 (20.0)	1
Source of infection												
Respiratory tract infection	19 (70.4)	82 (66.7)	0.12	8 (80.0)	19 (44.2)	0.08	11 (64.7)	63 (78.8)	0.22	12 (70.6)	7 (70.0)	1
Meningitis	0 (0.0)	14 (11.4)	0.08	0 (0)	5 (11.6)	0.57	0 (0.0)	9 (11.3)	0.35	0 (0.0)	0 (0.0)	1
Biliary tract infection	3 (11.1)	11 (8.9)	0.72	0 (0)	7 (16.3)	0.32	3 (17.6)	4 (5.0)	0.1	1 (5.9)	2 (20.0)	0.54
Primary bacteremia	3 (11.1)	6 (4.9)	0.21	1 (10.0)	5 (11.6)	1	2 (11.8)	1 (1.3)	0.08	2 (11.8)	1 (10.0)	1
Peritonitis	1 (3.7)	2 (1.6)	0.45	0 (0)	2 (4.7)	1	1 (5.9)	0 (0.0)	0.18	1 (5.9)	0 (0.0)	1
Other[c]	1 (3.7)	8 (6.5)	1	1 (10.0)	5 (11.6)	1	0 (0.0)	3 (3.8)	1	1 (5.9)	0 (0.0)	1
Charlson comorbidity index	5.7 ± 2.6	4.6 ± 2.4	**0.04**	5.2 ± 3.8	3.0 ± 2.7	**0.04**	5.9 ± 1.6	5.4 ± 1.8	0.26	5.3 ± 1.8	6.3 ± 3.6	0.34

[a]Myeloma, leukemia and lymphoma.

[b]Liver, bone marrow, kidney, and heart.

[c]Epiglotittis, liver abscess, endometritis, facial cellulites, septic shock of an aortic valve prosthetic origin, surgical wound infection and urinary tract infection.

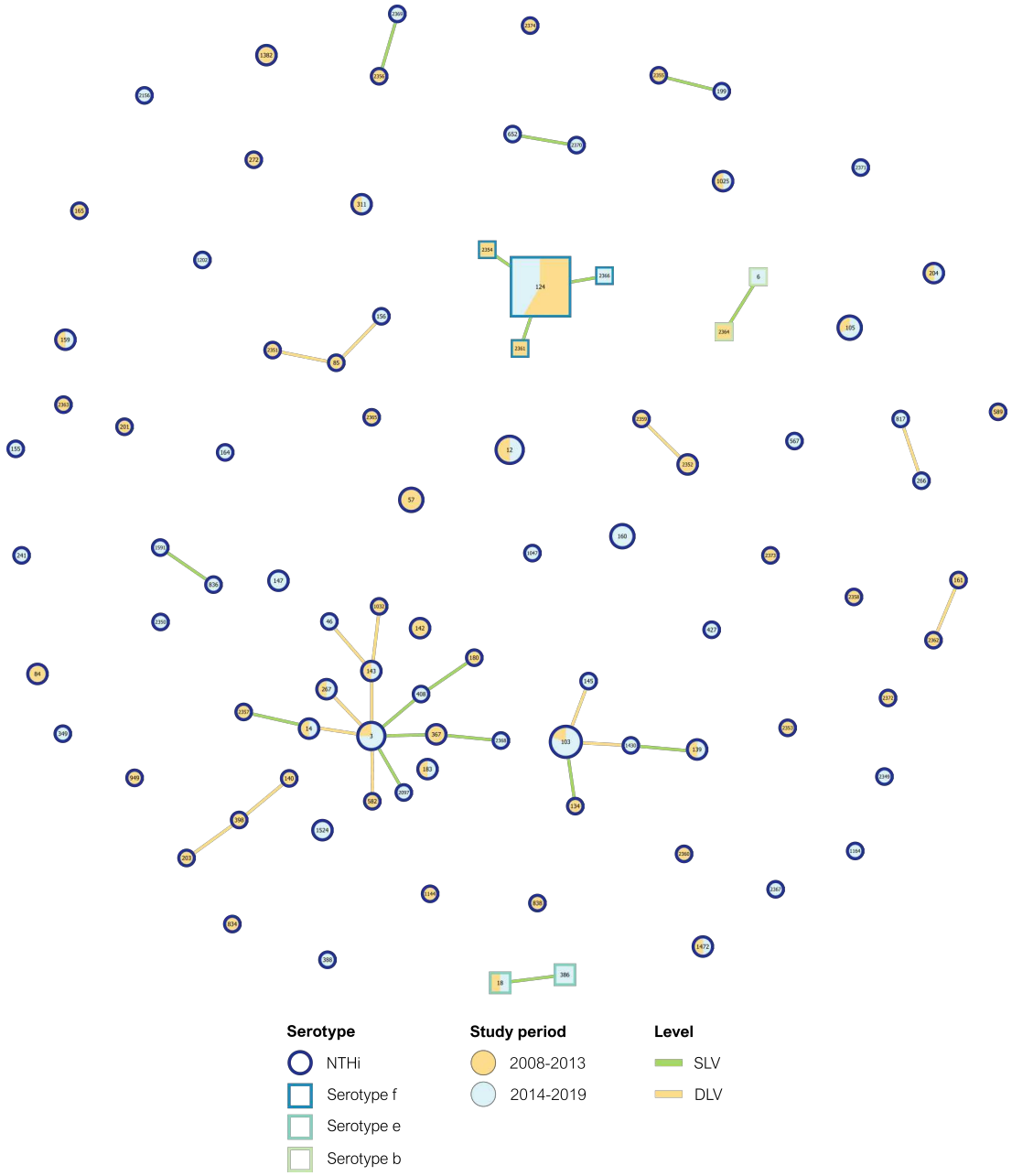


Figure S1. Relationship between the invasive *H. influenzae* isolates defined by PHYLOViZ. Circles for NTHi clones and squares for capsulated *H. influenzae* strains correspond to unique STs. The size of the circles/squares represents the number of isolates belonging to each ST. The inner color indicates the proportion of isolates for each period (2008-2013 [15] and 2014-2019 (this study)), while the outline color indicates the serotype. The relationship between the STs is defined as single- (SLV) or double-locus (DLV) variant, according to the number of different loci observed between the STs (one or two, respectively).

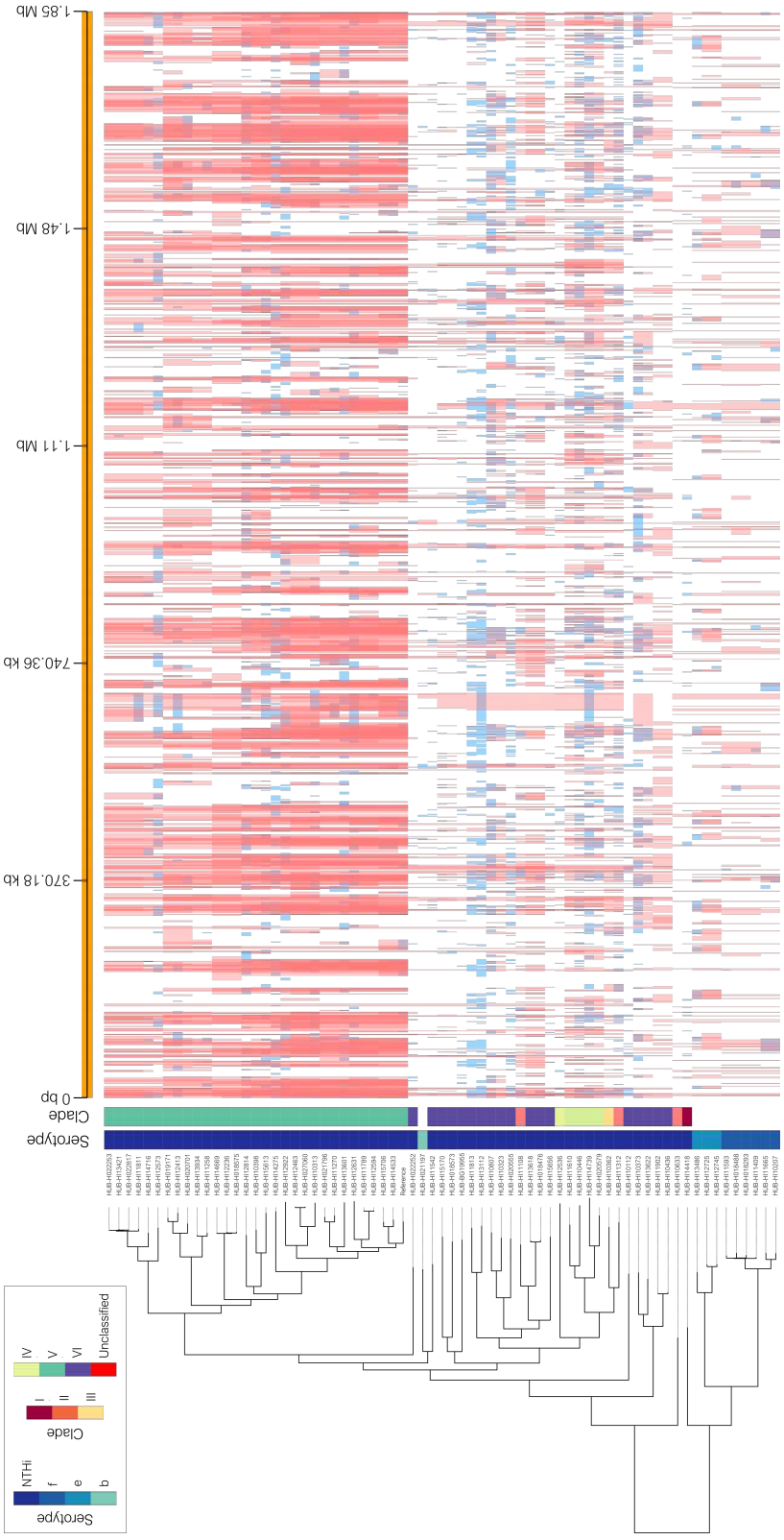


Figure S2. Sites affected by recombination in invasive *H. influenzae* isolates (2014-2019). Prediction of recombinant regions was performed using the Gubbins v2.3.1 software, and the core-SNP phylogenetic tree was built using RAxML-NG based only on shared positions after recombination removal. The Hi375 strain (CP009610) was used as a reference.

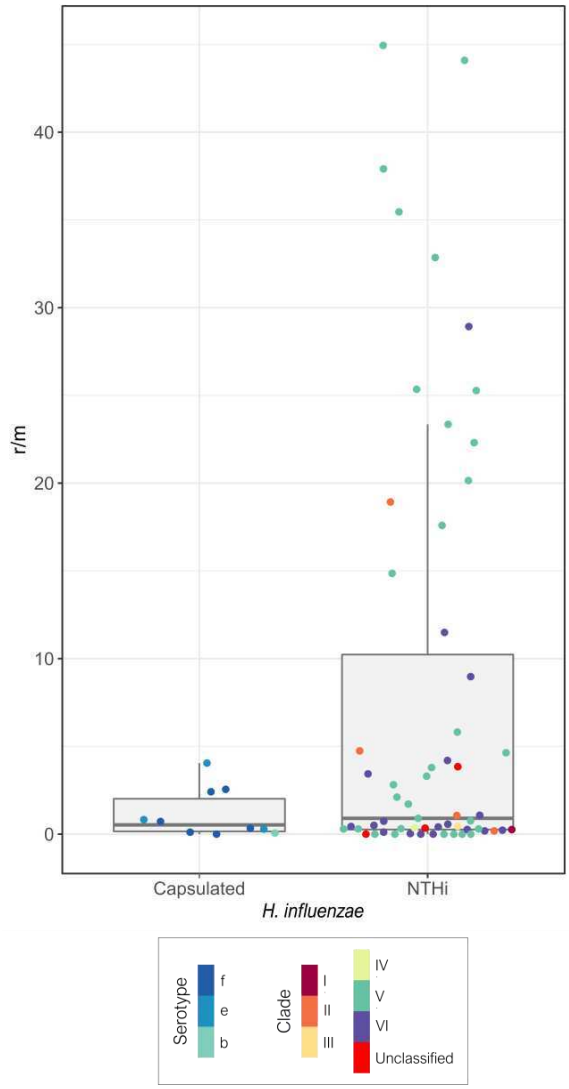


Figure S3. Relative impact of recombination and mutation (r/m) in invasive capsulated and NTHi strains (2014-2019). Prediction of recombinant regions was performed using the Gubbins v2.3.1 software, and the core-SNP phylogenetic tree was built using RAxML-NG based only on shared positions after recombination removal. The Hi375 strain (CP009610) was used as a reference.

Study 2

Comparative pangenome analysis of capsulated *Haemophilus influenzae* serotype f highlights their high genomic stability

Aida González-Díaz, Anna Carrera-Salinas, Miguel Pinto, Meritxell Cubero, Arie van der Ende, Jeroen D. Langereis, M. Ángeles Domínguez, Carmen Ardanuy, Paula Bajanca-Lavado, Sara Martí

Dr. Mª Ángeles Domínguez Luzón, Associate Professor of the Department of Pathology and Experimental Therapeutics of the Universitat de Barcelona and Head of the Microbiology Department, Hospital Universitari de Bellvitge (Barcelona), and Dr. Sara Martí Martí, Assistant Professor of the Department of Medicine of the Universitat de Barcelona and Senior Postdoctoral Researcher at the Centre for Biomedical Research Network on Respiratory Diseases (CIBERes),

DECLARE that the original article entitled "**Comparative pangenome analysis of capsulated *Haemophilus influenzae* serotype f highlights their high genomic stability**" has been published in Scientific Reports (Impact factor in 2021: 4.996). The authors Anna Carrera-Salinas and Aida González-Díaz have included this work in their respective doctoral theses (Aida González-Díaz used a preliminary version of this work in her doctoral thesis defended in 2020), and their contribution is detailed below:

Aida González-Díaz

- Library preparation for Whole-Genome Sequencing (WGS) and WGS (Illumina).
- Bioinformatic analysis and phylogeny (Linux).
- Preparation of figures: Figure 1, and Supplementary Figures S1-S4.
- Manuscript redaction.

Anna Carrera-Salinas

- Bioinformatic analysis and phylogeny (R-studio).
- SNP typing and pangenomic analysis.
- Preparation of figures: Figures 2-3, and Supplementary Figure S5.
- Manuscript redaction.

Signatures of thesis supervisors

MªÁngeles Domínguez Luzón **Sara Martí Martí**

scientific reports

Check for updates

OPEN

Comparative pangenome analysis of capsulated *Haemophilus influenzae* serotype f highlights their high genomic stability

Aida Gonzalez-Diaz[1,2], Anna Carrera-Salinas[1], Miguel Pinto[3], Meritxell Cubero[1,2], Arie van der Ende[4,5], Jeroen D. Langereis[6,7], M. Ángeles Domínguez[1,8,9], Carmen Ardanuy[1,2,8], Paula Bajanca-Lavado[10] & Sara Marti[1,2,11]✉

Haemophilus influenzae **is an opportunistic pathogen adapted to the human respiratory tract. Non-typeable *H. influenzae* are highly heterogeneous, but few studies have analysed the genomic variability of capsulated strains. This study aims to examine the genetic diversity of 37 serotype f isolates from the Netherlands, Portugal, and Spain, and to compare all capsulated genomes available on public databases. Serotype f isolates belonged to CC124 and shared few single nucleotide polymorphisms (SNPs) (n = 10,999), but a high core genome (> 80%). Three main clades were identified by the presence of 75, 60 and 41 exclusive genes for each clade, respectively. Multi-locus sequence type analysis of all capsulated genomes revealed a reduced number of clonal complexes associated with each serotype. Pangenome analysis showed a large pool of genes (n = 6360), many of which were accessory genome (n = 5323). Phylogenetic analysis revealed that serotypes a, b, and f had greater diversity. The total number of SNPs in serotype f was significantly lower than in serotypes a, b, and e ($p < 0.0001$), indicating low variability within the serotype f clonal complexes. Capsulated *H. influenzae* are genetically homogeneous, with few lineages in each serotype. Serotype f has high genetic stability regardless of time and country of isolation.**

Haemophilus influenzae is a Gram-negative coccobacillus of the *Pasteurellaceae* family that colonises the human nasopharynx and throat in more than 50% of children and 20–30% of adults, causing a wide range of infection from chronic respiratory disease to severe invasive diseases, such as bacteraemia and meningitis[1,2]. The capsule is an important virulence factor, though it is not present in all strains of *H. influenzae*. Six different capsular operons have been described that encode six unique polysaccharide capsules (a–f). Strains missing the genes for the capsular operon are known as non-typeable *H. influenzae* (NTHi)[3].

Prior to the introduction of a successful *H. influenzae* serotype b conjugate vaccine, invasive disease caused by the serotype b was leading significant cause of morbidity and mortality, especially in cases of meningitis in children under 5 years of age, while NTHi was almost exclusively associated with upper and lower respiratory tract infection[4,5]. However, widespread vaccine implementation has produced an epidemiological shift, and currently, most invasive infections occur in elderly patients with underlying conditions and are mainly caused by

[1]Microbiology Department, Hospital Universitari de Bellvitge, IDIBELL-UB, Feixa Llarga s/n, L'Hospitalet de Llobregat, 08907 Barcelona, Spain. [2]Research Network for Respiratory Diseases (CIBERES), ISCIII, Madrid, Spain. [3]Bioinformatics Unit, Department of Infectious Disease, National Institute of Health, Lisbon, Portugal. [4]Infection and Immunity Amsterdam, Department of Medical Microbiology and Infection Prevention, Amsterdam UMC, Amsterdam, The Netherlands. [5]Amsterdam UMC, Reference Laboratory for Bacterial Meningitis, Amsterdam, The Netherlands. [6]Section Pediatric Infectious Diseases, Laboratory of Medical Immunology, Radboud Institute for Molecular Life Sciences, Radboudumc, Nijmegen, The Netherlands. [7]Radboud Center for Infectious Diseases, Radboudumc, Nijmegen, The Netherlands. [8]Department of Pathology and Experimental Therapeutics, School of Medicine, University of Barcelona, Barcelona, Spain. [9]Research Network for Infectious Diseases (CIBERINFEC), ISCIII, Madrid, Spain. [10]Haemophilus Influenzae Reference Laboratory, Department of Infectious Disease, National Institute of Health, Lisbon, Portugal. [11]Department of Medicine, School of Medicine, University of Barcelona, Barcelona, Spain. ✉email: smartinm@bellvitgehospital.cat

NTHi, followed at some distance by non-type b serotypes[6]. Serotype f should be considered a leading non-type b serotype that causes adult invasive *H. influenzae* disease, such as bacteraemia, in Europe and the United States[7–10].

H. influenzae shares its ecological niche with many commensal bacteria and potential pathogens[11]. The local environment influences the ability to exchange genetic material; species that interact with other bacteria generally experience higher recombination rates than microorganisms living in less diverse settings[12]. Population structure analysis has shown that capsulated strains are highly clonal and have a limited number of serotype-associated lineages. By contrast, NTHi appears to have discrete subpopulation structures, but genetic diversity that is ten times greater. This clonality could be related to the presence of capsules and the fact that capsulated strains are more commonly found in invasive disease[13,14]. However, few studies have been performed in large datasets of capsulated genomes.

In this study, we examine the genomic diversity of *H. influenzae* serotype f through whole genome sequencing (WGS), using a multicentre collection of colonising and invasive isolates from the Netherlands, Portugal, and Spain. We also compare the population genetics of all capsulated *H. influenzae* genomes available in the National Center for Biotechnology Information (NCBI) and European Nucleotide Archive (ENA) databases.

Results

Pangenome variability in *H. influenzae* serotype f among countries and by sample origin. Of the 37 sequenced *H. influenzae* serotype f isolates, 33 were collected from invasive sites, including blood samples (n = 30) and cerebrospinal, joint, and pleural fluid samples (n = 1, each). The remaining four isolates were obtained from oropharyngeal colonisation of healthy children[15]. In the MLST profile, all isolates belonged to CC124, with most being ST124 (n = 31) and the rest being single-locus variants of ST124, such as ST1739 (n = 2), ST106, ST2390, ST2366, and ST2391 (n = 1, each) (Supplementary Table S1). The coloniser ES-HICOv-HILNM (ST106) and invasive ES-HUB-11665 (ST2366) strains were phylogenetically related but did not show any epidemiological association because they were isolated from two different regions of Spain (Oviedo and Barcelona) separated by twelve years.

The proportion of the genome shared by the 37 isolates was very high (> 80%), and only 10,999 SNPs were found in the core genome, with an average of 1297 SNPs (SD = 1799) compared to the reference genome. After core genome phylogenetic analysis, three clades (I–III) were distinguished by the presence of clade-specific allelic variants represented (Fig. 1A). Sub-clades of clade III did not present exclusive alleles that allow it to be considered an independent clade. Clades I and II were less common (n = 5 and n = 4, respectively) and included the four coloniser isolates. Clade I contained two invasive and three colonising strains, all of them isolated in Spain. Clade II included the remaining coloniser isolates as well as three invasive isolates, two from Spain and one from the Netherlands. By contrast, clade III grouped most serotype f genomes (n = 28) and was exclusively associated with invasive clones. No phylogenetic association was found with the geographical origin of clade III clones.

Pangenome analysis detected 1891 genes in the gene pool, of which 1571 were present in all genomes (core genes), 114 genes in 95–99% (soft-core genes), 67 in 15–94% (shell genes), and 139 genes in < 15% (cloud genes) (Fig. 1B). Country, source or year of isolation, as well as clade, were not related to the presence or absence of any gene. Among the core genes, 910 were monoallelic (same allele in all isolates) and 499 had alleles distributed indifferently across the isolates and clades. In addition, 162 genes had clade-specific alleles that allowed segregation of the genomes into the different clades: clade I had 75 specific alleles, clade II had 60 alleles, and clade III had 41 alleles. When these alleles were compared to other alleles of the same gene, the identity and number of SNPs showed low variability (Fig. 1): clade I segregating alleles had a mean identity of 97.1% (standard deviation [SD] = 2.7) and an average of 32.0 SNPs (SD = 33.1); clade II segregating alleles had an identity of 96.8% (SD = 2.9) and 31.4 SNPs (SD = 34.4); and clade III segregating alleles had an identity of 97.6% (SD = 2.5) and 19.0 SNPs (SD = 24.6). However, it should be noted that clade III alleles showed fewer SNPs compared to clade I (p-value = 0.0291) and clade II (p-value = 0.0538) alleles. The genes associated with clade segregating alleles, their protein products, and their functions are shown in Supplementary Table S2.

MLST and phylogenetic relationship of capsulated *H. influenzae*. A total of 800 genomes, comprising 763 of capsulated *H. influenzae* from the NCBI and ENA databases (Supplementary Table S3) and 37 of serotype f in this study, were included to determine the phylogenetic diversity of capsulated *H. influenzae*. The presence of capsular type in silico revealed 205 serotype a, 165 serotype b, 34 serotype c, 10 serotype d, 152 serotype e, and 234 serotype f genomes (197 from databases and 37 from this study).

The genomes of capsulated *H. influenzae* showed high homogeneity within each serotype. The phylogenetic analysis revealed three major branches, one containing the genomes of serotypes a, b, c, and d, one containing CC6 genomes of serotype b, and the other containing the genomes of serotypes e and f, as well as the CC464 genome of serotype b and genomes of serotype a associated to CC62 and CC372 (Supplementary Fig. S1). Each serotype had a distinct monophyletic lineage, except for serotypes a and b from the first branch, where CC4 and CC50, respectively, showed a different origin compared to other genomes of their serotypes.

Each serotype was classified into clades based on its phylogenetic origins and clonal complex distribution. Serotype a genomes were phylogenetically grouped into three different clades (Supplementary Fig. S2A). Those belonging to CC23 (n = 176) were grouped in a major clade that could be divided into two subclades, one mainly associated with ST23 and one that grouped single-, double- and triple-locus variants of ST23. CC1755 (n = 1), which only shared three loci of ST23, showed a close genetic relationship with this last subclade. The remaining genomes were grouped in two minor clades, one including CC62 (n = 18) and CC372 (n = 2) and another including CC4 (n = 8). Serotype b genomes mainly belonged to CC6 (n = 150) and formed one of the main groups based on phylogenetic analysis, together with an isolate from a new CC. Genomes from CC50 (n = 13) and CC464 (n = 1) constituted a minor phylogenetic group (Supplementary Fig. S2B). Serotype c and d genomes showed

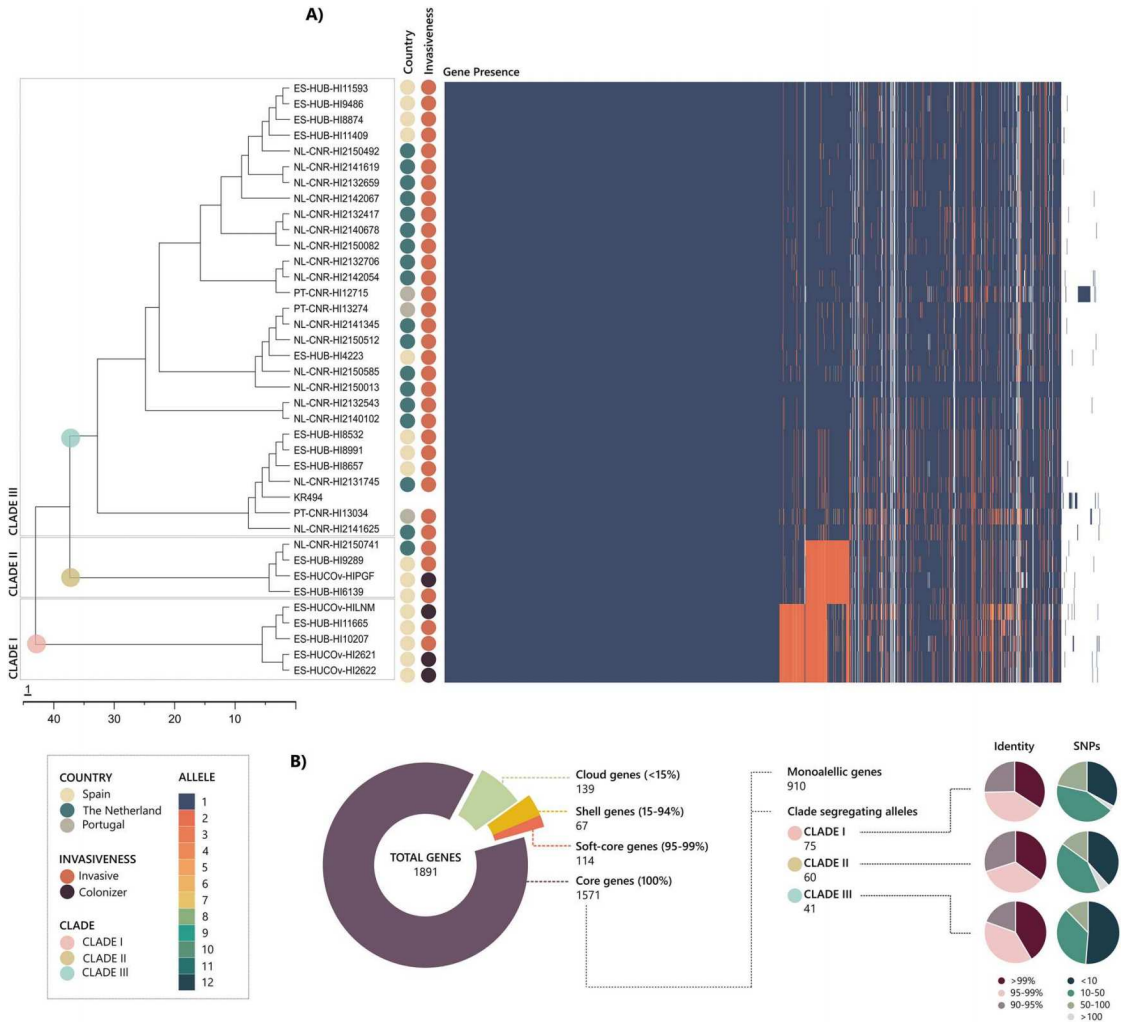


Figure 1. Pangenomic analysis of the 37 *H. influenzae* serotype f genomes. (**A**) Core-SNP phylogenetic tree, demographic data (country and invasiveness), genes detected, and the assigned allele. Clades I, II and III are indicated by coloured dots. The percentage of strains carrying each gene is presented graphically. (**B**) Distribution of genes detected in *H. influenzae* serotype f: core genes (100% of genomes), soft-core genes (95–99%), shell genes (15–94%), and cloud genes (< 15%). Core genes were classified as monoallelic (same allele in all the isolates) or clade segregating alleles (an allelic variant exclusive to one clade). The pie charts show the identity and number of SNPs for alleles of each clade in relation to the alleles of the same gene in other clades.

less genetic variability than other capsulated genomes, probably due to the low number of sequenced isolates. Serotype c genomes were linked to CC7, whereas serotype d genomes belonged to CC10 and were divided into two clades (Supplementary Fig. S3). Serotype e genomes were exclusively related to CC18, with ST18 being most abundant (n = 66) and were distributed in two different branches of the phylogenetic tree (Supplementary Fig. S4A). Serotype f genomes belonged to CC124 (n = 222) and CC16 (n = 12), which were distributed into major (clade I to III) and minor (clade IV) clusters, respectively. The 37 genomes from the Netherlands, Portugal, and Spain were distributed throughout the major cluster of the phylogenetic tree. The minor cluster showed a close phylogenetic relationship with two ST124 genomes (Supplementary Fig. S4B).

Pangenomic analysis of capsulated *H. influenzae*. The analysis of 800 pangenomes revealed that the gene pool of capsulated *H. influenzae* included 6360 genes (Fig. 2A). The proportion of genes present in all genomes (core genes) was very low across the capsulated *H. influenzae* population, accounting for only 5.1% (n = 322), whereas 11.2% (n = 715) were identified in 95–99% of the genomes (soft-core genes). The accessory genome included 5323 genes, distributed in shell genes (n = 1526; 24%), present in 15–94% of genomes, and cloud genes (n = 3797; 59.7%), present in < 15% of genomes.

A)

	Capsulated	Serotype A	Serotype B	Serotype C	Serotype D	Serotype E	Serotype F
N. GENOMES	800	205	165	34	10	152	234
CORE	322	1292	951	1598	1625	1430	1296
SOFT-CORE	715	178	552	23	0	197	346
SHELL	1526	321	435	191	209	342	122
CLOUD	3797	891	1579	123	62	697	1205
TOTAL	6360	2682	3517	1935	1896	2666	2969

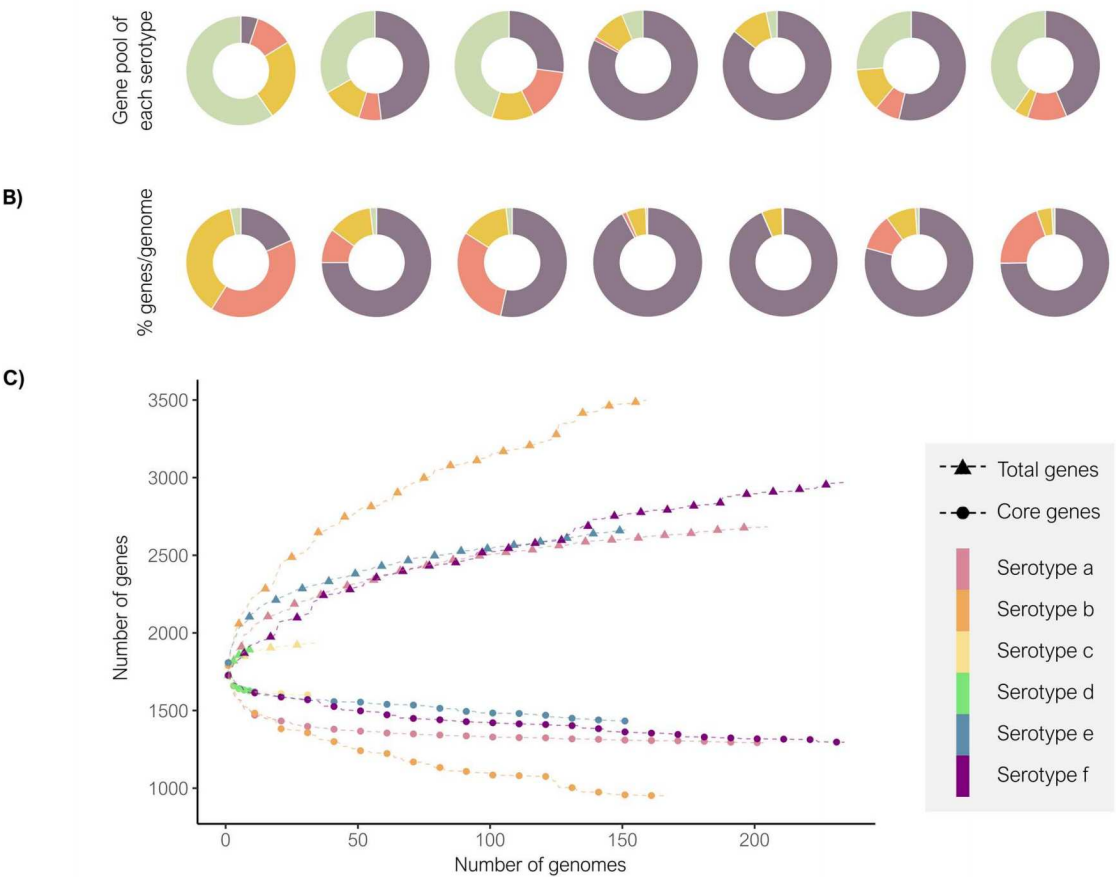


Figure 2. Pangenomic analysis of capsulated *H. influenzae*. (**A**) Gene pool of capsulated *H. influenzae* genomes included in this study. The number of core, soft-core, shell, cloud, and total genes of each serotype was determined using Roary, with a minimum identity percentage of 70% for BLASTp and the -cd parameter adjusted to 100. (**B**) Relative pangenome composition represented as a percentage of genes per genome of each serotype. Gene pool was defined as the set of all genes in a population. Donut charts indicate the distribution of core (100% of genomes), soft-core (95–99%), shell (15–94%), and cloud genes (< 15%). (**C**) Correlation between total and core genes in all capsulated *H. influenzae* genomes from this study and from the NCBI and ENA databases by serotype.

Differences in the gene pools were observed when serotypes were analysed separately (Fig. 2A). *H. influenzae* serotype b showed the largest gene pool (n = 3517) compared to the other serotypes, which oscillated between 1896 and 2969 genes. Moreover, serotype b had the lowest core genome proportion (42.7%) due to the large amount of cloud genes detected in the gene pool (44.9%). By contrast, serotypes c and d had a small gene pool

that was probably due to the low proportions of sequenced isolates (1935 and 1896 genes, respectively) and high proportions of core genes (83.7% and 85.7%, respectively). Finally, serotypes a, e, and f showed a better balance between the number of core and accessory genes present in the gene pool (54.8–45.2%, 61.0–39.0%, and 55.3–44.7%, respectively).

Despite the high number of genes detected within the gene pool of capsulated *H. influenzae* (n = 6360), each genome had an average of 1752 genes (SD = 51), ranging from 1725 (SD = 41) in serotype a genomes to 1805 (SD = 49) in serotype e genomes. Figure 2B depicts the pangenome composition for each serotype in an average genome. On average, serotype b had more accessory genes per genome than the other serotypes (mean = 46.6%, SD = 1.4), followed by serotype f (mean = 25.3%, SD = 1.1), serotype a (mean = 25.1%, SD = 1.8), and serotype e (mean = 20.7%, SD = 2.1). Serotype a and b isolates had subpopulations with fewer shell genes and more cloud genes per genome than the rest of the isolates for these serotypes. In serotype a, that subpopulation was associated with CC62 and CC372 (clade II), with an average of 43 shell genes (SD = 3) and 195 cloud genes (SD = 14); this contrasted with the 245 shell genes (SD = 43) and 15 cloud genes (SD = 28) observed in the other CCs. In serotype b, genomes related to CC50 and CC464 (clade II) had 134 shell genes (SD = 22) and 98 cloud genes (SD = 49), while CC6 isolates (clade I) had 264 shell genes (SD = 29) and 27 cloud genes (SD = 46).

Differences in gene pool composition between serotypes, especially between serotype b and serotypes c and d, could be due to a disparity in the number of genomes. The association between the number of genomes and the gene pool composition is shown in Supplementary Fig. S5. As more genomes were included, the number of total genes increased due to the introduction of more accessory genes. Core genes, however, fell considerably with the inclusion of the first genomes to reach a stable plateau. Although all serotypes showed the same tendency, it was notable that the variation between core and total genes was higher for serotype b than for the other serotypes, indicating that accessory gene acquisition was greater in serotype b (Fig. 2C).

A detailed analysis of the genetic composition revealed the presence and absence of genes associated with each serotype (Supplementary Table S4). The presence of 32 genes were associated with serotype f genomes but not with other serotypes. On the other hand, two genes, which code for the 30S ribosomal subunit protein S15 and a predicted nucleotide binding protein, were mostly absent in serotype f but present in other serotypes (Supplementary Table S4).

SNP typing of capsulated *H. influenzae*. The 800 capsulated *H. influenzae* genomes had 97,175 SNPs in the core genome, with an average of 26,626 SNPs (SD = 18,266) compared to the reference genome. Furthermore, serotype a (mean = 8499.5 SNPs, SD = 17,713.2), b (mean = 6048.2 SNPs, SD = 7472.1), and e (mean = 6849.4 SNPs, SD = 2895.3) genomes presented more SNPs than serotype f (mean = 2401.2, SD = 5206.9) genomes (p-value < 0.0001), whereas serotypes c (mean = 4037.1, SD = 4784.9) and d (mean = 6849.4 SNPs, SD = 2859.3) showed no significant differences despite having a greater number of SNPs than serotype f.

However, the genetic variability within each serotype and the reference genome used in each case should be considered (Fig. 3). Genomes of each serotype could be classified in CCs by the number of SNPs observed. For serotype a genomes, NML-Hia-1 (NZ_CP017811.1) of CC23 was used as reference. The isolates from CC23 showed an average of 1988 SNPs (SD = 1598) compared to the reference genome. CC4 genomes presented more genetic differences, with an average of 21,858 SNPs (SD = 1492), while CC62 and CC372 were the less related to the CC23 reference genome, with averages of 60,864 SNPs (SD = 1498) and 59,621 SNPs (SD = 394), respectively. In serotype b, using 10,810 (NC_016809.1) from CC6 as the reference strain, most genomes were grouped in two clusters, one related to CC6 that had 4418 SNPs (SD = 5045) and one linked to CC50 that had 21,122 SNPs (SD = 3364). Finally, strain KR494 (NC_022356.1) from CC124 was used as the reference for serotype f, showing 1238 SNPs (SD = 1450) for CC124 compared with the greater genetic distance of 23,906 SNPs for CC16 (SD = 552).

Discussion

The conjugate vaccine against serotype b has changed the global epidemiology of *H. influenzae*. NTHi is now the leading cause of invasive and non-invasive infection[16], while serotype b is decreasing and serotype f, the most common capsulated serotype in invasive infections, is increasing among adult populations[17,18]. The severity of invasive serotype f infection is particularly notable because it can affect immunocompetent patients and results in more than one-third of patients being admitted to intensive care units[17]. For these reasons, we provide a pangenome analysis of serotype f isolates associated with colonisation and invasive disease and perform a comparison with other capsular serotypes of *H. influenzae*.

The analysis of colonising and invasive serotype f isolates revealed minimal genetic diversity, with all being related to ST124 or a few single-locus variants of ST124. Consistent with prior reports, this was irrespective of the country, year, or source of isolation[19,20]. Bruun et al.[21], identified a long-term stable clone lineage in different countries for more than 50 years, and supporting their findings, our results suggest that this single clone (CC124) has persisted. Despite the low number of SNPs among the 37 isolates, which is consistent with the low number of SNPs found in serotype f genomes obtained from NCBI and ENA databases when compared to other serotypes, three distinct clades associated with CC124 could be identified. Clade III included most isolates and was exclusively associated with invasive strains and showed less variability than either clade I or II, which included the colonising strains. This suggests that the variability in serotype f genomes, despite being low, is mainly associated with colonising strains. According to a previous study, *Streptococcus pneumoniae*, which inhabits the same niche as *H. influenzae*, had high recombination rates that were linked to a longer colonisation status, favouring direct interactions with other bacteria of the same or different species[22]. For this reason, invasive clones may be more genetically homogeneous, while colonising strains may be more diverse due to higher genetic exchange with their environment.

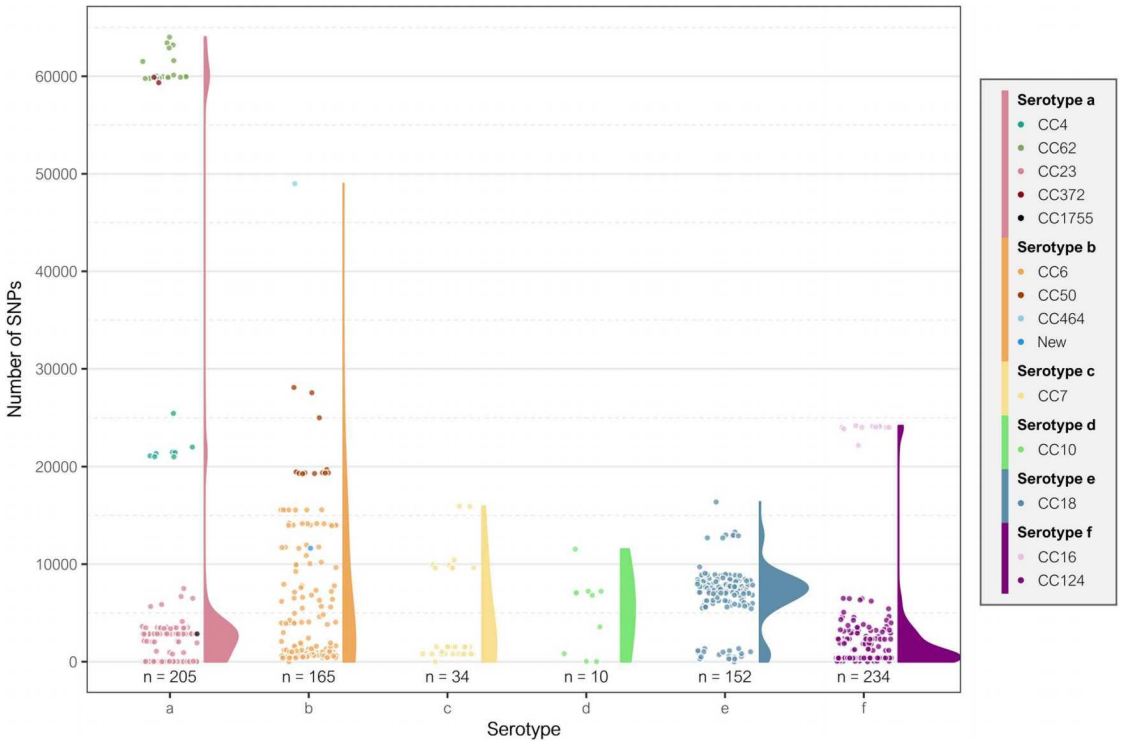


Figure 3. Core genome SNP typing of capsulated *H. influenzae* genomes. Each dot reflects the number of SNPs found in serotype a, b, c, d, e, and f genomes compared to the reference genomes NML-Hia-1 [CC23] (NZ_CP017811.1), 10810 [CC6] (NC_016809.1), M12125 [CC7] (SRR9847495), PTHi-10983 [CC10] (ERR2560729), M15895 [CC18] (NZ_CP031249.1), and KR494 [CC124] (NC_022356.1), respectively. Split violin plots show the distribution of the genomes based on the number of SNPs by each serotype.

The inclusion of all capsulated *H. influenzae* genomes available in the NCBI and ENA databases allowed comparison of the population structure for different serotypes and placed the genetic stability of serotype f in context. Moreover it provides an overview of the capsulated *H. influenzae* pangenome or supragenome, consisting of the entire set of genes available that is not contained by any particular isolate, but is available through a genetically diverse population. However, the lack of clinical data for the NCBI and ENA genomes precluded the identification of genetic differences between colonising and invasive strains. The pangenomic analysis of the capsulated genomes identified a pool of 6360 genes, with only 1037 being core or soft-core elements of genome. Pinto et al.[13] reported a smaller gene pool, which could be explained by the lower number of genomes in their study. Moreover, regardless of the serotype, all isolates carried roughly the same number of genes per genome (mean = 1752, SD = 51), again consistent with previous findings[20]. The maintenance of the overall number of genes per genome and the high fraction of accessory genes detected in the gene pool suggest that capsulated *H. influenzae* strains have a balance between gene acquisition and loss that could serve as a reservoir for DNA exchange. Hogg et al.[23], developed a supragenome model that predicted a species-level pangenome of 5000 or more genes, as well as a core-genome of about 1400 genes. However, models are estimations that require the analysis of many strains to be confirmed. Despite the fact that they used NTHi strains, the pangenomic richness observed in the capsulated strains of our study are consistent with their model. This diversity ensures the survival of the entire population in different environments, rather than the survival of an individual organism[24].

Despite the pangenomic differences in the overall capsulated population, each serotype included a phylogenetically highly clonal population related to a few STs. Similarly, previous studies demonstrated low genetic diversity within each serotype by multi-locus enzyme electrophoresis, pulsed-field gel electrophoresis, MLST, and WGS[25–28], suggesting that each serotype emerged once within the population[3]. However, despite the observed homogeneity of capsulated isolates, serotypes a and b displayed more variability in the MLST classification, with some STs differentiated by all seven loci, as Potts et al.[28] previously observed. According to the observed heterogeneity in serotypes a and b, the overall phylogenetic analysis revealed that these serotypes had distinct lineages compared to the monophyletic origins observed in the other serotypes. Serotype b also exhibited greater pangenomic diversity than the other serotypes, having the largest gene pool and the highest proportion of accessory genome. This diversity could be attributed to several advantages of serotype b isolates over other serotypes, including the production of haemocin, a bacteriocin active against non-type b serotypes[26], and the ability to evade the complement system[29]. These benefits would favour respiratory tract colonisation and genetic exchange

by serotype b strains, promoting genetic diversity. However, the introduction of the conjugate vaccine likely resulted in less colonisation by, and less invasive disease due to, these organisms[18].

Serotypes a, e, and f showed more diversity due to the accessory genome than serotypes c and d, but less than serotype b. This genetic diversity might promote bacterial survival in different ecological niches[24], potentially explaining the successful emergence of invasive disease caused by these serotypes since conjugate vaccine introduction[19,30]. Nevertheless, infections caused by serotypes a, e, and f are rare, suggesting that there is no strain replacement[16], probably due to their limitations in colonising the oropharynx. However, the genetic stability displayed by some clones, such as CC124 (serotype f) due to a low number of SNPs or CC23 (serotype a) due to a lower number of cloud genes, may be advantageous and may explain why these clones are more abundant and successful than other clones of these same serotypes[20,30,31]. Nevertheless, the addition of genomes from less abundant clonal complexes or changes in the used reference genomes could modify the overall number of SNPs and the genetic variability observed in each serotype, although the differences between clonal complexes of the same serotype would be conserved. Serotypes c and d, presented low variability and a more abundant core genome than the other serotypes. However, the pangenome composition of these serotypes is still unclear and further studies are needed to elucidate this question, because they are rarely isolated and the number of sequenced strains is low.

Distinguishing serotype f isolates from other serotypes, apart from capsular genes, could be useful in developing therapeutic strategies against this emerging serotype. This study provides a first approximation of the genetic determinants associated with each of the serotypes (Supplementary Table S4). However, the methodology used has certain limitations, as it is possible that truncated, duplicated, or those genes broken in different contigs would not be included. Thus, further studies are required to improve the identification of these genetic determinants.

In contrast to capsulated isolates, NTHi shows high genetic heterogeneity[32]. *H. influenzae* is a transformable bacterium for which homologous recombination enhances genetic diversity, affecting the commensal and the virulent behaviour of the microorganism[33]. Some studies have demonstrated that the level of recombination in NTHi was greater than in typeable isolates and that the polysaccharide capsule reduces the rate of gene transfer[27,34], probably due to its role as a physical barrier. In addition, colonisation is more commonly associated with NTHi than with capsulated isolates[35]. This would explain why NTHi clones have a high genetic heterogeneity while capsulated clones, which less frequently colonise the respiratory tract, have lower genetic heterogeneity.

Conclusion

Capsulated *H. influenzae* isolates present high genomic homogeneity with few lineages per serotype. The genetic stability of invasive serotype f genomes, regardless of time and country of isolation, highlights the importance of genetic homogeneity in the clonal expansion of this serotype.

