UNIVERSIDADE DE LISBOA
Faculdade de Medicina

Understanding pherotype specificity and its influence on the genetic structure of *Streptococcus pneumoniae*

Jorge Miguel Diamantino Miranda

Orientador: Prof. Doutor Mário Nuno Ramos de Almeida Ramirez

Tese especialmente elaborada para obtenção do grau de Doutor em Ciências e Tecnologias da Saúde, especialidade Microbiologia e Parasitologia

2019

UNIVERSIDADE DE LISBOA

Faculdade de Medicina

Understanding pherotype specificity and its influence on the genetic structure of *Streptococcus pneumoniae*

Jorge Miguel Diamantino Miranda

Orientador: Prof. Doutor Mário Nuno Ramos de Almeida Ramirez

Tese especialmente elaborada para obtenção do grau de Doutor em Ciências e Tecnologias da Saúde, especialidade Microbiologia e Parasitologia

Juri:

Presidente:

Doutor José Augusto Gamito Melo Cristino, Professor Catedrático e Presidente do Conselho Científico da Faculdade de Medicina da Universidade de Lisboa

Vogais:

- Doutor Paulo Jorge Pereira Cruz Paixão, Professor Auxiliar da Faculdade de Ciências Médicas da Universidade NOVA de Lisboa;

- Doutor Sérgio Joaquim Raposo Filipe, Professor Auxiliar da Faculdade de Ciências e Tecnologia da Universidade NOVA de Lisboa;

- Doutora Raquel de Sá Leão Domingues da Silva, Investigadora Principal do Instituto de Tecnologia Química e Biológica da Universidade NOVA de Lisboa;

- Doutor Mário Nuno Ramos de Almeida Ramirez, Professor Associado com Agregação da Faculdade de Medicina da Universidade de Lisboa (*Orientador*);

- Doutor Marcos Daniel Caetano Borges de Pinho, Professor Auxiliar da Faculdade de Medicina da Universidade de Lisboa.

Tese financiada pela Fundação para a Ciência e a Tecnologia
(SFRH/BD/81766/2011)

2019

A impressão desta tese foi aprovada pelo Conselho Científico da Faculdade de Medicina de Lisboa em reunião de 19 de março de 2019.

Acknowledgments

First of all I wish to thank my supervisor Professor Mário Ramirez for his guidance, for his constant support and for helping me whenever I needed. I also want to thank Professor José Melo Cristino, head of Instituto de Microbiologia, Faculdade de Medicina, Universidade de Lisboa.

I also want to thank Dr. Juan Hermoso for welcoming me in his laboratory and for his kindness. A special thanks to Dr. Sergio Galán for his guidance and fruitful discussions. Also thanks to all the persons from "Roca" who were there with me during my stay in Madrid.

I am extremely grateful to Dr. Sandra Aguiar. Besides guiding me in my first steps at working with pneumococcus, our scientific discussions were crucial for the publication of my last article. I am also thankful for all the knowledge I got about pneumococcal pherotypes from Dr. Margarida Carrolo besides the good moments that we shared in the lab. I am also grateful to my tutor Dr. João Carriço for all his support and willingness to help me in anything I needed.

I have to say a big thank you to all the members of our lab in iMM, especially to Andreia and Cat who were there and still are in the good and bad moments.

To my family, both Portuguese and Spanish, thank you for all your support (and patience!) during this thesis.

María Ángela, I have no words to describe your support. You are simply unique and thank you for never letting me give up.

Finally, I am deeply grateful to my parents and brothers who always support me no matter what.

To my family, both Portuguese and Spanish

Table of contents

Resumo i

Abstract vii

Abbreviations xi

Thesis Outline xiii

I. Introduction 1

1. Background 3

1.1. Pneumococcus discovery and identification 5

1.2. Colonization and disease 5

1.3. Antimicrobial resistance 8

1.3.1. ß-lactams 8

1.3.2. Macrolide, lincosamide and streptogramin B 9

1.3.3. Tetracycline 10

1.3.4. Chloramphenicol 10

1.3.5. Sulfonamide 11

1.3.6. Fluoroquinolone 11

1.3.7. Glycopeptide 12

1.3.8. Oxazolidinone 12

1.4. Vaccination 13

1.5. Genetic transformation 16

1.5.1. Competence regulation 17

1.5.2. Pherotypes determined by competence-stimulating peptide (CSP) 19

1.5.3. ComD and ComE, a histidine kinase receptor and a response regulator protein 20

1.5.4. ComX, the alternative sigma factor 21

1.5.5. Fratricide 22

2. Aims of the thesis 25

3. Pneumococcal isolates and typing methods 29

3.1. Pneumococcal isolates 31

3.2. Storing, identification and growing of isolates 31
3.3. Serotyping 33
3.4. Multilocus Sequence Typing (MLST) 34
3.5. Antimicrobial susceptibility 36
II. Pherotype diversity and abundance 37
4. Diversity of the *comCDE* locus 41
4.1. Materials and methods 44
4.1.1. Bacterial isolates 44
4.1.2. Sequencing of *comCDE* locus 44
4.1.3. Analysis of *comCDE* locus 46
4.2. Results 48
4.2.1. Pneumococcal invasive isolates and allele classification 48
4.2.2. Gene *comC* 49
4.2.3. Gene *comD* 51
4.2.4. Gene *comE* 60
4.2.5. Profiles of *comCDE* 65
4.3. Discussion 66
5. Pherotype abundance 71
5.1. Materials and methods 74
5.1.1. Bacterial isolates 74
5.1.2. Pherotype identification 74
5.1.3. Statistical analysis 76
5.2. Results 77
5.2.1. Pherotype abundance and evolution 77
5.2.2. ST393 serotype 25A/38 isolates 80
5.2.3. Serotype and pherotype 81
5.2.4. Sequence type and pherotype 84
5.2.5. Antibiotic resistance and pherotype 89
5.2.6. Comparing partitions 90
5.3. Discussion 92

III. Pherotype specificity and influence on the genetic structure 97

6. Structural study of pherotype specificity by X-ray crystallography 101

6.1. Designing of ComD constructs 104

6.2. Cloning of ComD sensor in vector pURI3-TEV 106

6.3. Expression tests of ComD sensor with a His-tag 108

6.3.1. Expression in E. coli BL21(DE3) 109

6.3.2. Expression in E. coli JM109(DE3) 110

6.4. Cloning of the fusion protein LSL_{150}-ComD in vector pKLSLt 112

6.5. Expression tests of the fusion protein LSL_{150}-ComD 114

6.6. Discussion 118

7. Serogroup 6 and influence of pherotype in genetic recombination .. 123

7.1. Materials and methods 128

7.1.1. Bacterial isolates, serotyping, multilocus sequence typing (MLST) and pherotype identification 128

7.1.2. Statistical analysis 128

7.1.3. Gene flow and genetic differentiation analyses 130

7.2. Results 131

7.2.2. Evolution of serogroup 6 serotypes in children and in adults 131

7.2.2. Serogroup 6 genetic diversity 133

7.2.3. Diversity of the *wciP* and *wciN* genes 137

7.2.4. Antimicrobial susceptibility 138

7.2.5. Pherotype and influence on genetic recombination 139

7.3. Discussion 148

IV. Final remarks and future perspectives 155

V. References 161

Appendix I. Published article 177

Resumo

Palavras-chave: *Streptococcus pneumoniae*, ferótipo, competência, transformação genética, estrutura genética.