Materials and methods

Study design and bacterial strains. Retrospective laboratory-based multicentre study on the genomic diversity of invasive *H. influenzae* serotype f isolates. Heterogeneity across countries was examined using serotype f isolates collected between 2009 and 2014 from National Reference Centres in the Netherlands (n = 18) and Portugal (n = 3) and from Bellvitge University Hospital in Spain (n = 12) (Supplementary Table S1). All were isolated from sterile sites, including blood, cerebrospinal fluid, joint fluid, and pleural fluid, and were serotyped according to the Centres for Disease Control and Prevention (CDC) guidelines (http://www.cdc.gov/meningitis/lab-manual/chpt10-pcr.html). Colonising *H. influenzae* serotype f strains (n = 4) from Spanish children attending a day care centre[15] were also included to establish the genetic differences between invasive and colonising isolates.

Whole genome sequencing of *H. influenzae* serotype f isolates. WGS was performed for 37 *H. influenzae* serotype f isolates. Strains were grown on chocolate agar plates (bioMérieux, Marcy l'Etoile, France) and incubated at 37 °C in 5% CO_2. Genomic DNA was extracted using the QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) and quantified with the QuantiFluor® dsDNA System (Promega, Wisconsin, USA). Libraries were prepared using Nextera XT and paired-end sequenced (2 × 150 base pairs) on a MiSeq Platform (Illumina Inc., San Diego, CA, USA), following the manufacturer's instructions.

Read quality assessment and genome assembly was done using the INNUca v4.2 pipeline (https://github.com/B-UMMI/INNUca). Briefly, a quality control of the reads was performed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc), followed by a read cleaning and trimming with Trimmomatic[36]. The genome was assembled using SPAdes[37] and was polished by Pilon[38]. In silico serotyping was conducted using hicap v1.0.3 (https://github.com/scwatts/hicap)[3]. The multi-locus sequence type (MLST) was determined in silico using the MLST v2.4 software (https://github.com/tseemann/mlst), and the new allele and sequence type (ST) numbers were registered in PubMLST (https://pubmlst.org). Genomes were classified into clonal complexes (CC) that included STs sharing at least five of the seven MLST alleles. The sequence reads were deposited at the ENA under the project accession number PRJEB45630 (Supplementary Table S1).

Core and accessory genome analysis of *H. influenzae* serotype f isolates. Core single nucleotide polymorphisms (SNPs) were extracted with Snippy's core module (snippy-core) for phylogenetic analysis, using the default parameters and the *H. influenzae* KR494 (NC_022356) genome as a reference. Subsequently, the whole genome alignment was subjected to the prediction and removal of recombinant regions using the Gubbins v2.3.1 software[39]. A novel core-SNP phylogenetic tree was built in RAxML-NG[40] based only on shared positions after recombination removal.

To characterise the genetic composition of the identified clades, the assembled genomes were annotated using Prokka v1.13.7[41] and pangenome analysis was done using Roary[42] with a minimum identity percentage of 70% for BLASTp, as previously described for this species[13], and the -cd parameter adjusted to 100. Allelic profiles were determined using roProfile (https://github.com/cimendes/roProfile) where alleles with size variation > 20% were discarded by default.

Analysis of population genetics for capsulated *H. influenzae* genomes. To better understand the phylogenetic diversity in capsulated *H. influenzae*, all capsulated genomes available in the NCBI and ENA databases were downloaded and selected (see Supplementary Fig. S6). Pre-selection of ENA genomes was performed by mapping reads against the *bexA* gene using Bowtie2[43]. A total of 763 genomes were identified as capsulated *H. influenzae* after in silico serotyping and MLST classification.

SNPs were studied with Snippy, using default parameters, and were visualised using the ggplot2 R package[44]. Phylogenetic analysis was performed using Snippy's core module and Gubbins v2.3.1 software, as described above. Pangenome analysis was done with Roary, and a gene pool was defined as the set of all genes detected in a population. A statistical analysis was performed based on the Roary results to determine the presence and absence of genes associated with serotype f. Thus, Scoary (https://github.com/AdmiralenOla/Scoary) was used for the analysis, and genes with specificity and sensitivity > 97.5% and < 2.5% were chosen to select the presence and absence of genes, respectively.

The genomes of strains NML-Hia-1 (NZ_CP017811.1, CC23)[45], 10810 (NC_016809.1, CC6)[20], M12125 (SRR9847495, CC7), PTHi-10983 (ERR2560729, CC10), M15895 (NZ_CP031249.1, CC18), and KR494 (NC_022356.1, CC124)[46], which belonged to the clinically most prevalent clonal complexes of each serotype, were used as references for serotypes a to f, respectively. In the overall analysis of capsulated *H. influenzae*, the genome of KR494 (NC_022356.1) was used as a reference.

Statistical analysis. Statistical analyses were performed in GraphPad Prism 5, using unpaired *t* test or one-way ANOVA (Newman–Keuls test), as appropriate. P-values < 0.05 were considered statistically significant.

Ethical approval. This study was in accordance with the Declaration of Helsinki from the World Medical Association. Written informed consent was not required as this was a retrospective and observational study with isolates obtained as part of routine microbiological tests, which was approved by the Clinical Research Ethics Committee of Bellvitge University Hospital (PR334/21). Patient confidentiality was always protected, and all personal data were anonymised following the current legal normative in Spain (LOPD 15/1999 and RD 1720/2007). Moreover, this project followed Law 14/2007 on Biomedical Research for the management of biological samples in clinical research.

Repositories. Sequence reads were deposited in the European Nucleotide Archive (ENA) under the project accession number PRJEB45630.

Received: 18 November 2021; Accepted: 11 February 2022
Published online: 24 February 2022

References

1. Nørskov-Lauritsen, N. Classification, identification, and clinical significance of *Haemophilus* and *Aggregatibacter* species with host specificity for humans. *Clin. Microbiol. Rev.* **27**, 214–240 (2014).
2. Slack, M. P. E. A review of the role of *Haemophilus influenzae* in community-acquired pneumonia. *Pneumonia* **6**, 26–43 (2015).
3. Watts, S. C. & Holt, K. E. hicap: In silico serotyping of the *Haemophilus influenzae* capsule locus. *J. Clin. Microbiol.* **57**, e00190-e219 (2019).
4. Le, P., Nghiem, V. T. & Swint, J. M. Post-GAVI sustainability of the *Haemophilus influenzae* type b vaccine program: The potential role of economic evaluation. *Hum. Vaccines Immunother.* **12**, 2403–2405 (2016).
5. Cerquetti, M. & Giufrè, M. Why we need a vaccine for non-typeable *Haemophilus influenzae*. *Hum. Vaccines Immunother.* **12**, 2357–2361 (2016).
6. Heinz, E. The return of Pfeiffer's bacillus: Rising incidence of ampicillin resistance in *Haemophilus influenzae*. *Microb. Genomics* **4**, 1–8 (2018).
7. Puig, C. *et al.* Clinical and molecular epidemiology of *Haemophilus influenzae* causing invasive disease in adult patients. *PLoS One* **9**, e112711 (2014).
8. Carrera-Salinas, A. *et al.* Epidemiology and population structure of *Haemophilus influenzae* causing invasive disease. *Microb. Genomics* **7**, 1–13 (2021).
9. Whittaker, R. *et al.* Epidemiology of invasive *Haemophilus influenzae* disease, Europe, 2007–2014. *Emerg. Infect. Dis.* **23**, 396–404 (2017).
10. Soeters, H. M. *et al.* Current epidemiology and trends in invasive *Haemophilus influenzae* disease—United States, 2009–2015. *Clin. Infect. Dis.* **67**, 881–889 (2018).
11. Bogaert, D. *et al.* Variability and diversity of nasopharyngeal microbiota in children: A metagenomic analysis. *PLoS One* **6**, e17035 (2011).
12. Georgiades, K. & Raoult, D. Defining pathogenic bacterial species in the genomic era. *Front. Microbiol.* **1**, 1–13 (2011).
13. Pinto, M. *et al.* Insights into the population structure and pan-genome of *Haemophilus influenzae*. *Infect. Genet. Evol.* **67**, 126–135 (2019).
14. De Chiara, M. *et al.* Genome sequencing of disease and carriage isolates of nontypeable *Haemophilus influenzae* identifies discrete population structure. *Proc. Natl. Acad. Sci.* **111**, 5439–5444 (2014).

15. Puig, C. *et al.* Oropharyngeal colonization by nontypeable *Haemophilus influenzae* among healthy children attending day care centers. *Microb. Drug Resist.* **20**, 450–455 (2014).
16. Agrawal, A. & Murphy, T. F. *Haemophilus influenzae* infections in the *H. influenzae* type b conjugate vaccine era. *J. Clin. Microbiol.* **49**, 3728–3732 (2011).
17. Resman, F. *et al.* Invasive disease caused by *Haemophilus influenzae* in Sweden 1997–2009; evidence of increasing incidence and clinical burden of non-type b strains. *Clin. Microbiol. Infect.* **17**, 1638–1645 (2011).
18. European Centre for Disease Prevention and Control (ECDC). *Haemophilus influenzae annual epidemiological report for 2018* (2020).
19. Ladhani, S. N. *et al.* Invasive *Haemophilus influenzae* serotype e and f disease, England and Wales. *Emerg. Infect. Dis.* **18**, 725–732 (2012).
20. Su, Y.-C., Resman, F., Hörhold, F. & Riesbeck, K. Comparative genomic analysis reveals distinct genotypic features of the emerging pathogen *Haemophilus influenzae* type f. *BMC Genomics* **15**, 38 (2014).
21. Bruun, B., Gahrn-Hansen, B., Westh, H. & Kilian, M. Clonal relationship of recent invasive *Haemophilus influenzae* serotype f isolates from Denmark and the United States. *J. Med. Microbiol.* **53**, 1161–1165 (2004).
22. Chaguza, C. *et al.* Recombination in *Streptococcus pneumoniae* lineages increase with carriage duration and size of the polysaccharide capsule. *MBio* **7**, e01053-e1116 (2016).
23. Hogg, J. S. *et al.* Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol.* **8**, 1–18 (2007).
24. Gilsdorf, J. R., Marrs, C. F. & Foxman, B. *Haemophilus influenzae*: Genetic variability and natural selection to identify virulence factors. *Infect. Immun.* **72**, 2457–2461 (2004).
25. Musser, J. M., Kroll, J. S., Moxonf, E. R. & Selandert, R. K. Evolutionary genetics of the encapsulated strains of *Haemophilus influenzae*. *Proc. Natl. Acad. Sci. U.S.A.* **85**, 7758–7762 (1988).
26. Omikunle, A. *et al.* Limited genetic diversity of recent invasive isolates of non-serotype b encapsulated *Haemophilus influenzae*. *J. Clin. Microbiol.* **40**, 1264–1270 (2002).
27. Meats, E. *et al.* Characterization of encapsulated and noncapsulated *Haemophilus influenzae* and determination of phylogenetic relationships by multilocus sequence typing. *J. Clin. Microbiol.* **41**, 1623–1636 (2003).
28. Potts, C. C. *et al.* Genomic characterization of *Haemophilus influenzae*: A focus on the capsule locus. *BMC Genomics* **20**, 1–9 (2019).
29. Hallström, T. & Riesbeck, K. *Haemophilus influenzae* and the complement system. *Trends Microbiol.* **18**, 258–265 (2010).
30. Ulanova, M. & Tsang, R. S. W. *Haemophilus influenzae* serotype a as a cause of serious invasive infections. *Lancet Infect. Dis.* **14**, 70–82 (2014).
31. Tsang, R. S. W. & Ulanova, M. The changing epidemiology of invasive *Haemophilus influenzae* disease: Emergence and global presence of serotype a strains that may require a new vaccine for control. *Vaccine* **35**, 4270–4275 (2017).
32. Staples, M., Graham, R. M. A. & Jennison, A. V. Characterisation of invasive clinical *Haemophilus influenzae* isolates in Queensland, Australia using whole-genome sequencing. *Epidemiol. Infect.* **145**, 1727–1736 (2017).
33. Power, P. M., Bentley, S. D., Parkhill, J., Moxon, E. R. & Hood, D. W. Investigations into genome diversity of *Haemophilus influenzae* using whole genome sequencing of clinical isolates and laboratory transformants. *BMC Microbiol.* **12**, 1–12 (2012).
34. Connor, T. R., Corander, J. & Hanage, W. P. Population subdivision and the detection of recombination in non-typable *Haemophilus influenzae*. *Microbiology* **158**, 2954–2964 (2012).
35. Deghmane, A.-E. *et al.* High diversity of invasive *Haemophilus influenzae* isolates in France and the emergence of resistance to third generation cephalosporins by alteration of *ftsI* gene. *J. Infect.* **79**, 7–14 (2019).
36. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
37. Bankevich, A. *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
38. Walker, B. J. *et al.* Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
39. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).
40. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).
41. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
42. Page, A. J. *et al.* Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
43. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
44. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).
45. Iskander, M., Hayden, K., Van Domselaar, G. & Tsang, R. First complete genome sequence of *Haemophilus influenzae* serotype a. *Genome Announc.* **5**, e01506-e1516 (2017).
46. Su, Y.-C., Hörhold, F., Singh, B. & Riesbeck, K. Complete genome sequence of encapsulated *Haemophilus influenzae* type f KR494, an invasive isolate that caused necrotizing myositis. *Genome Announc.* **1**, e00470-e513 (2013).

Acknowledgements

We would like to thank the staff of the Microbiology Laboratory of Bellvitge University Hospital who contributed daily to this project.

Author contributions

A.V.D.E., J.D.L., M.A.D., C.A., P.B.L. and S.M. supervised the study. A.G., M.P. and M.C. performed laboratory assays. A.G., A.C., M.P. and S.M. analysed and interpreted the data. A.C. and A.G. wrote the manuscript with the supervision of S.M. All authors read and approved the final manuscript.

Funding

This study was funded by Instituto de Salud Carlos III (ISCIII) through the Projects from the Fondo de Investigaciones Sanitarias "PI16/00977" to SM, and CIBER de Enfermedades Respiratorias (CIBERES–CB06/06/0037), co-funded by the European Regional Development Fund/European Social Fund (ERDF/ESF, "Investing in your future"), and CERCA Programme/Generalitat de Catalunya for institutional support. Bioinformatic analysis was supported by an Amazon Web Services (AWS) research grant to SM. AC was supported by FPU grant "FPU16/02202" (Formación de Profesorado Universitario, Ministerio de Educación, Spain), and SM was supported by Miguel Servet contract "CP19/00096" (ISCIII).

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-07185-5.

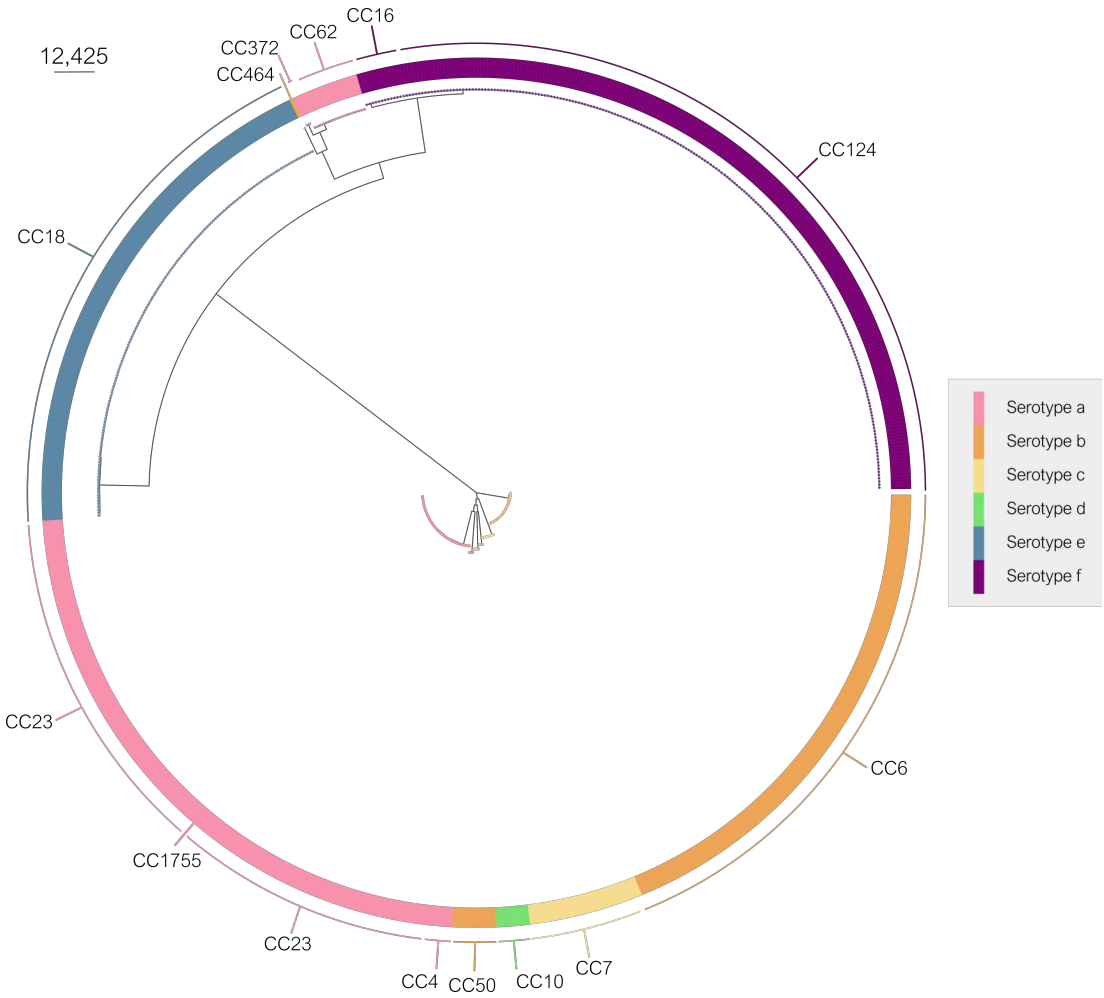
Correspondence and requests for materials should be addressed to S.M.

Reprints and permissions information is available at www.nature.com/reprints.

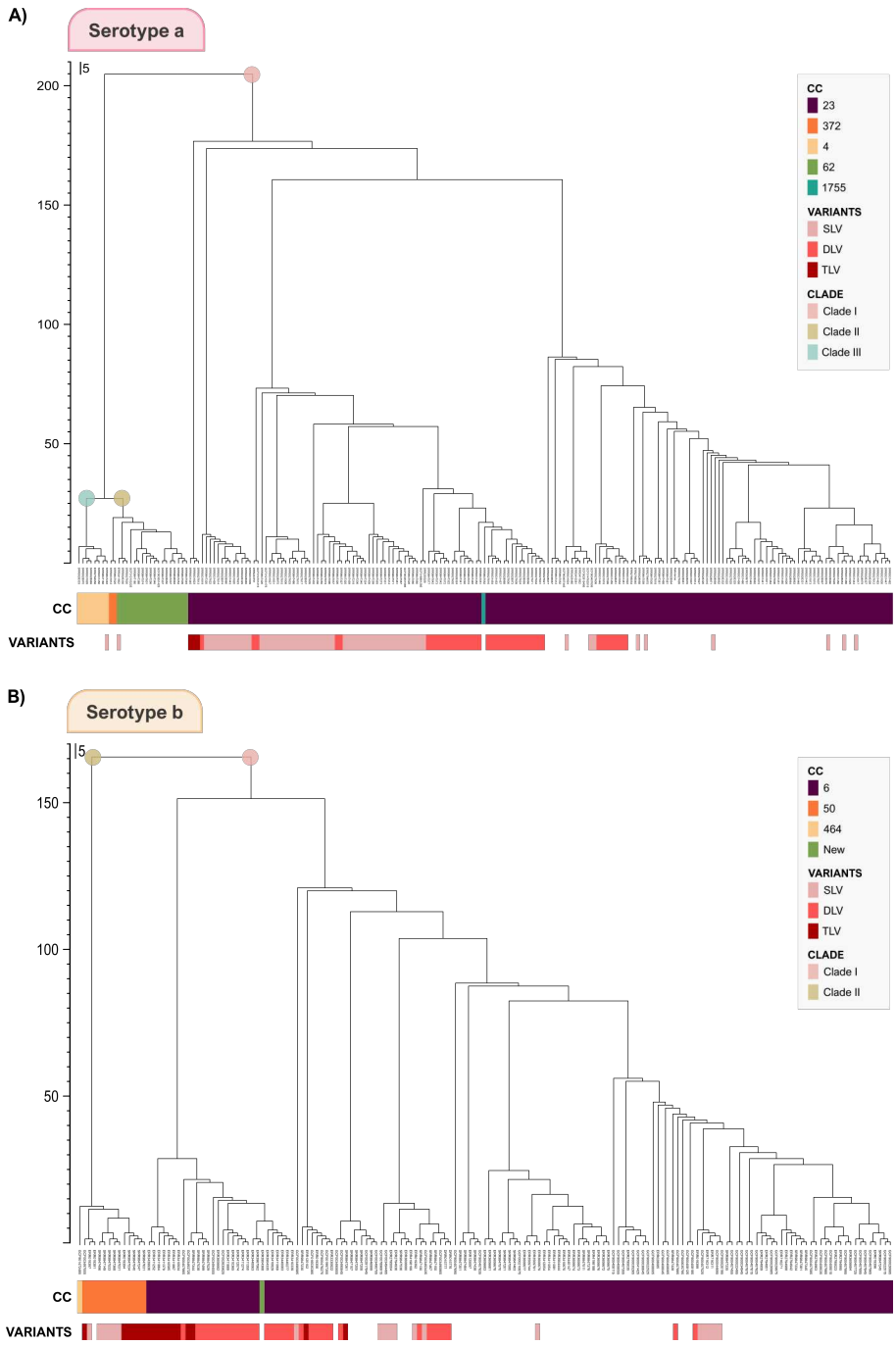
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

SUPPLEMENTARY MATERIAL

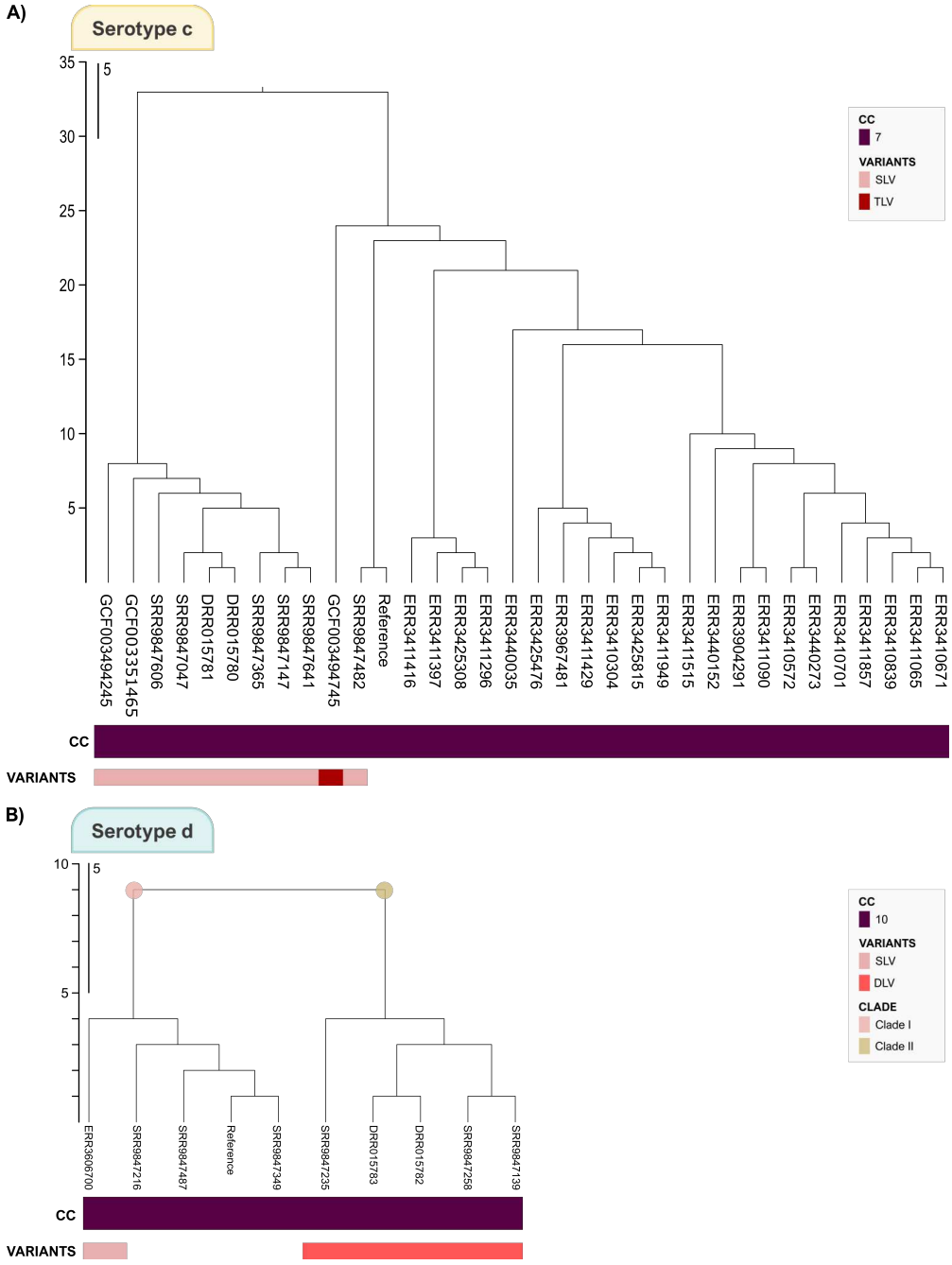
Supplementary Table S1, S2, S3, and S4 are available with the online version of this article.



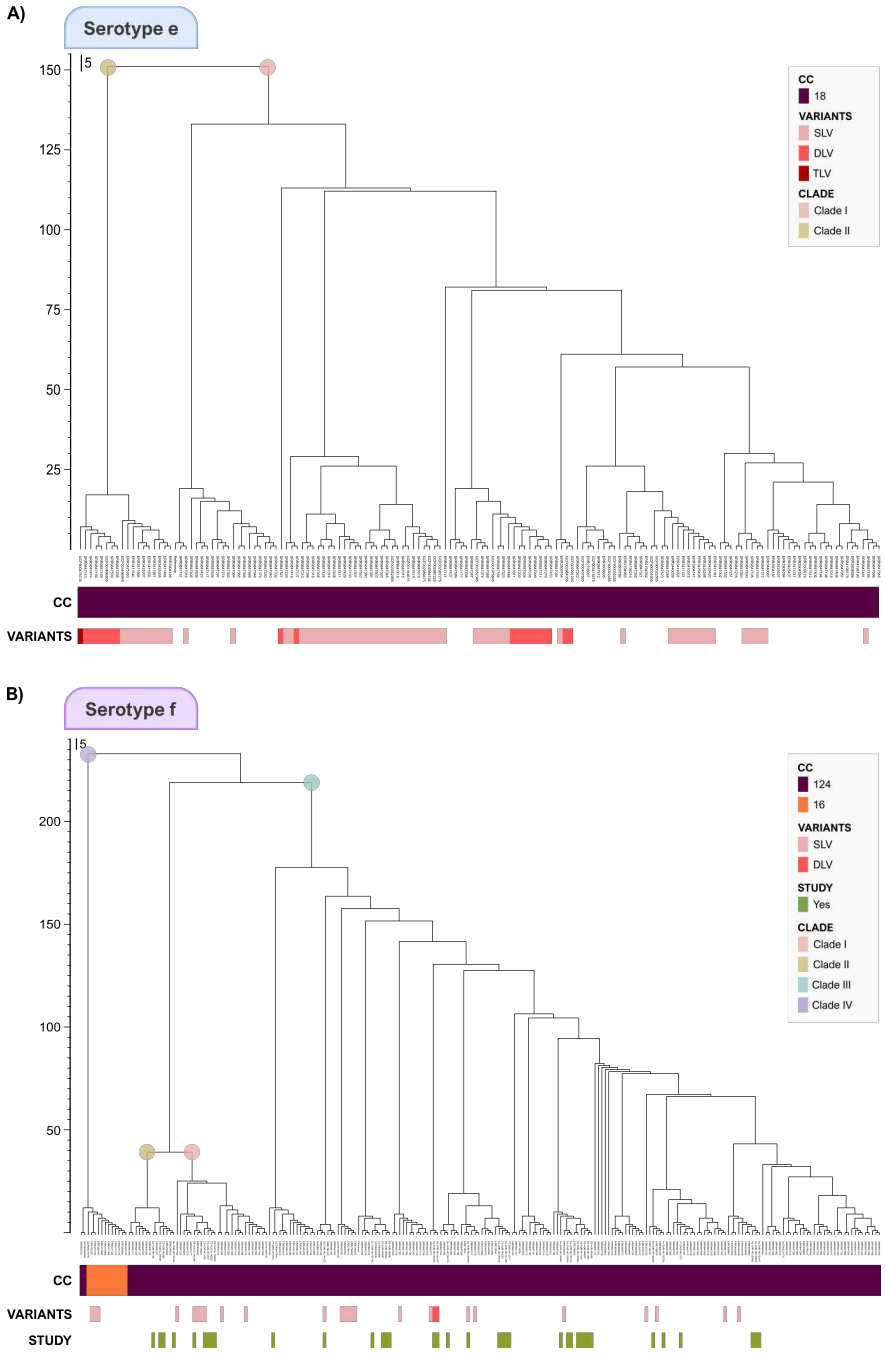
Supplementary Figure S1. Phylogenetic core-SNP tree of capsulated *H. influenzae* isolates.



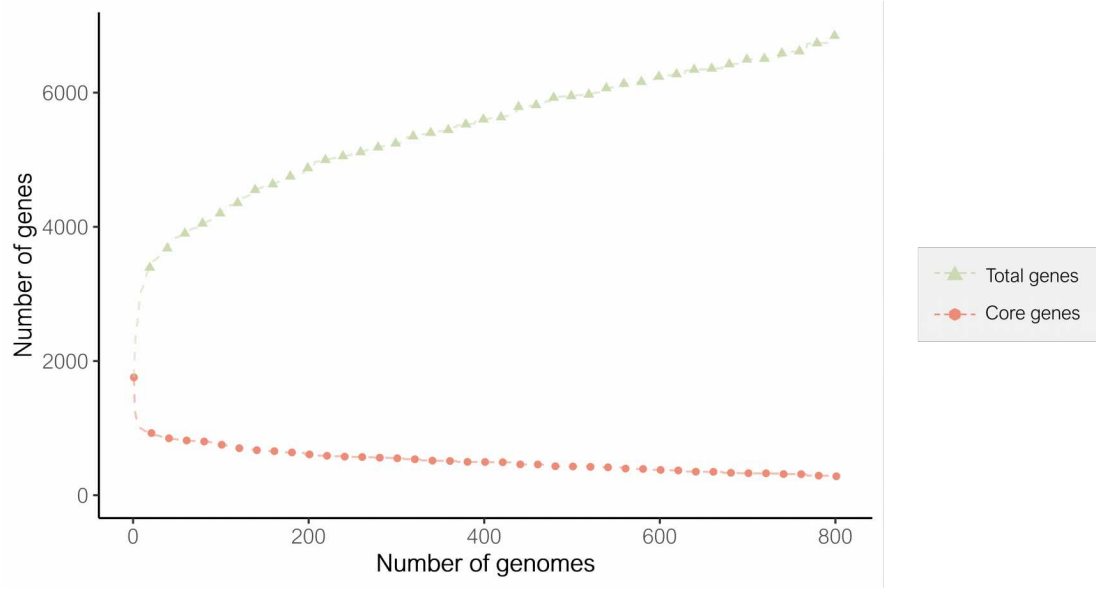
Supplementary Figure S2. Phylogenetic core-SNP tree of *H. influenzae* serotype a (A) and b (B). Genomes were classified into CCs, defined as sequence types sharing at least five of seven MLST alleles. STs differing at one, two, or three of the seven loci were defined as SLVs, DLVs and TLVs, respectively. Circles in the phylogenetic trees group genomes into clades. Abbreviations: CC, clonal complexes; MLST, multi-locus sequence type; SNP, single nucleotide polymorphism; ST, sequence type; SLV single-locus variant; DLV, double-locus variant; TLV, triple-locus variant.



Supplementary Figure S3. Phylogenetic core-SNP tree of *H. influenzae* serotype c (A) and d (B). Genomes were classified into CCs, defined as sequence types sharing at least five of seven MLST alleles. STs differing at one, two, or three of the seven loci were defined as SLVs, DLVs, and TLVs, respectively. Circles in the phylogenetic trees group genomes into clades. Abbreviations: CC, clonal complexes; MLST, multi-locus sequence type; SNP, single nucleotide polymorphism; ST, sequence type; SLV single-locus variant; DLV, double-locus variant; TLV, triple-locus variant.



Supplementary Figure S4. Phylogenetic core-SNP tree of *H. influenzae* serotype e (A) and f (B). Genomes were classified into CCs, defined as sequence types sharing at least five of seven MLST alleles. STs differing at one, two, or three of the seven loci were defined as SLVs, DLVs, and TLVs, respectively. Circles in the phylogenetic trees group genomes into clades. Abbreviations: CC, clonal complexes; MLST, multi-locus sequence type; SNP, single nucleotide polymorphism; ST, sequence type; SLV single-locus variant; DLV, double-locus variant; TLV, triple-locus variant.



Supplementary Figure S5. Correlation between total and core genes in all capsulated *H. influenzae* genomes from this study and from the NCBI and ENA databases.

Supplementary Figure S6. Flowchart of the study pipeline to select the genomes of capsulated *H. influenzae* available on the NCBI and ENA databases. Abbreviations: ENA, European Nucleotide Archive; MLST, multi-locus sequence type; NCBI, National Center for Biotechnology Information.

Study 3

Genetic adaptation and acquisition of macrolide resistance in *Haemophilus* spp during persistent respiratory tract colonization in chronic obstructive pulmonary disease (COPD) patients receiving long-term azithromycin treatment

Anna Carrera-Salinas, Aida González-Díaz, Rachel L. Ehrlich, Dàmaris Berbel, Fe Tubau, Xavier Pomares, Junkal Garmendia, M Ángeles Domínguez, Carmen Ardanuy, Daniel Huertas, Alicia Marín, Conchita Montón, Joshua Chang Mell, Salud Santos, Sara Martí

Dr. Mª Ángeles Domínguez Luzón, Associate Professor of the Department of Pathology and Experimental Therapeutics of the Universitat de Barcelona and Head of the Microbiology Department, Hospital Universitari de Bellvitge (Barcelona), and Dr. Sara Martí Martí, Assistant Professor of the Department of Medicine of the Universitat de Barcelona and Senior Postdoctoral Researcher at the Centre for Biomedical Research Network on Respiratory Diseases (CIBERes),

DECLARE that the original article entitled "**Genetic adaptation and acquisition of macrolide resistance in *Haemophilus* spp during persistent respiratory tract colonization in chronic obstructive pulmonary disease (COPD) patients receiving long-term azithromycin treatment**" has been submitted in the Microbiology Spectrum (Impact factor in 2021: 9.043). The contribution of Anna Carrera-Salinas is detailed below:

- Experimental work.
- Data analysis.
- Manuscript redaction.

Signatures of thesis supervisors

MªÁngeles Domínguez Luzón **Sara Martí Martí**

Genetic adaptation and acquisition of macrolide resistance in *Haemophilus* spp during persistent respiratory tract colonization in chronic obstructive pulmonary disease (COPD) patients receiving long-term azithromycin treatment

Anna Carrera-Salinas[1], Aida González-Díaz[1,2], Rachel L. Ehrlich[3], Dàmaris Berbel[1,2], Fe Tubau[1,2], Xavier Pomares[4], Junkal Garmendia[2,5], M Ángeles Domínguez[1,6,7], Carmen Ardanuy[1,2,7], Daniel Huertas[8], Alicia Marín[2,9], Conchita Montón[4], Joshua Chang Mell[3], Salud Santos[2,10,11#], Sara Marti[1,2,11#]

[1]Microbiology Department, Hospital Universitari de Bellvitge, IDIBELL-UB, Barcelona, Spain.

[2]Research Network for Respiratory Diseases (CIBERES), ISCIII, Madrid, Spain.

[3]Department of Microbiology and Immunology, Center for Genomic Sciences, Drexel University College of Medicine, Philadelphia, USA.

[4]Department of Respiratory Medicine, Hospital de Sabadell, Hospital Universitari Parc Taulí, Institut d'Investigació i Innovació Parc Taulí I3PT, Universitat Autònoma de Barcelona, Sabadell, Spain.

[5]Instituto de Agrobiotecnología, CSIC-Gobierno de Navarra, Mutilva, Spain.

[6]Research Network for Infectious Diseases (CIBERINFEC), ISCIII, Madrid, Spain.

[7]Department of Pathology and Experimental Therapeutics, School of Medicine, University of Barcelona, Barcelona, Spain.

[8]Department of Respiratory Medicine, Hospital Residència Sant Camil, Consorci Sanitari Alt Penedès-Garraf, Barcelona, Spain.

[9]Department of Respiratory Medicine, Hospital Universitari Germans Trias i Pujol, Barcelona, Spain.

[10]Department of Respiratory Medicine, Hospital Universitari de Bellvitge, IDIBELL-UB, Barcelona, Spain.

[11]Department of Medicine, School of Medicine, University of Barcelona, Barcelona, Spain.

Short running title: Evolution of *Haemophilus* spp in COPD patients with long-term azithromycin treatment

Keywords: *Haemophilus influenzae*, *Haemophilus parainfluenzae*, persistence, macrolide resistance, azithromycin, adaptation

#Corresponding authors: Sara Marti, smartinm@bellvitgehospital.cat; and Salud Santos, ssantosp@bellvitgehospital.cat.

ABSTRACT

Patients with chronic obstructive pulmonary disease (COPD) benefit from the immunomodulatory effect of azithromycin, but long-term administration may alter colonizing bacteria. Our goal was to identify changes in *Haemophilus influenzae* and *Haemophilus parainfluenzae* colonizing strains during azithromycin treatment.

H. influenzae and *H. parainfluenzae* isolates from COPD patients undergoing prolonged azithromycin treatment (Hospital Universitari de Bellvitge, Spain) were included to identify antimicrobial resistance changes and genetic variation accumulated during persistent respiratory colonization.

Four patients (P02, P08, P11, and P13) were persistently colonized by *H. influenzae*, and two (P04 and P11) by *H. parainfluenzae*.

All persistent lineages isolated before treatment were azithromycin susceptible, but developed resistance within the first months, apart from those belonging to P02, who discontinued the treatment. *H. influenzae* isolates from P08-ST107 acquired mutations in 23S rRNA, and those from P11-ST2480 and P13-ST165 had changes in the ribosomal proteins L4 and L22. In *H. parainfluenzae*, P04 persistent isolates acquired changes in *rlmC*; and P11 carried genes encoding MefE/MsrD efflux pumps in an integrative conjugative element, which was also identified in *H. influenzae* P11-ST147.

Other genetic variation mostly occurred in genes associated with cell wall and inorganic ion metabolism. Persistent *H. influenzae* strains all showed changes in *licA* and *hgpB* genes. Other genes (*lex1*, *lic3A*, *hgpC*, and *fadL)* had variation in multiple lineages. Furthermore, persistent strains showed loss, acquisition or genetic changes in prophage-associated regions.

Long-term azithromycin therapy results in macrolide resistance, as well as genetic changes that likely favor bacterial adaptation during persistent respiratory colonization.

INTRODUCTION

Chronic obstructive pulmonary disease (COPD) is a respiratory disorder characterized by airflow obstruction and inflammation, which results in chronically decreased lung function and respiratory failure. Although tobacco smoking is the main risk factor, other environmental and genetic factors increase the occurrence of COPD[1,2]. Acute exacerbations, characterized by increased airway inflammation and lower airway bacterial infection, represent an additional burden, requiring hospitalization, worsening comorbidities, and increasing mortality rates[3,4].

The prophylactic use of azithromycin in COPD has been demonstrated to reduce the frequency and severity of exacerbations and improve the quality of life due to its immunomodulatory and anti-inflammatory properties[3-5]. However, azithromycin is also a potent antibiotic that inhibits bacterial protein synthesis and interferes with the assembly of the 50S large ribosomal subunit[6]. Its antimicrobial properties may affect the respiratory microbiota, including selecting for bacterial resistance to macrolides, which can arise due to mutations in several genes (the 23S rRNA gene near the encoded macrolide-binding region (A2058); several genes encoding rRNA methylases; and genes encoding the L4 and L22 ribosomal proteins), or by acquisition of efflux pumps encoded by the *mef*(A) and *msr*(D) genes[7-9].

Bacterial colonization is an important stimulus for inflammation and plays an important role in modulating exacerbations. *Haemophilus influenzae* is an opportunistic pathogen that normally colonizes the nasopharynx but is also the main etiologic agent in COPD exacerbations, particularly non-capsulated nontypeable strains (NTHi)[10]. On the other hand, *H. parainfluenzae* is a frequent colonizer of the upper respiratory tract, although its role in COPD infection is unclear[11,12]. Different mechanisms in the bacteria, such as biofilm formation and intracellular survival, contribute to persistent colonization through evasion of the immune system and the action of antibiotics[13]. In-host bacterial evolution

during persistent COPD infections can occur through point mutations and DNA rearrangements, as well as through the acquisition or loss of mobile genetic elements, such as plasmids, bacteriophages, and pathogenicity islands that contain virulence factors[13,14]. Furthermore, *Haemophilus* spp. are naturally competent, and highly abundant DNA uptake signal sequences in their genomes (USSs: 5'-AAGTGCGGT-'3) promote exogenous DNA uptake and recombination among different strains[15].

Understanding the adaptive evolution of chronic bacterial pathogens to long-term drug selective pressure may improve our understanding of their role in COPD progression. Thus, the goals of this study were to characterize the *H. influenzae* and *H. parainfluenzae* population colonizing the respiratory tract in COPD patients that were treated with prolonged prophylactic azithromycin treatment to identify changes in azithromycin susceptibility and to detect genetic changes arising during persistent colonization.

MATERIALS AND METHODS

Study design and bacterial strains

Laboratory-based study including clinical respiratory isolates from a cohort of patients with severe COPD and frequent exacerbations (≥ 4 acute exacerbations, or ≥ 3 acute exacerbations with at least one hospitalization in the year prior to inclusion) who had initiated long-term oral azithromycin treatment (250 mg, 3 days/week) at the Hospital Universitari de Bellvitge (Spain). Patients were followed for at least one year, with sputum samples collected just prior to the start of azithromycin therapy (V1), three months after starting therapy (V2), twelve months after starting therapy (V3), and during any acute exacerbation episodes. Other respiratory bacterial isolates recovered during routine clinical microbiology laboratory work prior to initiating treatment were also included.

All the sputum samples were cultured on blood, chocolate, and MacConkey agar, and individual colonies were isolated and identified by mass spectrometry using MALDI-TOF

(MALDI Biotyper, Bruker). To account for potential polyclonal infection, at least 10 independent colonies identified as *H. influenzae* or *H. parainfluenzae* were chosen from each plated sample to study their clonal relatedness using pulse-field gel electrophoresis (PFGE), as previously described[32,43]. One isolate from each PFGE pattern found for each patient episode was selected for further analysis.

Antimicrobial susceptibility testing

Antimicrobial susceptibility was assessed by disk diffusion and microdilution using commercial panels (STRHAE2, Sensititre), following the recommended clinical breakpoints of the European Committee on Antimicrobial Susceptibility Testing (EUCAST) and the Clinical and Laboratory Standards Institute (CLSI). Manual microdilution was performed on strains with a minimum inhibitory concentration (MIC) for azithromycin greater than 4 mg/L, testing concentrations ranging from 1 to 512 mg/L.

Whole genome sequencing (WGS) and assembly

H. influenzae and *H. parainfluenzae* isolates were subjected to WGS, using both short- and long-read sequencing. Genomic DNA was extracted using the QIAamp DNA Mini Kit (Qiagen) and quantified by the QuantiFluor® dsDNA System (Promega). For short-read WGS, Nextera XT was used to prepare the libraries, followed by paired-end sequencing on a MiSeq platform (Illumina Inc.). For long-read WGS, Native Barcoding Expansion (EXP-NBD196) and the Ligation Sequencing Kit (SQK-LSK109) were used to prepare the libraries, followed by sequencing on FLO-MIN106D flow cells (R9.4.1) (Oxford Nanopore Technologies). A hybrid assembly of short and long reads was produced using Unicycler pipeline (github.com/rrwick/Unicycler)[44]. The assembly statistics were obtained using assembly-stats (github.com/sanger-pathogens/assembly-stats).

Phylogenetic analysis and antimicrobial resistance determinants

Phylogenetic analysis was conducted using Snippy (github.com/tseemann/snippy) and RAxML-NG[45], as previously described[46]. *H. influenzae* Hi375 (Genbank accession CP009610) and *H. parainfluenzae* T3T1 (Genbank accession NC_015964) genomes were used as references.

Multi-locus sequence typing (MLST) was determined in *H. influenzae* genomes using the MLST v2.4 software (github.com/tseemann/mlst). *In silico* serotyping was conducted using hicap (github.com/scwatts/hicap)[47], and NTHi were classified into clades based on the presence or absence of 17 accessory genes using patho_typing (github.com/B-UMMI/patho_typing)[48,49].

Acquired antimicrobial resistance determinants were screened using ResFinder[50]. The screening for mutations in genes involved in antibiotic resistance was performed using Geneious R9 (Biomatters), with the *H. influenzae* Hi375 (CP009610) and *H. parainfluenzae* T3T1 (NC_015964) genomes used as references. Short reads were mapped against 23S rRNA, and the percentage of altered sites was used to estimate the number of 23S rRNA copies with changes.

Bacterial persistence and genetic adaptation

Persistent lineages were defined as groups of isolates from the same patient collected over a span of more than three months. For *H. influenzae*, isolates were deemed part of the same lineage, if they had the same sequence type (ST) by *in silico* MLST or differed at only one of the MLST loci as single-locus variants (SLVs). Since they lack a specific MLST schema, *H. parainfluenzae* isolates were grouped into lineages if they differed by three or fewer PFGE bands.

Genome assemblies of persistent strains were annotated using Prokka v1.13.7[51] with orthologs initially checked in *H. influenzae* Hi375 (CP009610) or *H. parainfluenzae* T3T1 (NC_015964). To identify the genetic changes accumulated during persistent colonization, assemblies from each persistence episode were aligned using

progressiveMauve with Mauve 2.4.0[52], using the first isolate from each series as the complete genome reference for the lineage. Geneious R9 (Biomatters) was used to search for non-synonymous single-nucleotide polymorphisms (SNPs), insertions and deletions (indels), and changes in phase variation and promoter regions. Functional annotation of reference genomes was performed using EggNOG-mapper 2.1.9 (eggnog-mapper.embl.de)[53], and the biological function of the altered genes was assessed using OrthoDB[54].

Identification of genome-integrated prophage sequences used Phaster[55], and those corresponding to intact prophages (score > 90) were further considered, and named according to Addriaenssens *et al*.[56]. Phylogeny and classification of prophages was performed with VICTOR (victor.dsmz.de)[57], and prophages found in the genomes of persistent strains were aligned with progressiveMauve.

Statistical analysis

Statistical analyses were carried out using the GraphPad Prism 5 software, applying unpaired t-test or Fisher's exact test, when appropriate. P-values < 0.05 were considered statistically significant.

RESULTS

Identification of respiratory bacterial isolates in COPD patients

A total of 15 patients with severe COPD and frequent acute exacerbations were followed retrospectively prior to treatment and prospectively while receiving long-term azithromycin treatment (250 mg, 3 days/week) to characterize the opportunistic pathogens of their respiratory tract (**Supplementary Figure S1**). The most frequently isolated bacterial species in retrospective samples and those collected on the day of inclusion before starting the treatment were *Moraxella catarrhalis* (seven patients), followed by *H. influenzae* (six patients), and *Pseudomonas aeruginosa* (five patients).

On the other hand, cultured microbiota shifted after onset of azithromycin treatment (V2, V3, and acute exacerbations): *H. parainfluenzae* was detected at least once in 10 of the

15 patients, followed by *H. influenzae* and *P. aeruginosa* (isolated from seven patients, each), and *Stenotrophomonas maltophilia* (six patients).

Characterization of persistent colonization of the respiratory tract by *H. influenzae* and *H. parainfluenzae*

Persistent lineages were defined as those belonging to the same molecular type (differing by no more than one MLST locus or differing by three or fewer PFGE bands in *H. influenzae* and *H. parainfluenzae*, respectively) and were isolated in the same patient for periods of more than three months. Concurrent colonization by non-persisting strains was also observed; however, despite characterizing ten colonies per episode, it is possible that persistence rates were higher due to polyclonal infections.