Streptococcus pneumoniae é uma bactéria de Gram-positivo considerada um agente patogénico oportunista, pois coloniza de forma assintomática a nasofaringe humana, sendo a colonização um pré-requisito para causar uma infeção localizada ou sistémica. O uso disseminado de antimicrobianos diminuiu drasticamente a morbilidade e mortalidade a nível global da doença pneumocócica. No entanto, a resistência aos antibióticos surgiu rapidamente após o seu uso inicial. Atualmente, *S. pneumoniae* apresenta resistência a um grande número de antibióticos, incluindo β-lactâmicos e quinolonas. Assim, para prevenir infeções pneumocócicas e devido à emergência de estirpes resistentes aos antimicrobianos, foram desenvolvidas vacinas utilizando os polissacáridos capsulares pneumocócicos. Existem dois tipos de vacinas, as estritamente polissacarídicas e as conjugadas. Nas vacinas polissacarídicas apenas são incluídos os polissacáridos, enquanto nas vacinas conjugadas os polissacáridos estão unidos a uma proteína transportadora. Atualmente, em Portugal, uma vacina polissacarídica 23-valente (PPV23) e duas vacinas conjugadas pneumocócicas (PCV), uma com 10 serotipos (PCV10) e outra com 13 serotipos (PCV13) estão disponíveis no mercado. A PCV13 substituiu a vacina conjugada 7-valente (PCV7) aumentando a proteção conferida pela vacina para 13 serotipos. A PCV7 foi a primeira vacina conjugada introduzida em Portugal e causou uma alteração na distribuição dos serotipos, tanto na colonização como na doença invasiva pneumocócica (IPD). Os serotipos vacinais na IPD em crianças diminuíram e, de forma menos pronunciada, também diminuíram na IPD em adultos. A diminuição da incidência dos serotipos incluídos na PCV7 deu origem à expansão dos serotipos não-vacinais (NVT) que aumentaram em frequência na IPD. Os genotipos que apresentam serotipos incluídos nas vacinas também podem escapar à sua proteção alterando a sua cápsula através da recombinação genética dos seus genes capsulares, um processo denominado troca capsular.

A plasticidade do genoma bacteriano é essencial para a adaptação rápida a ambientes em mudança e para a colonização de novos nichos ecológicos. Os mecanismos moleculares responsáveis pela plasticidade do genoma podem ser mutações pontuais, rearranjos genómicos, elementos genéticos móveis ou a transferência horizontal de genes, tanto por transformação genética natural como

por transdução. A transferência horizontal de genes e a recombinação são importantes porque distribuem e fixam mutações benéficas e também são responsáveis pela aquisição de novos elementos genéticos, tendo um papel essencial no desenvolvimento de resistência a antibióticos, troca capsular e na aquisição de fatores de virulência. Entre os mecanismos de transferência genética horizontal, a transformação genética natural tem sido a mais estudada em *S. pneumoniae* e é crucial para a plasticidade genética desta bactéria. O estado de competência necessário para que a transformação genética natural ocorra nos pneumococos é rigidamente controlado por um sistema de *quorum-sensing*. O produto do gene *comC* é processado e secretado por um transportador ABC (ComAB) resultando na acumulação extracelular de um péptido de 17 aminoácidos, o péptido estimulador da competência (CSP). Um sistema de dois componentes, consistindo de uma histidina cinase (ComD) e o seu regulador de resposta (ComE), é responsável por detetar a concentração extracelular de CSP e despoletar a cascata de reações que induzem a competência através da ativação de um conjunto de genes precoces, incluindo o gene *comX*. O produto do gene *comX* é um fator sigma alternativo que desencadeia a expressão de genes tardios necessários à competência. Foram identificados vários alelos do gene *comC* em *S. pneumoniae*, dando origem a pelo menos 3 CSPs diferentes. No entanto, a maioria das estirpes apresenta apenas um dos dois principais alelos: CSP1 (≈70 % das estirpes) ou CSP2 (≈30 % das estirpes). Cada estirpe possui apenas uma cópia do gene *comC* e responde especificamente ao CSP produzido devido à especificidade mostrada pelo domínio sensor de ComD. Assim sendo, os pneumococos podem ser divididos em subpopulações tendo em conta o CSP a que respondem, definindo o ferótipo de uma estirpe (frequentemente o ferótipo CSP1 ou CSP2). A designação de ferótipo foi escolhida porque o CSP comporta-se como uma feromona.