Five of 15 patients had persistent *H. influenzae* or *H. parainfluenzae* colonization, and so isolates from these patients were included for WGS characterization (**Figure 1**): four patients (P02, P08, P11, and P13) were persistently colonized by *H. influenzae* (**Supplementary Dataset S1A**), and two (P04 and P11) were persistently colonized by *H. parainfluenzae* (**Supplementary Dataset S1B**). Despite persistent *H. influenzae* and *H. parainfluenzae* colonization, control of acute exacerbations in these patients at 1-year follow-up (1.25 ± 0.96) was comparable to those without persistent colonization (2.30 ± 1.42) (p-value = 0.2030). A brief description of persistent colonization in the five patients is described below.

Although patient P02 withdrew from azithromycin treatment early in the study due to adverse effects, the colonization pattern of *H. influenzae* isolates was examined, identifying 11 distinct PFGE patterns but only one possible case of persistence, in which two closely related strains were isolated more than three years apart (ST139 and single-locus variant ST2111).

P04 was colonized by five different *H. parainfluenzae* clones. One of these, belonging to PFGE type HPAR04.05, was isolated during an acute exacerbation one year after treatment began and continued to persistently

colonize the patient for at least 14 months after its initial detection.

Before treatment, P08 had two cases of persistent colonization by ST147 and ST2479 clones, but these were not further isolated during azithromycin treatment. ST107 clone was first identified three months before treatment and was repeatedly isolated during treatment, both in stable phases and acute exacerbations, for at least 570 days.

P11 was persistently colonized by two *H. influenzae* lineages (ST2480 and ST147) and one *H. parainfluenzae* (HPAR11.02). ST2480 was identified two months after beginning treatment and was isolated for over a year at stable and acute exacerbation episodes; ST147 was detected after 17 months of treatment and persisted for at least 5 months; and HPAR11.02 was detected one year after starting treatment and lasted at least 104 days.

Finally, P13 was persistently colonized by a single *H. influenzae* lineage belonging to ST165, which was isolated during the first year of treatment at the stable-phase control visits.

Phylogeny, population structure, and horizontal gene transfer

All the *H. influenzae* (n = 53) and *H. parainfluenzae* (n = 18) isolates sequenced were non-capsulated. Core-SNP phylogenetic analysis revealed the expected high diversity in both species, *H. influenzae* (Figure 2A) and *H. parainfluenzae* (Figure 2B). *H. parainfluenzae* showed greater genetic heterogeneity and phylogenetic distance between strains than *H. influenzae*. However, strains belonging to each persistent lineage formed well-defined clades with high genetic similarity. Classification of the NTHi strains based on the presence or absence of 17 accessory genes revealed that all were distributed among the six major clades in the phylogenetic tree with no evidence of large clonal expansions (Figure 2A).

MLST analysis also highlights the high heterogeneity in the NTHi population from these five patients, detecting 21 different STs among the 53 strains. Only ST147 was found in

different patients (P02, P08, and P11), which was also associated with two cases of persistence (in P08 and P11). The comparison of the ST147 strains isolated from these patients revealed that those from P11 had a 60 kilobase (kb) insertion in a gene for tRNA-Leu, between the genes encoding the methionine biosynthesis PLP-dependent protein and threonine synthase, which was not found in the ST147 strains from P02 and P08. This insertion corresponded to an integrative conjugative element (ICE) called ICE*HpaHUB5*, which contained 70 genes, including genes involved in replication, type IV secretion system and integration. This ICE also contained the macrolide efflux genetic assembly (MEGA) element, which carries *msr*(D) and *mef*(E) genes adjacent to *tet*(M) and was inserted into the *arsR* gene encoding an arsenical resistance repressor (**Figure 3**). Interestingly, this 60 kb region was also detected in *H. parainfluenzae* HPAR11.02 isolates (HUB-P11-HP02 and HUB-P11-HP03), inserted within distinct tRNA-Leu genes, and showing a high percentage of identity (100% and 99.997%, respectively) with the ICE found in the *H. influenzae* ST147 strains from P11, suggesting the probable and recent horizontal transference between the strains isolated from P11.

Macrolide resistance development during long-term azithromycin therapy

All strains isolated before treatment were susceptible to azithromycin but after starting the therapy, newly isolated clones had developed resistance within the first months of treatment (**Table 1**), except for those from patient P02, who discontinued the treatment (**Figure 1**).

In the persistent *H. parainfluenzae* HPAR04.05 isolates (P04), the initial strain was susceptible to azithromycin (MIC = 4 mg/L). The next isolate was resistant to azithromycin (MIC = 16 mg/L), and the last isolate had even higher resistance levels (MIC = 32 mg/L), along with predicted resistance mutations in genes encoding the ribosomal proteins L1 (D85G), S1 (T327A), and 23S rRNA (uracil(747)-C(5))-methyltransferase (RlmC) (F193L).

Two cases of persistence by *H. influenzae* ST147 and ST2479 in patient P08 were susceptible to azithromycin since they were isolated before treatment and thus not under macrolide pressure. However, serial isolates of ST107 – which was first isolated before treatment – showed serially increasing resistance, strongly suggesting adaptive evolution of a single persistent clone. The first strain isolated after starting treatment was susceptible, despite acquiring a premature stop codon in *acrR*, encoding the AcrAB-TolC efflux pump repressor. After 100 days of treatment, the ST107 clone was highly resistant (MIC = 128 mg/L), accompanied by 3 of the 6 copies of the 23S rRNA with an A2058G mutation, in addition to the premature stop in *acrR*. Later isolates of ST107 had even higher resistance (MIC = 256 mg/L), and all copies of 23S rRNA had acquired the A2058G mutation.

In patient P11, *H. influenzae* isolates of ST2480 were initially susceptible but then developed azithromycin resistance (MIC = 256 mg/L) within one year of treatment after an insertion in the gene encoding ribosomal protein L22. By contrast, a persistent azithromycin-resistant *H. parainfluenzae* HPAR11.02 clone (MIC = 8 mg/L) was isolated after a year of treatment, due to acquisition of the MEGA element with the *msr*(D) and *mef*(E) genes. The MEGA element was also found in the subsequent persistent azithromycin-resistant *H. influenzae* ST147 clones also isolated in patient P11 (MIC = 256 mg/L). It should be noted that ST147 clones isolated in other patients (P02 and P08) did not have the MEGA element. This suggests horizontal gene transfer of the MEGA element from an unknown donor into *H. parainfluenzae*, followed by subsequent transfer to ST147.

In the persistent clone ST165 from P13, the initial susceptible strain developed azithromycin resistance after three months of treatment (MIC = 128 mg/L), also due to substitutions in the L22 ribosomal protein (K90E) and the L4 ribosomal protein (G65D).

Genetic changes observed during persistent colonization

Genome-wide changes arising in serially collected persistent isolates were examined to determine additional changes taking place in *Haemophilus* spp. during prolonged azithromycin treatment (**Figure 4A** and **Supplementary Dataset S2**), in addition to those known to directly increase macrolide resistance. ST147 and ST2479 persistent isolates from P08 were excluded from this analysis since they were only isolated prior to treatment.

Most genomic variation among closely related persistent strains occurred in genes encoding proteins with unknown functions, followed by genes associated with membrane and cell wall structure, and inorganic ion transport and metabolism. **Figure 4B** depicts the biological functions of the genes undergoing genetic variation in each persistence case, and **Supplementary Dataset S3** contains the proportion of genes belonging to each biological function in reference genomes.

Among all the genetic changes, 16 genes stood out as having acquired genetic alterations across serially collected isolates over time in *H. influenzae* (**Supplementary Table S1**), while in *H. parainfluenzae* all the changes were unique to each case of persistence. In all cases of persistence by *H. influenzae*, mutations were observed for *hgpB* and *licA* genes, which encode hemoglobin-haptoglobin binding protein B and phosphorylcholine kinase LicA, respectively. These changes were mostly associated with phase variation at simple sequence repeats within the open-reading frames (*hgpB,* 5'-CCAA-3'; *licA,* 5'-CAAT-3'), which frameshifts protein translation and introduces premature stop codons. Non-synonymous SNPs were also found in the *hgpB* gene of the P02-ST139/ST2111 and P08-ST107 clones, and in the *licA* gene of P08-ST107 clone. Other genes that had allelic changes in independent lineages and patients were *hgpC*, coding for hemoglobin-haptoglobin binding protein C; *fadL* (*ompP1*) coding for an outer membrane transporter; *lic3A*, coding for CMP-

Neu5Ac--lipooligosaccharide α 2-3 sialyltransferase; *lex1*, coding for lipooligosaccharide biosynthesis protein Lex-1; and genes coding for glycosyltransferase family 2 and 8 proteins (**Supplementary Figure S2**). Most genetic changes in these genes were related to phase variation at simple-sequence repeats; however, changes to *fadL* were associated with distinct single-base indels that caused frameshifting and premature stop codons at different positions without exhibiting any truncation pattern throughout persistence (**Supplementary Figure S2D**).

Gain, loss, and evolution of genome-integrated prophages during persistent infection

Large insertion and deletion polymorphisms were discovered in some cases of persistence (**Figure 5**). An in-depth examination of these regions revealed gain and loss of genomic prophages, as well as, in some cases, accumulated mutations within the prophages over time.

H. influenzae strains associated with ST139/ST2111 from patient P02 had two genome-integrated prophages (*Haemophilus* phage HUB-P02-ST139-01, and *Haemophilus* phage HUB-P02-ST139-02), which had accumulated most of the SNPs (261/405) acquired over time, affecting 57 genes, most of them coding for proteins of unknown function (54.4%) and prophage structural proteins (29.8%). Integration of the *Haemophilus* phage HUB-P02-ST139-01 was accompanied by deletion of the downstream region, including genes coding for integrase, DNA cytosine methyltransferase, and 8 hypothetical proteins. This was followed by a region of high genetic variability, affecting 29 genes with respect to the reference strain (HUB-P02-HI01) (**Supplementary Figure S3A**). *Haemophilus* phage HUB-P02-ST139-02 had 260 SNPs and the last isolated strain (HUB-P02-HI19) showed 90.4% of identity to the reference strain (HUB-P02-HI01) (**Supplementary Figure S3B**). The location of this prophage within the genome is the same over time, suggesting the probable recombination

with another related prophage from an unknown donor.

In patient P04, persistent isolates of *H. parainfluenzae* HPAR04.05 had *the Haemophilus* phage HUB-P04-HPAR04.05-01 prophage, except for strain HUB-P04-HP10, which had lost 115,160 bp corresponding to the prophage and its surrounding regions (37,453 bp downstream and 41,640 bp upstream) (**Supplementary Dataset S4**). Thus, most of genetic change that accumulated in this persistence case was linked to loss of the prophage HUB-P04-HPAR04.05-01 in the HUB-P04-HP10 strain. The prophage genomes from serial isolates were nearly identical (**Supplementary Figure S3C**).

Acquisition and loss of genome-integrated prophages over time was detected in the patient P08 persistent *H. influenzae* ST107 isolates. The first two isolates (HUB-P08-HI09 and HUB-P08-HI10) contained the *Haemophilus* phage HUB-P08-ST107-02, which was absent from the later six strains, as seen by a 36.2 kb deletion in the genome alignment. On the other hand, the last six strains (HUB-P08-HI13 to HUB-P08-HI18) acquired two distinct prophages, *Haemophilus* phage HUB-P08-ST107-01, and *Haemophilus* phage HUB-P08-ST107-03, inserting 53.4 Kb and 40.5 Kb into these genomes in a gene for tRNA-Leu and in an intragenic non-coding region between genes encoding homoserine O-acetyltransferase and YcaO-like proteins, respectively. The prophages detected in these strains remained unchanged over time (**Supplementary Figure S3D-F**).

Persistence caused by *H. influenzae* ST2480 strains in patient P11 carried two prophages: *Haemophilus* phage HUB-P11-ST2480-01, which accumulated 3 synonymous SNPs over three serial isolations (**Supplementary Figure S3G**); and *Haemophilus* phage HUB-P11-ST2480-02, which did not acquire any genetic changes (**Supplementary Figure S3H**). ST147 persistent strains also had a genome-integrated prophage, *Haemophilus* phage HUB-P11-ST147-01, which remained unchanged in serial isolates (**Supplementary Figure S3I**).

The USS motif is found every ~1.2 kb of a *Haemophilus* genome, and these accumulate by mutation and biased uptake over long evolutionary timescales[16,17], therefore the frequency of USS within prophage insertions can be an indicator of how long a given phage has been resident in *Haemophilus* species. Similarly, although horizontally transferred phage from other species may have varying guanine-cytosine (GC) content from the new host, the GC content evolves towards a host-adapted GC content over time. Thus, deviations in prophage USS density or GC content are indicators of how recently phage-host coevolution began (**Supplementary Table S2**). On average, *Haemophilus* prophage sequences in the persistent isolates had a density of 0.41 USS/Kb (SD = 0.13), significantly lower than the whole genome density, which had 0.78 USS/Kb (SD = 0.03) (p-value < 0.0001), and in all cases USS density was lower in individual prophage compared to the genome average, indicating their relatively recent acquisition of *Haemophilus*. Different prophage sequences ranged from 0.26 to 0.69 USS per kb, suggesting some (*e.g. Haemophilus* phage HUB-P02-ST139-02) have been resident in *Haemophilus* much longer than others (*e.g. Haemophilus* phage HUB-P11-ST2480-02). Similarly, average GC content across prophages was 41.0% (SD = 2.03) compared to 38.3% (SD = 0.44) in whole genome sequences (p-value < 0.0001), suggesting their origin in other bacterial groups. In all prophages and genomes, GC content remained stable over time, as expected, though extensive variation added to *Haemophilus* phage HUB-P02-ST139-01 decreased GC content from 39.3% to 38.3%, comparable to the genome where it was integrated (38.1%).

Prophages identified in all the *H. influenzae* and *H. parainfluenzae* strains (n = 72) from the patients included in this study were analyzed phylogenetically (**Figure 6**): all of them belonged to the same family cluster and were classified into six genus (A-F) and 35 species clusters. There were no significant differences in GC content between genus clusters, though prophages from genus cluster A had fewer USS/kb (0.27

USS/kb, SD = 0.07) than prophages from other genus clusters, suggesting that it only more recently began infecting *Haemophilus* species. Moreover, in some cases, prophages identified in different patients and/or different clones belonged to the same species cluster, suggesting a possible transmission of prophages between microorganisms present in the same patient or between different strains: both *Haemophilus* phage HUB-P02-ST321-01 and *Haemophilus* phage HUB-P11-ST2480-02 belonged to species cluster A1, despite coming from different STs isolated from different patients; and *Haemophilus* phage HUB-P02-ST321-05 and *Haemophilus* phage HUB-P02-ST349-03, identified in different STs from the same patient, belonged to species cluster C5. Furthermore, the ST147 strains from P02, P08 and P11 carried a prophage belonging to the same species cluster F1, and the closely related F2, suggesting that this prophage may have higher specificity for ST147 clones.

DISCUSSION

Prolonged azithromycin treatment improves the quality of life in patients with severe COPD and recurrent exacerbations by modulating the immune system and decreasing inflammation in the respiratory tract[5,6]. However, our results indicate that long-term use of this antibiotic has major effects on the microbiota and favors the development of macrolide resistance through enhanced selection of resistant bacteria[18,19].

An in-depth analysis of respiratory tract samples from COPD patients identified differences in lung colonization by bacterial pathogens after long-term macrolide treatment. Before therapy, *M. catarrhalis* and *H. influenzae* were the most frequently isolated species during both stable phases and exacerbations, as previously reported[20]. However, once azithromycin treatment was initiated, *H. parainfluenzae*, *H. influenzae*, *P. aeruginosa*, and *S. maltophilia* became the most frequently isolated bacteria. In particular, we noted increased prevalence of *H. parainfluenzae* during

treatment. This might be because *H. parainfluenzae* is considered non-pathogenic and could be undersampled during routine clinical microbiology practice. Some studies support the role of *H. parainfluenzae* in respiratory tract colonization in both healthy subjects and during respiratory disease[11,12,21,22]. However, unlike *H. influenzae* colonization, *H. parainfluenzae* does not produce an increased inflammatory response[23], suggesting that the importance of this microorganism may be related to its capacity for antimicrobial resistance acquisition allowing it to act as a reservoir for resistance genes[24-26] that – under azithromycin pressure – may spread to other related bacteria sharing the same ecological niche.

Long-term azithromycin therapy is associated with the development of macrolide resistance[18]. A subtherapeutic but continuous dose is sufficient to induce macrolide resistance in both *H. influenzae* and *H. parainfluenzae* strains. Many of the genetic changes seen in persisting strains are consistent with changes to ribosomes that block macrolide binding. The A2058G mutation in the domain V of the 23S rRNA[7] was observed to arise in some persistent *H. influenzae* strains, and this change accumulated in the six different copies of the gene over long-term antibiotic therapy, resulting in multiple stepwise increases in macrolide resistance levels. Other genetic alterations linked to the acquisition of azithromycin resistance in persistently infecting *H. influenzae* strains were changes in the ribosomal proteins L4 and L22, which have been described[6,7]. Alterations in genes encoding ribosome-related proteins, including S1, L1, and RlmC, were also observed in some *H. parainfluenzae* macrolide-resistant strains. S1 is involved in translation, transcription, and RNA stability[27], whereas L1 is a component of the ribosomal 50S subunit that binds to 23S rRNA and participates in translation[28]. RlmC is a methyltransferase that methylates U747 of 23S rRNA, promoting RlmA methylation of G748, and facilitating the binding of macrolides[29]. Thus, changes in RlmC could indirectly affect macrolide binding to the ribosome, as previously suggested

by Shoji *et al*., who observed that disrupted RlmCD mutants had increased resistance to telithromycin in *S. pneumoniae*[30]. However, additional research is required to confirm the role of the specific RlmC mutation observed here (F193L) in mediating macrolide resistance. These results indicate that the microbiota in patients undergoing prolonged use of azithromycin should be monitored, since the treatment promotes the development and acquisition of resistance by respiratory tract colonizing pathogens. Macrolides are commonly used to treat respiratory infections, and the presence of colonizing bacteria with acquired and transmissible resistance mechanisms may complicate their treatment.

Genetic changes were also seen in persistent *Haemophilus* species that could confer macrolide resistance via efflux pumps. The 23S rRNA gene mutations was observed in strains that had previously acquired a premature stop codon in AcrR (Q11), a negative regulator that results in the overexpression of the AcrAB efflux pump. The change in AcrR produced no change in azithromycin susceptibility, indicating no direct association with resistance, although this change could enable subsequent mutations in other genes to provide resistance, as previously reported for *acrB*[31]. Previous studies have described transferrable transposons in *Streptococcus* spp. and *H. parainfluenzae* that carry *tet*(M) and the MEGA element[8,9], which encodes for the Mef(E) and Msr(D) efflux pumps that can confer macrolide resistance. In this study, an integrative conjugative element carrying the *tet*(M)-MEGA element was found to arise in *H. parainfluenzae* strains and then later *H. influenzae* strains in the same patient. This pattern suggests the possibility of horizontal transmission between species sharing the same ecological niche within the patient, further supporting the role of *H. parainfluenzae* as a reservoir for resistance genes[24,32].

During persistent colonization, genetic differences that arose were also usually seen in successive serial isolates, suggesting selective sweeps and bacterial adaptation. Most of the changes affecting genes of known

function encode proteins involved in the structure of the bacterial cell surface. In fact, most of *H. influenzae* persistence cases showed alterations in *licA*, *lex1*, and *lic3A*. These changes likely reflect host immune pressures selecting for *Haemophilus* strains to alter their surface antigens or evade the effects of complement-mediated killing. The *licA* gene encodes a phosphorylcholine kinase, which adds phosphorylcholine to lipooligosaccharide (LOS) and interacts with C-reactive protein (CRP), activating the complement system. Phase variation of *licA* cause changes in its expression and, as a result, the amount of phosphorylcholine that coats the bacterial surface, which may interfere with bactericidal antibody access to the surface and affect strain virulence and sensitivity to serum bactericidal activity[33,34]. On the other hand, Lex1 is involved in LOS synthesis and *lic3A*, which encodes the α-2,3-sialyltransferase responsible for the addition of N-acetylneuraminic acid (Neu5Ac) to the LOS, may have phase variation changes, affecting epithelial cell adhesion, invasion and serum resistance[35,36].

Genes involved in inorganic ion metabolism and transport also had a high number of changes across persistently infecting isolates, and some genes coding for TonB-dependent receptors stand out, particularly *hgpB* and *hgpC*, which are involved in hemoglobin and hemoglobin-haptoglobin uptake to supply the requirement of *H. influenzae* for heme[37]. Another gene that showed recurrent alterations was *fadL*, as observed by Moleres *et al*.[38], who described truncated variants, albeit in different positions than those in our study, which could affect the interaction with host cells and increase resistance to the bactericidal effects of long-chain fatty acids found in high abundance in the lungs of COPD patients. Other studies on the adaptation of *H. influenzae* to disease have also found changes in these genes[38-40], suggesting that rapid genetic change at these gene could provide a selective advantage that favours bacterial survival and colonization over long periods of time.

Genome-integrated prophages within *H. influenzae* and *H. parainfluenzae* genomes were common, and their acquisition and loss was frequently observed in serial isolates. Although phage can debilitate a bacterial population in the lytic phase, lysogenic prophage are potent vehicles of horizontal gene transfer, which may improve the ecological fitness of bacteria in certain conditions and persist in a latent state while co-evolving with the bacterial chromosome[41]. Some prophages showed high genetic changes, which could be due to recombination or selective pressures exerted by the bacteria themselves and can direct the prophage adaptation to the genomes where they have been integrated. These changes could potentially be used to track phage-host coevolution. Szafrański *et al.*[42] indicate that USSs have a higher density of core genes than accessory genes. Therefore, prophages in the lysogenic phase would accumulate more USSs as they evolved with the host, while also adapting the GC content to that of the genome into which they had been integrated[17,42]. Our prophages showed a wide range of USS densities, indicating that some of them have recently been acquired by *H. influenzae*, and others have been within the species for a long time. However, additional research involving more prophages and also prophages identified in other bacteria is required.

In conclusion, the antibiotic pressure exerted by prolonged azithromycin treatment in COPD patients, even at low doses, leads to the development of macrolide resistance through the accumulation of mutations in 23S rRNA, mutations in ribosomal proteins, or the acquisition of efflux pumps through transmission of mobile genetic elements. Furthermore, during persistent colonization of the respiratory tract, *H. influenzae* and *H. parainfluenzae* acquire prophages and other genetic changes, particularly in genes associated with membrane and cell wall, as well as inorganic ion metabolism and transport proteins, which likely improves their adaptation to host selective pressures and allows bacterial survival for long periods of time. To prevent the spread of resistant and highly host-adapted

clones, further monitoring of bacterial pathoadaptive evolution is required.

ETHICAL STATEMENT

This study was in accordance with the Declaration of Helsinki from the World Medical Association. Written informed consent was not required as this was a retrospective and observational study with isolates obtained as part of routine microbiological tests, which was approved by the Clinical Research Ethics Committee of Bellvitge University Hospital (PR252/22). The patient cohort has the approval of the Spanish Health Department for its performance as a post-authorization observational study (catalogued as such by the Spanish Medicines Agency, SPP-AZT-2016-01) and the Clinical Research Ethics Committee of Bellvitge University Hospital (EPA019/16). Patient confidentiality was always protected, and all personal data were anonymized following the current legal normative in Spain (LOPD 15/1999 and RD 1720/2007). Moreover, this project followed Law 14/2007 on Biomedical Research for the management of biological samples in clinical research.

ACKNOWLEDGMENTS

We would like to thank the staff of the Microbiology Laboratory of Bellvitge University Hospital who contributed daily to this project.

This study was funded by the Fundación Española del Pulmón SEPAR (88/2016 to DH and 1116/2020 to SM); Fondo de Investigaciones Sanitarias (PI16/00977 to SM); Fundació Catalana de Pneumologia, FUCAP (Beca Albert Agustí 2017 to DH); and from Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES, CB06/06/0037), an initiative of the Instituto de Salud Carlos III (ISCIII). The European Regional Development Fund/European Social Fund (ERDF/ESF, “Investing in your future”) also provided financial support, and CERCA Programme/Generalitat de Catalunya provided institutional support. Bioinformatic analysis was supported by an Amazon Web Services (AWS) research grant to SM. SS received

financial support by Menarini. AC was supported by Formación de Profesorado Universitario from the Ministerio de Educación of Spain (FPU16/02202), and SM by Miguel Servet contract (CP19/00096) (ISCIII).

DATA AVAILABILITY

Sequence reads were deposited in the European Nucleotide Archive (ENA) under the project accession number PRJEB53905. The accession numbers of the reads used in this study are listed in Supplementary Dataset S1A and B.

REFERENCES

1. López-Campos, J. L., Tan, W. & Soriano, J. B. Global burden of COPD. *Respirology* 21, 14–23 (2016).

2. Raherison, C. & Girodet, P. O. Epidemiology of COPD. *Eur. Respir. Rev.* 18, 213–221 (2009).

3. Ko, F. W. *et al.* Acute exacerbation of COPD. *Respirology* 21, 1152–1165 (2016).

4. MacLeod, M. *et al.* Chronic obstructive pulmonary disease exacerbation fundamentals: Diagnosis, treatment, prevention and disease impact. *Respirology* 26, 532–551 (2021).

5. Albert, R. K. *et al.* Azithromycin for prevention of exacerbations of COPD. *N. Engl. J. Med.* 365, 689–698 (2011).

6. Parnham, M. J. *et al.* Azithromycin: Mechanisms of action and their relevance for clinical applications. *Pharmacol. Ther.* 143, 225–245 (2014).

7. Roberts, M. C. Update on macrolide-lincosamide-streptogramin, ketolide, and oxazolidinone resistance genes. *FEMS Microbiol. Lett.* 282, 147–159 (2008).

8. Varaldo, P. E., Montanari, M. P. & Giovanetti, E. Genetic elements responsible for erythromycin resistance in streptococci. *Antimicrob. Agents Chemother.* 53, 343–353 (2009).

9. Endimiani, A., Allemann, A., Wüthrich, D., Lupo, A. & Hilty, M. First report of the macrolide efflux genetic assembly (MEGA) element in *Haemophilus parainfluenzae*. *Int. J. Antimicrob. Agents* 49, 265–266 (2017).

10. Sethi, S. & Murphy, T. F. Infection in the pathogenesis and course of chronic obstructive pulmonary disease. *N. Engl. J. Med.* 359, 2355–2365 (2008).

11. Hill, S. L., Mitchell, J. L., Stockley, R. A. & Wilson, R. The role of *Haemophilus parainfluenzae* in COPD. *Chest* 117, 293S (2000).

12. Patel, I. S., Seemungal, A. R., Wilks, M., Lloyd-Owen, S. J. & Donaldson, G. C. Relationship between bacterial colonisation and the frequency, character, and severity of COPD exacerbations. *Thorax* 57, 759–764 (2002).

13. Finney, L. J., Ritchie, A., Pollard, E., Johnston, S. L. & Mallia, P. Lower airway colonization and inflammatory response in COPD: a focus on *Haemophilus influenzae*. *Int. J. Chron. Obstruct. Pulmon. Dis.* 9, 1119–1132 (2014).

14. Dobrindt, U., Zdziarski, J., Salvador, E. & Hacker, J. Bacterial genome plasticity and its impact on adaptation during persistent infection. *Int. J. Med. Microbiol.* 300, 363–366 (2010).

15. Mell, J. C., Hall, I. M. & Redfield, R. J. Defining the DNA uptake specificity of naturally competent *Haemophilus influenzae* cells. *Nucleic Acids Res.* 40, 8536–8549 (2012).

16. Redfield, R. J. *et al.* Evolution of competence and DNA uptake specificity in the *Pasteurellaceae*. *BMC Evol. Biol.* 6, 82 (2006).

17. Mell, J. C. & Redfield, R. J. Natural competence and the evolution of DNA uptake specificity. *J. Bacteriol.* 196, 1471–1483 (2014).

18. Serisier, D. J. Risks of population antimicrobial resistance associated with chronic macrolide use for inflammatory airway diseases. *Lancet Respir. Med.* 1, 262–274 (2013).

19. Yamaya, M. *et al.* Macrolide effects on the prevention of COPD exacerbations. *Eur. Respir. J.* 40, 485–494 (2012).

20. Leung, J. M. *et al.* The role of acute and chronic respiratory colonization and infections in the pathogenesis of COPD. *Respirology* 22, 634–650 (2017).

21. Kuklinska, D. & Kilian, M. Relative proportions of *Haemophilus* species in the throat of healthy children and adults. *Eur. J. Clin. Microbiol.* 3, 249–252 (1984).

22. Kosikowska, U., Biernasiuk, A., Rybojad, P., Łoś, R. & Malm, A. *Haemophilus parainfluenzae* as a marker of the upper respiratory tract microbiota changes under the influence of preoperative prophylaxis with or without postoperative treatment in patients with lung cancer. *BMC Microbiol.* 16, 62 (2016).

23. Marin, A. *et al.* Variability and effects of bronchial colonisation in patients with moderate COPD. *Eur. Respir. J.* 35, 295–302 (2010).

24. Sierra, Y. *et al.* Genome-wide analysis of urogenital and respiratory multidrug-resistant *Haemophilus parainfluenzae*. *J. Antimicrob. Chemother.* 76, 1741–1751 (2021).

25. Gromkova, R. C., Mottalini, T. C. & Dove, M. G. Genetic transformation in *Haemophilus parainfluenzae* clinical isolates. *Curr. Microbiol.* 37, 123–126 (1998).

26. Juhas, M. *et al.* Sequence and functional analyses of *Haemophilus spp.* genomic islands. *Genome Biol.* 8, 237 (2007).

27. Sengupta, J., Agrawal, R. K. & Frank, J. Visualization of protein S1 within the 30S ribosomal subunit and its interaction with messenger RNA. *Proc. Natl. Acad. Sci. U. S. A.* 98, 11991–11996 (2001).

28. Nevskaya, N. *et al.* Ribosomal protein L1 recognizes the same specific structural motif in its target sites on the autoregulatory mRNA and 23S rRNA. *Nucleic Acids Res.* 33, 478–485 (2005).

29. Desmolaize, B. *et al.* A single methyltransferase YefA (RlmCD) catalyses both m5U747 and m5U1939 modifications in *Bacillus subtilis* 23S rRNA. *Nucleic Acids Res.* 39, 9368–9375 (2011).

30. Shoji, T. *et al.* RlmCD-mediated U747 methylation promotes efficient G748 methylation by methyltransferase RlmAII in 23S rRNA in *Streptococcus pneumoniae*; interplay between two rRNA methylations responsible for telithromycin susceptibility. *Nucleic Acids Res.* 43, 8964–8972 (2015).

31. Seyama, S., Wajima, T., Nakaminami, H. & Noguchi, N. Amino acid substitution in the major multidrug efflux transporter protein AcrB contributes to low susceptibility to azithromycin in *Haemophilus influenzae*. *Antimicrob. Agents Chemother.* 61, e01337-17 (2017).

32. Sierra, Y. *et al.* Emergence of multidrug resistance among *Haemophilus parainfluenzae* from respiratory and urogenital samples in Barcelona, Spain. *Eur. J. Clin. Microbiol. Infect. Dis.* 39, 703–710 (2019).

33. Clark, S. E., *et al*. Phosphorylcholine allows for evasion of bactericidal antibody by *Haemophilus influenzae*. *PLOS Pathog.* 8, e1002521 (2012).

34. Humphries, H. E. & High, N. J. The role of *licA* phase variation in the pathogenesis of invasive disease by *Haemophilus influenzae* type b. *FEMS Immunol. Med. Microbiol.* 34, 221–230 (2002).

35. Zhou, Q. *et al.* Two glycosyltransferase genes of *Haemophilus parasuis* SC096 implicated in lipooligosaccharide biosynthesis, serum resistance, adherence, and invasion. *Front. Cell. Infect. Microbiol.* 6, 100 (2016).

36. Fox, K. L. *et al.* Identification of a bifunctional lipopolysaccharide sialyltransferase in *Haemophilus influenzae*. *J. Biol. Chem.* 281, 40024–40032 (2006).

37. Harrison, A. *et al.* Genomic sequence of an otitis media isolate of nontypeable *Haemophilus influenzae*: comparative study with *H. influenzae* serotype d, strain KW20. *J. Bacteriol.* 187, 4627–4636 (2005).

38. Moleres, J. *et al.* Antagonistic pleiotropy in the bifunctional surface protein FadL (OmpP1) during adaptation of *Haemophilus influenzae* to chronic lung infection associated with chronic obstructive pulmonary disease. *MBio* 9, 1–23 (2018).

39. Garmendia, J. *et al.* Characterization of nontypable *Haemophilus influenzae* isolates recovered from adult patients with underlying chronic lung disease reveals genotypic and phenotypic traits associated with persistent infection. *PLoS One* 9, e97020 (2014).

40. Harrison, A. *et al.* Continuous microevolution accelerates disease progression during sequential episodes of infection. *Cell Rep.* 30, 2978-2988.e3 (2020).

41. Koskella, B. & Brockhurst, M. A. Bacteria–phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiol. Rev.* 38, 916–931 (2014).

42. Szafrański, S. P. *et al.* Diversity patterns of bacteriophages infecting *Aggregatibacter* and *Haemophilus* species across clades and niches. *ISME J.* 13, 2500–2522 (2019).

43. Puig, C., Calatayud, L., Martí, S., Tubau, F. & Garcia-Vidal, C. Molecular epidemiology of nontypeable *Haemophilus influenzae* causing community-acquired pneumonia in adults. *PLoS One* 8, e82515 (2013).

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

45. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35, 4453–4455 (2019).

46. Carrera-Salinas, A. *et al.* Epidemiology and population structure of *Haemophilus influenzae* causing invasive disease. *Microb. Genomics* 7, 000723 (2021).

47. Watts, S. C. & Holta, K. E. HICAP: *In silico* serotyping of the *Haemophilus influenzae* capsule locus. *J. Clin. Microbiol.* 57, e00190-19 (2019).

48. De Chiara, M. *et al.* Genome sequencing of disease and carriage isolates of nontypeable *Haemophilus influenzae* identifies discrete population structure. *Proc. Natl. Acad. Sci.* 111, 5439–5444 (2014).

49. Pinto, M. *et al.* Insights into the population structure and pan-genome of *Haemophilus influenzae*. *Infect. Genet. Evol.* 67, 126–135 (2019).

50. Bortolaia, V. *et al.* ResFinder 4.0 for predictions of phenotypes from genotypes. *J. Antimicrob. Chemother.* 75, 3491–3500 (2020).

51. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069 (2014).

52. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14, 1394–1403 (2004).

53. Cantalapiedra, C. P., Hernąndez-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* 38, 5825–5829 (2021).

54. Zdobnov, E. M. *et al.* OrthoDB in 2020: evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* 49, D389–D393 (2021).

55. Arndt, D. *et al.* PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* 44, W16–W21 (2016).

56. Adriaenssens, E. M. & Rodney Brister, J. How to name and classify your phage: an informal guide. *Viruses* 9, 70 (2017).

57. Meier-Kolthoff, J. P. & Göker, M. VICTOR: genome-based phylogeny and classification of prokaryotic viruses. *Bioinformatics* 33, 3396–3404 (2017).

Table 1. Azithromycin resistance and macrolide resistance determinants in persistent *H. influenzae* and *H. parainfluenzae* strains. Negative values for treatment time indicate days before starting azithromycin treatment. ST, sequence type; S, susceptible; R, resistant. *SLV, single-locus variant.

Patient	Microorganism	PFGE pattern	ST	Treatment time (days)	Azithromycin resistance			Macrolide resistance determinants	
					S/R	Disk diffusion (mm)	Microdilution (mg/L)	Genetic determinant	Alteration
P02	*H. influenzae*	HINF02.05	ST139	-518	S	27	2	-	-
		HINF02.05	ST2111*	558	S	24	2	-	-
P04	*H. parainfluenzae*	HPAR04.05	-	364	S	21	4	-	-
		HPAR04.05	-	385	R	11	16	-	-
		HPAR04.05	-	800	R	0	32	L1	D85G
								S1	T327A
								RlmC	F193L
P08	*H. influenzae*	HINF08.04	ST2479	-610; -477	S	23	2	-	-
	H. influenzae	HINF08.06	ST147	-217; -119	S	19	2	-	-
	H. influenzae	HINF08.01	ST107	-109; -21	S	24	4	-	-
				27	S	19	4	*acrR*	Q11STOP
				100	R	0	128	*acrR*	Q11STOP
								23S rRNA	A2058G (50% copies)
				272; 364; 461	R	0	256	*acrR*	Q11STOP
								23S rRNA	A2058G (100% copies)
P11	*H. influenzae*	HINF11.01	ST2480	78; 110	S	19	4	-	-
				364; 398; 468	R	0	256	L22	78-GPSMKRVMPRAK-79
	H. parainfluenzae	HPAR11.02	-	364; 468	R	0	8	MsrD/MefE	-
	H. influenzae	HINF11.02	ST147	531; 679	R	0	256	MsrD/MefE	-
P13	*H. influenzae*	HINF13.01	ST165	0	S	18	4	-	-
				92	R	0	128	L22	K90E
				342	R	0	128	L22	K90E
								L4	G65D

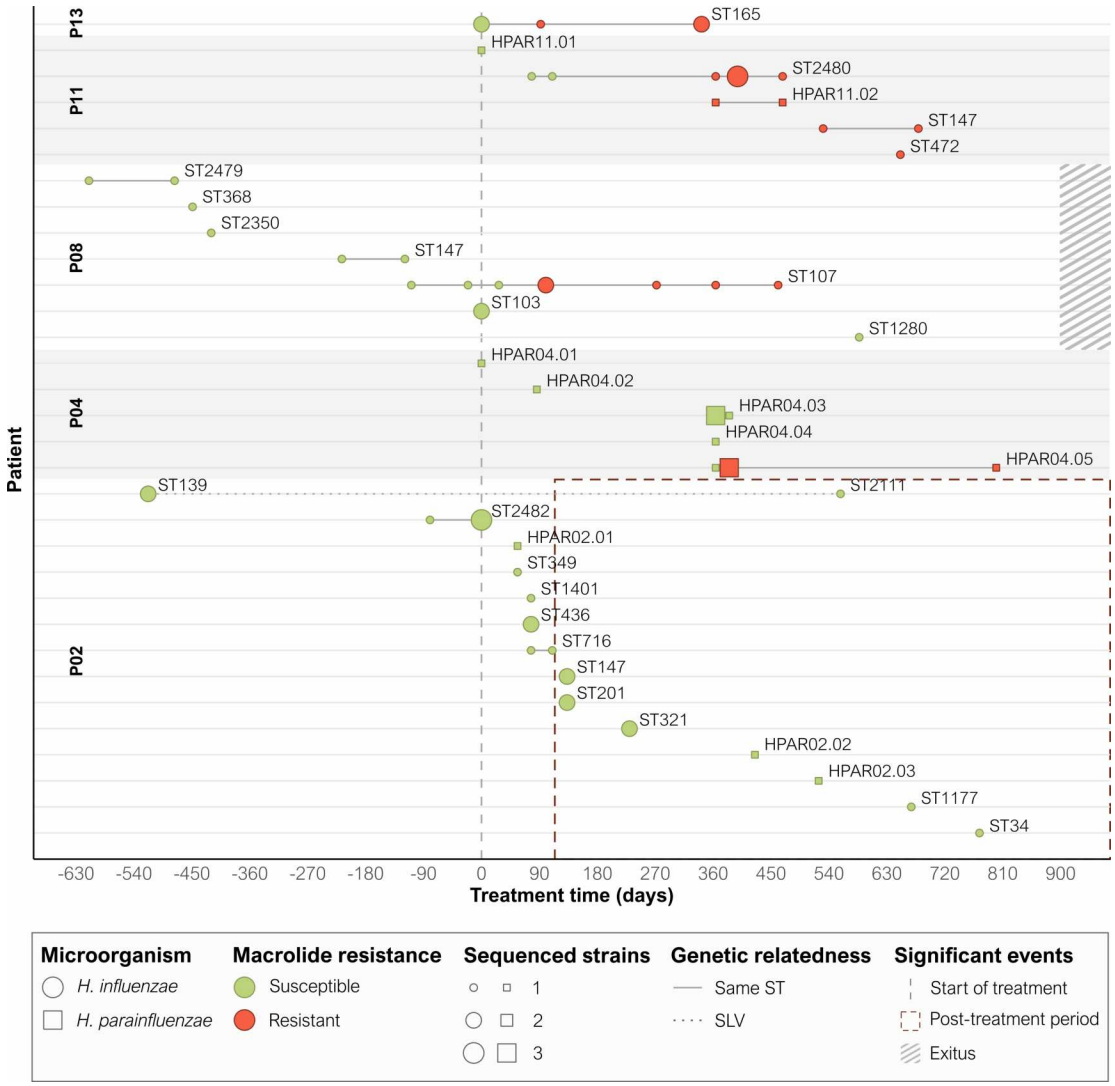


Figure 1. Timeline illustrating *H. influenzae* and *H. parainfluenzae* strains isolated from respiratory samples in COPD patients with long-term azithromycin therapy. Persistence was defined as the isolation of the same clone in a patient for more than three months based on the ST and PFGE pattern in *H. influenzae* and *H. parainfluenzae*, respectively. Persistence is depicted as a grey line connecting the isolates. ST, sequence type; SLV, single-locus variant.

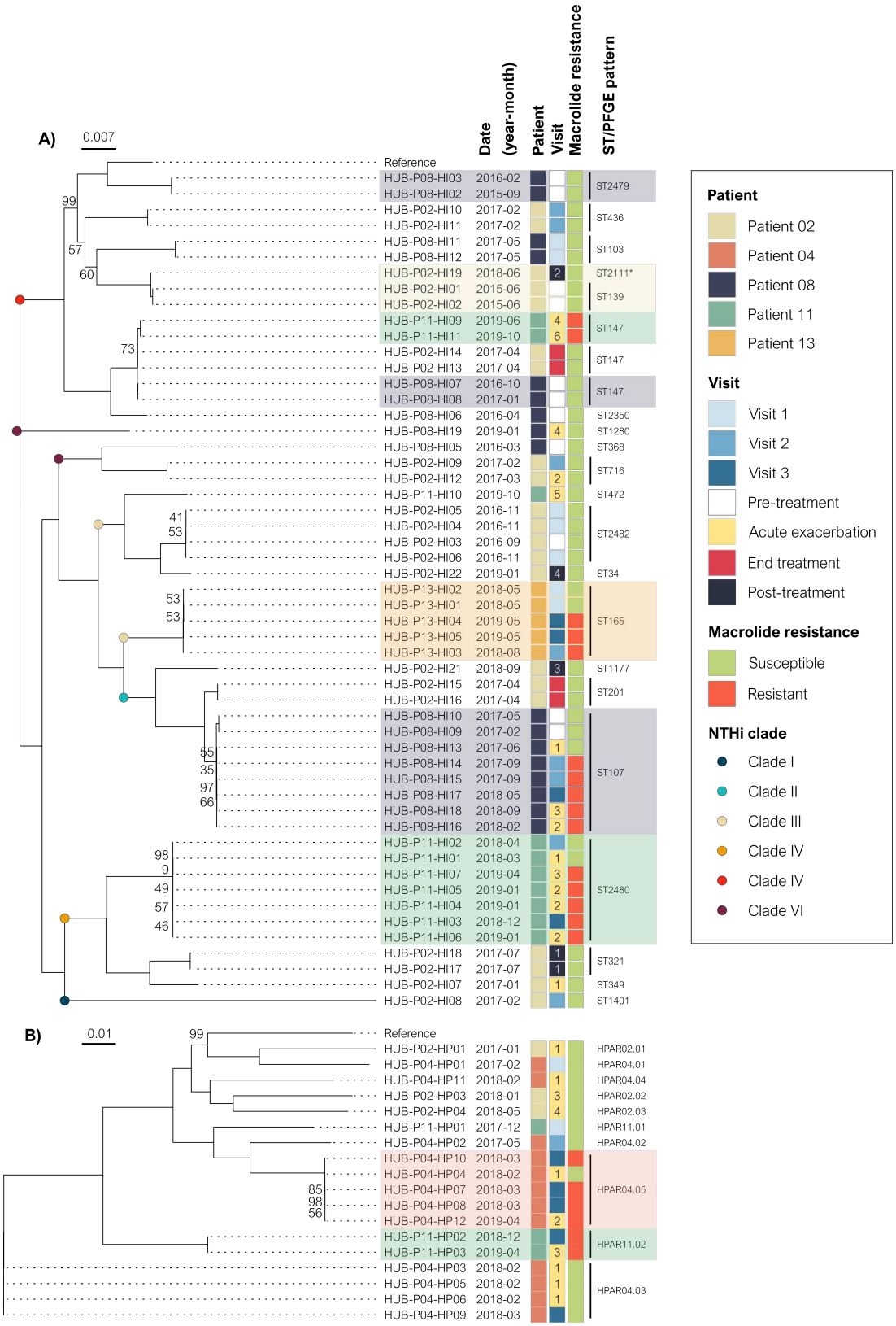

A)
0.007
Date (year-month)
Patient
Visit
Macrolide resistance
ST/PFGE pattern
Reference
HUB-P08-HI03 2016-02 ST2479
HUB-P08-HI02 2015-09
HUB-P02-HI10 2017-02 ST436
HUB-P02-HI11 2017-02
HUB-P08-HI11 2017-05 ST103
HUB-P08-HI12 2017-05
HUB-P02-HI19 2018-06 ST2111*
HUB-P02-HI01 2015-06 ST139
HUB-P02-HI02 2015-06
HUB-P11-HI09 2019-06 ST147
HUB-P11-HI11 2019-10
HUB-P02-HI14 2017-04 ST147
HUB-P02-HI13 2017-04
HUB-P08-HI07 2016-10 ST147
HUB-P08-HI08 2017-01
HUB-P08-HI06 2016-04 ST2350
HUB-P08-HI19 2019-01 ST1280
HUB-P08-HI05 2016-03 ST368
HUB-P02-HI09 2017-02 ST716
HUB-P02-HI12 2017-03
HUB-P11-HI10 2019-10 ST472
HUB-P02-HI05 2016-11 ST2482
HUB-P02-HI04 2016-11
HUB-P02-HI03 2016-09
HUB-P02-HI06 2016-11
HUB-P02-HI22 2019-01 ST34
HUB-P13-HI02 2018-05 ST165
HUB-P13-HI01 2018-05
HUB-P13-HI04 2019-05
HUB-P13-HI05 2019-05
HUB-P13-HI03 2018-08
HUB-P02-HI21 2018-09 ST1177
HUB-P02-HI15 2017-04 ST201
HUB-P02-HI16 2017-04
HUB-P08-HI10 2017-05 ST107
HUB-P08-HI09 2017-02
HUB-P08-HI13 2017-06
HUB-P08-HI14 2017-09
HUB-P08-HI15 2017-09
HUB-P08-HI17 2018-05
HUB-P08-HI18 2018-09
HUB-P08-HI16 2018-02
HUB-P11-HI02 2018-04 ST2480
HUB-P11-HI01 2018-03
HUB-P11-HI07 2019-04
HUB-P11-HI05 2019-01
HUB-P11-HI04 2019-01
HUB-P11-HI03 2018-12
HUB-P11-HI06 2019-01
HUB-P02-HI18 2017-07 ST321
HUB-P02-HI17 2017-07
HUB-P02-HI07 2017-01 ST349
HUB-P02-HI08 2017-02 ST1401
B)
0.01
Reference
HUB-P02-HP01 2017-01 HPAR02.01
HUB-P04-HP01 2017-02 HPAR04.01
HUB-P04-HP11 2018-02 HPAR04.04
HUB-P02-HP03 2018-01 HPAR02.02
HUB-P02-HP04 2018-05 HPAR02.03
HUB-P11-HP01 2017-12 HPAR11.01
HUB-P04-HP02 2017-05 HPAR04.02
HUB-P04-HP10 2018-03 HPAR04.05
HUB-P04-HP04 2018-02
HUB-P04-HP07 2018-03
HUB-P04-HP08 2018-03
HUB-P04-HP12 2019-04
HUB-P11-HP02 2018-12 HPAR11.02
HUB-P11-HP03 2019-04
HUB-P04-HP03 2018-02 HPAR04.03
HUB-P04-HP05 2018-02
HUB-P04-HP06 2018-02
HUB-P04-HP09 2018-03
Patient
Patient 02
Patient 04
Patient 08
Patient 11
Patient 13
Visit
Visit 1
Visit 2
Visit 3
Pre-treatment
Acute exacerbation
End treatment
Post-treatment
Macrolide resistance
Susceptible
Resistant
NTHi clade
Clade I
Clade II
Clade III
Clade IV
Clade IV
Clade VI

Figure 2. Phylogenetic tree of *Haemophilus spp.* isolated from COPD patients during long-term azithromycin treatment. (A) *H. influenzae* phylogenetic tree. NTHi were classified into six clades (I-VI) as previously described[48,49]. Strain Hi375 (CP009610) was used as the reference. **(B)** *H. parainfluenzae* phylogenetic tree. Strain T3T1 (NC_015964) was used as the reference. Persistence was defined as the isolation of the same lineage in a patient for more than three months based on the ST in *H. influenzae* isolates or the PFGE pattern in *H. parainfluenzae* isolates. Shadowed areas highlight persistent lineages. The numbers in the acute exacerbation and post-treatment boxes are used to distinguish between episodes. Bootstrap values other than 100 are given at branch nodes. *SLV, single-locus variant.

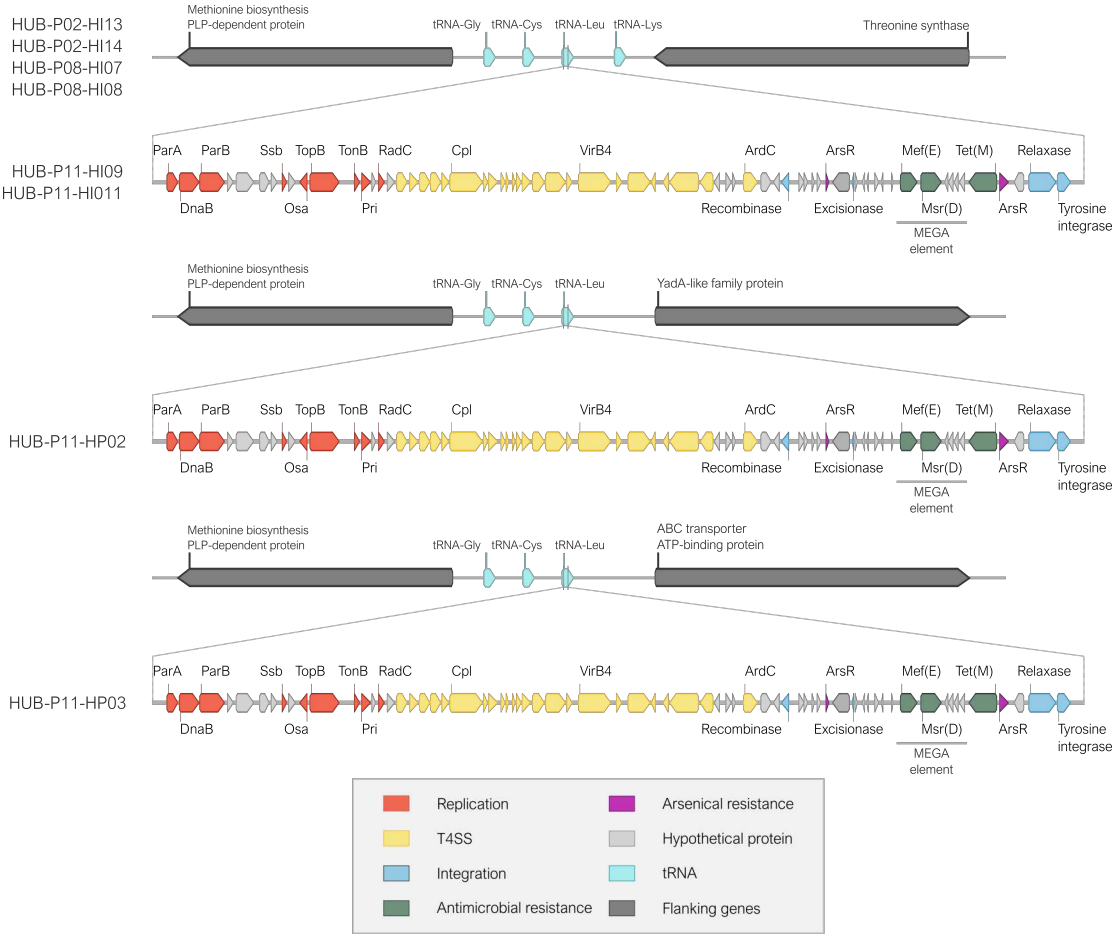


Figure 3. Schematic representation of ICE*HpaHUB5* structure containing *tet*(M)-MEGA element observed in *H. influenzae* ST147 (HUB-P11-HI09/011) and *H. parainfluenzae* HPAR11.02 (HUB-P11-HP02/03) clones from P11. The detected MEGA element was the same as that found in the *H. parainfluenzae* strain AE-2096513 (KJ545575).

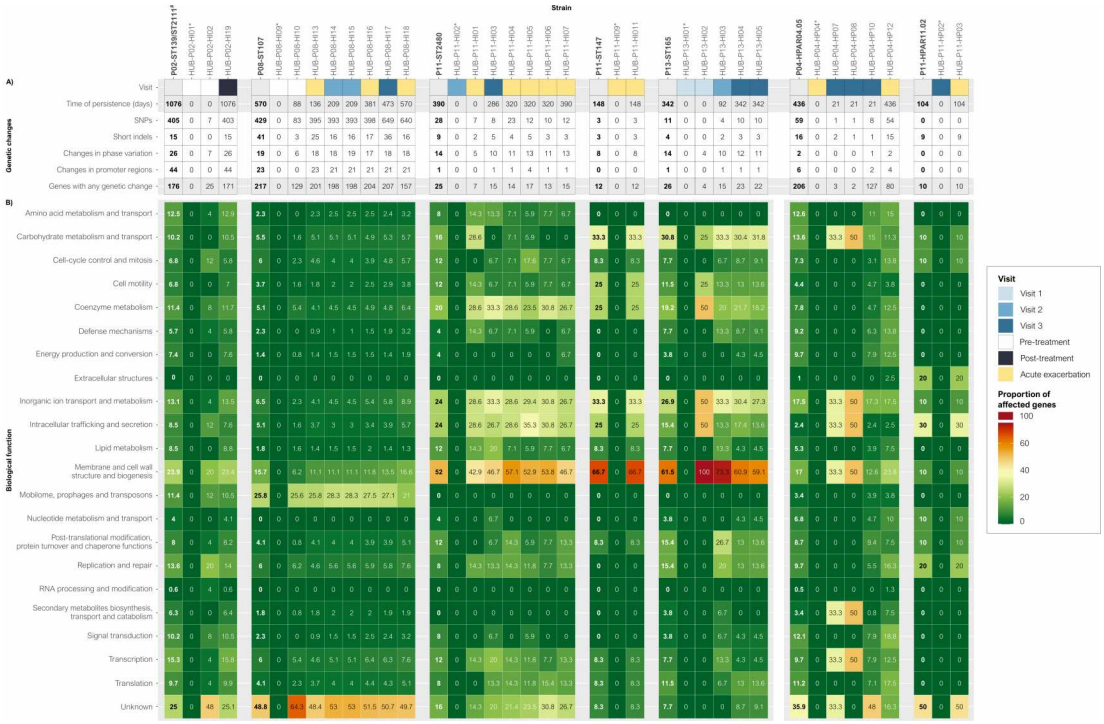


Figure 4. Genetic changes and biological functions of the affected genes in each persistent case. **(A)** Number of genetic changes, including non-synonymous SNPs, short indels (≤ 10 bp), and changes in phase variation and promoter regions, taking place during persistent colonization and the number of genes that presented any of these changes. **(B)** Biological functions of the genes affected in each persistent case. The numbers inside the boxes represent the percentage (%) of genes with any genetic change that had each biological function. Each gene may be annotated with more than one biological function, so that the sum of all the percentages may be greater than 100%. In the first column of each persistence case, the overall proportions of genes that were altered are highlighted in bold. *The first persistent strain isolated over time with a closed genome was used as the reference in each persistence case.

Figure 5. Alignment of whole-genome sequences of persistent *Haemophilus* spp. isolates. Assemblies from each persistence episode were aligned using the first isolate from each series as the complete genome reference* for the lineage. Prophage-associated regions were identified by Phaster. Acquisition date is indicated in brackets (year-month). #SLV, single-locus variant.

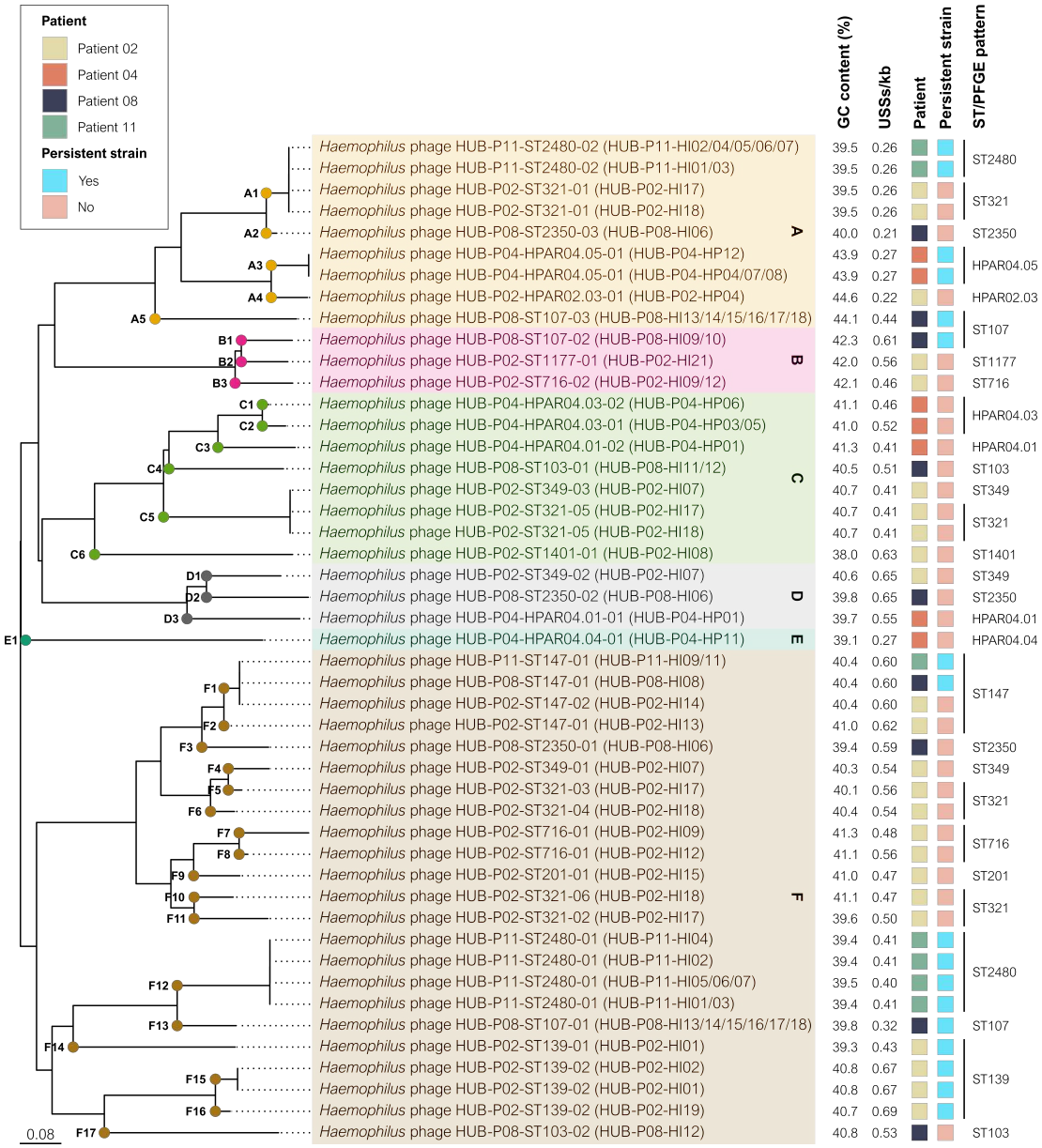


Figure 6. Phylogenetic tree of genome-integrated prophages detected in *Haemophilus* spp. strains. VICTOR software was used to calculate genome-BLAST distance phylogeny (GBDP) using D0 formula to estimate phylogeny from nucleotide sequences. The name of the prophages are colored according to the genus clusters, and the dots arranged on the branches of the tree represent the different species clusters identified by VICTOR software. GC: guanine-cytosine; USS: DNA uptake signal sequences.

SUPPLEMENTARY MATERIAL

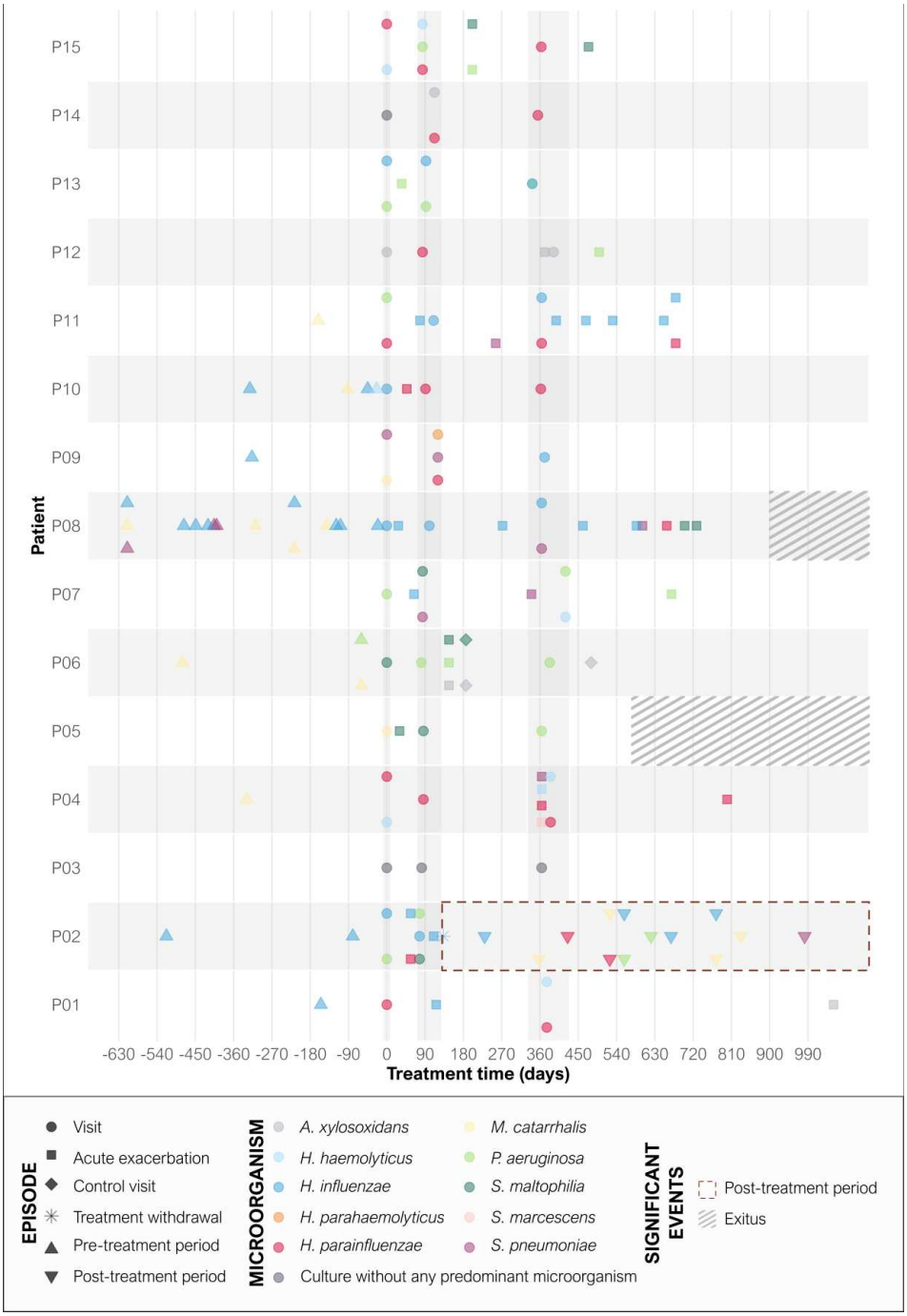
Supplementary Table S1. Genetic determinants with genetic changes in three or more cases of *H. influenzae* persistence during azithromycin treatment. Asterisks indicate that the strain had two genetically altered copies of the gene. ST, sequence type; #SLV, single-locus variant.

Genetic determinant	Locus tag	Accession number	Persistence case				
			P02-ST139/ ST2111#	P08-ST107	P11-ST2480	P11-ST147	P13-ST165
Hemoglobin-haptoglobin binding protein (*hgpB*)	NF38_01815	WP_164927859.1	X*	X	X	X	X
Phosphorylcholine kinase (*licA*)	NF38_07490	WP_005693562.1	X	X	X	X	X
Hemoglobin-haptoglobin binding protein C (*hgpC*)	NF38_01570	WP_010869038.1	X	X	X		X
OmpP1/FadL family transporter (*fadL*)	NF38_03050	WP_038440615.1	X		X	X	X
CMP-Neu5Ac--lipooligosaccharide alpha 2-3 sialyltransferase (*lic3A*)	NF38_03620	WP_038440754.1	X*	X	X		X*
Lipooligosaccharide biosynthesis protein (*lex1*)	NF38_02325	WP_038440349.1	X*	X		X	X
Glycosyltransferase family 2 protein	NF38_07685	WP_038441301.1	X	X	X		X
Glycosyltransferase family 8 protein	NF38_06815	WP_038441170.1	X	X		X	X
Glycosyltransferase family 8 protein	NF38_04110	WP_038441445.1	X	X	X		X
Glycosyltransferase	NF38_06750	WP_005688459.1	X			X	X
Immunoglobulin A1 protease autotransporter precursor	NF38_00020	WP_005693332	X			X	X
Acyltransferase family protein	NF38_03095	WP_038440625.1	X	X			X
Outer membrane protein 2 (*ompP2*)	NF38_04725	WP_005694426	X			X	X
Two-partner secretion (TPS) system translocator (*hmw1B*)	NF38_05835	WP_038441016.1	X*	X*	X*		
Filamentous hemagglutinin N-terminal domain-containing protein (*hmw1A*)	NF38_07770	WP_038441018.1	X*	X*	X		
Transferrin-binding protein-like solute binding protein (*tbpB*)	NF38_09000	WP_038441419.1		X		X	X

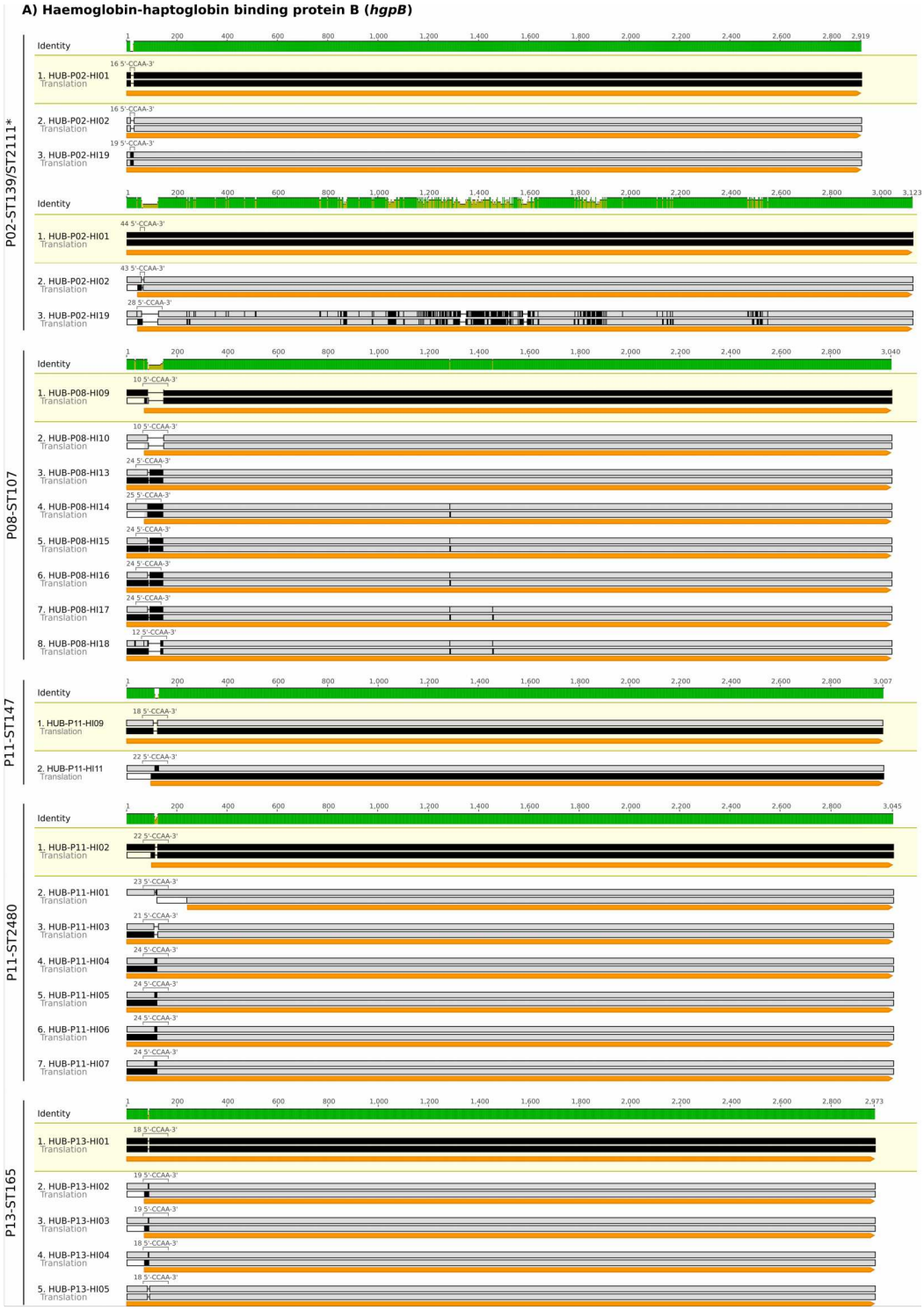
Supplementary Table S2. Comparison of GC content and USS/Kb between prophage and genome sequences. GC: Guanine-cytosine; USS: DNA uptake signal sequences. *Reference strain.

Prophage	Strain	GC content prophage (%)	GC content genome (%)	USS/Kb prophage	USS/Kb genome
Haemophilus phage HUB-P02-ST139-01	HUB-P02-HI01*	39.3	38.1	0.43	0.8
	HUB-P02-HI02	39.3	38.2	0.43	0.8
	HUB-P02-HI19	38.3	38.1	0.54	0.81
Haemophilus phage HUB-P02-ST139-02	HUB-P02-HI01*	40.8	38.1	0.67	0.8
	HUB-P02-HI02	40.8	38.2	0.67	0.8
	HUB-P02-HI19	40.7	38.1	0.69	0.81
Haemophilus phage HUB-P04-HPAR04.05-01	HUB-P04-HP04*	43.9	39.6	0.28	0.71
	HUB-P04-HP07	43.9	39.6	0.28	0.71
	HUB-P04-HP08	43.9	39.6	0.28	0.71
	HUB-P04-HP12	43.9	39.7	0.28	0.71
Haemophilus phage HUB-P08-ST107-01	HUB-P08-HI13	39.8	38.2	0.32	0.79
	HUB-P08-HI14	39.8	38.2	0.32	0.79
	HUB-P08-HI15	39.8	38.2	0.32	0.79
	HUB-P08-HI16	39.8	38.2	0.32	0.79
	HUB-P08-HI17	39.8	38.2	0.32	0.79
	HUB-P08-HI18	39.8	38.2	0.32	0.79
Haemophilus phage HUB-P08-ST107-02	HUB-P08-HI09*	42.3	38.1	0.61	0.8
	HUB-P08-HI10	42.3	38.1	0.61	0.74
Haemophilus phage HUB-P08-ST107-03	HUB-P08-HI13	44.4	38.2	0.44	0.79
	HUB-P08-HI14	44.4	38.2	0.44	0.79
	HUB-P08-HI15	44.4	38.2	0.44	0.79
	HUB-P08-HI16	44.4	38.2	0.44	0.79
	HUB-P08-HI17	44.4	38.2	0.44	0.79
	HUB-P08-HI18	44.4	38.2	0.44	0.79
Haemophilus phage HUB-P11-ST2480-01	HUB-P11-HI01	39.4	38.2	0.41	0.77
	HUB-P11-HI02*	39.4	38.2	0.41	0.77
	HUB-P11-HI03	39.4	38.2	0.41	0.77
	HUB-P11-HI04	39.4	38.2	0.41	0.77
	HUB-P11-HI05	39.5	38.2	0.41	0.77
	HUB-P11-HI06	39.5	38.2	0.41	0.77
	HUB-P11-HI07	39.5	38.2	0.41	0.77
Haemophilus phage HUB-P11-ST2480-02	HUB-P11-HI01	39.5	38.2	0.26	0.77
	HUB-P11-HI02*	39.5	38.2	0.26	0.77
	HUB-P11-HI03	39.5	38.2	0.26	0.77
	HUB-P11-HI04	39.5	38.2	0.26	0.77
	HUB-P11-HI05	39.5	38.2	0.26	0.77
	HUB-P11-HI06	39.5	38.2	0.26	0.77
	HUB-P11-HI07	39.5	38.2	0.26	0.77
Haemophilus phage HUB-P11-ST147-01	HUB-P11-HI09*	40.4	38.1	0.6	0.78
	HUB-P11-HI011	40.4	38.1	0.6	0.78

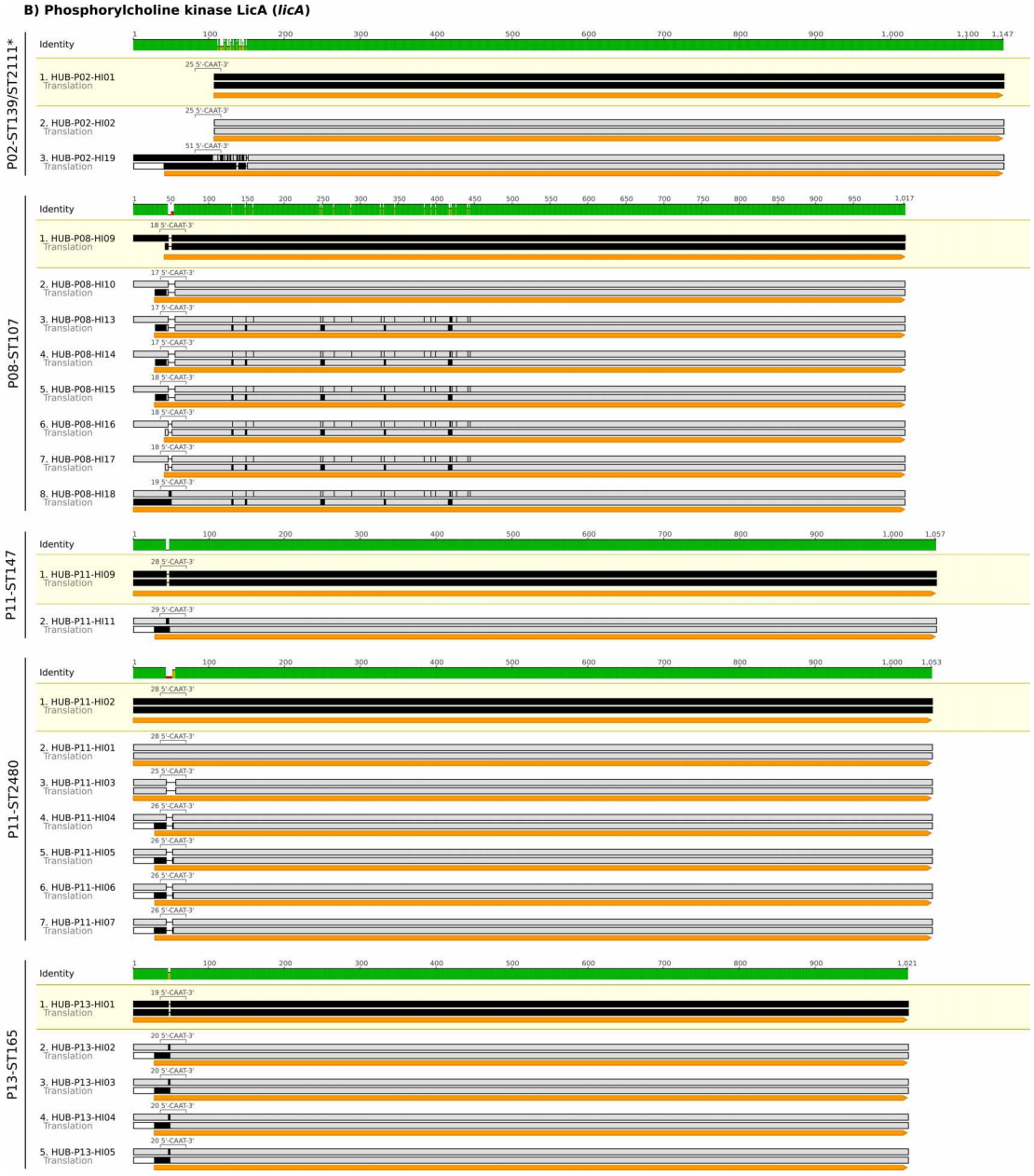
Supplementary Dataset S1, S2, S3, and S4 are available on request.



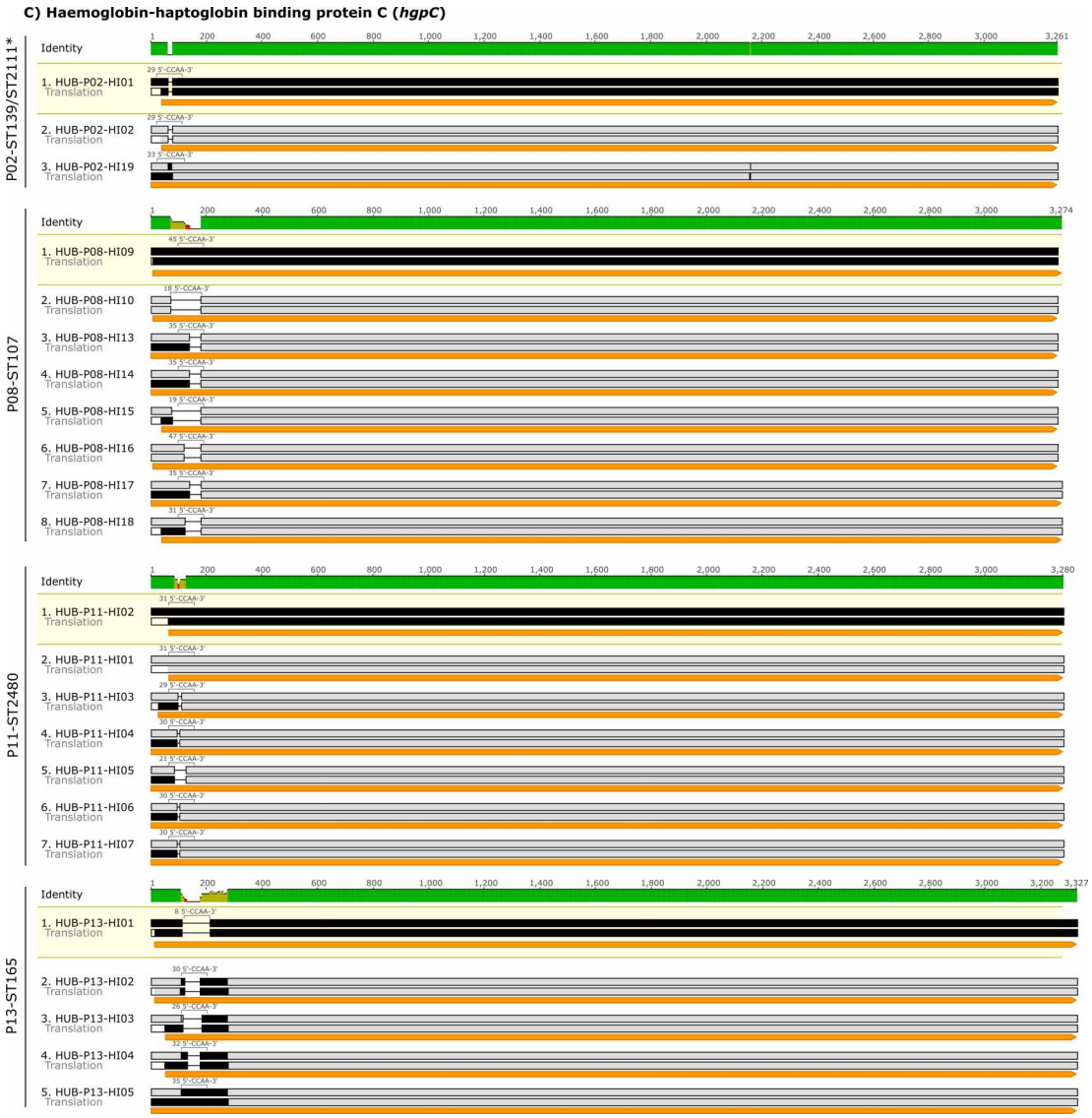
Supplementary Figure S1. Timeline illustrating the microorganisms isolated from respiratory samples in the severe COPD patients with long-term azithromycin therapy included in this study. The vertical shaded regions represent the time intervals between programmed visits: prior to start azithromycin therapy (V1), three (V2), and twelve (V3) months after starting therapy.



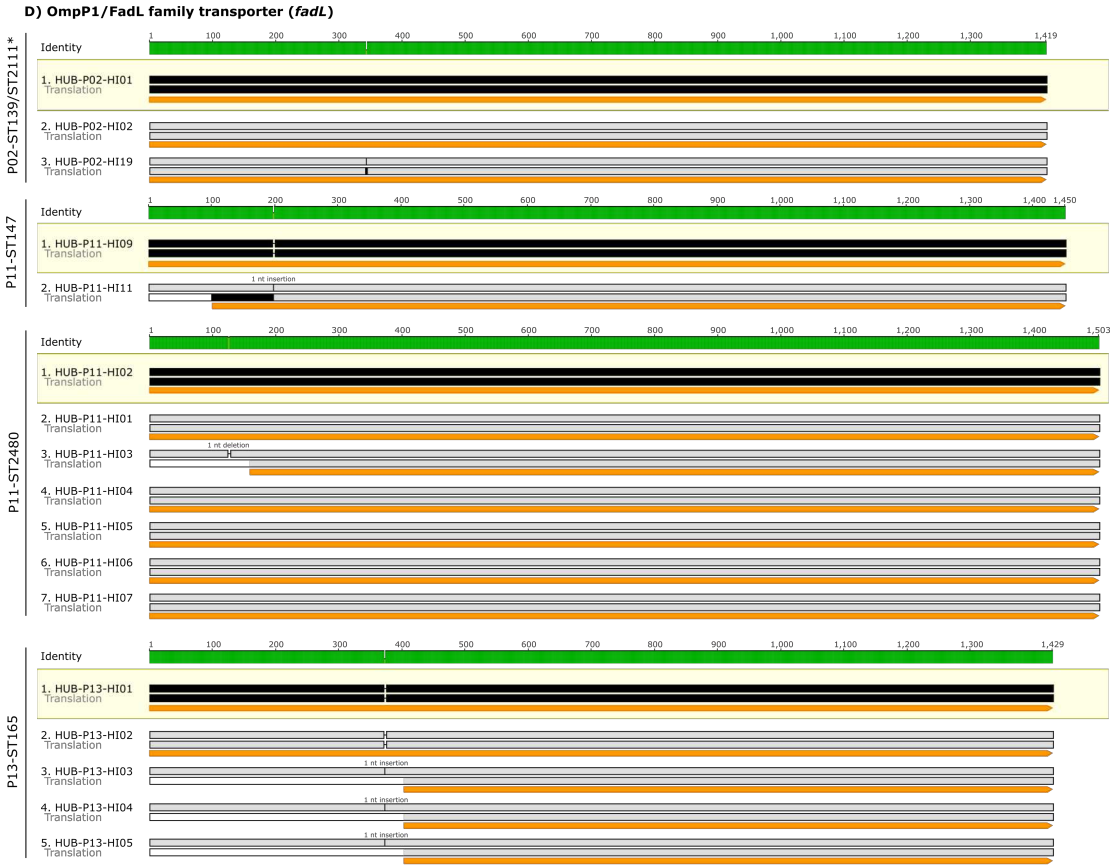
Supplementary Figure S2 (continued to next page).



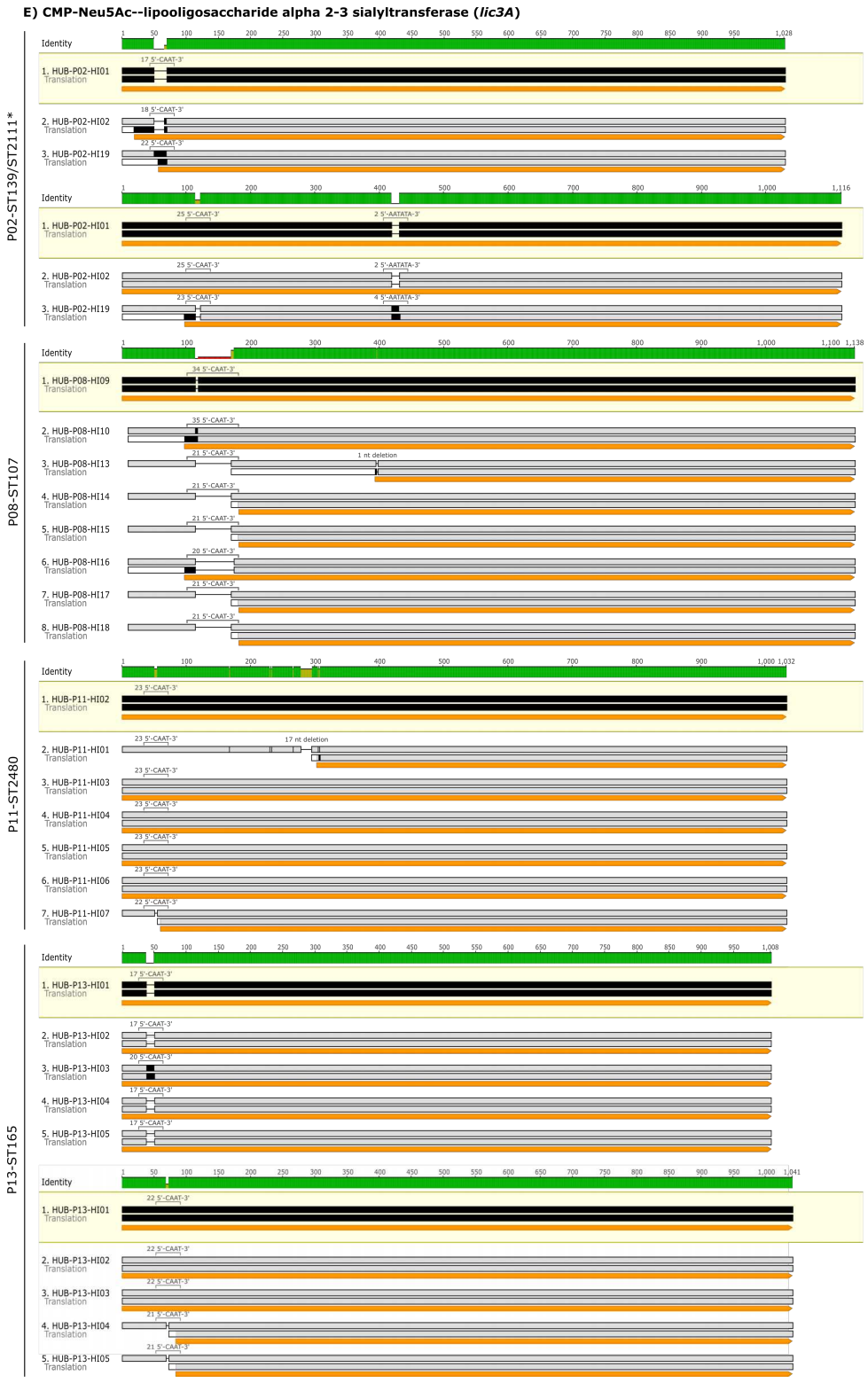
Supplementary Figure S2 (continued to next page).



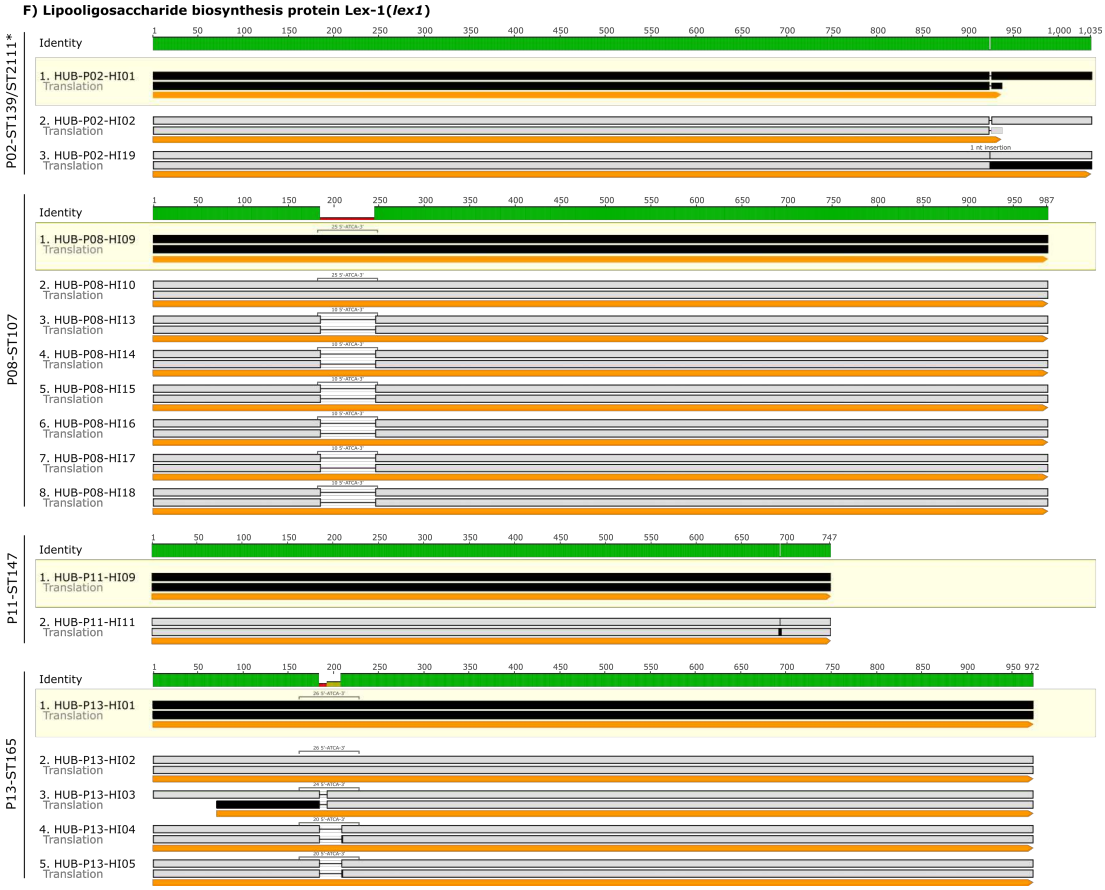
Supplementary Figure S2 (continued to next page).



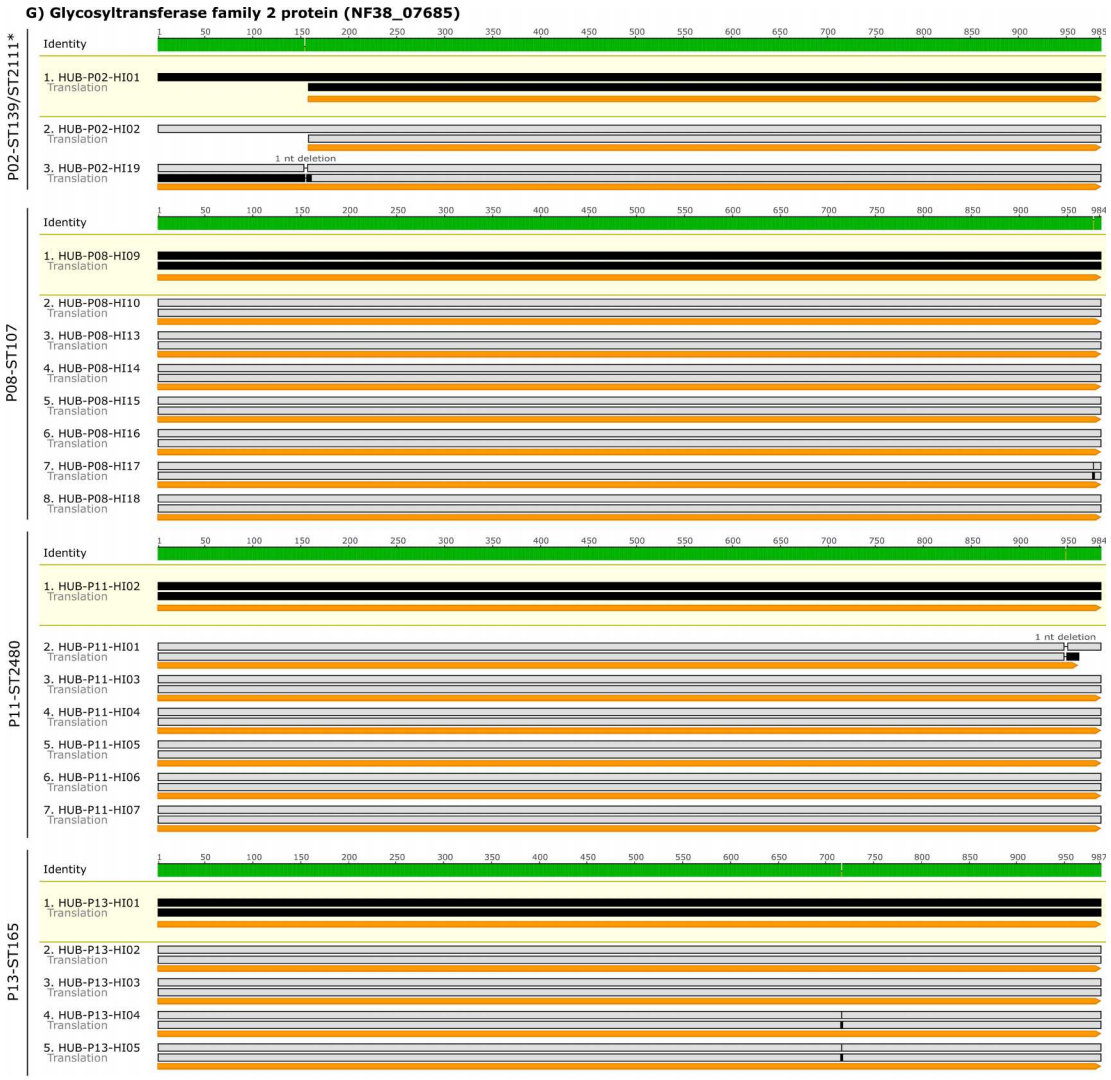
Supplementary Figure S2 (continued to next page).



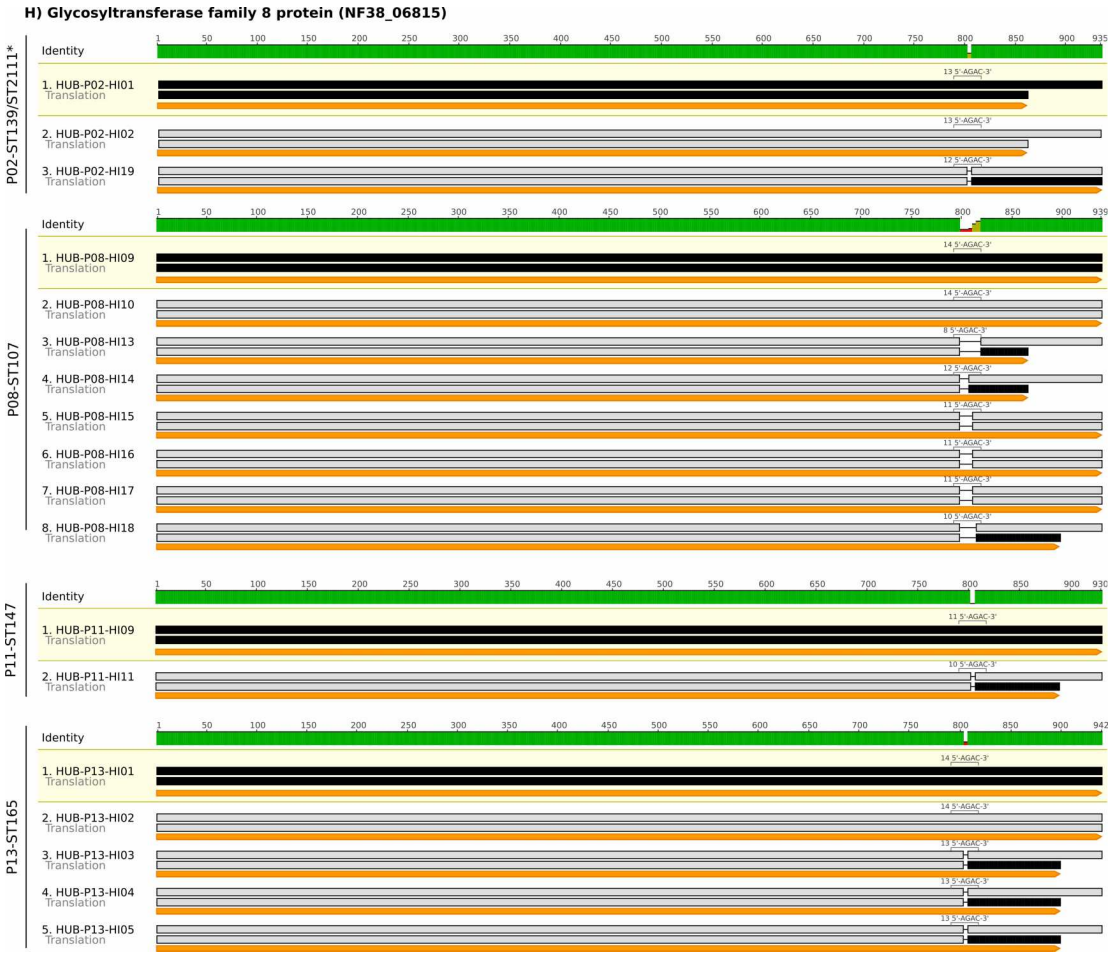
Supplementary Figure S2 (continued to next page).



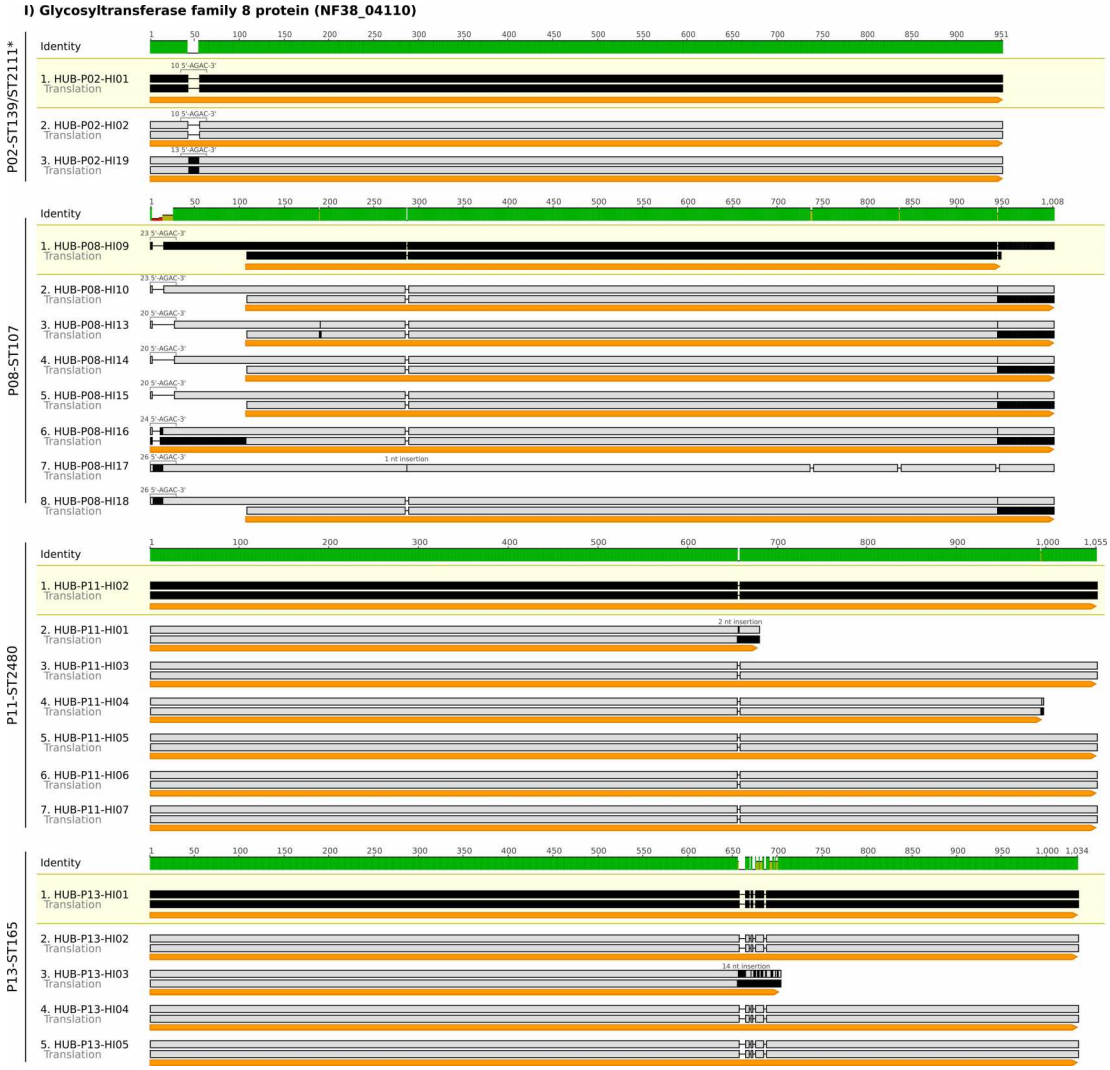
Supplementary Figure S2 (continued to next page).



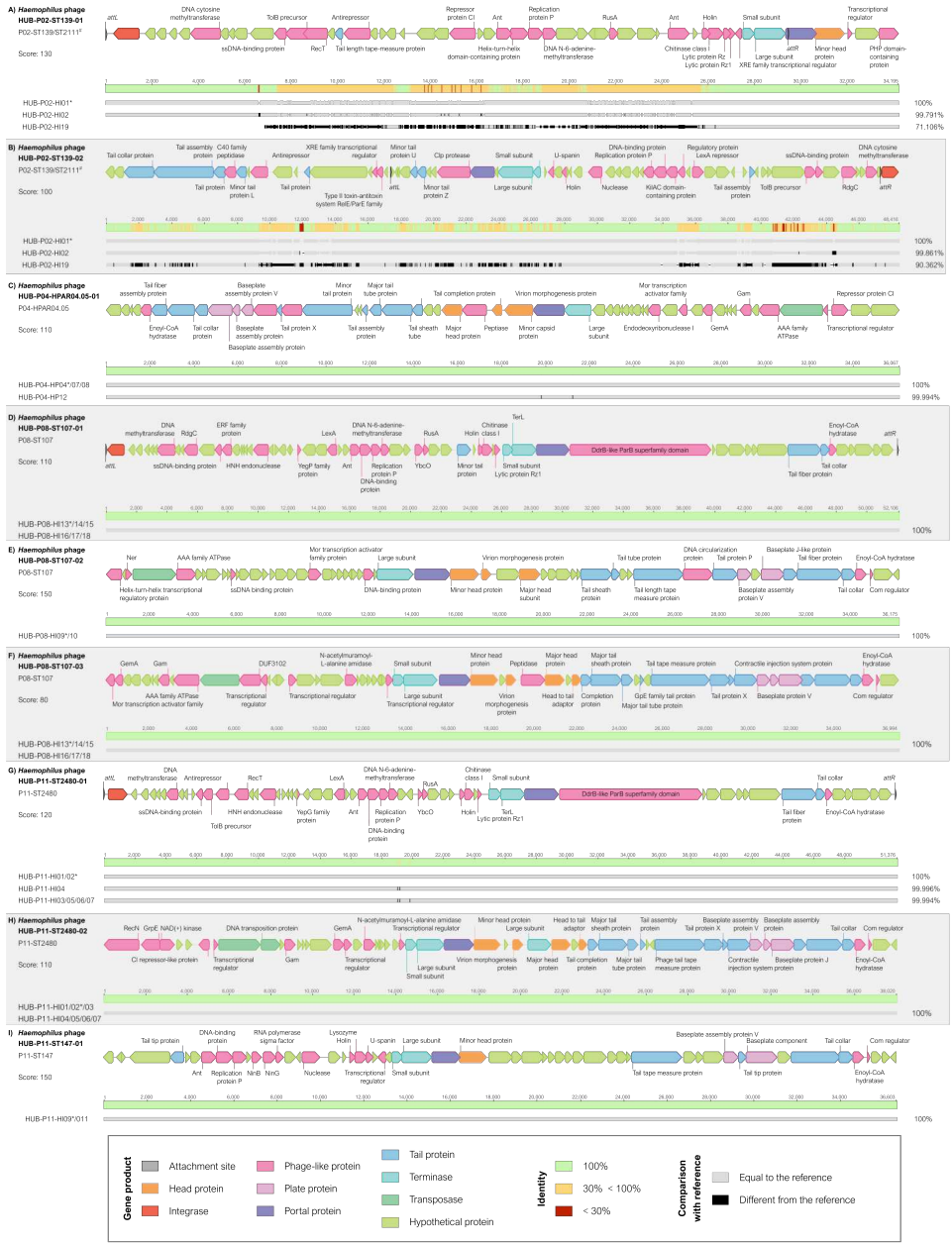
Supplementary Figure S2 (continued to next page).



Supplementary Figure S2 (continued to next page).



Supplementary Figure S2. Alignment of genes that showed genetic changes in four or more cases of *H. influenzae* persistence. (A) *hgpB*, coding by hemoglobin-haptoglobin binding protein B; **(B)** *licA*, coding by phosphorylcholine kinase LicA; **(C)** *hgpC*, coding by hemoglobin-haptoglobin binding protein C; **(D)** *fadL* (*ompP 1*), coding by an outer membrane transporter ; **(E)** *lic 3A*, coding by CMP-Neu5Ac--lipooligosaccharide alpha 2-3 sialyltransferase ; **(F)** *lex 1*, coding for lipooligosaccharide biosynthesis protein Lex-1; and **(G-I)** genes coding by glycosyltransferase family 2 and 8 proteins. The reference strains of each persistence case are highlighted in yellow. The numbers above the identity bar show the gene position (nt). The aligned gene, its translation and potential open reading frames (orange arrows) are shown for each strain. Changes from the reference strain are indicated in black. *SLV, single-locus variant.



Supplementary Figure S3. Schematic representation of genome-integrated prophages detected in persistent *Haemophilus* spp. Intact prophages detected by Phaster (score ≥ 90) were included. *Haemophilus* phage HUB-P08-ST107-03 was included despite having a score < 90 because strains HUB-P08-HI13 to HUB-P08-HI18 showed a 36 kb insertion compared to the reference strain (HUB-P08-HI09). The direction of the arrows indicates the direction of the reading frame of the genes. The first prophage found in a closed genome over time in each case was used as a reference (*), and the percentage of identity of each sequence with respect to the reference is indicated to the right of each horizontal grey line. #SLV, single-locus variant.

Study 4

Staphylococcus aureus surface protein G (*sasG*) allelic variants: correlation between biofilm formation and their prevalence in methicillin-resistant *S. aureus* (MRSA) clones

Anna Carrera-Salinas, Aida González-Díaz, Daniel Antonio Vázquez-Sánchez, Mariana Camoez, Jordi Niubó, Jordi Càmara, Carmen Ardanuy, Sara Martí, M. Ángeles Domínguez

Dr. Mª Ángeles Domínguez Luzón, Associate Professor of the Department of Pathology and Experimental Therapeutics of the Universitat de Barcelona and Head of the Microbiology Department, Hospital Universitari de Bellvitge (Barcelona), and Dr. Sara Martí Martí, Assistant Professor of the Department of Medicine of the Universitat de Barcelona and Senior Postdoctoral Researcher at the Centre for Biomedical Research Network on Respiratory Diseases (CIBERes),

DECLARE that the original article entitled "***Staphylococcus aureus* surface protein G (*sasG*) allelic variants: correlation between biofilm formation and their prevalence in methicillin-resistant *S. aureus* (MRSA) clones**" has been published in Research in Microbiology (Impact factor in 2021: 3.946). The contribution of Anna Carrera-Salinas is detailed below:

- Experimental work.
- Data analysis.
- Manuscript redaction.

Signatures of thesis supervisors

MªÁngeles Domínguez Luzón **Sara Martí Martí**

Research in Microbiology 173 (2022) 103921

Contents lists available at ScienceDirect

Research in Microbiology

journal homepage: www.elsevier.com/locate/resmic

Original Article

Staphylococcus aureus surface protein G (*sasG*) allelic variants: correlation between biofilm formation and their prevalence in methicillin-resistant *S. aureus* (MRSA) clones

Anna Carrera-Salinas [a], Aida González-Díaz [a, b], Daniel Antonio Vázquez-Sánchez [a, c], Mariana Camoez [a, c], Jordi Niubó [a], Jordi Càmara [a, b], Carmen Ardanuy [a, b, d], Sara Martí [a, b, *], M Ángeles Domínguez [a, c, d, **], on behalf of REIPI/GEIH Study Groups

[a] Department of Microbiology, Hospital Universitari de Bellvitge, Institut D'Investigació Biomèdica de Bellvitge (IDIBELL), Barcelona, Spain
[b] Research Network for Respiratory Diseases (CIBERES), ISCIII, Madrid, Spain
[c] Research Network for Infectious Diseases (CIBERINFEC), ISCIII, Madrid, Spain
[d] Department of Pathology and Experimental Therapeutics, College of Medicine, Universitat de Barcelona, Barcelona, Spain

ARTICLE INFO

Article history:
Received 11 November 2021
Accepted 14 January 2022
Available online 26 January 2022

Keywords:
SasG protein
Staphylococcus aureus
Biofilms
Sepsis

ABSTRACT

Methicillin-resistant *Staphylococcus aureus* (MRSA) may persist for long periods due to biofilm formation. The objective of this study was to describe biofilm formation in association with the presence of *S. aureus* surface protein G (*sasG*) and its allelic variants in MRSA bacteraemia isolates from endemic (CC5, CC8, CC22) and sporadic clones in Spain (2008–2015). Crystal violet staining was used to assess biofilm formation; DNA microarray, RT-qPCR, and long-read whole genome sequencing were applied to determine the presence, expression and structure of *sasG*, respectively. The endemic CC5 and CC8 clones produced more biofilm than the sporadic clones; these endemic clones carried *sasG* allelic variant 1. Otherwise, sporadic clones, with less biofilm formation, showed either an absence of *sasG* (65%) or the presence of allelic variant 2 (35%). Variants 1 and 2 differed in the expression of *sasG* (1.56 ± 1.20 and 0.37 ± 0.32, respectively). The analysis of a large cohort of closed *S. aureus* genomes available on the NCBI database confirmed the distribution of the two allelic variants with low amino acid identity (68.1%) among endemic and sporadic clones. SasG variant 1 present in the major CC5 and CC8 clones was correlated with increased biofilm formation and may represent an important virulence determinant.

1. Introduction

Staphylococcus aureus is a major cause of nosocomial bloodstream infections worldwide. Compared to other pathogens, *S. aureus* causes infections that are associated with longer hospitalisations, higher treatment costs, and greater morbidity and mortality rates. The high level of methicillin resistance detected in *S. aureus* bacteraemia isolates represents a therapeutic challenge that is directly related to poorer outcomes [1,2]. Although the prevalence of methicillin resistance in *S. aureus* bacteraemia isolates has decreased over time, it remains high in Southern Europe, accounting for more than 25% of the invasive isolates reported in Spain [3].

Protective mechanisms, such as biofilm formation, help bacteria to persist in the environment or within the human host [4]. Biofilm provides an optimal environment for the exchange of genetic materials and confers tolerance to many disinfectants, antimicrobials, and host defences [5]. The attachment of biofilms to hospital surfaces or medical equipment together with the ineffectiveness of biocides contribute to the transmission of hospital-acquired infections [6]. In addition, biofilm attachment to indwelling devices, such as prosthetic implants, catheters, and cardiac pacemakers,

* Corresponding author. Department of Microbiology, Hospital Universitari de Bellvitge, Institut D'Investigació Biomèdica de Bellvitge (IDIBELL), Barcelona, Spain.
** Corresponding author. Spanish Network for Research in Infectious Diseases (REIPI), ISCIII, Madrid, Spain.
E-mail addresses: acarrera@idibell.cat (A. Carrera-Salinas), agonzalezd@bellvitgehospital.cat (A. González-Díaz), dvazquez@idibell.cat (D.A. Vázquez-Sánchez), mariana.camoez81@gmail.com (M. Camoez), jniubo@bellvitgehospital.cat (J. Niubó), jcamara@bellvitgehospital.cat (J. Càmara), c.ardanuy@bellvitgehospital.cat (C. Ardanuy), smartinm@bellvitgehospital.cat (S. Martí), adominguez@bellvitgehospital.cat (M.Á. Domínguez).

https://doi.org/10.1016/j.resmic.2022.103921
0923-2508/

serves as a reservoir for the spread of bacteria throughout the human body, promoting bloodstream infections and increasing the complexity of biofilm eradication, which usually requires surgical removal of the foreign body [7,8].

S. aureus expresses a wide range of virulence factors that lead to biofilm formation and, as a result, successful colonisation, survival, and the development of subsequent infections. Polysaccharide intercellular adhesins (PIA) synthesised by *ica* operon-encoded enzymes have been associated with biofilm formation [9]. However, an *ica*-independent pathway mediated by staphylococcal cell wall-anchored proteins, such as biofilm-associated protein (Bap), clumping factor B (ClfB), fibronectin-binding proteins (FnBPs), and *S. aureus* surface protein G (SasG), has also been reported to promote biofilm production [10]. In particular, SasG, which is related to biofilm formation and nasal colonization, has an A domain in the N-terminus, followed by a B domain formed by G5-E repeats, and a cell wall anchoring domain in the C-terminal [11]. The A domain contains a highly conserved region as well as a variable region that allows for the differentiation of two different variants [12,13].

Since biofilm-associated methicillin-resistant *S. aureus* (MRSA) bloodstream infections represent a therapeutic burden, in-depth studies are needed to identify and understand the mechanisms underlying biofilm formation, which will be critical in the development of new therapeutic strategies. The MRSA population has been extensively studied using multilocus sequence typing (MLST), which assigns sequence types (ST) based on the allelic profile obtained for seven housekeeping genes. In this context, strains that share at least five of the seven loci belong to the same clonal complex (CC). Based on an MRSA collection of clinical isolates from patients with bloodstream infections, this study aimed to link the differences in biofilm formation between the most prevalent clones (CC5, CC8, and CC22) considered to be endemic and the less frequent or sporadic MRSA bacteraemia clones in Spain [14] to the *sasG* allelic variants. Our results showed that SasG, in particular variant 1, is a virulence factor that showed a positive correlation with biofilm production in the most common Spanish MRSA clones. Its role should be considered when developing measures to prevent the spread of these clones.

2. Materials and methods

2.1. Study design and bacterial characterisation

This was a retrospective laboratory-based multicentre study that included MRSA strains from a cohort of patients with bloodstream infections admitted to hospitals belonging to the Spanish Network for Research in Infectious Diseases (REIPI) between 2008 and 2015.

Clinical MRSA isolates were selected based on MLST, including representative endemic (n = 60) and sporadic (n = 20) clones (Supplementary Dataset 1). MRSA clones with higher prevalence in bloodstream infections were considered endemic (CC5, CC8, and CC22), whereas clones from other clonal complexes (CC1, CC30, CC45, CC72, CC88, and CC398) were classified as sporadic because they were rarely isolated from bacteraemia [14].

Microarray genotyping was used to identify the virulence factors associated with adhesion and biofilm formation. DNA extraction was performed with the QIAamp DNA Mini Kit (Qiagen, Germany), followed by genotyping with the *S. aureus* Genotyping Kit 2.0 (Alere Technologies GmbH, Germany) according to the manufacturer's instructions. Isolated and biotin-labelled ssDNA was analysed using the ArrayMate device (Alere Technologies GmbH, Germany) to determine the presence or absence of *S. aureus* genes [15].

2.2. Biofilm assay

Biofilm formation was assessed by crystal violet staining. Bacterial cultures were adjusted to a turbidity of 1 McFarland, diluted 1:1000 in 200 μL of BHI medium supplemented with 1% (w/v) glucose, and incubated statically at 37 °C for 24 h in 96-well advanced TC surface polystyrene plates (Greiner Bio-One, Austria). All experiments were performed in triplicate and at least at two independent time points. Before biofilm staining, the A_{600} was measured to determine bacterial growth. The culture broth was removed, the wells were washed twice with distilled water to eliminate planktonic bacteria, and the biofilm was dried at 50 °C for 30 min. The biofilm was stained with 150 μL of 0.5% crystal violet at room temperature for 20 min and washed with distilled water twice to remove unbound dye. Finally, the dye was dissolved in 150 μL of 95% ethanol and the A_{570} was measured. The cut-offs for biofilm formation were A_{570} values that were 5, 10 or 20 times the A_{570} value of the negative control (BHI medium). The mean biofilm formation values of each strain were used to classify them as very strong (A_{570} value > 20 times that of the negative control), strong (A_{570} value 10 to 20 times that of the negative control), weak (A_{570} value 5 to 10 times that of the negative control), or non-biofilm producers (A_{570} value < 5 times that of the negative control); to compare biofilm formation between endemic and sporadic clones; and to compare biofilm production in strains with or without *sasG* and its allelic variants.

Biofilm detachment was assessed in biofilms grown for 24 h, as previously described [16]. Briefly, biofilms were washed with distilled water, treated with 100 μg/mL proteinase K or 10 mM sodium metaperiodate, at 37 °C for 2 h. Treated biofilms were washed and stained with crystal violet.

2.3. sasG gene expression

sasG expression was determined by quantitative reverse transcription PCR (RT-qPCR) in 18 clinical MRSA isolates selected to represent the two *sasG* allelic variants (nine of each) previously defined [12,13,17] and with different degrees of biofilm formation. Briefly, bacterial cultures grown in BHI for 24 h at 37 °C and 400 rpm were stabilised using RNAlater (Invitrogen, Thermo Fisher Scientific, USA). Total RNA purification was performed using the RNeasy Mini Kit (Qiagen, Germany) and the RNase-Free DNase Set (Qiagen, Germany), following the manufacturer's instructions. Reverse transcription was performed with the QuantiTect Reverse Transcription Kit (Qiagen, Germany), followed by qPCR on a QuantStudio 5 Real-Time PCR system (Applied Biosystems, Thermo Fisher Scientific, USA) prepared with the QuantiTect SYBR Green PCR Kit (Qiagen, Germany), using *sasG* (*sasG*-F: 5′-GAGCCTTCATCAACAGAACGA-3'; *sasG*-R: 5′-AACAAAGACACTGCACCACAA-3′) and 16S rRNA primers (16SrRNA-F: 5′-CTCGTGTCGTGAGATGTTGG-3'; 16SrRNA-R: 5′-TTCGCTGCCCTTTGTATTGT-3′). Expression was determined using the $2^{-\Delta\Delta Ct}$ method and the results are expressed as the logarithmic value of $2^{-\Delta\Delta Ct}$. 16S rRNA was used as the housekeeping gene and the strain with the lowest *sasG* expression was selected as the reference.

2.4. sasG gene sequencing

sasG allele and structure were characterised in all the clinical MRSA bacteraemia isolates by long-read whole genome sequencing using MinION (Oxford Nanopore Technologies, UK) according to the manufacturer's instructions. Libraries were prepared using the Native Barcoding Expansion (EXP-NBD196) and the Ligation Sequencing Kit (SQK-LSK109) and loaded into FLO-MIN106D flow cells (R9.4.1) for a 48-h run. Basecalling was performed with

MinKNOW (version 4.1.22) using the option to trim the sequencing adapters. Assembly was conducted using the Unicycler pipeline [18]. *sasG* sequences were analysed with Geneious R9 (Biomatters, New Zealand), using NCTC8325 (NC_007795) and 24117-WT (NZ_LT996891) as references for allelic variants 1 and 2, respectively. The *sasG* sequences were aligned with the reference sequences, and the allelic variant was determined based on the identity with the A domain's variable region, as previously described [12,13]. G5-E repeats were defined based on the Foster classification [10].

2.5. Closed genome analysis

A total of 608 closed *S. aureus* genomes deposited on the NCBI database were downloaded on the 24th of March 2021. *In silico* MLST was undertaken using the MLST software v2.4 (github.com/tseemann/mlst). Strains were classified into clonal complexes, defined as sequence types (STs) sharing at least five of seven alleles. Genomes from undescribed clonal complexes and those not identified among the Spanish endemic and sporadic MRSA bacteraemia isolates (n = 137) were excluded from the study. In-depth analysis of *sasG* was performed with Geneious R9 (Biomatters, New Zealand), using reference strains for the two different alleles: the NCTC8325 strain (NC_007795) for allelic variant 1, and the 24117-WT strain (HYPERLINK "https://www.ncbi.nlm.nih.gov/nuccore/NZ_LT996891" \o "https://www.ncbi.nlm.nih.gov/nuccore/NZ_LT996891"NZ_LT996891) for allelic variant 2. Sequence alignment of NCTC8325 (HYPERLINK "https://www.ncbi.nlm.nih.gov/nuccore/NC_007795" \o "https://www.ncbi.nlm.nih.gov/nuccore/NC_007795"NC_007795) and 24117-WT (NZ_LT996891) for allelic variants 1 and 2, respectively, was used to determine the percentage of identity between the two allelic variants using Clustal Omega [19].

2.6. Statistical analysis

Statistical analyses were carried out with the GraphPad Prism 5 software, using non-parametric Mann–Whitney test or Fisher's exact test, when appropriate. P-values < 0.05 were considered statistically significant.

2.7. Ethical statement

This study was in accordance with the Declaration of Helsinki from the World Medical Association. It was approved by the Clinical Research Ethics Committee of Bellvitge University Hospital (PR183/21). Written informed consent was not required as this was a retrospective and observational study with isolates obtained as part of routine microbiological tests. Patient confidentiality was always protected, and all personal data were anonymised following the current legal normative in Spain (LOPD 15/1999 and RD 1720/2007). Moreover, this project followed Law 14/2007 on Biomedical Research for the management of biological samples in clinical research.

3. Results

3.1. Characterisation of MRSA bacteraemia isolates

DNA microarray provided information on the presence or absence of *S. aureus* target sequences, including species-specific, antimicrobial resistance, and virulence-associated genes, along with other genes involved in adhesion and biofilm formation (Fig. 1). The CC8, CC22, CC45, CC72, and CC398 clones were related to *agr* type I, the CC5 clones were *agr* type II, while the CC1, CC30, and CC88 clones were *agr* type III. All the strains tested carried the staphylococcal accessory regulator gene (*sarA*), the intracellular adhesion (*ica*) operon, the clumping factor A and B genes (*clfA* and *clfB*), the fibronectin-binding protein A gene (*fnbA*), the serine-aspartate repeat protein C gene (*sdrC*), and the von Willebrand binding protein gene (*vwb*). All of them were negative for the gene encoding the biofilm-associated protein (*bap*). The *S. aureus* surface protein G gene (*sasG*) was present in all the CC1, CC5, CC8, CC22, CC72, and CC88 strains.

3.2. Biofilm formation among endemic and sporadic MRSA clones

Twenty-two strains produced very strong biofilms ($mean_{A570} = 3.7$; SD = 0.2), 45 strains produced strong biofilms ($mean_{A570} = 2.7$; SD = 0.4), and 13 formed weak biofilms ($mean_{A570} = 1.2$; SD = 0.2), according to the established cut-offs for biofilm formation (5, 10 or 20 times the A_{570} value of the negative control).

After 24 h of growth on a static solid support, no significant differences in absorbance (A_{600}) were observed between endemic and sporadic clones (Supplementary Fig. S1), indicating that the variation in biofilm formation was not caused by differential bacterial growth. Despite this, when the clones were compared based on their clinical prevalence, different values of biofilm formation were obtained. Isolates from endemic clones (CC5, CC8, and CC22) produced significantly more biofilm than sporadic clones (CC1, CC30, CC45, CC72, CC88, and CC398), which had a lower biofilm production (p-value <0.0001) (Fig. 2A). Most of the endemic clones were strong and very strong biofilm producers (61.7% and 33.3%), in contrast to the sporadic clones (45% and 5%, respectively). Isolates belonging to CC22, an endemic clonal complex, showed biofilm production comparable to that of sporadic clones (p-value = 0.1255), producing less biofilm than the CC5 (p-value = 0.0328) and CC8 isolates (p-value = 0.0060) (Fig. 2B). Furthermore, strains were classified in STs based on the allelic profile of the seven housekeeping genes in the MLST scheme. Those belonging to ST22 and ST217 (CC22), ST30, ST435 and ST1870 (CC30), and ST398 (CC398) had the lowest biofilm production.

To determine whether biofilm production was protein- or sugar-dependent, biofilms were treated with proteinase K and sodium metaperiodate, respectively. The biofilms of all the tested isolates were sensitive to degradation by proteinase K but slightly more resistant to sodium metaperiodate (Fig. 3A and E), indicating that biofilm formation occurred primarily through an *ica*-independent mechanism.

3.3. Correlation between sasG and biofilm formation

The analysis of the 80 MRSA bacteraemia isolates showed that all the endemic strains and 35% of the sporadic strains (CC1, CC72, and CC88) carried *sasG* (Fig. 1). In addition, as shown in Fig. 4A, clones carrying *sasG* developed stronger biofilms than those without *sasG* (p-value = 0.0001).

Two *sasG* allelic variants have already been described: allelic variant 1 in the strains NCTC8325 (CC8), COL (CC8), and Mu50 (CC5); and allelic variant 2 in the strains MW2 (CC1) and SANGER-476 (CC1) [12,13,17]. Analysis of the *sasG* allelic variants revealed that the CC5 and CC8 strains carried variant 1, while the CC1, CC22, CC72, and CC88 strains carried variant 2 (Fig. 1). Allelic variant 1 was the most prevalent among the endemic clones (85%), while allelic variant 2 was the only one identified in the sporadic clones. When the *sasG* allelic variant was considered (Fig. 4B), strains with allelic variant 1 formed stronger biofilms than strains with allelic variant 2 (p-value = 0.0009) or strains without the *sasG* gene (p-value < 0.0001). Strains with allelic variant 2 tended to form stronger biofilms than strains without the gene (p-value = 0.0624). The presence of allelic variant 2 in CC22 could explain its low

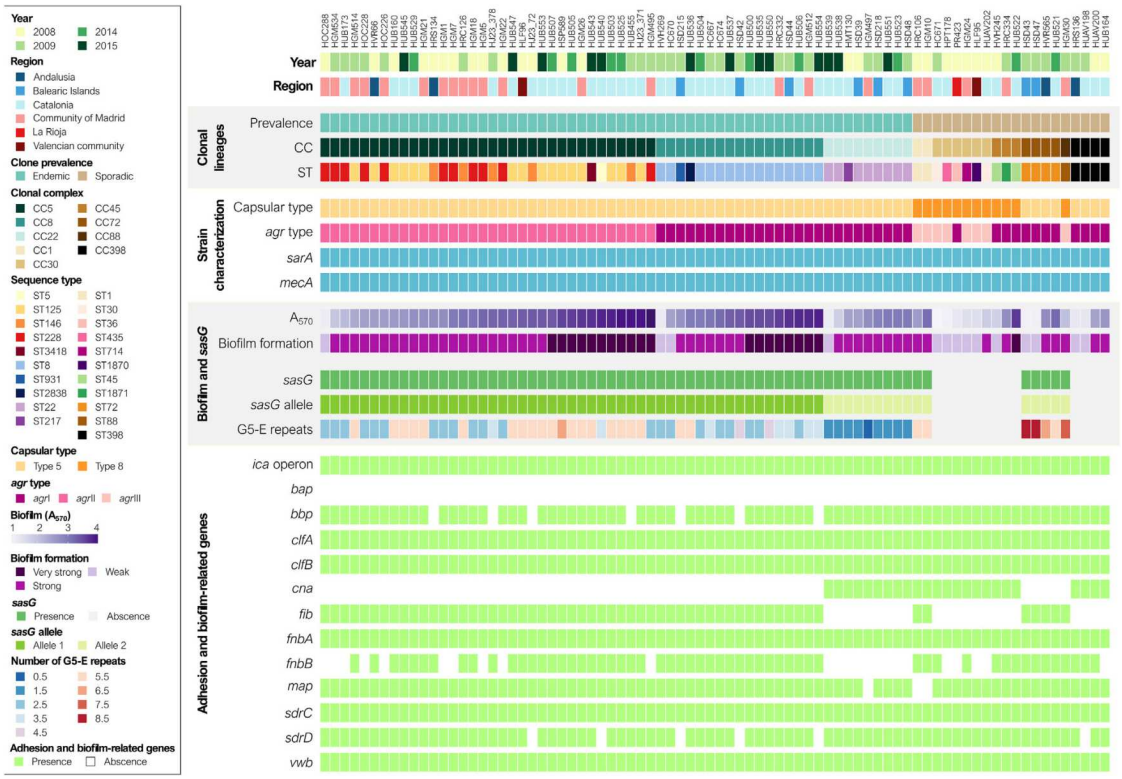


Fig. 1. Characterisation of MRSA bacteraemia isolates. Strains were grouped by the prevalence of the clone (endemic or sporadic) and by the clonal complex (CC). Biofilm formation after 24 h was measured by crystal violet absorbance at 570 nm (A_{570}). According to the established cut-offs, the strains were classified as very strong (A_{570} value > 20 times that of the negative control), strong (A_{570} value 10 to 20 times that of the negative control), weak (A_{570} value 5 to 10 times that of the negative control), or non-biofilm producers (A_{570} value < 5 times that of the negative control). Abbreviations: *cap*, capsular operon; *agr*, accessory gene regulator; *sarA*, staphylococcal accessory regulator; *mecA*, methicillin-resistance gene; *ica*, intracellular adhesion operon; *bap*, biofilm-associated protein gene; *bbp*, bone sialo-binding protein gene; *clfA*, clumping factor A gene; *clfB*, clumping factor B gene; *cna*, collagen adhesion gene; *fib*, fibrinogen-binding protein gene; *fnbA*, fibronectin-binding protein A gene; *fnbB*, fibronectin-binding protein B gene; *map*, extracellular adhesion protein gene; *sasG*, *S. aureus* surface protein G gene; *sdrC*, serine-aspartate repeat protein C gene; *sdrD*, serine-aspartate repeat protein D gene; and *vwb*, von Willebrand binding protein gene. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

biofilm production compared to the other endemic clonal complexes, which carried allelic variant 1. Furthermore, biofilms were found to be more sensitive to degradation by proteinase K than to sodium metaperiodate in all strains, regardless of the presence of absence of *sasG* or allelic variant, indicating protein-dependent biofilm formation (Fig. 3B–D and F-H). The number of G5-E repeats did not show any correlation with biofilm formation.

The expression of the *sas*G gene was assessed in 18 MRSA bacteraemia isolates, nine for each allele, to determine if the observed differences in biofilm formation could be explained by different levels of expression of the allelic variants. Despite strain-specific variation, gene expression was higher in strains with allelic variant 1 (mean = 1.6; SD = 1.2) than in those with allelic variant 2 (mean = 0.4; SD = 0.3) (p-value = 0.0104) (Supplementary Fig. S2). Among those expressing variant 1, gene expression was higher in the CC5 strains (mean = 2.2; SD = 1.3) than in the CC8 strains (mean = 0.8; SD = 0.2) (p-value = 0.0666).

3.4. Detection of sasG and its allelic variants in closed S. aureus genomes

Our collection included a small number of sporadic strains due to the low prevalence of these clonal complexes in bacteraemia. For this reason, an analysis of the closed *S. aureus* genomes available on the NCBI database, regardless of methicillin susceptibility, was conducted to confirm the differences in the distribution of *sasG* allelic variants between endemic and sporadic strains (Supplementary Datasets 2 and 3). Strains belonging to clonal complexes with a high prevalence in bloodstream infections in Spain were classified as endemic (CC5, CC8, and CC22), whereas those belonging to less frequent clonal complexes were classified as sporadic (CC1, CC30, CC45, CC72, and CC398). As a result, 309 strains (65.6%) were assigned to endemic clonal complexes and 162 (34.4%) were assigned to the sporadic group. The *sasG* gene was detected in 74.9% of the strains (n = 353), specifically in 99.0% of the endemic clones and in 29.0% of the sporadic clones. When considering the allelic variants, variant 1 was predominant in the endemic clones (94.2%), whereas allelic variant 2 was present in all the sporadic clones. All the CC5 and CC8 clones were associated with allelic variant 1, except for three ST6 (CC5) strains that carried allelic variant 2 and three CC8 strains in which the gene was not detected. All the endemic CC22 and the sporadic CC72 and CC88 clones carried allelic variant 2 (Fig. 5). These data confirmed the *sasG* distribution observed in our MRSA bloodstream collection.

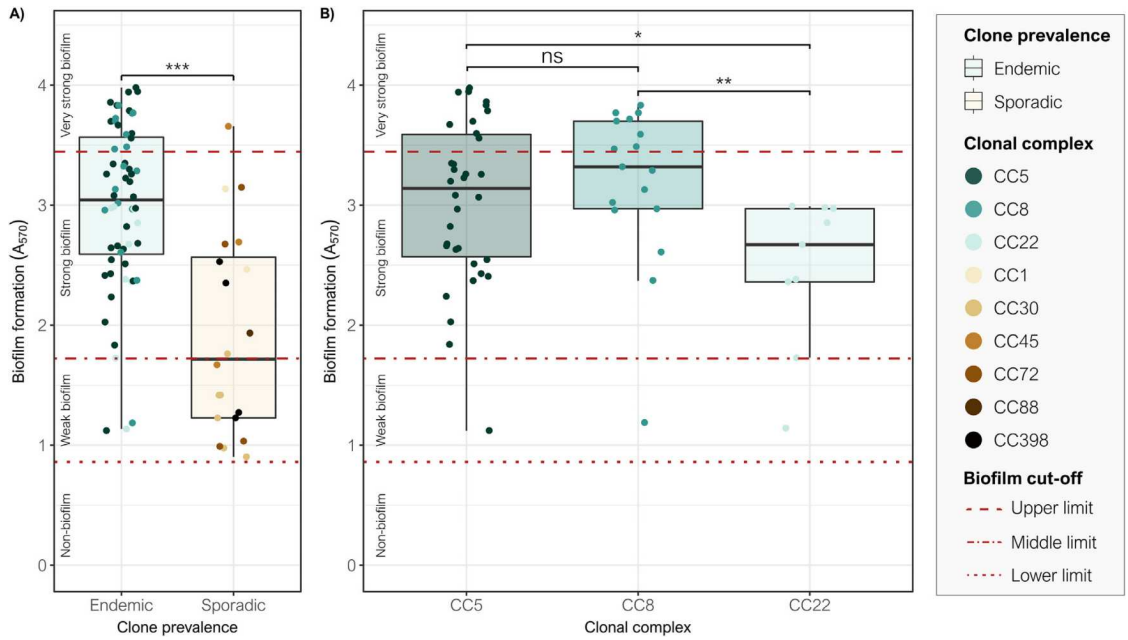


Fig. 2. Biofilm formation after 24 h assessed by crystal violet absorbance at 570 nm (A_{570}). (A) Comparison of biofilm formation between endemic and sporadic MRSA bacteraemia isolates. (B) Comparison of biofilm formation between endemic clonal complexes of MRSA bacteraemia isolates. Each dot represents the mean growth or biofilm formation of each strain obtained after at least three replicates at two independent time points. Each box plot shows the median, upper and lower quartiles (boxes) and the 1.5 x IQR (interquartile range) values (vertical lines) for each group. Horizontal dashed, dashed-dotted and dotted red lines represent the established cut-offs to classify the strains as very strong (A_{570} value > 20 times that of the negative control), strong (A_{570} value 10 to 20 times that of the negative control), weak (A_{570} value 5 to 10 times that of the negative control), or non-biofilm producers (A_{570} value < 5 times that of the negative control). ns, p-value > 0.05; * p-value < 0.05; ** p-value < 0.01; *** p-value < 0.001. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

3.5. Comparison between the sasG allelic variants

Gene analysis showed that, regardless of the allelic variant, *sasG* from all the strains of the Spanish MRSA bloodstream collection and the NCBI closed genomes had the same structure: an N-terminal secretion signal, followed by an A domain, a B domain with G5-E repeats, and a C-terminal cell wall anchor domain (Fig. 6). However, the length of the gene differed between the strains due to the number of G5-E repeats, which ranged from 0 to 9.5.

The percentage of identity for allelic variants 1 and 2 was assessed using two representative strains (NCTC8325 and 24117-WT, respectively), with the same number of G5-E repeats. The overall percentage of identity between their protein sequences was 68.1%. The N-terminal secretion signal and the C-terminal cell wall anchor domain were highly conserved between the two variants, showing 100% and 98.3% identity, respectively (Fig. 6). The A domain included a well-conserved region (residues 51–181) with an identity of 100% and a variable region (residues 182–430) with an identity of 61.8% between the two alleles. In this case, the B domain had seven variable G5-E repeats of 128 residues each and a truncated eighth G5-E repeat of only 68 residues. The percentage of identity of the repeats ranged from 55.5% to 56.3%. The final partial G5-E repeat showed an identity of 60.3% between the SasG variants.

Furthermore, the percentage of identity of the A domain from all the NCBI closed genomes was also determined, as this is the area used to identify the two allelic variants. The results showed that the A domain sequences of allelic variant 1 had an identity of 86.9%, while variant 2 had an identity of 99.6% (Table 1). The low percentage identity observed in the A domain of allelic variant 1 was explained by its variable region, which had an identity of 78.2%. Further investigation revealed that the drop in the percentage of identity was caused primarily by the presence of a premature codon stop in a few strains (n = 39), the majority of which belonged to ST239.

4. Discussion

The persistence of *S. aureus* due to biofilm formation contributes to the transmission and spread of hospital-acquired infections [20,21]. The goal of this genomic study is to correlate the differences in biofilm formation between endemic and sporadic MRSA clones in Spain with the distribution of *sasG* allelic variants. As a result, it was established a connection between the increased biofilm production in endemic CC5 and CC8 clones and the presence of *sasG* allelic variant 1.

The endemic CC5 and CC8 MRSA clones, which are the most frequent in bloodstream infections in Spain [14], developed stronger biofilms than the sporadic isolates, such as CC1, CC30, CC45, CC72, CC88, and CC398. Previous studies have supported this observation by demonstrating that the CC5 and CC8 strains form stronger biofilms than strains from other clonal complexes [22–25]. This suggests that the ability to form strong biofilms in endemic clones may be an important virulence factor, helping the bacteria survive in the host or in the hospital environment and transforming them into successful pathogens. However, despite being an endemic clone, CC22 formed biofilms at the same rate as sporadic clones, a trend that Smith and colleagues [26] had previously observed. They found that biofilm production in the most prevalent MRSA clone in Scotland, EMRSA-15 (CC22), was stronger

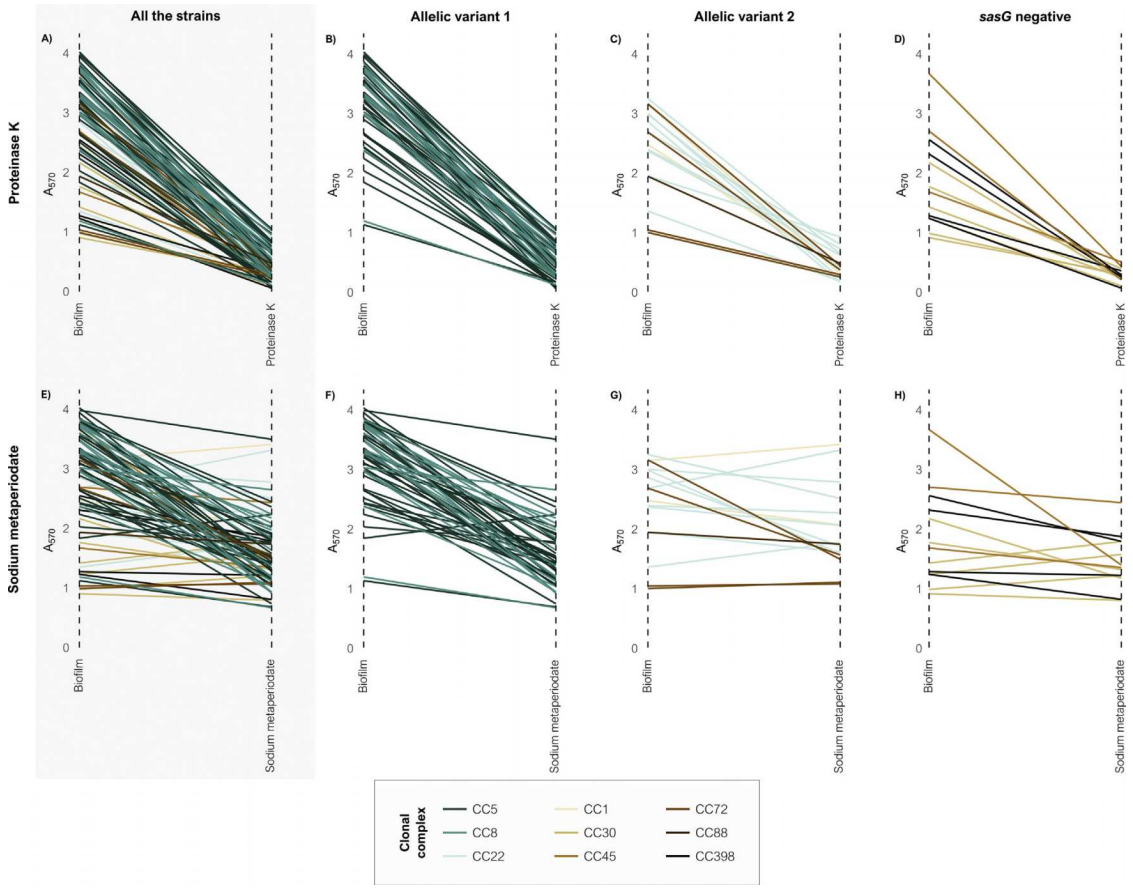


Fig. 3. Biofilm response to a 2-h proteinase K or sodium metaperiodate treatment measured by crystal violet absorbance at 570 nm (A_{570}). (A and E) All MRSA isolates. (B and F) Isolates carrying allelic variant 1. (C and G) Isolates carrying allelic variant 2. (D and H) Isolates without *sasG*. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

than that of EMRSA-16 (CC30), but similar to that of other sporadic or less frequent isolates.

Our results indicated that biofilm formation was protein-dependent, as the biofilm was more sensitive to the action of proteinase K than sodium metaperiodate, a sugar degrading agent. These findings suggest that, as previously shown in other studies [27,28], MRSA clones primarily use an *ica*-independent biofilm formation pathway. Together with the tendency for clone-specific biofilm formation, microarray genotyping was used to search for biofilm-associated virulence factors that could be used to differentiate between endemic and sporadic clones. Digging into the *ica*-independent pathway, each clonal complex showed a combination of surface, capsule and regulatory genes, as previously described by Lindsay et al. [29]. This could explain the variability in the biofilm formation capacities observed between clonal complexes. In particular, our study revealed that SasG, a cell wall-anchored protein related to biofilm formation and nasal colonisation [30,31], was carried by all the MRSA isolates belonging to endemic clonal complexes and only by a few from sporadic clonal complexes. Some studies have shown that SasG is strongly associated with invasive disease rather than nasal carriage [12,32], suggesting that it may play a role in invasive disease. These findings could explain the presence of the gene in the endemic bacteraemia strains, giving them some advantage to cause bloodstream infections against less prevalent strains that lack the gene. This advantage could be associated with increased biofilm production, creating an environment that promotes the persistence of these clones and further genetic adaptation that could lead to invasive disease.

Several studies have proposed mechanisms by which SasG can be responsible for biofilm production. Geoghegan et al. [31] showed that SasG promoted biofilm production through the cleavage of the G5-E repeats in the B domain, releasing the N-terminal region (A domain and some G5-E repeats). The remaining G5-E repeats attached to the cell wall dimerised in a zinc-dependent manner with other exposed G5-E repeats, promoting cell-to-cell interaction and biofilm formation. In addition, Yonemoto et al. [33] observed that *sasG* interacted with extracellular DNA, conferring protection against nucleases and enhancing biofilm stability. The coexistence of these biofilm-associated mechanisms would explain the higher production of biofilm in strains carrying this gene.

Previous studies identified two *sasG* allelic variants, defined by the variable region of the A domain located between residues 182 and 429 [12,13]. The high percentage of homology in the secretion signal and wall anchor domains indicates the importance of these regions for the protein function. Furthermore, the A domain has a high percentage of identity, even though most ST239 strains had a

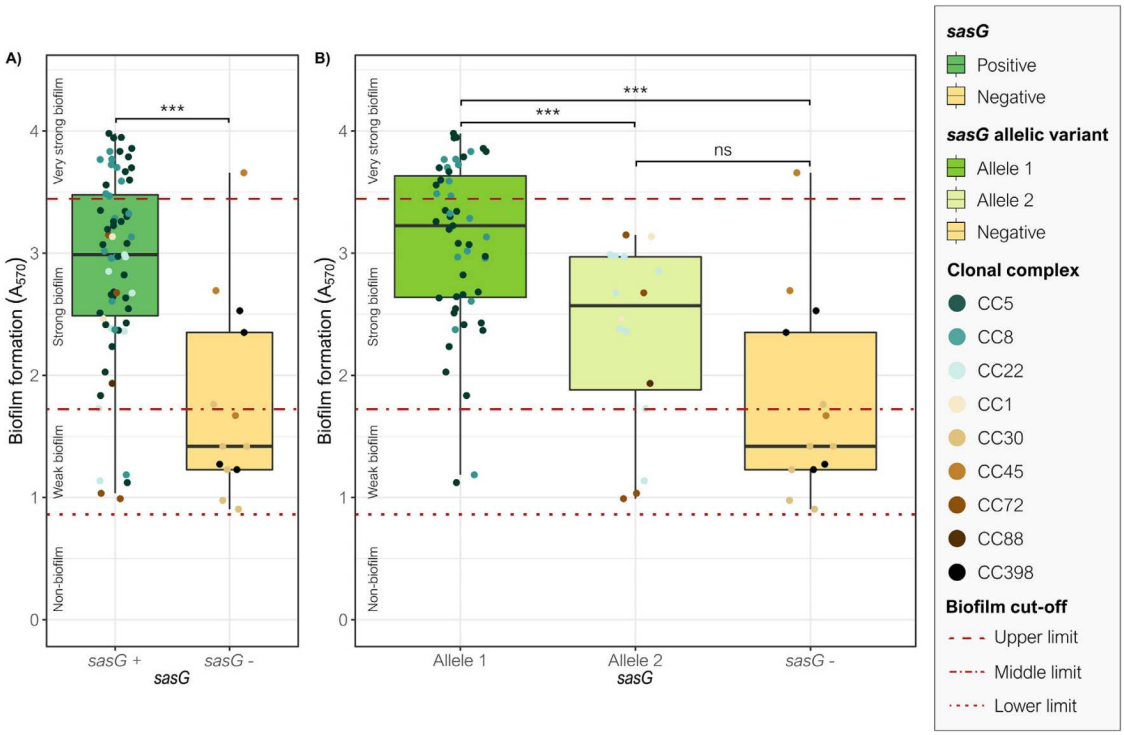


Fig. 4. Differences in biofilm formation in MRSA strains with or without *sasG* and its allelic variants. (A) Biofilm formation in clones with (*sasG* +) or without (*sasG* -) *sasG*. (B) Biofilm formation in clones carrying allelic variant 1 (described in the COL and Mu50 strains), allelic variant 2 (described in the MW2 strain), or without *sasG* (*sasG* -). Each dot represents the mean biofilm formation value for each strain obtained after at least three replicates at two independent time points. Each box plot shows the median, upper and lower quartiles (boxes) and the 1.5 x IQR (inter-quartile range) values (vertical lines) for each group. Horizontal dashed, dashed-dotted and dotted red lines represent the established cut-offs to classify the strains as very strong (A_{570} value > 20 times that of the negative control), strong (A_{570} value 10 to 20 times that of the negative control), weak (A_{570} value 5 to 10 times that of the negative control), or non-biofilm producers (A_{570} value < 5 times that of the negative control). ns, p-value > 0.05; * p-value < 0.05; *** p-value < 0.001. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

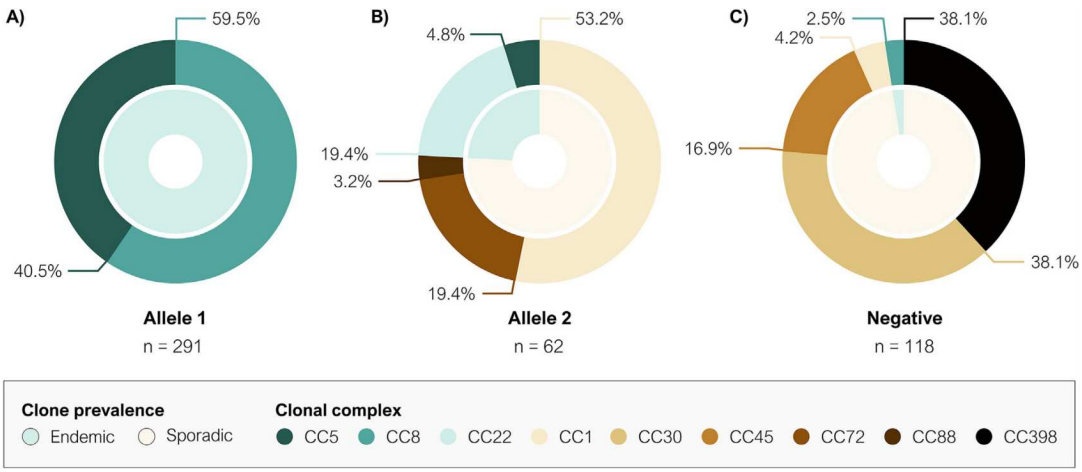


Fig. 5. Distribution of *sasG* allelic variants in the closed genomes from the NCBI database. (A) Clones carrying allelic variant 1. (B) Clones carrying allelic variant 2. (C) Clones without the *sasG* gene. The inner circle indicates the prevalence, while the outer circle indicates the clonal complex of the clones.

premature stop codon in the variable region, which reduced this percentage in the variable region of allele 1. However, we were unable to determine the impact of this mutation on biofilm formation because none of the strains in our collection belonged to ST239. Otherwise, the variability observed in the variable region of the A domain and the B domain, would indicate their importance in

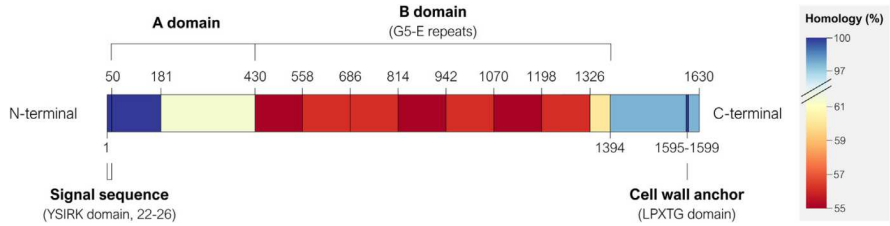


Fig. 6. SasG structure and the identity of its allelic variants. Percentage of identity between the SasG amino acid sequences of the NCTC8325 and 24117-WT strains, representative of allelic variants 1 and 2, respectively. Adapted from Corrigan et al. [11].

Table 1
SasG A domain percentage of identity from all the NCBI closed genomes.

Region (residues)	Allele 1	Allele 2
A-domain (51–429)	86.90%	99.60%
Conserved region (51–181)	98.50%	99.80%
Variable region (182–429)	78.20%	99.40%

the differential functions of both variants. We observed that the allelic variant 1 was mostly identified in the endemic CC5 and CC8 clones, which had higher biofilm production, while variant 2 was present in lower biofilm producing sporadic clones. Although CC22 is frequently isolated in invasive disease, it has been associated with allelic variant 2 and lower biofilm production when compared to other endemic clones. The success of CC22 could be associated with the epidemic spread of EMRSA-15 across Europe, probably due to the acquisition of fluoroquinolone resistance and some adaptive genetic changes [34] that favoured the success of these clones in the hospital environment. This proves that the analysis of *sasG* alone cannot fully define the biofilm phenotype, as many other genes may be involved in biofilm formation.

Cassat et al. [17] also observed differences in biofilm production between strains carrying different allelic variants of *sasG*, suggesting that the variants have different functions or that one of them is inactive or has reduced function. Our results confirmed this hypothesis since variant 1 was associated with higher gene expression compared to variant 2. On the other hand, Corrigan et al. [11] described that the number of G5-E repeats in the B domain was related to biofilm formation, in particular, they found that variants with five, six and eight repeats had higher biofilm formation. However, we did not observe any correlation between G5-E repeats and biofilm formation, probably because we have few strains representing each number of G5-E repeats in each allele.

In summary, starting with a small collection of Spanish MRSA bacteraemia isolates, a correlation was observed between CC5 and CC8 endemic clones and the *sasG* variant 1, which was confirmed using all publicly available *S. aureus* genomes in the NCBI database. The genomic analysis performed on the NCBI collection give a clear distribution of the *sasG* alleles among *S. aureus* clonal complexes. These results highlight the SasG variant 1 as an important virulence factor that may be associated to bacteraemia and increased biofilm formation, promoting bacterial survival both in the host and in the hospital environment. Therefore, the presence of allelic variant 1 could confer an adaptative benefit over strains without or those with allelic variant 2. The development of control measures is necessary to reduce and prevent the spread of MRSA clones. Further studies are needed to determine the *in* vivo significance of these findings, and if the inhibition of variant 1 could have a positive impact on preventing both biofilm formation and the spread of the most common MRSA clones.

Declaration of competing interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Acknowledgments

We would like to thank the staff of the Microbiology Laboratory of Bellvitge University Hospital who contributed daily to this project. This study was funded by Instituto de Salud Carlos III through the Project from the Fondo de Investigaciones Sanitarias "PI16/01382" to MD; by Plan Nacional de I+D+I 2013-2016 and Instituto de Salud Carlos III, Subdirección General de Redes y Centros de Investigación Cooperativa, by Ministerio de Ciencia, Innovación y Universidades, by CIBER de Enfermedades Infecciosas (CIBERINFEC-CB21/13/00009), and co-financed by European Development Regional Fund "A way to achieve Europe", Operative program Intelligent Growth 2014-2020; and funded by CIBER de Enfermedades Respiratorias (CIBERES–CB06/06/0037), co-funded by the European Regional Development Fund/European Social Fund (ERDF/ESF, "Investing in your future"), and CERCA programme/Generalitat de Catalunya for institutional support. Bioinformatic analysis was supported by an Amazon Web Services research grant to SM. AC was supported by FPU grant "FPU16/02202" (Formación de Profesorado Universitario, Ministerio de Educación, Spain), and SM was supported by Miguel Servet contract "CP19/00096" (ISCIII).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.resmic.2022.103921.

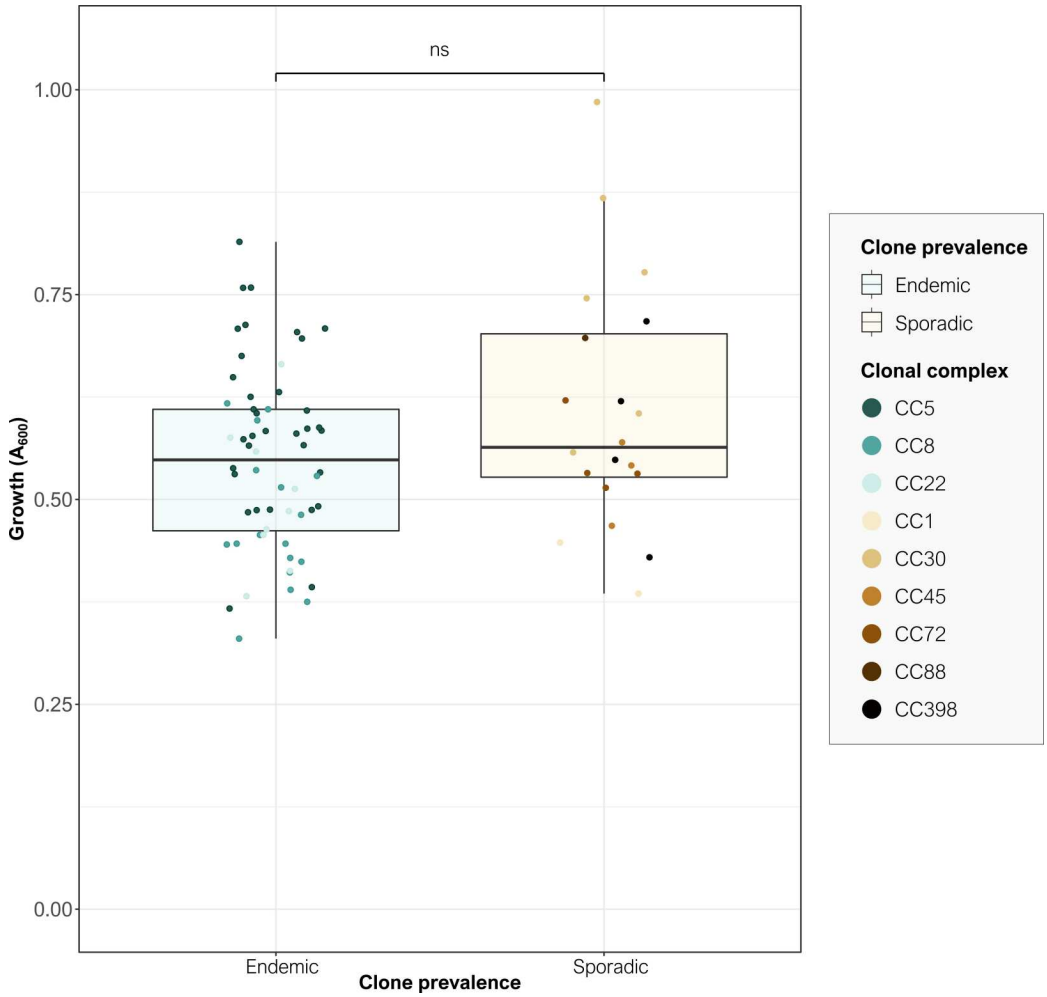
References

[1] Naber CK. *Staphylococcus aureus* bacteremia: epidemiology, pathophysiology, and management strategies. Clin Infect Dis 2009;48:S231–7. https://doi.org/10.1086/598189.
[2] Hassoun A, Linden PK, Friedman B. Incidence, prevalence, and management of MRSA bacteremia across patient populations - a review of recent developments in MRSA management and treatment. Crit Care 2017;21:211. https://doi.org/10.1186/s13054-017-1801-3.
[3] ECDC. Surveillance of antimicrobial resistance in Europe. https://doi.org/10.2900/230516; 2017.
[4] Archer NK, Mazaitis MJ, Costerton JW, Leid JG, Powers ME, Shirtlif ME. *Staphylococcus aureus* biofilms properties, regulation and roles in human disease. Virulence 2011;2(5):445–59. https://doi.org/10.4161/viru.2.5.17724.
[5] Donlan RM. Biofilms: microbial life on surfaces. Emerg Infect Dis 2002;8:881–90. https://doi.org/10.3201/eid0809.020063.
[6] Smith K, Hunter IS. Efficacy of common hospital biocides with biofilms of multi-drug resistant clinical isolates. J Med Microbiol 2008;57:966–73. https://doi.org/10.1099/jmm.0.47668-0.
[7] Craft KM, Nguyen JM, Berg LJ, Townsend SD. Methicillin-resistant *Staphylococcus aureus* (MRSA): antibiotic-resistance and the biofilm phenotype. Med Chem Commun 2019;10:1231–41. https://doi.org/10.1039/c9md00044e.

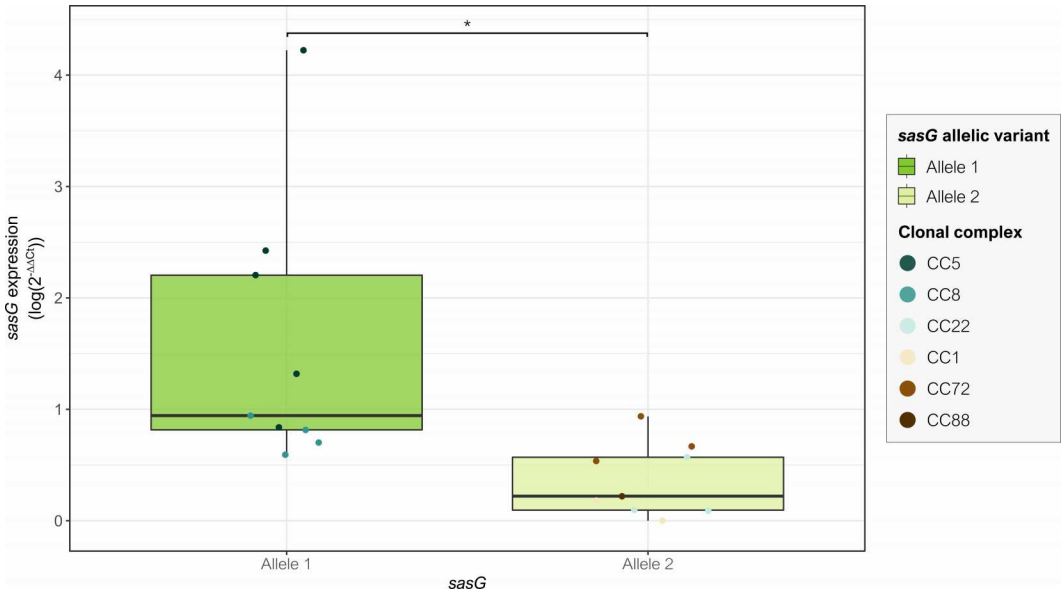
[8] Otto M. Staphylococcal biofilms. Curr Top Microbiol Immunol 2008;322:207–28. https://doi.org/10.1007/978-3-540-75418-3_10.
[9] O'Gara JP. *Ica* and beyond: biofilm mechanisms and regulation in *Staphylococcus epidermidis* and *Staphylococcus aureus*. FEMS Microbiol Lett 2007;270:179–88. https://doi.org/10.1111/j.1574-6968.2007.00688.x.
[10] Foster TJ, Geoghegan JA, Ganesh VK, Höök M. Adhesion, invasion and evasion: the many functions of the surface proteins of *Staphylococcus aureus*. Nat Rev Microbiol 2014;12:49–62. https://doi.org/10.1038/nrmicro3161.
[11] Corrigan RM, Rigby D, Handley P, Foster TJ. The role of *Staphylococcus aureus* surface protein SasG in adherence and biofilm formation. Microbiology 2007;153:2435–46. https://doi.org/10.1099/mic.0.2007/006676-0.
[12] Roche FM, Massey R, Peacock SJ, Day NPJ, Visai L, Speziale P, et al. Characterization of novel LPXTG-containing proteins of *Staphylococcus aureus* identified from genome sequences. Microbiology 2003;149:643–54. https://doi.org/10.1099/mic.0.25996-0.
[13] Kuroda M, Ito R, Tanaka Y, Yao M, Matoba K, Saito S, et al. *Staphylococcus aureus* surface protein SasG contributes to intercellular autoaggregation of *Staphylococcus aureus*. Biochem Biophys Res Commun 2008;377:1102–6. https://doi.org/10.1016/j.bbrc.2008.10.134.
[14] Gasch O, Camoez M, Dominguez MA, Padilla B, Pintado V, Almirante B, et al. Predictive factors for mortality in patients with methicillin-resistant *Staphylococcus aureus* bloodstream infection: impact on outcome of host, microorganism and therapy. Clin Microbiol Infect 2013;19:1049–57. https://doi.org/10.1111/1469-0691.12108.
[15] Monecke S, Coombs G, Shore AC, Coleman DC, Akpaka P, Borg M, et al. A field guide to pandemic, epidemic and sporadic clones of methicillin-resistant *Staphylococcus aureus*. PLoS One 2011;6:e17936. https://doi.org/10.1371/journal.pone.0017936.
[16] Puig C, Domenech A, Garmendia J, Langereis JD, Mayer P, Calatayud L, et al. Increased biofilm formation by nontypeable *Haemophilus influenzae* isolates from patients with invasive disease or otitis media versus strains recovered from cases of respiratory infections. Appl Environ Microbiol 2014;80:7088–95. https://doi.org/10.1128/AEM.02544-14.
[17] Cassat JE, Dunman PM, McAleese F, Murphy E, Projan SJ, Smeltzer MS. Comparative genomics of *Staphylococcus aureus* musculoskeletal isolates. J Bacteriol 2005;187:576–92. https://doi.org/10.1128/JB.187.2.576-592.2005.
[18] Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 2017;13:e1005595. https://doi.org/10.1371/journal.pcbi.1005595.
[19] Madeira F, Mi Park Y, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Web Serv Issue 2019;47:W636–41. https://doi.org/10.1093/nar/gkz268.
[20] Bhattacharya M, Wozniak DJ, Stoodley P, Hall-Stoodley L. Prevention and treatment of *Staphylococcus aureus* biofilms. Expert Rev Anti Infect Ther 2015;13:1499–516. https://doi.org/10.1586/14787210.2015.1100533.
[21] Costa DM, Johani K, Melo DS, Lopes LKO, Lopes Lima LKO, Tipple AFV, et al. Biofilm contamination of high-touched surfaces in intensive care units: epidemiology and potential impacts. Lett Appl Microbiol 2019;68:269–76. https://doi.org/10.1111/lam.13127.
[22] Naicker PR, Karayem K, Hoek KGP, Harvey J, Wasserman E. Biofilm formation in invasive *Staphylococcus aureus* isolates is associated with the clonal lineage. Microb Pathog 2016;90:41–9. https://doi.org/10.1016/j.micpath.2015.10.023.
[23] Jotić A, Božić DD, Milovanović J, Pavlović B, Ješić S, Pelemiš M, et al. Biofilm formation on tympanostomy tubes depends on methicillin-resistant *Staphylococcus aureus* genetic lineage. Eur Arch Oto-Rhino-Laryngol 2016;273:615–20. https://doi.org/10.1007/s00405-015-3607-8.
[24] Côrtes MF, Beltrame CO, Ramundo MS, Ferreira FA, Figueiredo AMS. The influence of different factors including *fnbA* and *mecA* expression on biofilm formed by MRSA clinical isolates with different genetic backgrounds. Int J Med Microbiol 2015;305:140–7. https://doi.org/10.1016/j.ijmm.2014.11.011.
[25] Tasse J, Trouillet-Assant S, Josse J, Martins-Simões P, Valour F, Langlois-Jacques C, et al. Association between biofilm formation phenotype and clonal lineage in *Staphylococcus aureus* strains from bone and joint infections. PLoS One 2018;13(8):e020. https://doi.org/10.1371/journal.pone.0200064.
[26] Smith K, Perez A, Ramage G, Lappin D, Gemmell CG, Lang S. Biofilm formation by Scottish clinical isolates of *Staphylococcus aureus*. J Med Microbiol 2008;57:1018–23. https://doi.org/10.1099/jmm.0.2008/000968-0.
[27] O'Neill E, Pozzi C, Houston P, Smyth D, Humphreys H, Robinson DA, et al. Association between methicillin susceptibility and biofilm regulation in *Staphylococcus aureus* isolates from device-related infections. J Clin Microbiol 2007;45:1379–88. https://doi.org/10.1128/JCM.02280-06.
[28] Fitzpatrick F, Humphreys H, O'Gara JP. Evidence for *icaADBC*-independent biofilm development mechanism in methicillin-resistant *Staphylococcus aureus* clinical isolates. J Clin Microbiol 2005;43:1976. https://doi.org/10.1128/JCM.43.4.1973-1976.2005.
[29] Lindsay JA, Moore CE, Day NP, Peacock SJ, Witney AA, Stabler RA, et al. Microarrays reveal that each of the ten dominant lineages of *Staphylococcus aureus* has a unique combination of surface-associated and regulatory genes. J Bacteriol 2006;188:669–76. https://doi.org/10.1128/JB.188.2.669-676.2006.
[30] Roche FM, Meehan M, Foster TJ. The *Staphylococcus aureus* surface protein SasG and its homologues promote bacterial adherence to human desquamated nasal epithelial cells. Microbiology 2003;149:2759–67. https://doi.org/10.1099/mic.0.26412-0.
[31] Geoghegan JA, Corrigan RM, Gruszka DT, Speziale P, O'Gara JP, Potts JR, et al. Role of surface protein SasG in biofilm formation by *Staphylococcus aureus*. J Bacteriol 2010;192:5663–73. https://doi.org/10.1128/JB.00628-10.
[32] Rasmussen G, Monecke S, Ehricht R, Söderquist B. Prevalence of clonal complexes and virulence genes among commensal and invasive *Staphylococcus aureus* isolates in Sweden. PLoS One 2013;8:e77477. https://doi.org/10.1371/journal.pone.0077477.
[33] Yonemoto K, Chiba A, Sugimoto S, Sato C, Saito M, Kinjo Y, et al. Redundant and distinct roles of secreted protein Eap and cell wall-anchored protein SasG in biofilm formation and pathogenicity of *Staphylococcus aureus*. Infect Immun 2019;87. https://doi.org/10.1128/IAI.00894-18. e00894-18.
[34] Holden MTG, Hsu LY, Kurt K, Weinert LA, Mather AE, Harris SR, et al. A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic. Genome Res 2013;23:653–64. https://doi.org/10.1101/gr.147710.112.

SUPPLEMENTARY MATERIAL

Supplementary Dataset S1, S2, and S3 are available with the online version of this article.



Supplementary Figure S1. Bacterial growth after 24 hours assessed by absorbance at 600 nm (A_{600}). Comparison of bacterial growth on a static solid support between endemic and sporadic MRSA bacteraemia isolates. Each dot represents the mean bacterial value for each strain obtained after at least three replicates at two independent time points. Each box plot shows the median, upper and lower quartiles (boxes) and the 1.5 x IQR (inter-quartile range) values (vertical lines) for each group. ns, p-value > 0.05.



Supplementary Figure S2. Expression of *sasG* allelic variants in MRSA bacteraemia isolates. *sasG* expression was determined using the $2^{-\Delta\Delta Ct}$ method and the results are expressed as the logarithmic value of $2^{-\Delta\Delta Ct}$. 16S rRNA was used as the housekeeping gene and the sample with the lowest *sasG* expression (HRC106) was selected as the reference sample. Each dot represents the mean *sasG* expression value for each strain at two independent time points. Each box plot shows the median, upper and lower quartiles (boxes) and the 1.5 x IQR (inter-quartile range) values (vertical lines) for each group. * p-value < 0.05.

Additional results

Host adaptive changes of *Staphylococcus aureus* through respiratory colonisation and bloodstream infection

Anna Carrera-Salinas, Aida González-Díaz, Ileana Paula Salto, Maria Mrakovcic, Daniel Antonio Vázquez-Sánchez, Meritxell Cubero, Mª Ángeles Domínguez, Silke Niemann, Sara Martí

Host adaptive changes of *Staphylococcus aureus* through respiratory colonisation and bloodstream infection

Anna Carrera-Salinas[1], Aida González-Díaz[1,2], Ileana Paula Salto[3], Maria Mrakovcic[3], Daniel Antonio Vázquez-Sánchez[1], Meritxell Cubero[1,2], Mª Ángeles Domínguez[1,4], Silke Niemann[3], Sara Martí[1,2]

[1]Microbiology Dept. Bellvitge University Hospital. IDIBELL-UB, Barcelona, Spain.

[2]Research Network for Respiratory Diseases (CIBERES), ISCIII, Madrid, Spain.

[3]Institute of Medical Microbiology, University Hospital of Münster, Münster, Germany.

[4]Research Network for Infectious Diseases (CIBERINFEC), ISCIII, Madrid, Spain.

INTRODUCTION

Staphylococcus aureus is an important opportunistic human pathogen that usually colonises the anterior nares and skin[1]. Although antimicrobials, the immune response, or competition with other microorganisms may resolve this colonisation, approximately 20% of adults are persistent carriers[2,3]. Intracellular colonisation of epithelial, endothelial, and inflammatory cells provides the perfect niche for bacteria to avoid the hostile upper respiratory tract environment and survive for long periods[4,5]. However, *S. aureus* colonisation increases the risk of subsequent infection, including pneumonia, bacteraemia, and skin and soft tissue infections (SSTIs), with evidence that 80% of bloodstream infections in nasal carriers are caused by a colonising strain[6-8]. During colonisation, the microbial genome may evolve through point mutations, insertions, deletions, or DNA rearrangements, and may also acquire exogenous DNA from the environment and other pathogens by horizontal gene transfer[9,10]. These genetic changes may affect a broad range of virulence factors (e.g., surface proteins) that promote bacterial adhesion, biofilm formation, iron acquisition, or immune evasion, leading to changes in host-bacterial interactions[11].

Given that *S. aureus* colonisation represents a risk factor for severe disease, this study aimed to explore the

genetic changes associated with intracellular host cell invasion in respiratory and invasive clinical isolates of *S. aureus*.

METHODS

Study design and bacterial strain selection

This retrospective laboratory-based study was based on *S. aureus* strains from patients with respiratory colonisation who subsequently developed a bloodstream infection caused by the same clone at Hospital Universitari de Bellvitge (Spain). Medical records were reviewed to obtain demographic and clinical data, including age, gender, and underlying conditions.

Analysis of the clonal relatedness of respiratory and bacteraemic strains was by pulse-field gel electrophoresis (PFGE), as previously described[12]. PFGE band patterns were analysed with Fingerprinting II Software 3.0 (BioRad), using the Dice coefficient and setting optimisation and tolerance to 1%. Respiratory and bacteraemic isolates from the same patient with a Dice index ≥80% were considered genetically related and selected for further analysis. Persistence was defined as the isolation of the same clone, including the same sequence type (ST) and its single-locus variants (SLVs), from the same patient over at least 3 months. Finally, antimicrobial susceptibility was assessed by disk diffusion, according to the recommended European Committee on Antimicrobial Susceptibility Testing (EUCAST) breakpoints[13].

Whole genome sequencing

Genomic DNA was extracted using the QIAamp DNA Mini Kit (Qiagen) and quantifed by QuantiFluor® dsDNA System (Promega). Nextera XT was used to prepare the libraries, followed by paired-end sequencing on a MiSeq platform (Illumina Inc.). Sequences were assembled using default parameters in the INNUca v4.2 pipeline (github.com/BUMMI/INNUca). MLST was performed using the MLST v2.4 software (github.com/tseemann/mlst).

Phylogenetic analysis was conducted using Snippy and RAxML-

NG. Briefly, core SNPs were extracted and aligned with Snippy's core module (github.com/tseemann/snippy), using default parameters. The core genome SNP alignments were used to build a core-SNP phylogenetic tree in RAxML-NG[15], applying a discrete gamma model of rate heterogeneity. Trees were visualised in R package ggtree[16]. The *S. aureus* NCTC 8325 (NC_007795) genome was used as the reference.

To evaluate the accumulated genetic changes between respiratory and bacteraemic strains, the genome obtained from the first respiratory isolate of each patient was concatenated with Mauve[17], and annotated with Prokka v1.13.7[18], using *S. aureus* genomes N315 (NC_002745), ST228 (NC_020529), and JKD6159 (NC_017338) as the references for clones belonging to ST125, ST228, and ST3207, respectively. The contigs of each patient were aligned in Mauve using the concatenated genome as the reference. To characterise the genetic changes, alignments were analysed in Geneious R9 (Biomatters).

Host cell invasion assay

Bacterial invasion of host cells was assessed with the human lung epithelial A549 cell line, as previously described[19]. Firstly, bacterial stocks (adjusted to OD_{540} = 1) were prepared after growing bacterial cultures in tryptic soy broth for 3 hours at 37ºC under shaking conditions, before storage at -20ºC until use. Adjusted cultures were serially diluted and plated on blood agar plates to determine the number of colony-forming units (CFU) per bacterial stock.

A549 cells were seeded in 12-well tissue culture-treated plates (Corning, United States) at a density of 40,000 cells/cm^2 in Dulbecco's modified Eagle's Medium with stable glutamine, with 3.7 g/L $NaHCO_3$, 4.5 g D-glucose (DMEM, Biochrom GmbH, Germany), and 10% fetal calf serum (FCS, Sigma-Aldrich, Germany). These were incubated for 3 days at 37ºC in 5% CO_2 to form confluent cell monolayers. The A549 cells were then washed with phosphate buffered saline (PBS, Sigma-Aldrich, Germany) before being exposed to invasion media containing DMEM

supplemented with 1% human serum albumin (Kedrion, Italy) and 10 mmol/L HEPES (Biochrom GmbH, Germany). Subsequently, A549 cells were infected with the previously prepared bacterial stock adjusted at a multiplicity of infection (MOI) of 50. Infected cells were incubated for 3 hours at 37ºC in 5% CO_2. The extracellular and adherent bacteria were then removed, and incubated for 30 minutes with lysostaphin (20 mg/mL). A549 cells were detached by trypsin/EDTA (Biochrom GmbH, Germany) and counted before lysis with cold distilled water and sonication to release the intracellular bacteria. Cell lysates were plated on blood agar plates in serial dilutions to determine the intracellular bacteria, which were expressed as intracellular bacteria per A549 cell (CFU/A549 cells).

Statistical analysis

Statistical analyses were carried out using GraphPad Prism version 5, applying one-way analysis of variance (Kruskal-Wallis test). P-values<0.05 were considered statistically significant.

RESULTS

Clinical data and strain characterisation

Three patients (HUB-P01, HUB-P02, and HUB-P03) showed respiratory colonisation by *S. aureus* that progressed to a bloodstream infection caused by the same clone. Table 1 shows the demographics, underlying conditions, and clinical characteristics of these patients. They had several underlying conditions, the most common being chronic obstructive pulmonary disease (COPD) (66.7%), and had Charlson comorbidity indexes ranging from 3 to 8 (mean, 6; SD, 2.6).

In all cases, the focus of the bloodstream infection was a previous respiratory tract infection, including bronchiectasis-related infection, pneumonia, and sinusitis. Furthermore, all patients died within 30 days of the first *S. aureus* isolation in the blood (mean, 9.7 days; SD, 8.4).

The *S. aureus* clone ST228, together with its SLVs (ST7667 and ST7668) (CC5), colonised the upper and lower respiratory tract of HUB-P01 for more than 350 days and were isolated from

the blood within 429 days after its initial detection in the upper respiratory tract (Figure 1). Clone ST125 (CC5) colonised the upper respiratory tract of patient HUB-P02 and was detected in both lower and blood samples 35 days later. Finally, clone ST3207 (CC398) was found colonising the upper respiratory tract of HUB-P03 and was isolated in both lower respiratory tract and blood samples 28 days later.

No changes in the antimicrobial susceptibility of these strains were observed during follow-up.

Core SNP phylogenetic analysis revealed low genetic diversity within clonal complexes, with strains in each patient forming well-defined phylogenetic clusters with high genetic similarity between respiratory and bacteraemic isolates (Figure 2).

Host cell intracellular invasion and genetic differences between respiratory and bacteraemic isolates

Host cell invasion assays revealed that bacteraemic isolates had significant lower capability to invade A549 cells than respiratory isolates (Figure 3). Furthermore, the respiratory isolate in HUB-P02, obtained at the same time as the bacteraemic isolate, showed less intracellular invasion than previously isolated respiratory strains.

Genetic analysis revealed some differences between the respiratory and blood isolates. Changes were found in genes among isolates with less intracellular invasion ability from two patients. These encoded bacteria-host cell binding proteins, such as extracellular matrix binding protein gene (*ebh*), clumping factor B gene (*clfB*), and fibronectin-binding protein A gene (*fnbA*) (Figure 4). Bacteraemic isolates from HUB-P01 (HUB-P01-SA07 and HUB-P01-SA08) presented two genetic differences compared to respiratory isolates: the *ebh* gene had a mutation associated with a premature stop codon, which could result in a truncated protein; and the *clfB* gene had several punctual mutations in the serine-aspartate repeat region. Furthermore, although not statistically significant, the bacteraemic isolate HUB-P01-SA07, had a premature stop codon in the *fnbA* gene and showed less intracellular invasion of A549 host

cells than HUB-P01-SA08. Bacteraemic isolates from HUB-P03 (HUB-P03-SA03 and HUB-P03-SA04) showed an insertion in the *clfB* gene in the proline rich region. No genetic differences were observed between the respiratory and bacteraemic isolates from HUB-P02, despite observing differences in the intracellular invasion of host cells.

CONCLUSIONS

S. aureus can colonise the respiratory tract and survive intracellularly for long periods. During colonisation, genetic changes may occur that alter bacterial environmental adaptation and alter the bacterial-host interaction, regardless of host conditions.

Bacteraemic isolates showed a lower capacity to invade and colonise lung epithelial A549 cells intracellularly than respiratory isolates. The alterations in intracellular invasion may be result from alterations in bacterial-host cell binding. In turn, this could be explained by genetic changes observed in three genes: the *ebh* gene codes for the giant surface extracellular matrix binding protein, which mediates fibronectin, plasminogen, and albumin binding[20,21]; the *fnbA* gene codes for the fibronectin-binding protein A and is associated with fibronectin, fibrinogen, and elastin binding[22]; and the *clfB* gene codes for clumping factor B and binds to fibrinogen and cytokeratin[23,24].

These findings show that the phenotypic and genotypic changes that occur during colonisation may lead to a bacterial shift from respiratory colonisation to invasive disease. Nevertheless, further analysis, including transcriptomic or proteomic analysis, is necessary to explain the alterations in intracellular invasion that could not be determined by genome sequencing.

ACKNOWLEDGEMENTS

We would like to thank the staff of the Microbiology Laboratory of Bellvitge University who contributed daily to this project. We also thank the staff of the Institute of Medical Microbiology from University Hospital of Münster for their technical assistance.

FUNDING

This study was funded by grants from Fondo de Investigaciones Sanitarias "PI16/00977" and from Sociedad Española de Neumología y Cirugía Torácica (SEPAR 418/2017) to SM. It was also funded by a grant from Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES, CB06/06/0037), an initiative of the Instituto de Salud Carlos III (ISCIII). The European Regional Development Fund/European Social Fund (ERDF/ESF, "Investing in your future") provided financial support, and CERCA Programme/Generalitat de Catalunya provided institutional support. Bioinformatic analysis was supported by an Amazon Web Services (AWS) research grant to SM. AC was supported by FPU grant "FPU16/02202" (Formación de Profesorado Universitario, Ministerio de Educación, Spain), and SM by Miguel Servet contract "CP19/00096" (ISCIII).

REFERENCES

1. Sakr, A., Brégeon, F., Mège, J.-L., Rolain, J.-M. & Blin, O. ***Staphylococcus aureus* nasal colonization: an update on mechanisms, epidemiology, risk factors, and subsequent infections.** Front. Microbiol. 8, 2419 (2018).

2. Wertheim, H. F. L. *et al*. **The role of nasal carriage in *Staphylococcus aureus* infections.** Lancet Infect. Dis. 5, 751–62 (2005).

3. Laux, C., Peschel, A. & Krismer, B. ***Staphylococcus aureus* colonization of the human nose and interaction with other microbiome members.** Microbiol. Spectr. 7, 723–730 (2019).

4. Turner, N. A. *et al*. **Methicillin-resistant *Staphylococcus aureus*: an overview of basic and clinical research.** Nat. Rev. Microbiol. 17, 203–218 (2019).

5. Fraunholz, M. & Sinha, B. **Intracellular *Staphylococcus aureus*: live-in and let die.** Front. Cell. Infect. Microbiol. 2, 1–10 (2012).

6. Kao, K. C. *et al*. **Risk factors of methicillin-resistant *Staphylococcus aureus* infection and correlation with nasal colonization based on molecular genotyping in medical intensive care units.** Medicine (Baltimore). 94, e1100 (2015).

7. Tilahun, B., Faust, A. C., McCorstin, P. & Ortegon, A. **Nasal colonization and lower respiratory tract infections with methicillin-resistant *Staphylococcus aureus*.** Am. J. Crit. Care 24, 8–12 (2015).

8. von Eiff, C. *et al*. **Nasal carriage as a source of *Staphylococcus aureus* bacteremia.** N Engl J Med 344, 11–6 (2001).

9. Foley, J. **Mini-review: Strategies for variation and evolution of bacterial antigens.** Comput. Struct. Biotechnol. J. 13, 407–416 (2015).

10. Munita, J. M. & Arias, C. A. **Mechanisms of antibiotic resistance.** Microbiol. Spectr. 4, 1–37 (2016).

11. Foster, T. J., Geoghegan, J. A., Ganesh, V. K. & Höök, M. **Adhesion, invasion and evasion: the many functions of the surface proteins of *Staphylococcus aureus*.** Nat. Rev. Microbiol. 12, 49–62 (2014).

12. Puig, C., Calatayud, L., Martí, S., Tubau, F. & Garcia-Vidal, C. **Molecular epidemiology of nontypeable *Haemophilus influenzae* causing community-acquired pneumonia in adults.** PLoS One 8, 82515 (2013).

13. The European Committee on Antimicrobial Susceptibility Testing. **Breakpoint tables for interpretation of MICs and zone diameters.** (2022).

14. Bortolaia, V. *et al*. **ResFinder 4.0 for predictions of phenotypes from genotypes.** J. Antimicrob. Chemother. 75, 3491–3500 (2020).

15. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. **RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference.** Bioinformatics 35, 4453–4455 (2019).

16. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. Y. **ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data.** Methods Ecol. Evol. 8, 28–36 (2017).

17. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. **Mauve: Multiple alignment of conserved genomic sequence with rearrangements.** Genome Res. 14, 1403 (2004).

18. Seemann, T. **Prokka: Rapid prokaryotic genome annotation.** Bioinformatics 30, 2068–2069 (2014).

19. Strobel, M. *et al*. **Post-invasion events after infection with *Staphylococcus aureus* are strongly dependent on both the host cell type and the infecting *S. aureus* strain.** Clin. Microbiol. Infect. 22, 799–809 (2016).

20. Cheng, A. G., Missiakas, D. & Schneewind, O. **The giant protein Ebh is a determinant of *Staphylococcus aureus* cell size and complement resistance.** J. Bacteriol. 196, 971–981 (2014).

21. Kuroda, M. *et al*. ***Staphylococcus aureus* giant protein Ebh is involved in tolerance to transient hyperosmotic pressure.** Biochem. Biophys. Res. Commun. 374, 237–241 (2008).

22. Loughman, A. *et al*. **Sequence diversity in the A domain of *Staphylococcus aureus* fibronectin-binding protein A.** BMC Microbiol. 8, 74 (2008).

23. Perkins, S. *et al*. **Structural organization of the fibrinogen-binding region of the clumping factor B MSCRAMM of *Staphylococcus aureus*.** J. Biol. Chem. 276, 44721–44728 (2001).

24. Walsh, E. J., O'Brien, L. M., Liang, X., Hook, M. & Foster, T. J. **Clumping factor B, a fibrinogen-binding MSCRAMM (microbial surface components recognizing adhesive matrix molecules) adhesin of *Staphylococcus aureus*, also binds to the tail region of type I cytokeratin 10.** J. Biol. Chem. 279, 50691–9 (2004).

Table 1. Clinical characteristics and underlying conditions of included patients.

	Overall	Patients included in this study		
		HUB-P01	HUB-P02	HUB-P03
Characteristics [no. (%)]				
Age (mean ± SD; range)	64.7 ± 9.0; 56-74	56	74	64
Sex[a]	1-2 (33.3-66.7)	Female	Male	Male
Underlying conditions				
Immunosuppressive therapy	1 (33.3)	Yes	No	No
Diabetes	1 (33.3)	Yes	No	No
Heart disease	1 (33.3)	No	Yes	No
COPD	2 (66.7)	No	Yes	Yes
Bronchiectasis	1 (33.3)	Yes	No	No
Chronic liver disease	1 (33.3)	Yes	No	No
Organ transplant[b]	1 (33.3)	Yes	No	No
Chronic kidney disease	1 (33.3)	Yes	No	No
Charlson comorbidity index	6 ± 2.6	8	7	3
Source of bloodstream infection				
Respiratory tract infection	3 (100)	Bronchiectasis	Pneumonia	Sinusitis
30-day mortality[c]	3 (100)	Yes	Yes	Yes

[a]Female-Male; [b]Liver trasplant; [c]After first bacteraemic isolate.

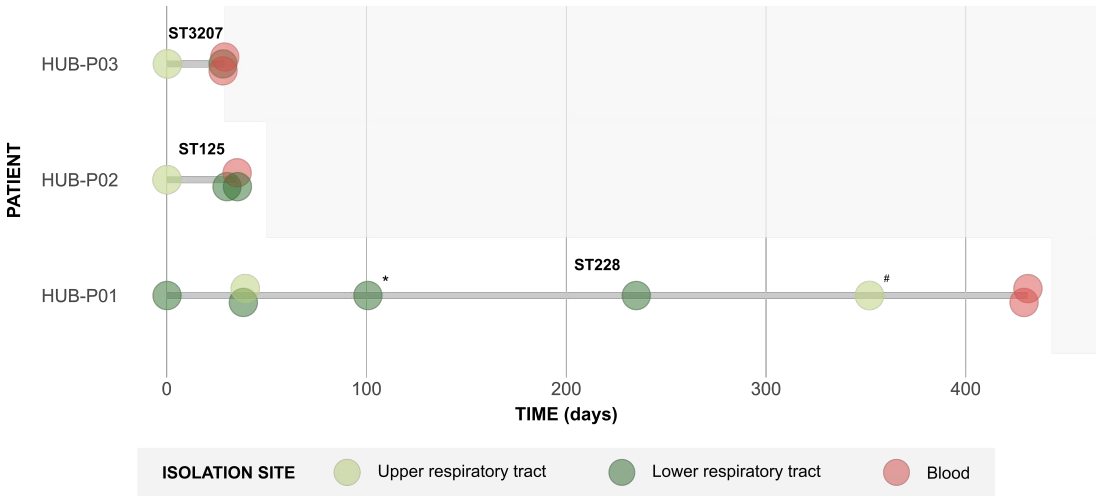


Figure 1. Timeline of strains isolated from patients with *S. aureus* respiratory colonisation and subsequent bloodstream infection due to the same clone. The sequence type is listed at the top of each line, per patient. Single-locus variants of ST228 are denoted by an asterisk and a hash mark (ST7667* and ST7668#). Death is indicated by the grey shaded area. Upper respiratory tract samples included nasal swabs and nasal exudates, while lower respiratory tract samples included sputum, bronchoalveolar lavages, and bronchotracheal aspirates, and transtracheal aspirates.

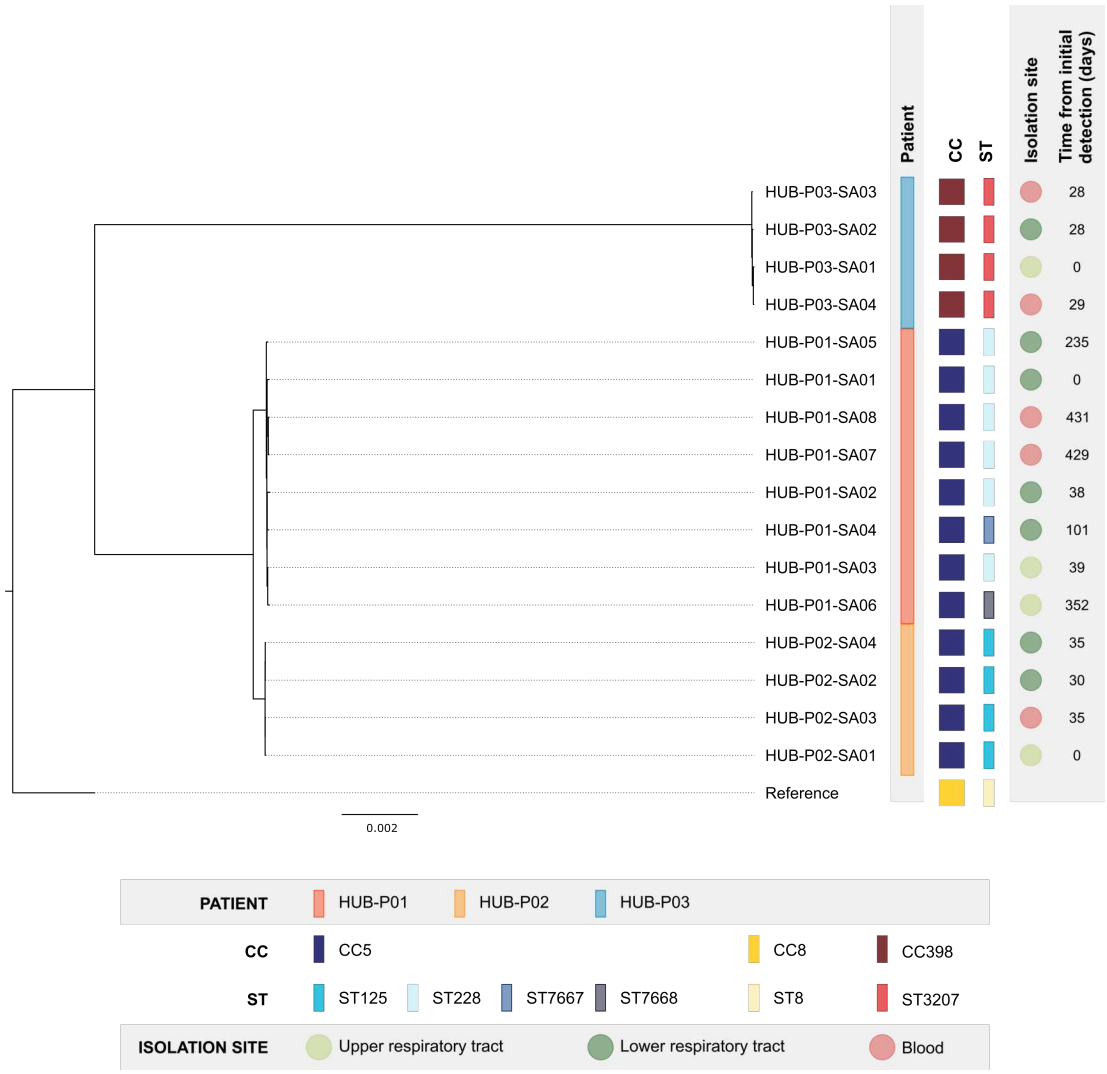


Figure 2. Core SNP phylogenetic tree of *S. aureus* strains isolated from patients with respiratory colonisation and subsequent bloodstream infection. Strain NCTC 8325 (NC_007795) was used as the reference.

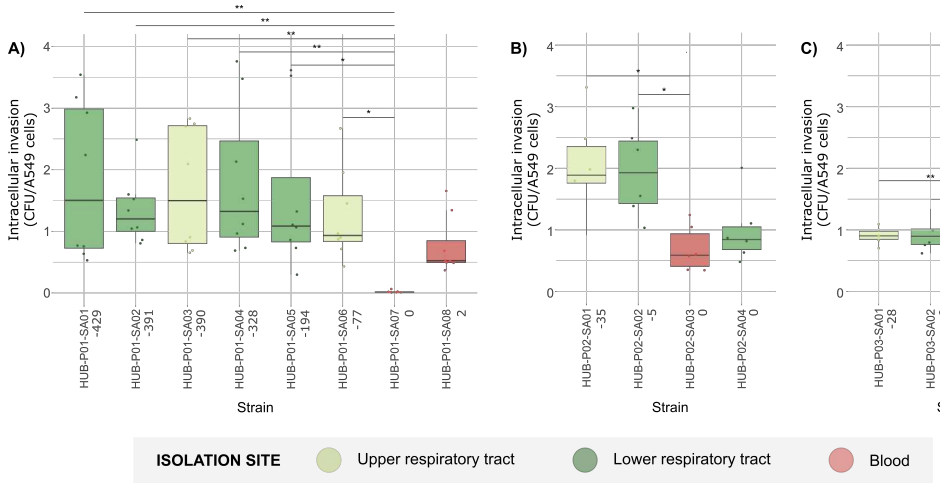


Figure 3. Intracellular invasion of colonising and bacteraemic *S. aureus* isolates after 3 hours of A549 cells infection. Comparison of intracellular invasion between isolates from HUB-P01 **(A)**, HUB-P02 **(B)**, and HUB-P03 **(C)**. The number of days between each isolate and the first bacteraemic isolate is plotted next to the strain name, with negative values representing days before isolation. Each dot represents the intracellular invasion (CFU/A549 cells) obtained at each replicate. Each box plot shows the median, upper and lower quartiles (boxes) and the 1.5 x IQR (inter-quartile range) values (vertical lines) for each strain. CFU, colony-forming units; * p-value<0.05; ** p-value<0.01.

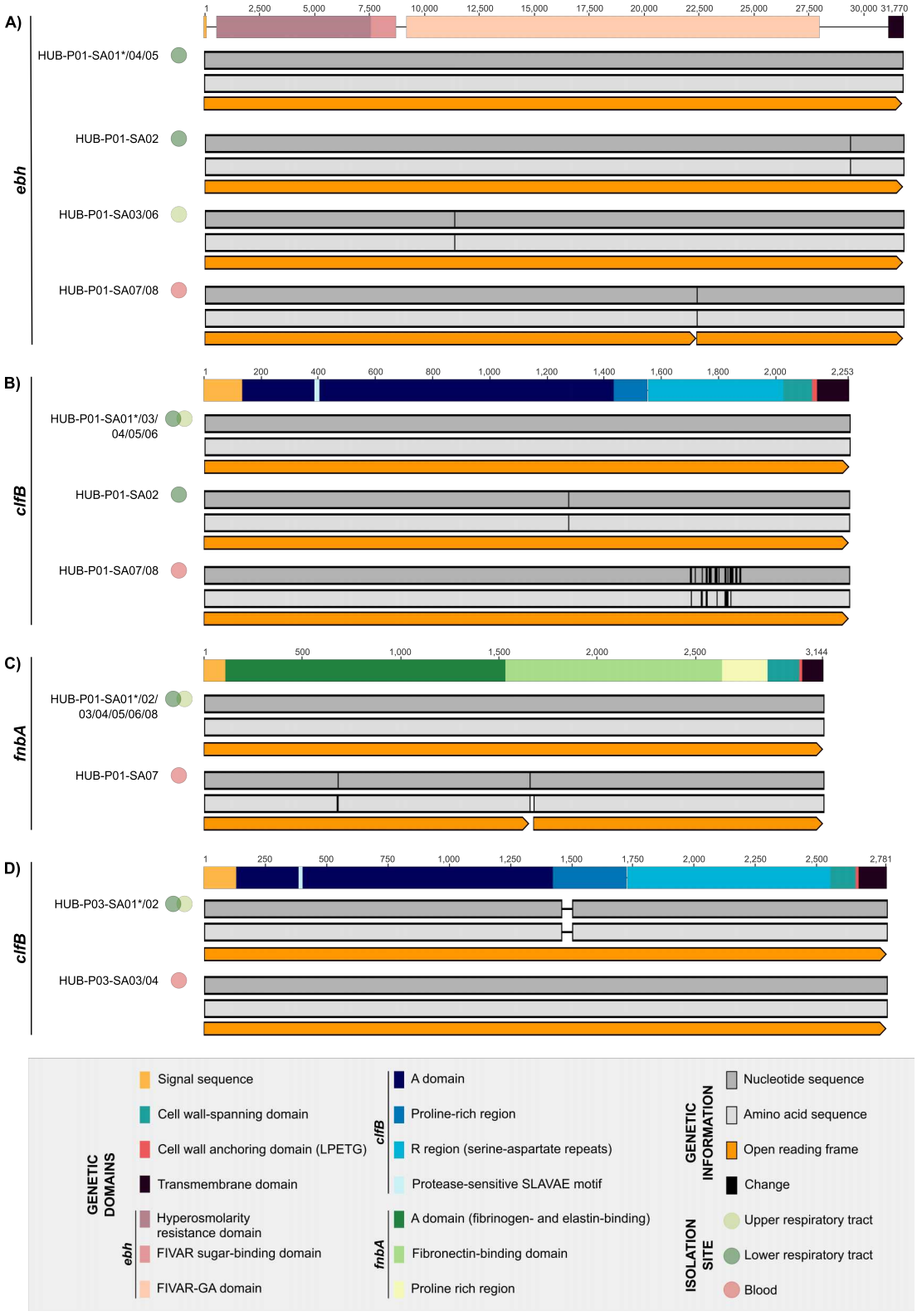

A)
ebh
HUB-P01-SA01*/04/05
HUB-P01-SA02
HUB-P01-SA03/06
HUB-P01-SA07/08
1
2,500
5,000
7,500
10,000
12,500
15,000
17,500
20,000
22,500
25,000
27,500
30,000
31,770
B)
clfB
HUB-P01-SA01*/03/
04/05/06
HUB-P01-SA02
HUB-P01-SA07/08
1
200
400
600
800
1,000
1,200
1,400
1,600
1,800
2,000
2,253
C)
fnbA
HUB-P01-SA01*/02/
03/04/05/06/08
HUB-P01-SA07
1
500
1,000
1,500
2,000
2,500
3,144
D)
clfB
HUB-P03-SA01*/02
HUB-P03-SA03/04
1
250
500
750
1,000
1,250
1,500
1,750
2,000
2,250
2,500
2,781
GENETIC DOMAINS
Signal sequence
Cell wall-spanning domain
Cell wall anchoring domain (LPETG)
Transmembrane domain
ebh
Hyperosmolarity resistance domain
FIVAR sugar-binding domain
FIVAR-GA domain
clfB
A domain
Proline-rich region
R region (serine-aspartate repeats)
Protease-sensitive SLAVAE motif
fnbA
A domain (fibrinogen- and elastin-binding)
Fibronectin-binding domain
Proline rich region
GENETIC INFORMATION
Nucleotide sequence
Amino acid sequence
Open reading frame
Change
ISOLATION SITE
Upper respiratory tract
Lower respiratory tract
Blood

Figure 4. Alignment of genes showing genetic changes in respiratory and bacteraemic *S. aureus* isolates. Genetic changes in HUB-P01 were observed in **(A)** the *ebh* gene coding for the extracellular matrix binding protein; **(B)** the *clfB* gene coding for the clumping factor B; and **(C)** the *fnbA* gene coding for fibronectin-binding protein A. The genetic changes in HUB-P03 were observed in **(D)** *clfB* gene. Asterisks indicate the reference strain used in each case. The numbers above each gene show the position. Gene structures were adapted from Cheng *et al*.[20] for *ebh*, Perkins *et al*.[23] for *clfB*, and Loughman *et al*.[22] for *fnbA*. FIVAR, found in various architecture; GA, albumin-binding module.

4. Discussion

The respiratory tract microbiota is essential for the maintenance of a healthy respiratory system. Among these, however, the opportunistic pathogens *H. influenzae* and *S. aureus* often colonise the respiratory tract[69,70,129]. To survive in the harsh environment of the respiratory tract for long periods, these bacteria use genetic adaptation, biofilm formation, and intracellular invasion to deal with the selective pressures imposed by the immune system and antimicrobial therapies[6,8,12,156]. Although most interactions with humans are asymptomatic, changes in an individual's health or adaptive changes in the bacteria can lead to serious infections, such as pneumonia, meningitis, or bacteraemia[42,115]. In addition, antimicrobial agents, the immune response, and vaccines can alter colonisation patterns, modifying the epidemiology of infections caused by these microorganisms[157,158]. Epidemiological surveillance, along with a better comprehension of the mechanisms involved in colonisation, are essential to control and prevent severe infections.

This thesis focused on the microbiological factors that influence colonisation and subsequent infection by *H. influenzae* and *S. aureus*. In the first part, we studied the epidemiology of invasive *H. influenzae* disease, focusing on the population structure of NTHi and the pangenome of emerging capsulated serotypes. The second part focused on the genetic adaptive changes and pathophysiological mechanisms involved in *H. influenzae* and *S. aureus* colonisation.

H. influenzae epidemiology and population structure

H. influenzae was a major cause of invasive disease, with serotype b the leading cause of meningitis in children. However, introduction of the serotype b conjugate vaccine changed the global epidemiology of *H. influenzae*, such that today, most cases are caused by NTHi, followed by non-type b serotypes, and affect elderly adults with underlying conditions[80,159,160]. As the epidemiology of *H. influenzae* disease has evolved, attention has moved to non-vaccine preventable serotypes and the increasing prevalence of β-lactam resistance[55,161]. Thus, **Studies 1 and 2** in this thesis focused on

the epidemiological surveillance, antimicrobial resistance profile, and genetic diversity of *H. influenzae*. Understanding the population structure can help prevent the emergence of clones that cause severe infections.

Incidence

The aim of **Study 1** was to provide an update on invasive *H. influenzae* disease at Bellvitge University Hospital for the 2014-2019 period compared to the 2008-2013 period. The overall incidence of invasive *H. influenzae* disease in adults was stable at 2.07 cases per 100,000 population between the two study periods, consistent with the incidence in other countries[62,64,162]. The introduction of the serotype b conjugate vaccine in the national immunisation programme has provided direct and herd immunity thanks to its high effectiveness and coverage, explaining the observed low incidence and reduction in serotype b infections[60]. Although the incidence remained stable and low over time in our data, some studies have reported a slight increase in invasive *H. influenzae* disease due to non-type b clones[61,163,164]. Epidemiological surveillance should therefore continue.

Invasive disease primarily affects patients at both extremes of age, especially neonates or older people with underlying comorbidity[61,65]. Most patients with invasive disease in **Study 1** had an underlying condition, with immunosuppressive therapy, solid organ malignancy, and diabetes most common. Severe underlying conditions were the leading cause of death, particularly in young adults. However, more cases occurred in people older than 65 years with a high Charlson comorbidity index and underlying disease (cardiac, diabetes, and COPD) than in younger adults, indicating that these are potential risk factors for invasive disease.

The overall 30-day mortality rate was 18.0%, within the 10-20% range found in other studies[165,166]. Of note, the 30-day mortality decreased between the two periods from 20.7% to 14.7%, most likely due to improved diagnosis and management of bloodstream infection.

Antimicrobial susceptibility

H. influenzae is predominantly susceptible to clinically relevant antibiotics, but ampicillin resistance is a matter of concern due to β-lactamase activity[167].

Antimicrobial resistance rates in Bellvitge University Hospital remained constant between 2008 and 2019, but ampicillin resistance increased slightly from 10% to 17.6%, consistent with other studies indicating ampicillin resistance rates of 13%-28%[167–169]. β-lactamase activity in our strains explained the ampicillin resistance, with TEM-1 β-lactamase. This was detected in all but one strain that carried a pB1000 plasmid with ROB-1 β-lactamase. San Millan *et al*.[170] reported that the presence of this plasmid carrying the ROB-1 enzyme reduced the fitness of *H. influenzae*, potentially explaining the low prevalence of this β-lactamase type. It should be noted that β-lactamase production was mainly associated with ST103, ST160, and ST836 clones, as observed elsewhere in Europe[67,161,169].

The accumulation of mutations in *fstI*, which codes for PBP3, is another mechanism of resistance to β-lactams, though this appeared less frequently than β-lactamases and BLNAR isolates remain uncommon[167]. In **Study 1**, some strains had amino acid substitutions in PBP3, and most were classified as Dabernat *et al.*[81] group II strains. These show reduced susceptibility to ampicillin (minimum inhibitory concentration, 1 mg/L), and are the most common and widespread among isolates with PBP3 alterations[171,172]. By contrast, group III strains were not found in our study. These are associated with high-level resistance to ampicillin and cephalosporins, and although mainly distributed in Asia[173,174], some cases have been reported in northern Europe[171]. Continued surveillance of BLNAR resistance is therefore required.

Study 1 also revealed the emergence of fluoroquinolone resistance between 2014 and 2019, with no ciprofloxacin-resistant strains identified in the 2008-2013 period[175]. Fluoroquinolone resistance is explained by the accumulation of alterations in the quinolone resistance determining regions (QRDRs) of GyrA (S84L and D88G/N) and ParC (S84I)[176], the main targets of these antimicrobials. Despite the low worldwide

prevalence of fluoroquinolone resistance in *H. influenzae*[83,177], surveillance must still monitor its progression. In fact, Tanaka *et al.*[178] reported the emergence of an ST422 clone in the community associated with therapeutic failure. This showed reduced quinolone susceptibility due to alterations in the QRDRs of GyrA and ParC. The frequent use of fluoroquinolones to treat respiratory infections may expedite the development of resistance and the widespread of resistant clones.

Serotype distribution

Changes in the epidemiology of *H. influenzae* disease have been observed since the introduction of the serotype b vaccine. This is evidenced by the significant drop in cases of invasive disease, particularly among children younger than 5 years, from 8.13 million of cases in 2000 to 340,000 cases in 2015[60].

Our study showed that NTHi was the most prevalent cause of invasive *H. influenzae* disease, accounting for 84.8% of cases. Capsulated strains followed by some distance. Other studies agree that NTHi strains are the most frequent cause of invasive disease, ranging from 54.6% to 82%, and suggest a 3% annual increase[64–67]. Despite the observed increase in the prevalence of invasive NTHi disease, the mechanisms underlying this change are poorly understood. Langereis *et al*.[165] suggested a few explanations. First, strain replacement could have occurred due to the serotype b and pneumococcal conjugate vaccines, both of which have been linked to increased NTHi colonisation, leading to increased transmission among people susceptible to invasive infection. Second, invasive NTHi strains could be more virulent due to their genetic competence, which allows for a high frequency of DNA exchange between strains and facilitates the acquisition of new virulence factors. Finally, demographic changes, such as the ageing population and increase in predisposing factors could favour NTHi infection through favourable changes in immune status.

Serotype f was the second leading cause of invasive disease, accounting for 10.9% of cases in Bellvitge University Hospital between 2014 and 2019. According to the European Centre for Disease Prevention and Control[63], serotype f isolates cause 9% of

invasive infections, show an annual increase in prevalence, and represent the most common serotype in invasive disease across Europe. The severity of invasive serotype f infection is particularly notable because it can affect immunocompetent patients and leads to more than one-third of patients being admitted to intensive care units[164]. In **Study 2**, we therefore provided a pangenome analysis of the serotype f isolates associated with colonisation and invasive disease, comparing them with other capsular serotypes of *H. influenzae*.

Only 1.4% of patients with invasive *H. influenzae* disease had serotype b isolates. According to the European Centre for Disease Prevention and Control[63], serotype b isolates caused 7% of invasive infections, with marked reductions since the serotype b vaccination was introduced. However, in some countries with high vaccine coverage, such as Italy and Portugal, serotype b (10.4% and 13.5%, respectively) has been responsible for more cases of invasive disease than serotype f (5.2% and 3.1%, respectively)[67,161], highlighting the importance of epidemiological surveillance.

Although rare in the pre-vaccine era, serotype a is currently emerging as a cause of severe invasive infection, causing serious morbidity and mortality in the USA, Brazil, Australia, and Canada, where it accounts for up to 23.1% of invasive infections[66,68,165]. Population genetics, environmental factors, and socioeconomic factors may contribute to the increased colonisation and high transmission of this serotype in certain populations, although the exact reasons are unknown[55,68]. Several studies indicate that most of the serotype a infections occurred in children, explaining why it was not detected in our study of adults.

Population structure

Continuous surveillance of the population structure is required to control the evolution of *H. influenzae* and develop effective prevention strategies over time. Therefore, **Study 1** emphasized the NTHi population structure, whereas **Study 2** was focused on the capsulated clones.

The NTHi strains constituted a heterogeneous genetic population compared to the capsulated strains. Homologous recombination allows continuous evolution and adaptation through the exchange of genetic material within bacteria of the same ecological niche and affect their commensal and virulent behaviours. The environment influences this exchange: bacteria able to interact with each other may experience a higher rate of recombination than those living in less diverse environments[179]. Some studies have shown higher levels of homologous recombination in NTHi than in capsulated isolates, which combined with the absence of the physical barrier produced by the polysaccharide capsule, could facilitate genetic exchange and thus diversity in NTHi population[180–183]. In fact, the recombination analysis performed with strains in **Study 1** revealed that many sites were affected by recombination, especially in NTHi (r/m = 7.85) compared with capsulated isolates (r/m = 1.14).

Since most NTHi strains belong to unique STs, MLST classification cannot adequately study the heterogeneous NTHi population and might miss details of the genetic diversity, particularly if the sample size is small and the clones do not belong to the most prevalent STs. Therefore, to identify NTHi subpopulations, we propose combining phylogenetic analysis with a clade classification based on the presence or absence of 17 accessory genes coding for surface associated proteins and virulence determinants[28,33].

We discovered that most NTHi strains isolated in invasive disease (51.7%) belonged to clade V and had the same monophyletic origin, followed by clades VI (31.0%) and II (6.9%), with clades I, III, and IV observed rarely. Furthermore, clade V contained the most common STs, including ST3, ST103, and ST160, known to have the highest prevalence[67,161], and increased in frequency over time. Clade V components also stood out for being associated with the production of β-lactamases, with 9 of the 12 strains containing this enzyme found in this clade. It is worth considering whether the rise in β-lactam resistance reflects the emergence of these clones or whether β-lactamase production confers an advantage over less common clones. In any case, the evolution

of clade V clones in invasive disease should be monitored. By contrast, the population structure of strains colonising the respiratory tract in patients with COPD, included in Study 3, showed a different clade distribution. Most of the strains belonged to clade V (39.1%), but this was in a lower proportion than in isolates from invasive disease, followed by clades III (17.3%) and II/IV (13.1%, each) at higher proportions. These findings suggest that the population structure of NTHi differs between clones that cause invasive disease and respiratory tract colonisation, though larger studies with more strains from different geographic regions are needed to confirm this observation.

Each capsulated serotype conforms a highly clonal and phylogenetically related population, with only a few clonal complexes associated with each capsular serotype. For example, serotype a was associated with CC23 (n = 176), CC62 (n = 18), CC4 (n = 8), CC372 (n = 2), and CC1755 (n = 1); serotype b, with CC6 (n = 150), CC50 (n = 13), and CC464 (n = 1); serotype c, with CC7 (n = 34); serotype d, with CC10 (n = 10); serotype e, with CC18 (n = 152); and serotype f, with CC124 (n = 222) and CC16 (n =12). Similarly, previous studies using multi-locus enzyme electrophoresis, pulsed-field gel electrophoresis, MLST, and whole genome sequencing[32,184–186] have revealed low genetic diversity within each serotype, implying that each emerged once within the population[31]. However, serotypes a and b had more phylogenetic variability, showing multiple lineages that contrasted to the monophyletic origins of the other serotypes. The phylogenetic analysis of capsulated strains revealed two major phylogenetic clades, one including serotypes a, b, c, and d, and another including serotypes a, b, e, and f, consistent with previous findings[32]. Due to the increasing importance of serotype a in invasive disease, studies have recently focused on the structure of its population[62,187]. Topaz *et al*.[188] defined four serotype a clades, with clades I (CC23) and II (CC4) related to serotypes b, c, and d, and clades III (CC60) and IV (CC62) related to serotypes b, e, and f. These clades correspond with those identified in Study 2, though only three could be defined because of the lack of CC60 strains.

Pangenome of capsulated *H. influenzae*

Understanding the genetic background of *H. influenzae* can improve our comprehension of its epidemiological evolution. Although NTHi is the most common cause of both invasive and non-invasive disease, information on the genomes of non-type b capsulated strains is needed because of their involvement in invasive infection. Few studies have been performed using large datasets of capsulated genomes, so in Study 2, we aimed to fill this knowledge gap and further characterise the pangenome of serotype f, which is a rising cause of invasive disease[61,189,190].

The inclusion of all capsulated *H. influenzae* genomes available in the National Centre for Biotechnology Information (NCBI) and European Nucleotide Archive (ENA) (n = 800) enabled population structure comparison between different serotypes and provided an overview of their pangenome or supragenome, comprising the entire set of genes not contained by any particular isolate but available across a genetically diverse population. The pangenomic analysis of capsulated genomes revealed 6,360 genes, with only 1,037 belonging to the core or soft-core genome (i.e., genes present in ≥95% of all genomes). Pinto *et al*.[28] reported a smaller gene pool, which could be explained by the lower number of genomes included in their study. Furthermore, consistent with previous findings[191] and regardless of the serotype, all isolates carried roughly the same number of genes per genome (mean = 1,752). Maintenance of the overall number of genes per genome and the high fraction of accessory genes detected in the gene pool suggest that capsulated *H. influenzae* strains have a balance between gene acquisition and loss that could serve as a reservoir for DNA exchange. Hogg *et al.*[27], developed a supragenome model that predicted a species-level pangenome of 5,000 or more genes with a core genome of about 1,400 genes. Despite the differences caused by the inclusion of NTHi strains in their analysis, the pangenomic richness observed in the capsulated strains of our study is consistent with their model. This diversity ensures the survival of the entire population in different environments, rather than the survival of an individual organism[29].

Despite the observed pangenomic diversity in the overall capsulated population, the proportion of core genes increased when each serotype was considered individually. As previously observed, this reflected the fact that each serotype comprised a highly clonal population[184–186].

Serotype b had the largest gene pool (n = 3,517 genes) and the largest accessory genome (57.3%, i.e., genes present in <95% of the genomes). This higher diversity in pangenome composition compared to the other serotypes could be explained by its high prevalence prior to conjugate vaccine introduction, which provided more opportunities to acquire mutations and transfer genetic material to other microorganisms. In addition, its ability to escape the complement system and produce haemocin, a bacteriocin active against non-type b serotypes, favour respiratory colonisation and thus genetic diversity over other serotypes[185,192].

By contrast, serotypes a, e, and f had 2,682 (45.2% accessory), 2,666 (39.0% accessory), and 2,969 (44.7% accessory) genes in their gene pools, respectively. This reflected greater genetic diversity than seen with serotypes c and d, but less than with serotype b. Genetic diversity may promote bacterial survival in different ecological niches[29], explaining the successful emergence of invasive disease caused by serotypes a, e, and f since the introduction of the conjugate vaccines[68,190]. Indeed, serotypes c and d presented low variability, having 1,935 (16.2% accessory) and 1,896 (14.3% accessory) genes. Nevertheless, the pangenome composition of serotypes c and d is still unclear and require further research because they are not only isolated rarely but also have a low number of sequenced strains.

Genetic stability in *H. influenzae* serotype f

Genetic diversity may benefit overall survival of the species, but genetic stability may confer advantages in the expansion of specific clones. This appears to be the case for CC23 from serotype a and CC124 from serotype f, which are more abundant and successful than other clones of the same serotypes[68,191,193]. CC23 has been identified as the main cause of invasive disease among serotype a clones, causing severe infections

that require admission to intensive care in 47.5% of cases[194]. Despite primarily affecting North America, cases of invasive disease due to CC23 are emerging in Europe[195,196]. Conversely, CC124 prevails among serotype f isolates, irrespective of the country, year, or source of isolation[190,191]. In 2004, Bruun *et al*.[197] studied serotype f clones from Denmark and the United States, as well as non-invasive and historical isolates from 1945 to the 1970s. This revealed a long-term stable clonal lineage in different countries over more than 50 years. In conjunction with the findings of Study 2, CC124 appears to have been genetically stable over time, indicating the importance of genetic homogeneity in the clonal expansion of this serotype. Capsular switching, in which serotype b recombines with a non-type b serotype to exchange capsular operon genes and thus evade vaccine-mediated immunity, could explain the increased prevalence of CC23 and CC124. Another explanation could be capsular replacement, with non-type b serotypes exploiting the decrease in colonisation by serotype b due to vaccine-mediated immunity. Phylogenetic analysis suggests that the increases in CC23 and CC124 most likely reflects capsular replacement rather than capsular switching, which also appears to be uncommon in this species[190,194].

Invasive disease due to serotype a is still uncommon in our setting, as evidenced in Study 1, where no cases were observed. Therefore, Study 2 focused on the genomic diversity of *H. influenzae* serotype f, using a multicentre collection of colonising and invasive isolates from the Netherlands, Portugal, and Spain. Serotype f was shown to have few SNPs and a large core genome. Despite this, CC124 strains were classified in three distinct clades by phylogenetic analysis and allelic variant specificity: clades I and II included colonising and some invasive strains, while clade III was exclusively associated with invasive strains. Although more colonising strains should be included, the results suggest that genetic variability in serotype f, despite being low, is introduced by colonising strains. A previous study found a positive association between colonisation duration and recombination rates in *S. pneumoniae*, which inhabits the same niche as *H. influenzae*[198]. Therefore, invasive clones may be more genetically homogeneous,

while colonising strains that have the opportunity to interact with other bacteria may take part in more genetic exchange, increasing their relative genetic diversity.

Distinguishing serotype f isolates from other serotypes, apart from capsular genes, could facilitate efforts to monitor the evolution of this emerging serotype. Study 2 provides a first approximation of the genetic determinants associated with each serotype. It would be interesting for future research to identify genetic variations between colonising and invasive capsulated strains, because the lack of clinical data for the NCBI and ENA genomes hampered this characterisation.

Adaptive evolution and colonisation in *H. influenzae* and *S. aureus*

Bacterial colonisation is often the first step before the development of severe invasive infection, and as such, can affect the epidemiology of these diseases[38,132,133]. Characterising the mechanisms that underlie bacterial adaptation and colonisation, such as biofilm formation and intracellular invasion, may not only improve our understanding of the role of opportunistic pathogens in disease progression but also enable the development of strategies to prevent the spread of infection. Thus, the second objective of this thesis focused on the adaptive evolution that occurs during persistent colonisation and the mechanisms that enable persistent colonisation by *H. influenzae* (Study 3) and *S. aureus* (Studies 4 and 5), important colonising bacteria of the respiratory tract that can cause severe disease.

Genetic evolution in persistent *H. influenzae*

NTHi colonises the lower respiratory tract of up to 50% of patients with COPD, one of the most prevalent chronic respiratory diseases[199,200], and is the primary cause of acute exacerbations[72,73]. Bacteria must adapt to selection pressures, co-occurring species, and antimicrobial therapies to maintain persistent colonisation. Common adaptation mechanisms include the development of antimicrobial resistance, increased mutation rates, and changes in virulence factors[6].

Study 3 focused on the adaptive evolution that takes place during persistent *H. influenzae* colonisation. Most of the genetic changes that occurred during persistent colonisation affected genes that encode proteins involved in cell wall synthesis and structure. The expression of many of the genes involved in LOS synthesis is regulated by phase variation, which allows for rapid adaptation to the environment and increases the ability to colonise persistently[201]. In fact, most cases of *H. influenzae* persistence showed changes in phase variation of the genes *licA*, *lex1*, and *lic3A*.

The *licA* gene, also known as *lic1A*, encodes a PCho kinase involved in the addition of PCho to LOS. It has been described that PCho activates the complement system through its interaction with C-reactive protein. Changes in *licA* expression result in altered PCho levels, which in turn, affects sensitivity to serum bactericidal activity[202,203]. Several studies indicate that strains expressing *licA* gene are selected during nasopharyngeal colonisation, while strains not expressing *licA* are cleared faster[204,205]. Swords *et al*.[206], reported the presence of PCho on the NTHi cell surface promotes adhesion and bacterial invasion of bronchial cells via its interaction with the PAF receptor, and Langereis *et al*.[207] reported that invasive NTHi isolates had lower PCho levels than nasopharyngeal isolates. Thus, *licA* seems relevant to both complement response regulation and persistent respiratory tract colonisation via adhesion and intracellular invasion.

The *lex1* gene, also known as *lic2A*, encodes a galactosyltransferase that enables the attachment of a galactose residue to the LOS and further hexose extensions in the presence of the *lgtC* gene. Clark *et al.*[208] demonstrated that the presence of galactose on the bacterial surface contributes to its survival in the presence of human serum.

The *lic3A* gene encodes α-2,3-sialyltransferase responsible for the addition of N-acetylneuraminic acid (Neu5Ac) to the LOS. Hosking *et al*.[209] suggested that this gene's expression is important for colonisation, based on the finding that most *H. influenzae* strains colonising the nasopharynx had the ability to express *lic3A* whereas bloodstream and cerebrospinal fluid populations lacked this ability. Other sialyltransferases, such as

Lic3B, SiaA, and LsgB, are involved in decorating the LOS with sialic acids, suggesting the importance of this process[210,211]. It has been demonstrated that the sialic acid on the surface of *H. influenzae* mimics the glycosphingolipid antigens found in human cells, which helps to avoid recognition by the immune system, increase serum resistance[212,213], and favour biofilm formation[214].

Further research should determine whether the changes in phase variation observed in the persistent strains in Study 3 affect the expression of these genes. Given that most of the strains had open reading frames compatible with gene expression, we could speculate that their expression favours immune evasion, adhesion, and intracellular invasion, and in doing so, the maintenance of persistent colonisation.

Genes involved in inorganic ion metabolism and transport also showed multiple changes during persistent colonisation, particularly *hgpB* and *hgpC* encoding TonB-dependent receptors. *H. influenzae* requires heme for aerobic growth and has multiple mechanisms, including *hgpA*, *hgpB*, and *hgpC*, to use haemoglobin and haemoglobin-haptoglobin complexes as heme sources, with phase variation also able to modify their expression[215]. Complete deletion of the three *hgp* genes renders the bacteria unable to use haemoglobin-haptoglobin complexes, while the expression of only one is sufficient to recover haemoglobin-haptoglobin uptake[216]. Although the effects of genetic variation in these iron transporters are unknown, the presence of multiple *hgp* genes suggests that a population expressing a specific Hgp may be selected to overcome the host immune response directed against another Hgp[217].

Another gene that showed recurrent alterations was *fadL*, a gene coding for a long-chain fatty acid outer membrane transporter. Changes in this gene were associated with the insertion or deletion of one nucleotide, resulting in the introduction of premature stop codons at different positions and the appearance of alternative open reading frames that could potentially code for shorter variants of the protein. Moleres *et al.*[45] also observed these alterations within multiple lineages and patients. They suggested that loss of FadL function impairs the entry of NTHi into epithelial cells while increasing resistance to the

bactericidal effects of long-chain fatty acids, which are abundant inflammatory mediators in the damaged lung tissue of patients with COPD.

Other studies of the pathological adaptation of *H. influenzae* have also found changes in these genes [45,218,219], suggesting that they frequently mediate genetic adaptation and could provide a selective advantage that favours bacterial survival and colonisation over long periods. Nevertheless, genome regions associated with integrated prophages showed several genetic changes. These could have resulted from selective pressures exerted by the bacteria and could direct prophage adaptation to the genomes where they have been integrated[220,221].

Impact of azithromycin on antimicrobial resistance

Patients with chronic respiratory diseases receive many antibiotic, and all can significantly affect the microbiota, favouring the development of antimicrobial resistance and enhancing the selection of resistant bacteria[222,223]. Participants in Study 3 received long-term azithromycin therapy, which has been shown to reduce the frequency of exacerbations and improve the quality of life due to its immunomodulatory and anti-inflammatory properties[224–226]. However, azithromycin also inhibits bacterial protein synthesis by interfering with the assembly of the 50S large ribosomal subunit[227]. Long-term therapy can therefore lead to macrolide resistance, as observed in our strains. All *H. influenzae* strains isolated before azithromycin treatment were susceptible, but they developed resistance within the first months of treatment due to mutations, as previously described in *H. influenzae*[228]. These included mutations in the macrolide-binding region of the 23S rRNA (A2058), as well as in the genes coding for the L4 (G65D) and L22 (K90E and an insertion between residues 78 and 79) ribosomal proteins. Antibiotic pressures triggered the gradual acquisition of the A2058G mutation in 23S rRNA by the different copies, resulting in a progressive increase in the minimum inhibitory concentration to macrolides in some strains. Before this change, strains had a premature stop codon in *acrR*, a negative regulator that causes the overexpression of the AcrAB efflux pump. The *acrR* mutation had no effect on azithromycin susceptibility,

implying that it is not directly related to resistance, but instead, may favour subsequent mutations in other genes, as previously reported for *acrB*[229].

The selection pressure exerted by long-term azithromycin treatment resulted in the development of these chromosome mutations that could not be transmitted and spread readily to other microorganisms. However, the acquisition of an integrative conjugative element carrying the MEGA element, which provides resistance due to the presence of MefE and MsrD expulsion pumps, was also observed in one case of persistent *H. influenzae*. Of note, *H. parainfluenzae* strains carrying this integrative conjugative element were found in the patient from whom these *H. influenzae* strains were isolated. This suggests the possibility of horizontal transmission between species that share the same ecological niche, supporting previous studies that indicated *H. parainfluenzae* could serve as a reservoir of resistance genes[230,231].

Azithromycin therapy also altered the pattern of colonising bacteria in the respiratory tract. Prior to treatment, *H. influenzae* and *M. catarrhalis* were most abundant, as already reported in both stable phases and acute exacerbations[232]. However, after azithromycin treatment started, *H. parainfluenzae*, *H. influenzae*, *P. aeruginosa*, and *S. maltophilia* became the most frequently isolated bacteria. Thus, azithromycin treatment appears to eradicate the carriage of azithromycin-susceptible microorganisms, like *M. catarrhalis*[233], and promotes the selection of intrinsically resistant ones, such as *P. aeruginosa*[234] and *S. maltophilia*[235]. Furthermore, the selection of *H. parainfluenzae* is likely related to two factors. First, its high genetic plasticity, which may favour adaptation. Second, its intrinsic antimicrobial resistance, which may facilitate a role as a reservoir for resistance genes[231,236,237]; under selective pressure by azithromycin, these may be spread to other bacteria sharing the ecological niche.

Overall, the horizontal acquisition and transmission of mobile genetic elements is of greatest concern because it allows macrolide resistance to spread among colonising pathogens in the respiratory tract. Macrolides are widely used to treat respiratory

infections[238], so the presence of colonising bacteria with acquired and transmissible resistance mechanisms may complicate their treatment.

Adaptive evolution in *S. aureus*

Biofilm formation and intracellular invasion of host cells provides the perfect niche for bacteria to avoid the hostile environment of the upper respiratory tract and survive for long periods[131,239]. *S. aureus* colonisation increases the risk of serious infection and both the transmission and spread of clones among patients[240,241]. Thus, **Study 4 and the Additional Results** focused on the genetic characterisation of biofilm formation and intracellular invasion, respectively. These showed the importance of different variants of, or alterations in, cell wall protein-coding genes for regulating bacterial-host interactions.

In **Study 4**, we compared differences in biofilm formation between endemic and sporadic MRSA clones associated with bloodstream infections. The endemic CC5 and CC8 MRSA clones, the most frequent strains associated with bloodstream infection in Spain[242], formed stronger biofilms than strains from other clonal complexes. This observation, which is consistent with previous studies[243–246], suggests that biofilm formation in endemic clones may be an important physiological process that enhances bacterial survival in the host or hospital environment, allowing these bacteria to become successful pathogens. Furthermore, the *sasG* gene encoding a cell wall anchored protein, which is associated with biofilm formation and nasal colonisation[111,247], was found in all endemic clonal complex isolates and only a few sporadic ones. The presence of this gene may give endemic strains an advantage over sporadic strains, allowing them to form better biofilms and thus create environments that favour bacterial persistence and adaptation. In addition, *sasG* has two allelic variants defined by the variable region of the A domain[248,249] that also correlated with biofilm formation. Allelic variant 1 was associated with the endemic CC5 and CC8 clones, higher biofilm production, and higher gene expression, while variant 2 was associated with sporadic clones, lower biofilm production, and lower gene expression. These results support the findings by Cassat *et al*.[113], suggesting that *sasG* variants have different functions. Therefore, **Study 4**

highlights the importance of *sasG* allelic variant 1 as a virulence factor, suggesting that it confers an adaptive benefit over strains with allelic variant 2 or without *sasG*.

S. aureus can also colonise the host as an intracellular pathogen, being able to survive within neutrophils, epithelial cells, and endothelial cells, thereby conferring protection against host defences and antimicrobial agents[108,156]. The bacteria can persist in this state and provide a reservoir for relapsing infection that could be associated with therapeutic failure[250].

In the **Additional Results** we explored the ability of the bacteria to invade human cells and studied the genetic profile of near-isogenic bacterial pairs isolated from colonising and bacteraemic periods in single patients. When compared to respiratory isolates, *S. aureus* isolated from blood samples had lost its ability to enter host cells. Genetic signature analysis of the bacteraemic isolates revealed changes in three surface protein-coding genes that regulate bacteria-host binding interactions: *clfB*, coding for clumping factor B (ClfB); *fnbA*, coding for fibronectin-binding protein A (FnBPA); and *ebh*, coding for extracellular matrix binding protein (Ebh).

ClfB is expressed during the exponential growth phase and is essential for successful nasal colonisation. It facilitates adherence to desquamated epithelial cells by the binding of its N-terminal region to fibrinogen and cytokeratin 10 exposed on the host cell surface. It also adheres to blood clots, biomaterial, and damaged heart valves, promoting bacterial cell aggregation in the presence of fibrinogen[251,252]. FnBPA mediates the aggregation of platelets and promotes colonisation by binding to fibronectin, fibrinogen, and elastin. The ligand binding properties of FnBPA were conserved despite antigenic variation, which probably enhances immune system evasion[253]. Finally, Ebh is the largest protein (~1.1 MDa) encoded by the *S. aureus* genome. This bacterial cell wall protein that mediates fibronectin, plasminogen, and albumin binding, provides resistance to the complement system, and contributes to the structural stability in the presence of osmotic pressures[254,255]. Furthermore, Walker *et al*.[256] suggest that *S. aureus* can aggregate with itself to increase its pathogenesis in some *in vivo*

environments, such as in the bloodstream. They observed that Ebh prevented *S. aureus* self-aggregation and concluded that the repression of *ebh* expression could facilitate *S. aureus* progressing to invasive disease.

The results obtained in **Study 4** and the **Additional Results** indicate that SasG, ClfB, FnBPA, and Ebh represent important virulence factors that could increase bacterial survival and persistence in hostile environments through biofilm formation, adherence, or intracellular invasion.

5. Conclusions

H. influenzae epidemiology and population structure

1. Nowadays, invasive *H. influenzae* disease is mainly associated with non-typeable (NTHi) strains, followed by serotype f strains. Its incidence increases with age, especially in patients with underlying comorbidities.

2. The NTHi population is genetically diverse, but phylogenetic analysis using a clade classification based on the accessory genome revealed that most invasive disease belongs to clade V and is associated with β-lactamase production.

3. Each serotype represents a genetically homogeneous population. However, the pangenome composition of serotype b is more diverse than the other serotypes, which could be explained by its high prevalence before the conjugate vaccine was introduced.

4. Serotype f strains associated with invasive disease exhibit high genetic stability, regardless of country or time of isolation, highlighting the importance of genetic homogeneity to the clonal expansion of this serotype.

Adaptive evolution and colonisation in *H. influenzae* and *S. aureus*

1. Changes associated with phase variation in genes encoding proteins involved in cell wall structure and inorganic ion transport favour *H. influenzae* adaptation during persistent respiratory tract colonisation.

2. Antibiotic pressure exerted by long-term azithromycin treatment in patients with chronic obstructive pulmonary disease can lead to macrolide resistance in *H. influenzae* due to mutations in the 23S rRNA, modifications in the ribosomal proteins L4 and L22, and the acquisition of MsrD and MefE efflux pumps.

3. Horizontal transmission of genetic elements containing antimicrobial resistance determinants is a cause for concern, and *H. parainfluenzae* may act as a reservoir for other species in the same niche.

4. Methicillin-resistant *S. aureus* (MRSA) isolates from the most prevalent clones that caused bacteraemia in Spain between 2008 and 2015 (CC5 and CC8) developed stronger biofilms than isolates belonging to sporadic MRSA clones. Therefore, biofilms may enhance bacterial persistence in the human host or environment.

5. Allelic variant 1 of *sasG*, encoding *S. aureus* surface protein G, was present in the most prevalent *S. aureus* clones (CC5 and CC8) and correlated with increased biofilm formation. This may be considered an important virulence factor that favours bacterial survival.

6. During respiratory colonisation by *S. aureus*, changes may occur in the bacteria-host interaction. Respiratory isolates had a greater ability to invade host cells, compared with blood isolates, which showed changes in genes associated with binding to the host extracellular matrix, such as clumping factor B (*clfB*), extracellular matrix binding protein (*ebh*), and fibronectin-binding protein A (*fnbA*).

6. References

1. Man, W. H., De Steenhuijsen Piters, W. A. A. & Bogaert, D. The microbiota of the respiratory tract: gatekeeper to respiratory health. *Nat. Rev. Microbiol.* 15, 259–270 (2017).

2. Welp, A. L. & Bomberger, J. M. Bacterial community interactions during chronic respiratory disease. *Front. Cell. Infect. Microbiol.* 10, 213 (2020).

3. Wypych, T. P., Wickramasinghe, L. C. & Marsland, B. J. The influence of the microbiome on respiratory health. *Nat. Immunol.* 20, 1279–1290 (2019).

4. Toews, G. B. Impact of bacterial infections on airway diseases. *Eur. Respir. Rev.* 14, 62–68 (2005).

5. Bakkeren, E., Diard, M. & Hardt, W. D. Evolutionary causes and consequences of bacterial antibiotic persistence. *Nat. Rev. Microbiol.* 18, 479–490 (2020).

6. Cullen, L. & McClean, S. Bacterial adaptation during chronic respiratory infections. *Pathogens* 4, 89 (2015).

7. Munita, J. M. & Arias, C. A. Mechanisms of antibiotic resistance. *Microbiol. Spectr.* 4, 1–37 (2016).

8. Foley, J. Mini-review: Strategies for variation and evolution of bacterial antigens. *Comput. Struct. Biotechnol. J.* 13, 407–416 (2015).

9. Saunders, N. J. *et al.* Repeat-associated phase variable genes in the complete genome sequence of *Neisseria meningitidis* strain MC58. *Mol. Microbiol.* 37, 207–215 (2000).

10. Deitsch, K. W., Lukehart, S. A. & Stringer, J. R. Common strategies for antigenic variation by bacterial, fungal and protozoan pathogens. *Nat. Rev. Microbiol.* 7, 503 (2009).

11. Dhital, S., Deo, P., Stuart, I. & Naderer, T. Bacterial outer membrane vesicles and host cell death signaling. *Trends Microbiol.* 29, 1106–1116 (2021).

12. Koo, H., Allan, R. N., Howlin, R. P., Stoodley, P. & Hall-Stoodley, L. Targeting microbial biofilms: Current and prospective therapeutic strategies. *Nat. Rev. Microbiol.* 15, 740–755 (2017).

13. Donlan, R. M. Biofilms: microbial life on surfaces. *Emerg. Infect. Dis.* 8, 881–90 (2002).

14. Proctor, R. A. *et al.* Small colony variants: a pathogenic form of bacteria that facilitates persistent and recurrent infections. *Nat. Rev. Microbiol.* 4, 295–305 (2006).

15. Johns, B. E., Purdy, K. J., Tucker, N. P. & Maddocks, S. E. Phenotypic and genotypic characteristics of small colony variants and their role in chronic infection. *Microbiol. Insights* 8, 23 (2015).

16. Prat, C. & Lacoma, A. Bacteria in the respiratory tract—how to treat? Or do not treat? *Int. J. Infect. Dis.* 51, 113–122 (2016).

17. Morens, D. M., Taubenberger, J. K., Folkers, G. K. & Fauci, A. S. Pandemic Influenza's 500th Anniversary. *Clin. Infect. Dis.* 51, 1442–1444 (2010).

18. Nørskov-Lauritsen, N. *Haemophilus*. in *Bergey's Manual of Systematics of Archaea and*

Bacteria 1–22 (Wiley Online Library, 2021).

19. Christensen, H. *et al. Pasteurellaceae*. in *Bergey's Manual of Systematics of Archaea and Bacteria* 1–26 (Wiley Online Library, 2020).

20. Nørskov-Lauritsen, N. Classification, identification, and clinical significance of *Haemophilus* and *Aggregatibacter* species with host specificity for humans. *Clin. Microbiol. Rev.* 27, 214–240 (2014).

21. Inzana, T. J., Johnson, J. L., Shell, L., Möller, K. & Kilian, M. Isolation and characterization of a newly identified *Haemophilus* species from cats: '*Haemophilus felis*'. *J. Clin. Microbiol.* 30, 2108–2112 (1992).

22. Zheng, M. L. *et al. Haemophilus seminalis* sp. nov., isolated from human semen. *Int. J. Syst. Evol. Microbiol.* 70, 2588–2595 (2020).

23. Public Health England. UK Standards for Microbiology Investigations: X and V factor test. *Bacteriol. Test Proced.* 38, 1–15 (2019).

24. Ledeboer, N. A. & Doern, G. V. *Haemophilus*. in *Manual of Clinical Microbiology* 667–684 (Wiley Online Library, 2015).

25. Lâm, T. T., Claus, H., Frosch, M. & Vogel, U. Sequence analysis of serotype-specific synthesis regions II of *Haemophilus influenzae* serotypes c and d: evidence for common ancestry of capsule synthesis in *Pasteurellaceae* and *Neisseria meningitidis*. *Res. Microbiol.* 162, 483–487 (2011).

26. National Center Biotechnology Information (NCBI). *Haemophilus influenzae* (ID 165) - Genome. *NCBI* www.ncbi.nlm.nih.gov/genome/165 (2021).

27. Hogg, J. S. *et al.* Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol.* 8, 1–18 (2007).

28. Pinto, M. *et al.* Insights into the population structure and pan-genome of *Haemophilus influenzae*. *Infect. Genet. Evol.* 67, 126–135 (2019).

29. Gilsdorf, J. R., Marrs, C. F. & Foxman, B. *Haemophilus influenzae*: Genetic variability and natural selection to identify virulence factors. *Infect. Immun.* 72, 2457–2461 (2004).

30. LaCross, N. C., Marrs, C. F. & Gilsdorf, J. R. Population structure in nontypeable *Haemophilus influenzae*. *Infect. Genet. Evol.* 14, 125–136 (2013).

31. Watts, S. C. & Holta, K. E. HICAP: *In silico* serotyping of the *Haemophilus influenzae* capsule locus. *J. Clin. Microbiol.* 57, e00190-19 (2019).

32. Potts, C. C. *et al.* Genomic characterization of *Haemophilus influenzae*: A focus on the capsule locus. *BMC Genomics* 20, 1–9 (2019).

33. De Chiara, M. *et al.* Genome sequencing of disease and carriage isolates of nontypeable *Haemophilus influenzae* identifies discrete population structure. *Proc. Natl. Acad. Sci.* 111, 5439–5444 (2014).

34. Riesbeck, K. Complement evasion by the human respiratory tract pathogens *Haemophilus influenzae* and *Moraxella catarrhalis*. *FEBS Lett.* 594, 2586–2597 (2020).

35. Kapogiannis, B. G., Satola, S., Keyserling, H. L. & Farley, M. M. Invasive infections with *Haemophilus influenzae* serotype a containing an IS*1016-bexA* partial deletion: possible association with virulence. *Clin. Infect. Dis.* 41, e97–e103 (2005).

36. Ohkusu, K., Nash, K. A. & Inderlied, C. B. Molecular characterisation of *Haemophilus influenzae* type a and untypeable strains isolated simultaneously from cerebrospinal fluid and blood: novel use of quantitative real-time PCR based on the cap copy number to determine virulence. *Clin. Microbiol. Infect.* 11, 637–643 (2005).

37. St. Geme, J. W. Molecular and cellular determinants of non-typeable *Haemophilus influenzae* adherence and invasion. *Cell. Microbiol.* 4, 191–200 (2002).

38. Garmendia, J. *et al.* Genotypic and phenotypic diversity in the noncapsulated *Haemophilus influenzae*: adaptation and pathogenesis in the human airways. *Int. Microbiol.* 15, 157–170 (2012).

39. St. Geme, J. W. The pathogenesis of nontypable *Haemophilus influenzae* otitis media. *Vaccine* 19, S41–S50 (2000).

40. Langereis, J. D. & De Jonge, M. I. Unraveling *Haemophilus influenzae* virulence mechanisms enable discovery of new targets for antimicrobials and vaccines. *Curr. Opin. Infect. Dis.* 33, 231–237 (2020).

41. Ahearn, C. P., Gallo, M. C. & Murphy, T. F. Insights on persistent airway infection by non-typeable *Haemophilus influenzae* in chronic obstructive pulmonary disease. *Pathog. Dis.* 75, 42 (2017).

42. Jalalvand, F. & Riesbeck, K. Update on non-typeable *Haemophilus influenzae*-mediated disease and vaccine development. *Expert Rev. Vaccines* 17, 503–512 (2018).

43. Duell, B. L., Su, Y.-C. & Riesbeck, K. Host-pathogen interactions of nontypeable *Haemophilus influenzae* : from commensal to pathogen. *FEBS Lett.* 590, 3840–3853 (2016).

44. Singh, B., Al-Jubair, T., Mörgelin, M., Thunnissen, M. M. & Riesbeck, K. The unique structure of *Haemophilus influenzae* protein E reveals multiple binding sites for host factors. *Infect. Immun.* 81, 801–814 (2013).

45. Moleres, J. *et al.* Antagonistic pleiotropy in the bifunctional surface protein FadL (OmpP1) during adaptation of *Haemophilus influenzae* to chronic lung infection associated with chronic obstructive pulmonary disease. *MBio* 9, 1–23 (2018).

46. Gilsdorf, J. R., Mccrea, K. W., & Marrs, C. F. Role of pili in *Haemophilus influenzae* adherence and colonization. *Infect. Immun.* 65, 2997–3002 (1997).

47. Carruthers, M. D. *et al.* Biological roles of nontypeable *Haemophilus influenzae* type IV pilus proteins encoded by the *pil* and *com* operons. *J. Bacteriol.* 194, 1933 (2012).

48. St. Geme, J. W. & Yeo, H. J. A prototype two-partner secretion pathway: the *Haemophilus influenzae* HMW1 and HMW2 adhesin systems. *Trends Microbiol.* 17, 355–360 (2009).

49. Spahich, N. A. & St. Geme, J. W. Structure and function of the *Haemophilus influenzae* autotransporters. *Front. Cell. Infect. Microbiol.* 1, 5 (2011).

50. Morton, D. J., Hempel, R. J., Seale, T. W., Whitby, P. W. & Stull, T. L. A functional *tonB* gene is required for both virulence and competitive fitness in a chinchilla model of *Haemophilus influenzae* otitis media. *BMC Res. Notes* 5, 327 (2012).

51. Rodríguez-Arce, I. *et al.* Moonlighting of *Haemophilus influenzae* heme acquisition systems contributes to the host airway-pathogen interplay in a coordinated manner. *Virulence* 10, 315–333 (2019).

52. Wong, S. M. S. & Akerley, B. J. Genome-Scale approaches to identify genes essential for *Haemophilus influenzae* pathogenesis. *Front. Cell. Infect. Microbiol.* 2, 23 (2012).

53. Swords, W. E. Quorum signaling and sensing by nontypeable *Haemophilus influenzae*. *Front. Cell. Infect. Microbiol.* 2, 100 (2012).

54. Peltola, H. Worldwide *Haemophilus influenzae* type b disease at the beginning of the 21st century: global analysis of the disease burden 25 years after the use of the polysaccharide vaccine and a decade after the advent of conjugates. *Clin. Microbiol. Rev.* 13, 317 (2000).

55. Wen, S., Feng, D., Chen, D., Yang, L. & Xu, Z. Molecular epidemiology and evolution of *Haemophilus influenzae*. *Infect. Genet. Evol.* 80, 104205 (2020).

56. Levine, O., Schwartz, B., Pierce, N. & Kane, M. Development, evaluation and implementation of *Haemophilus influenzae* type b vaccines for young children in developing countries: current status and priority actions. *Pediatr. Infect. Dis. J.* 17, S95-113 (1998).

57. Vella, M. & Pace, D. Glycoconjugate vaccines: an update. *Expert Opin. Biol. Ther.* 15, 529–546 (2015).

58. Tsang, R. S. W. A narrative review of the molecular epidemiology and laboratory surveillance of vaccine preventable bacterial meningitis agents: *Streptococcus pneumoniae*, *Neisseria meningitidis*, *Haemophilus influenzae* and *Streptococcus agalactiae*. *Microorganisms* 9, 1–19 (2021).

59. Consejo Interterritorial & Sistema Nacional de Salud. Calendario común de vacunación a lo largo de toda la vida. www.mscbs.gob.es/profesionales/saludPublica/prevPromocion/vacunaciones/calendario-y-coberturas/docs/CalendarioVacunacion_Todalavida.pdf (2021).

60. Slack, M. P. E., Cripps, A. W., Grimwood, K., Mackenzie, G. A. & Ulanova, M. Invasive *Haemophilus influenzae* infections after 3 decades of Hib protein conjugate vaccine use. *Clin. Microbiol. Rev.* 34, e00028-21 (2021).

61. MacNeil, J. R. *et al.* Current epidemiology and trends in invasive *Haemophilus influenzae* disease-United States, 1989-2008. *Clin. Infect. Dis.* 53, 1230–1236 (2011).

62. Soeters, H. M. *et al.* Current epidemiology and trends in invasive *Haemophilus influenzae* disease - United States, 2009-2015. *Clin. Infect. Dis.* 67, 881–889 (2018).

63. European Centre for Disease Prevention and Control. *Haemophilus influenzae* - Annual Epidemiological Report for 2018. *ECDC* www.ecdc.europa.eu/en/publications-data/haemophilus-influenzae-annual-epidemiological-report-2018 (2020).

64. Whittaker, R. *et al.* Epidemiology of invasive *Haemophilus influenzae* disease, Europe, 2007–2014. *Emerg. Infect. Dis.* 23, 396–404 (2017).

65. Blain, A. *et al.* Invasive *Haemophilus influenzae* disease in adults ≥65 years, United States, 2011. *Open Forum Infect. Dis.* 1, ofu044 (2014).

66. Tsang, R. S. W. *et al.* Antibiotic susceptibility and molecular analysis of invasive *Haemophilus influenzae* in Canada, 2007 to 2014. *J. Antimicrob. Chemother.* 72, 1314–1319 (2017).

67. Giufrè, M. *et al.* Increasing trend in invasive non-typeable *Haemophilus influenzae* disease and molecular characterization of the isolates, Italy, 2012-2016. *Vaccine* 36, 6615–6622 (2018).

68. Ulanova, M. & Tsang, R. S. W. *Haemophilus influenzae* serotype a as a cause of serious invasive infections. *Lancet Infect. Dis.* 14, 70–82 (2014).

69. Dabernat, H. *et al. Haemophilus influenzae* carriage in children attending French day care centers: A molecular epidemiological study. *J. Clin. Microbiol.* 41, 1664–1672 (2003).

70. Giufrè, M. *et al.* Nasopharyngeal carriage of *Haemophilus influenzae* among adults with co-morbidities. *Vaccine* 40, 826–832 (2022).

71. Verhagen, L. M., Luesink, M., Warris, A., De Groot, R. & Hermans, P. W. M. Bacterial respiratory pathogens in children with inherited immune and airway disorders: Nasopharyngeal carriage and disease risk. *Pediatr. Infect. Dis. J.* 32, 399–404 (2013).

72. Finney, L. J., Ritchie, A., Pollard, E., Johnston, S. L. & Mallia, P. Lower airway colonization and inflammatory response in COPD: a focus on *Haemophilus influenzae*. *Int. J. Chron. Obstruct. Pulmon. Dis.* 9, 1119–1132 (2014).

73. Van Eldere, J., Slack, M. P. E., Ladhani, S. & Cripps, A. W. Non-typeable *Haemophilus influenzae*, an under-recognised pathogen. *Lancet Infect. Dis.* 14, 1281–1292 (2014).

74. Clementi, C. F. & Murphy, T. F. Non-typeable *Haemophilus influenzae* invasion and persistence in the human respiratory tract. *Front. Cell. Infect. Microbiol.* 1, 1 (2011).

75. Khattak, Z. E. & Anjum, F. *Haemophilus influenzae*. *StatPearls* www.ncbi.nlm.nih.gov/books/NBK562176/ (2022).

76. Bakaletz, L. O. & Novotny, L. A. Nontypeable *Haemophilus influenzae* (NTHi). *Trends Microbiol.* 26, 727–728 (2018).

77. Kingery, J., Taylor, W. & Braun, D. Acute otitis media. *Osteopath. Fam. Physician* 8, 22–25 (2021).

78. Domínguez, M. Á. & Sopena Galindo, N. Infecciones causadas por *Haemophilus* y otros bacilos gramnegativos. in *Farreras-Rozman. Medicina Interna* 2129–2136 (Elsevier, 2020).

79. Shoar, S. & Musher, D. M. Etiology of community-acquired pneumonia in adults: a systematic review. *Pneumonia* 12, 1–10 (2020).

80. Heinz, E. The return of Pfeiffer's bacillus: Rising incidence of ampicillin resistance in *Haemophilus influenzae*. *Microb. Genomics* 4, 1–8 (2018).

81. Dabernat, H. *et al.* Diversity of β-lactam resistance-conferring amino acid substitutions in penicillin-binding protein 3 of *Haemophilus influenzae*. *Antimicrob. Agents Chemother.* 46, 2208–2218 (2002).

82. García-Cobos, S. *et al.* Ampicillin-resistant non-β-lactamase-producing *Haemophilus influenzae* in Spain: Recent emergence of clonal isolates with increased resistance to cefotaxime and cefixime. *Antimicrob. Agents Chemother.* 51, 2564–2573 (2007).

83. Puig, C. *et al.* Molecular characterization of fluoroquinolone resistance in nontypeable *Haemophilus influenzae* clinical isolates. *Antimicrob. Agents Chemother.* 59, 461–466 (2015).

84. Chen, D. *et al.* Microbial virulence, molecular epidemiology and pathogenic factors of fluoroquinolone-resistant *Haemophilus influenzae* infections in Guangzhou, China. *Ann. Clin. Microbiol. Antimicrob.* 17, 41 (2018).

85. Peric, M., Bozdogan, B., Jacobs, M. R. & Appelbaum, P. C. Effects of an efflux mechanism and ribosomal mutations on macrolide susceptibility of *Haemophilus influenzae* clinical isolates. *Antimicrob. Agents Chemother.* 47, 1017–1022 (2003).

86. Licitra, G. Etymologia: *Staphylococcus*. *Emerg. Infect. Dis.* 19, 1553 (2013).

87. Götz, F., Bannerman, T. & Schleifer, K. H. The Genera *Staphylococcus* and *Macrococcus*. *The Prokaryotes* 4, 75 (2006).

88. Lory, S. The Family *Staphylococcaceae*. in *The Prokaryotes: Firmicutes and Tenericutes* 363–366 (Springer, 2014).

89. Madhaiyan, M., Wirth, J. S. & Saravanan, V. S. Phylogenomic analyses of the *Staphylococcaceae* family suggest the reclassification of five species within the genus *Staphylococcus* as heterotypic synonyms, the promotion of five subspecies to novel species. *Int. J. Syst. Evol. Microbiol.* 70, 5926–5936 (2020).

90. Schleifer, K. Phylum XIII. Firmicutes. in *Bergey's Manual of Systematic Bacteriology* 19–1317 (Wiley Online Library, 2009).

91. Gherardi, G., Di Bonaventura, G. & Savini, V. *Staphylococcal Taxonomy*. in *Pet-to-Man Travelling Staphylococci: A World in Progress* 1–10 (Academic Press, 2018).

92. Foster, T. *Staphylococcus*. in *Medical Microbiology* 309–320 (Galveston, 1996).

93. Becker, K., Heilmann, C. & Peters, G. Coagulase-negative staphylococci. *Clin. Microbiol. Rev.* 27, 870–926 (2014).

94. Li, Y. *et al. Corticicoccus populi* gen. nov., sp. nov., a member of the family *Staphylococcaceae*, isolated from symptomatic bark of *Populus* × *euramericana* canker. *Int. J. Syst. Evol. Microbiol.* 67, 789–794 (2017).

95. Peacock, S. *Staphylococcus aureus*. in *Principles and Practice of Clinical Bacteriology* 73–98 (Wiley Online Library, 2006).

96. Public Health England & National Health Service. Identification of *Staphylococcus* species, *Micrococcus* species and *Rothia* species. *UK Stand. Microbiol. Investig.* 4, 1–26 (2014).

97. National Center for Biotechnology Information (NCBI). *Staphylococcus aureus* (ID 154) - Genome. *NCBI* www.ncbi.nlm.nih.gov/genome/154 (2022).

98. Bosi, E. *et al.* Comparative genome-scale modelling of *Staphylococcus aureus* strains identifies strain-specific metabolic capabilities linked to pathogenicity. *Proc. Natl. Acad. Sci. U. S. A.* 113, E3801–E3809 (2016).

99. John, J., George, S., Nori, S. R. C. & Nelson-Sathi, S. Phylogenomic analysis reveals the evolutionary route of resistant genes in *Staphylococcus aureus*. *Genome Biol. Evol.* 11, 2917–2926 (2019).

100. Everitt, R. G. *et al.* Mobile elements drive recombination hotspots in the core genome of *Staphylococcus aureus*. *Nat. Commun.* 5, 1–9 (2014).

101. Vos, M. & Didelot, X. A comparison of homologous recombination rates in bacteria and archaea. *ISME J.* 3, 199–208 (2009).

102. Planet, P. J. *et al.* Architecture of a species: Phylogenomics of *Staphylococcus aureus*. *Trends Microbiol.* 25, 153–166 (2017).

103. Aanensen, D. M. *et al.* Whole-genome sequencing for routine pathogen surveillance in public health: A population snapshot of invasive *Staphylococcus aureus* in Europe. *MBio* 7, e00444-16 (2016).

104. Dayan, G. H. *et al. Staphylococcus aureus*: the current state of disease, pathophysiology and strategies for prevention. *Expert Rev. Vaccines* 15, 1373–1392 (2016).

105. Lee, A. S. *et al.* Methicillin-resistant *Staphylococcus aureus*. *Nat. Rev. Dis. Prim.* 4, 1–23 (2018).

106. O'Riordan, K. & Lee, J. C. *Staphylococcus aureus* capsular polysaccharides. *Clin. Microbiol. Rev.* 17, 218–34 (2004).

107. Lowy, F. D. *Staphylococcus aureus* infections. *N. Engl. J. Med.* 339, 520–532 (1998).

108. Foster, T. J., Geoghegan, J. A., Ganesh, V. K. & Höök, M. Adhesion, invasion and evasion: the many functions of the surface proteins of *Staphylococcus aureus*. *Nat. Rev. Microbiol.* 12, 49–62 (2014).

109. Edwards, A. M., Massey, R. C. & Clarke, S. R. Molecular mechanisms of *Staphylococcus aureus* nasopharyngeal colonization. *Mol. Oral Microbiol.* 27, 1–10 (2012).

110. Foster, T. J. The MSCRAMM family of cell-wall-anchored surface proteins of Gram-positive cocci. *Trends Microbiol.* 27, 927–941 (2019).

111. Geoghegan, J. A. *et al.* Role of surface protein SasG in biofilm formation by

Staphylococcus aureus. *J. Bacteriol.* 192, 5663–73 (2010).

112. Naber, C. K. *Staphylococcus aureus* bacteremia: epidemiology, pathophysiology, and management strategies. *Clin. Infect. Dis.* 48, S231–S237 (2009).

113. Cassat, J. E. *et al.* Comparative genomics of *Staphylococcus aureus* musculoskeletal isolates. *J. Bacteriol.* 187, 576–92 (2005).

114. Leung, A. K. C., Barankin, B. & Leong, K. F. Staphylococcal-scalded skin syndrome: evaluation, diagnosis, and management. *World J. Pediatr.* 14, 116–120 (2018).

115. Thomer, L., Schneewind, O. & Missiakas, D. Pathogenesis of *Staphylococcus aureus* bloodstream infections. *Annu. Rev. Pathol.* 11, 364 (2016).

116. Tam, K. & Torres, V. J. *Staphylococcus aureus* secreted toxins and extracellular enzymes. *Microbiol. Spectr.* 7, GPP3-0039–2018 (2019).

117. Painter, K. L., Krishna, A., Wigneshweraraj, S. & Edwards, A. M. What role does the quorum-sensing accessory gene regulator system play during *Staphylococcus aureus* bacteremia? *Trends Microbiol.* 22, 676–685 (2014).

118. Balasubramanian, D., Harper, L., Shopsin, B. & Torres, V. J. *Staphylococcus aureus* pathogenesis in diverse host environments. *Pathog. Dis.* 75, 5 (2017).

119. Novick, R. P. & Geisinger, E. Quorum sensing in staphylococci. *Annu. Rev. Genet.* 42, 541–564 (2008).

120. Jenul, C. & Horswill, A. R. Regulation of *Staphylococcus aureus* virulence. *Microbiol. Spectr.* 7, GPP3-0031–2018 (2019).

121. Harkins, C. P. *et al.* Methicillin-resistant *Staphylococcus aureus* emerged long before the introduction of methicillin into clinical practice. *Genome Biol.* 18, 130 (2017).

122. Monaco, M., Pimentel de Araujo, F., Cruciani, M., Coccia, E. M. & Pantosti, A. Worldwide epidemiology and antibiotic resistance of *Staphylococcus aureus*. in *Current Topics in Microbiology and Immunology* 21–56 (Springer, 2017).

123. Thurlow, L. R., Joshi, G. S. & Richardson, A. R. Virulence strategies of the dominant USA300 lineage of community associated methicillin resistant *Staphylococcus aureus* (CA-MRSA). *FEMS Immunol. Med. Microbiol.* 65, 22 (2012).

124. O'Brien, F. G. *et al.* Population dynamics of methicillin-susceptible and -resistant *Staphylococcus aureus* in remote communities. *J. Antimicrob. Chemother.* 64, 693 (2009).

125. Mick, V. *et al.* Molecular characterization of resistance to rifampicin in an emerging hospital-associated Methicillin-resistant *Staphylococcus aureus* clone ST228, Spain. *BMC Microbiol.* 10, 68 (2010).

126. Turner, N. A. *et al.* Methicillin-resistant *Staphylococcus aureus*: an overview of basic and clinical research. *Nat. Rev. Microbiol.* 17, 203–218 (2019).

127. Aires de Sousa, M. *et al.* Two international methicillin-resistant *Staphylococcus aureus*

clones endemic in a university hospital in Patras, Greece. *J. Clin. Microbiol.* 41, 2032 (2003).

128. Camoez, M. *et al.* Prevalence and molecular characterization of methicillin-resistant *Staphylococcus aureus* ST398 resistant to tetracycline at a Spanish hospital over 12 years. *PLoS One* 8, e72828 (2013).

129. Laux, C., Peschel, A. & Krismer, B. *Staphylococcus aureus* colonization of the human nose and interaction with other microbiome members. *Microbiol. Spectr.* 7, 723–730 (2019).

130. Wertheim, H. F. L. *et al.* The role of nasal carriage in *Staphylococcus aureus* infections. *Lancet Infect. Dis.* 5, 751–62 (2005).

131. Sakr, A., Brégeon, F., Mège, J.-L., Rolain, J.-M. & Blin, O. *Staphylococcus aureus* nasal colonization: an update on mechanisms, epidemiology, risk factors, and subsequent infections. *Front. Microbiol.* 8, 2419 (2018).

132. von Eiff, C., Becker, K., Machka, K., Stammer, H. & Peters, G. Nasal carriage as a source of Staphylococcus aureus bacteremia. *N. Engl. J. Med.* 344, 11–6 (2001).

133. Wertheim, H. *et al.* Risk and outcome of nosocomial *Staphylococcus aureus* bacteraemia in nasal carriers versus non-carriers. *Lancet* 364, 703–705 (2004).

134. Botelho-Nevers, E. *et al.* Decolonization of *Staphylococcus aureus* carriage. *Médecine Mal. Infect.* 47, 305–310 (2017).

135. Mulcahy, M. E. & Mcloughlin, R. M. Host-bacterial crosstalk determines *Staphylococcus aureus* nasal colonization. *Trends Microbiol.* 24, 872–886 (2016).

136. Mehraj, J. *et al.* Epidemiology of *Staphylococcus aureus* nasal carriage patterns in the community. in *How to overcome the antibiotic crisis* 55–87 (Springer, 2016).

137. Lacoma, A. *et al.* Cigarette smoke exposure redirects *Staphylococcus aureus* to a virulence profile associated with persistent infection. *Sci. Rep.* 9, 1–15 (2019).

138. Krismer, B., Weidenmaier, C., Zipperer, A. & Peschel, A. The commensal lifestyle of *Staphylococcus aureus* and its interactions with the nasal microbiota. *Nat. Rev. Microbiol.* 15, 675–687 (2017).

139. Morgene, M. F. *et al. Staphylococcus aureus* colonization and non-influenza respiratory viruses: Interactions and synergism mechanisms. *Virulence* 9, 1354–1363 (2018).

140. Centers for Disease Control and Prevention (CDC). *Staphylococcus aureus* in healthcare settings. www.cdc.gov/hai/organisms/staph.html (2011).

141. Van Wamel, W. J. B. *Staphylococcus aureus* infections, some second thoughts. *Curr. Opin. Infect. Dis.* 30, 303–308 (2017).

142. Silversides, J. A., Lappin, E. & Ferguson, A. J. Staphylococcal Toxic Shock Syndrome: mechanisms and management. *Curr. Infect. Dis. Rep.* 12, 392–400 (2010).

143. Tong, S. Y. C. *et al. Staphylococcus aureus* infections: epidemiology, pathophysiology,

clinical manifestations, and management. *Clin. Microbiol. Rev.* 28, 603 (2015).

144. Tande, A. J. & Patel, R. Prosthetic joint infection. *Clin. Microbiol. Rev.* 27, 345 (2014).

145. Bergin, S. P., Holland, T. L., Fowler, V. G. & Tong, S. Y. C. Bacteremia, sepsis, and infective endocarditis associated with *Staphylococcus aureus*. in *Current Topics in Microbiology and Immunology* 263–296 (Springer, 2015).

146. David, M. Z. & Daum, R. S. Treatment of *Staphylococcus aureus* infections. in *Staphylococcus aureus* 325–383 (Springer, 2017).

147. Ito, T. *et al.* Staphylococcal cassette chromosome *mec* (SCC*mec*). Analysis of MRSA. in Methicillin-resistant *Staphylococcus aureus* (MRSA) protocols 131–148 (Springer, 2013).

148. Guo, Y., Song, G., Sun, M., Wang, J. & Wang, Y. Prevalence and therapies of antibiotic-resistance in *Staphylococcus aureus*. *Front. Cell. Infect. Microbiol.* 10, 107 (2020).

149. Hiramatsu, K. *et al.* Genomic basis for methicillin resistance in *Staphylococcus aureus*. *Infect. Chemother.* 45, 136 (2013).

150. McGuinness, W. A., Malachowa, N. & DeLeo, F. R. Vancomycin resistance in *Staphylococcus aureus*. *Yale J. Biol. Med.* 90, 281 (2017).

151. Silverman, J. A., Mortin, L. I., VanPraagh, A. D. G., Li, T. & Alder, J. Inhibition of daptomycin by pulmonary surfactant: in vitro modeling and clinical impact. *J. Infect. Dis.* 191, 2149–2152 (2005).

152. Roch, M. *et al.* Daptomycin resistance in clinical MRSA strains is associated with a high biological fitness cost. *Front. Micr* 8, 2303 (2017).

153. Watkins, R. R., Lemonovich, T. L. & File, T. M. An evidence-based review of linezolid for the treatment of methicillin-resistant *Staphylococcus aureus* (MRSA): place in therapy. *Core Evid.* 7, 131–143 (2012).

154. García, M. S. *et al.* Clinical outbreak of linezolid-resistant *Staphylococcus aureus* in an intensive care unit. *J. Am. Med. Assoc.* 303, 2260–2264 (2010).

155. Escolà-Vergé, L., Los-Arcos, I. & Almirante, B. New antibiotics for the treatment of infections by multidrug-resistant microorganisms. *Med. Clin. (Barc).* 154, 351–357 (2020).

156. Strobel, M. *et al.* Post-invasion events after infection with *Staphylococcus aureus* are strongly dependent on both the host cell type and the infecting *S. aureus* strain. *Clin. Microbiol. Infect.* 22, 799–809 (2016).

157. Giufrè, M. *et al.* Carriage of *Haemophilus influenzae* in the oropharynx of young children and molecular epidemiology of the isolates after fifteen years of *H. influenzae* type b vaccination in Italy. *Vaccine* 33, 6227–6234 (2015).

158. Tacconelli, E. *et al.* Antibiotic usage and risk of colonization and infection with antibiotic-resistant bacteria: a hospital population-based study. *Antimicrob. Agents Chemother.* 53, 4269 (2009).

159. Wang, S., Tafalla, M., Hanssens, L. & Dolhain, J. A review of *Haemophilus influenzae* disease in Europe from 2000–2014: challenges, successes and the contribution of hexavalent combination vaccines. *Expert Rev. Vaccines* 16, 1095–1105 (2017).

160. Agrawal, A. & Murphy, T. F. *Haemophilus influenzae* infections in the *H. influenzae* type b conjugate vaccine era. *J. Clin. Microbiol.* 49, 3728–3732 (2011).

161. Heliodoro, C. I. M., Bettencourt, C. R. & Bajanca-Lavado, M. P. Molecular epidemiology of invasive *Haemophilus influenzae* disease in Portugal: an update of the post-vaccine period, 2011–2018. *Eur. J. Clin. Microbiol. Infect. Dis.* 39, 1471–1480 (2020).

162. McTaggart, L. R. *et al.* Increased incidence of invasive *Haemophilus influenzae* disease driven by non-type b isolates in Ontario, Canada, 2014 to 2018. *Microbiol. Spectr.* 9, e00803-21 (2021).

163. Desai, S. *et al.* The epidemiology of invasive *Haemophilus influenzae* non-serotype B disease in Ontario, Canada from 2004 to 2013. *PLoS One* 10, e0142179 (2015).

164. Resman, F. *et al.* Invasive disease caused by *Haemophilus influenzae* in Sweden 1997–2009; evidence of increasing incidence and clinical burden of non-type b strains. *Clin. Microbiol. Infect.* 17, 1638–1645 (2011).

165. Langereis, J. D. & De Jonge, M. I. Invasive disease caused by nontypeable *Haemophilus influenzae*. *Emerg. Infect. Dis.* 21, 1711–1718 (2015).

166. Chiappini, E., Inturrisi, F., Orlandini, E., de Martino, M. & de Waure, C. Hospitalization rates and outcome of invasive bacterial vaccine-preventable diseases in Tuscany: A historical cohort study of the 2000-2016 period. *BMC Infect. Dis.* 18, 396 (2018).

167. Potts, C. C. *et al.* Antimicrobial susceptibility survey of invasive *Haemophilus influenzae* in the United States in 2016. *Microbiol. Spectr.* 10, e02579-21 (2022).

168. León, M. E. *et al.* Epidemiologic study of *Haemophilus influenzae* causing invasive and non-invasive disease in Paraguay (1999-2017). *Enferm. Infecc. Microbiol. Clin.* 39, 59–64 (2021).

169. McElligott, M. *et al.* Epidemiology of *Haemophilus influenzae* in the Republic of Ireland, 2010–2018. *Eur. J. Clin. Microbiol. Infect. Dis.* 39, 2335–2344 (2020).

170. San Millan, A. *et al. Haemophilus influenzae* clinical isolates with plasmid pB1000 bearing *bla*ROB-1: Fitness cost and interspecies dissemination. *Antimicrob. Agents Chemother.* 54, 1506–1511 (2010).

171. Skaare, D. *et al.* Multilocus sequence typing and *ftsI* sequencing: a powerful tool for surveillance of penicillin-binding protein 3-mediated beta-lactam resistance in nontypeable *Haemophilus influenzae*. *BMC Microbiol.* 14, 131 (2014).

172. Ubukata, K. *et al.* Association of amino acid substitutions in penicillin-binding protein 3 with β-lactam resistance in β-lactamase-negative ampicillin-resistant *Haemophilus influenzae*. *Antimicrob. Agents Chemother.* 45, 1693–1699 (2001).

173. Honda, H. *et al.* Multiclonal expansion and high prevalence of β-lactamase-negative *Haemophilus influenzae* with high-level ampicillin resistance in Japan and susceptibility to

quinolones. *Antimicrob. Agents Chemother.* 62, e00851-18 (2018).

174. Park, C. *et al.* Genetic diversity of the *ftsI* gene in β-lactamase-nonproducing ampicillin-resistant and β-lactamase-producing amoxicillin/clavulanic acid-resistant nasopharyngeal *Haemophilus influenzae* strains isolated from children in South Korea. *Microb. Drug Resist.* 19, 224–230 (2013).

175. Puig, C. *et al.* Clinical and molecular epidemiology of *Haemophilus influenzae* causing invasive disease in adult patients. *PLoS One* 9, e112711 (2014).

176. Shoji, H. *et al.* A molecular analysis of quinolone-resistant *Haemophilus influenzae*: Validation of the mutations in quinolone resistance-determining regions. *J. Infect. Chemother.* 20, 250–255 (2014).

177. Biedenbach, D. J. & Jones, R. N. Five-year analysis of *Haemophilus influenzae* isolates with reduced susceptibility to fluoroquinolones: prevalence results from the SENTRY antimicrobial surveillance program. *Diagn. Microbiol. Infect. Dis.* 46, 55–61 (2003).

178. Tanaka, E., Wajima, T., Hirai, Y., Nakaminami, H. & Noguchi, N. Dissemination of quinolone low-susceptible *Haemophilus influenzae* ST422 in Tokyo, Japan. *J. Infect. Chemother.* 27, 962–966 (2021).

179. Georgiades, K. & Raoult, D. Defining pathogenic bacterial species in the genomic era. *Front. Microbiol.* 1, 1–13 (2011).

180. Staples, M., Graham, R. M. A. & Jennison, A. V. Characterisation of invasive clinical *Haemophilus influenzae* isolates in Queensland, Australia using whole-genome sequencing. *Epidemiol. Infect.* 145, 1727–1736 (2017).

181. Power, P. M., Bentley, S. D., Parkhill, J., Moxon, E. R. & Hood, D. W. Investigations into genome diversity of *Haemophilus influenzae* using whole genome sequencing of clinical isolates and laboratory transformants. *BMC Microbiol.* 12, 1–12 (2012).

182. Connor, T. R., Corander, J. & Hanage, W. P. Population subdivision and the detection of recombination in non-typable *Haemophilus influenzae*. *Microbiology* 158, 2954–2964 (2012).

183. Deghmane, A.-E. *et al.* High diversity of invasive *Haemophilus influenzae* isolates in France and the emergence of resistance to third generation cephalosporins by alteration of *ftsI* gene. *J. Infect.* 79, 7–14 (2019).

184. Musser, J. M., Kroll, J. S., Moxonf, E. R. & Selandert, R. K. Evolutionary genetics of the encapsulated strains of *Haemophilus influenzae*. *Proc. Nati. Acad. Sci. USA* 85, 7758–7762 (1988).

185. Omikunle, A. *et al.* Limited genetic diversity of recent invasive isolates of non-serotype b encapsulated *Haemophilus influenzae*. *J. Clin. Microbiol.* 40, 1264–1270 (2002).

186. Meats, E. *et al.* Characterization of encapsulated and noncapsulated *Haemophilus influenzae* and determination of phylogenetic relationships by Multilocus Sequence Typing. *J. Clin. Microbiol.* 41, 1623–1636 (2003).

187. Tsang, R. S. W., Bruce, M. G., Lem, M., Barreto, L. & Ulanova, M. A review of invasive

Haemophilus influenzae disease in the Indigenous populations of North America. *Epidemiol. Infect.* 142, 1354 (2014).

188. Topaz, N. *et al.* Phylogenetic structure and comparative genomics of multi-national invasive *Haemophilus influenzae* serotype a isolates. *Front. Microbiol.* 13, 856884 (2022).

189. Adam, H. J. *et al.* Changing epidemiology of invasive *Haemophilus influenzae* in Ontario, Canada: Evidence for herd effects and strain replacement due to Hib vaccination. *Vaccine* 28, 4073–4078 (2010).

190. Ladhani, S. N. *et al.* Invasive *Haemophilus influenzae* serotype e and f disease, England and Wales. *Emerg. Infect. Dis.* 18, 725–732 (2012).

191. Su, Y.-C., Resman, F., Hörhold, F. & Riesbeck, K. Comparative genomic analysis reveals distinct genotypic features of the emerging pathogen *Haemophilus influenzae* type f. *BMC Genomics* 15, 38 (2014).

192. Hallströ, T. & Riesbeck, K. *Haemophilus influenzae* and the complement system. *Trends Microbiol.* 18, 258–265 (2010).

193. Tsang, R. S. W. & Ulanova, M. The changing epidemiology of invasive *Haemophilus influenzae* disease: Emergence and global presence of serotype a strains that may require a new vaccine for control. *Vaccine* 35, 4270–4275 (2017).

194. Shuel, M., Knox, N. & Tsang, R. S. W. Global population structure of *Haemophilus influenzae* serotype a (Hia) and emergence of invasive Hia disease: capsule switching or capsule replacement? *Can. J. Microbiol.* 67, 875–884 (2021).

195. Giufrè, M., Cardines, R., Brigante, G., Orecchioni, F. & Cerquetti, M. Emergence of invasive *Haemophilus influenzae* type A disease in Italy. *Clin. Infect. Dis.* 64, 1626–1628 (2017).

196. López-Olaizola, M. *et al. Haemophilus influenzae* type a sequence type 23, Northern Spain. *Emerg. Infect. Dis.* 27, 2506 (2021).

197. Bruun, B., Gahrn-Hansen, B., Westh, H. & Kilian, M. Clonal relationship of recent invasive *Haemophilus influenzae* serotype f isolates from Denmark and the United States. *J. Med. Microbiol.* 53, 1161–1165 (2004).

198. Chaguza, C. *et al.* Recombination in *Streptococcus pneumoniae* lineages increase with carriage duration and size of the polysaccharide capsule. *MBio* 7, e01053-16 (2016).

199. World Health Organization. Chronic respiratory diseases. in *Global surveillance, prevention and control of chronic respiratory diseases: A comprehensive approach* 12-37 (2007).

200. GBD Chronic Respiratory Disease Collaborators. Prevalence and attributable health burden of chronic respiratory diseases, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Respir. Med.* 8, 585–596 (2020).

201. Phillips, Z. N. *et al.* Analysis of invasive nontypeable *Haemophilus influenzae* isolates reveals selection for the expression state of particular phase-variable lipooligosaccharide

biosynthetic genes. *Infect. Immun.* 87, e00093-19 (2019).

202. Clark, S. E., Snow, J., Li, J., Zola, T. A. & Weiser, J. N. Phosphorylcholine allows for evasion of bactericidal antibody by *Haemophilus influenzae*. *PLOS Pathog.* 8, e1002521 (2012).

203. Humphries, H. E. & High, N. J. The role of *licA* phase variation in the pathogenesis of invasive disease by *Haemophilus influenzae* type b. *FEMS Immunol. Med. Microbiol.* 34, 221–230 (2002).

204. Poole, J. *et al.* Analysis of nontypeable *Haemophilus influenzae* phase-variable genes during experimental human nasopharyngeal colonization. *J. Infect. Dis.* 208, 727 (2013).

205. Tong, H. H., Blue, L. E., James, M. A., Chen, Y. P. & Demaria, T. F. Evaluation of phase variation of nontypeable *Haemophilus influenzae* lipooligosaccharide during nasopharyngeal colonization and development of otitis media in the chinchilla model. *Infect. Immun.* 68, 4597 (2000).

206. Swords, W. E. *et al.* Non-typeable *Haemophilus influenzae* adhere to and invade human bronchial epithelial cells via an interaction of lipooligosaccharide with the PAF receptor. *Mol. Microbiol.* 37, 13–27 (2002).

207. Langereis, J. D. *et al.* Nontypeable *Haemophilus influenzae* invasive blood isolates are mainly phosphorylcholine negative and show decreased complement-mediated killing that is associated with lower binding of IgM and CRP in comparison to colonizing isolates from the oropharynx. *Infect. Immun.* 87, e00604-18 (2019).

208. Clark, S. E., Eichelberger, K. R. & Weiser, J. N. Evasion of killing by human antibody and complement through multiple variations in the surface oligosaccharide of *Haemophilus influenzae*. *Mol. Microbiol.* 88, 603 (2013).

209. Hosking, S. L., Craig, J. E. & High, N. J. Phase variation of *lic1A*, *lic2A* and *lic3A* in colonization of the nasopharynx, bloodstream and cerebrospinal fluid by *Haemophilus influenzae* type b. *Microbiology* 145, 3005–3011 (1999).

210. Jones, P. A. *et al. Haemophilus influenzae* type b strain A2 has multiple sialyltransferases Involved in lipooligosaccharide sialylation. *J. Biol. Chem.* 277, P14598 (2002).

211. Fox, K. L. *et al.* Identification of a bifunctional lipopolysaccharide sialyltransferase in *Haemophilus influenzae*. *J. Biol. Chem.* 281, 40024–40032 (2006).

212. Apicella, M. A. Nontypeable *Haemophilus influenzae*: the role of N-acetyl-5-neuraminic acid in biology. *Front. Cell. Infect. Microbiol.* 2, 19 (2012).

213. Hood, D. W. *et al.* Identification of a lipopolysaccharide alpha-2,3-sialyltransferase from *Haemophilus influenzae*. *Mol. Microbiol.* 39, 350 (2001).

214. Greiner, L. L. *et al.* Nontypeable *Haemophilus influenzae* strain 2019 produces a biofilm containing N-acetylneuraminic acid that may mimic sialylated O-linked glycans. *Infect. Immun.* 72, 4249–4260 (2004).

215. Whitby, P. W., Seale, T. W., VanWagoner, T. M., Morton, D. J. & Stull, T. L. The iron/heme regulated genes of *Haemophilus influenzae*: Comparative transcriptional profiling as a tool

to define the species core modulon. *BMC Genomics* 10, 1–19 (2009).

216. Stull, T. L., Morton, D. J., Whitby, P. W., Jin, H. & Ren, Z. Effect of multiple mutations in the hemoglobin- and hemoglobin-haptoglobin-binding proteins, HgpA, HgpB, and HgpC, of *Haemophilus* type b. *Infect. Immun.* 67, 2729–2739 (1999).

217. Morton, D. J., VanWagoner, T. M., Seale, T. W., Whitby, P. W. & Stull, T. L. Differential utilization by *Haemophilus influenzae* of haemoglobin complexed to the three human haptoglobin phenotypes. *FEMS Immunol. Med. Microbiol.* 46, 426–432 (2006).

218. Garmendia, J. *et al.* Characterization of nontypable *Haemophilus influenzae* isolates recovered from adult patients with underlying chronic lung disease reveals genotypic and phenotypic traits associated with persistent infection. *PLoS One* 9, e97020 (2014).

219. Harrison, A. *et al.* Continuous microevolution accelerates disease progression during sequential episodes of infection. *Cell Rep.* 30, 2978-2988.e3 (2020).

220. Koskella, B. & Brockhurst, M. A. Bacteria–phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiol. Rev.* 38, 916–931 (2014).

221. Szafrański, S. P. *et al.* Diversity patterns of bacteriophages infecting *Aggregatibacter* and *Haemophilus* species across clades and niches. *ISME J.* 13, 2500–2522 (2019).

222. Serisier, D. J. Risks of population antimicrobial resistance associated with chronic macrolide use for inflammatory airway diseases. *Lancet Respir. Med.* 1, 262–274 (2013).

223. Yamaya, M. *et al.* Macrolide effects on the prevention of COPD exacerbations. *Eur. Respir. J.* 40, 485–494 (2012).

224. Ko, F. W. *et al.* Acute exacerbation of COPD. *Respirology* 21, 1152–1165 (2016).

225. MacLeod, M. *et al.* Chronic obstructive pulmonary disease exacerbation fundamentals: Diagnosis, treatment, prevention and disease impact. *Respirology* 26, 532–551 (2021).

226. Albert, R. K. *et al.* Azithromycin for prevention of exacerbations of COPD. *N. Engl. J. Med.* 365, 689–698 (2011).

227. Parnham, M. J. *et al.* Azithromycin: Mechanisms of action and their relevance for clinical applications. *Pharmacol. Ther.* 143, 225–245 (2014).

228. Roberts, M. C. Update on macrolide-lincosamide-streptogramin, ketolide, and oxazolidinone resistance genes. *FEMS Microbiol. Lett.* 282, 147–159 (2008).

229. Seyama, S., Wajima, T., Nakaminami, H. & Noguchi, N. Amino acid substitution in the major multidrug efflux transporter protein AcrB contributes to low susceptibility to azithromycin in *Haemophilus influenzae*. *Antimicrob. Agents Chemother.* 61, e01337-17 (2017).

230. Sierra, Y. *et al.* Emergence of multidrug resistance among *Haemophilus parainfluenzae* from respiratory and urogenital samples in Barcelona, Spain. *Eur. J. Clin. Microbiol. Infect. Dis.* 39, 703–710 (2019).

231. Sierra, Y. *et al.* Genome-wide analysis of urogenital and respiratory multidrug-resistant *Haemophilus parainfluenzae*. *J. Antimicrob. Chemother.* 76, 1741–1751 (2021).

232. Leung, J. M. *et al.* The role of acute and chronic respiratory colonization and infections in the pathogenesis of COPD. *Respirology* 22, 634–650 (2017).

233. Hare, K. M. *et al.* Antimicrobial susceptibility and impact of macrolide antibiotics on *Moraxella catarrhalis* in the upper and lower airways of children with chronic endobronchial suppuration. *J. Med. Microbiol.* 68, 1140–1147 (2019).

234. Chalmers, J. D. Macrolide resistance in *Pseudomonas aeruginosa*: implications for practice. *Eur. Respir. J.* 49, 1700689 (2017).

235. Sánchez, M. B. Antibiotic resistance in the opportunistic pathogen *Stenotrophomonas maltophilia*. *Front. Microbiol.* 6, 658 (2015).

236. Gromkova, R. C., Mottalini, T. C. & Dove, M. G. Genetic transformation in *Haemophilus parainfluenzae* clinical isolates. *Curr. Microbiol.* 37, 123–126 (1998).

237. Juhas, M. *et al.* Sequence and functional analyses of *Haemophilus spp.* genomic islands. *Genome Biol.* 8, 237 (2007).

238. Wierzbowski, A. K., Hoban, D. J., Hisanaga, T., DeCorby, M. & Zhanel, G. G. The use of macrolides in treatment of upper respiratory tract infections. *Curr. Allergy Asthma Rep.* 6, 171–181 (2006).

239. Fraunholz, M. & Sinha, B. Intracellular *Staphylococcus aureus*: live-in and let die. *Front. Cell. Infect. Microbiol.* 2, 1–10 (2012).

240. Kao, K. C. *et al.* Risk factors of methicillin-resistant *Staphylococcus aureus* infection and correlation with nasal colonization based on molecular genotyping in medical intensive care units: A prospective observational study. *Medicine (Baltimore).* 94, 1–7 (2015).

241. Costa, D. M. *et al.* Biofilm contamination of high-touched surfaces in intensive care units: epidemiology and potential impacts. *Lett. Appl. Microbiol.* 68, 269–276 (2019).

242. Gasch, O. *et al.* Predictive factors for mortality in patients with methicillin-resistant *Staphylococcus aureus* bloodstream infection: impact on outcome of host, microorganism and therapy. *Clin. Microbiol. Infect.* 19, 1049–1057 (2013).

243. Naicker, P. R., Karayem, K., Hoek, K. G. P., Harvey, J. & Wasserman, E. Biofilm formation in invasive *Staphylococcus aureus* isolates is associated with the clonal lineage. *Microb. Pathog.* 90, 41–49 (2016).

244. Jotić, A. *et al.* Biofilm formation on tympanostomy tubes depends on methicillin-resistant *Staphylococcus aureus* genetic lineage. *Eur. Arch. Oto-Rhino-Laryngology* 273, 615–620 (2016).

245. Côrtes, M. F., Beltrame, C. O., Ramundo, M. S., Ferreira, F. A. & Figueiredo, A. M. S. The influence of different factors including *fnbA* and *mecA* expression on biofilm formed by MRSA clinical isolates with different genetic backgrounds. *Int. J. Med. Microbiol.* 305, 140–147 (2015).

246. Tasse, J. *et al.* Association between biofilm formation phenotype and clonal lineage in *Staphylococcus aureus* strains from bone and joint infections. *PLoS One* 13, e020 (2018).

247. Roche, F. M., Meehan, M. & Foster, T. J. The *Staphylococcus aureus* surface protein SasG and its homologues promote bacterial adherence to human desquamated nasal epithelial cells. *Microbiology* 149, 2759–2767 (2003).

248. Roche, F. M. *et al.* Characterization of novel LPXTG-containing proteins of *Staphylococcus aureus* identified from genome sequences. *Microbiology* 149, 643–654 (2003).

249. Kuroda, M. *et al. Staphylococcus aureus* surface protein SasG contributes to intercellular autoaggregation of *Staphylococcus aureus*. *Biochem. Biophys. Res. Commun.* 377, 1102–1106 (2008).

250. Peyrusson, F. *et al.* Intracellular *Staphylococcus aureus* persisters upon antibiotic exposure. *Nat. Commun.* 11, 1–14 (2020).

251. Perkins, S. *et al.* Structural Organization of the fibrinogen-binding region of the clumping factor B MSCRAMM of *Staphylococcus aureus*. *J. Biol. Chem.* 276, 44721–44728 (2001).

252. Walsh, E. J., O'Brien, L. M., Liang, X., Hook, M. & Foster, T. J. Clumping factor B, a fibrinogen-binding MSCRAMM (microbial surface components recognizing adhesive matrix molecules) adhesin of *Staphylococcus aureus*, also binds to the tail region of type I cytokeratin 10. *J. Biol. Chem.* 279, 50691–9 (2004).

253. Loughman, A. *et al.* Sequence diversity in the A domain of *Staphylococcus aureus* fibronectin-binding protein A. *BMC Microbiol.* 8, 74 (2008).

254. Cheng, A. G., Missiakas, D. & Schneewind, O. The giant protein Ebh is a determinant of *Staphylococcus aureus* cell size and complement resistance. *J. Bacteriol.* 196, 971–981 (2014).

255. Kuroda, M. *et al. Staphylococcus aureus* giant protein Ebh is involved in tolerance to transient hyperosmotic pressure. *Biochem. Biophys. Res. Commun.* 374, 237–241 (2008).

256. Walker, J. N., Crosby, H. A., Spaulding, A. R., Salgado-Pabó N, W. & Malone, C. L. The *Staphylococcus aureus* ArlRS two-component system is a novel regulator of agglutination and pathogenesis. *PLoS Pathog.* 9, e1003819 (2013).