O objetivo principal desta tese foi estudar os ferótipos de *S. pneumoniae*, que são definidos pela resposta específica de uma estirpe a uma variante do CSP, e o impacto da diversidade do ferótipo na estrutura genética da população pneumocócica. Para atingir este objetivo, a diversidade e a abundância dos ferótipos foram determinadas em estirpes recolhidas de infeções pneumocócicas invasivas ocorridas em Portugal. Para identificar os determinantes estruturais da especificidade dos ferótipos, tentámos obter a estrutura 3D do domínio sensor de ComD mediante cristalografia de raios X e, posteriormente, cristalizá-lo em complexo com o CSP. As interações específicas entre ComD e CSP de cada um dos

ferótipos poderiam ser observadas em detalhe usando esta técnica. Para aplicar esta técnica fiz um estágio no laboratório do Dr. Juan Hermoso no Instituto Química Física Rocasolano, CSIC, Madrid. Também foi escolhido o serogrupo 6, cujos serotipos apresentam ferótipos diferentes, para avaliar se a recombinação entre as estirpes foi influenciada ou não. O estudo do serogrupo 6 também teve como objetivo avaliar a evolução clonal e dos serotipos durante 14 anos (1999-2012) e verificar se os novos serotipos identificados recentemente no serogrupo 6 estavam presentes na nossa coleção.

Foi realizado um estudo detalhado da diversidade genética do operão *comCDE* de uma amostra de estirpes responsáveis por infeções pneumocócicas invasivas em Portugal. Foram identificados 30 perfis únicos dos alelos *comCDE*. Três variantes de CSP foram identificadas correspondendo aos ferótipos CSP1, CSP2 e CSP3. No entanto, os CSPs encontrados neste estudo já eram conhecidos, não tendo sido identificadas novas variantes do CSP. A diversidade do gene *comD* foi relacionada com o ferótipo, não acontecendo o mesmo com o gene *comE*. A maior parte da divergência genética estava localizada na região que codifica o CSP e o domínio sensor de ComD. O gene *comD* apresentou a maior diversidade, enquanto o *comE* era mais conservado, apresentando apenas 5 variantes proteicas. Foi a primeira vez que o alelo *comC*-3 e, portanto, o ferótipo CSP3 foram identificados em pneumococos isolados em Portugal. *ComC*-1.1 (CSP1), *ComD*-1.1 e *ComE*-1 foram as variantes proteicas mais frequentemente identificadas nestas estirpes.

Utilizou-se uma extensa coleção de estirpes (n=903) responsáveis por infeções pneumocócicas invasivas em crianças (<18 anos) durante um período de 14 anos para identificar a abundância de cada um dos ferótipos e avaliar a sua evolução, usando um novo esquema de PCR para identificar o ferótipo. Também foram realizadas associações entre o ferótipo e outras características epidemiológicas. O ferótipo mais abundante foi o CSP1 (n=681, 75,4 %), seguido pelo CSP2 (n=192, 21,3 %) e depois pelo CSP3 (n=20, 2,2 %). As proporções dos ferótipos permaneceram estáveis, apesar das mudanças ocorridas na distribuição dos serotipos. Os ferótipos CSP1 e CSP2 apresentaram o mesmo grau de diversidade em relação ao serotipo, ST e CC. Este trabalho confirmou que o ferótipo CSP3 não é frequente mas está distribuído em diversas linhagens genéticas.

Outro objetivo desta tese foi estudar a especificidade do domínio sensor de ComD em relação ao CSP através da sua caracterização estrutural por cristalografia de raios X. Quatro construções de partes do sensor de ComD foram

desenhadas e produzidas com as etiquetas de afinidade 6x His e LSL_{150} e inseridas nas estirpes *E. coli* BL21(DE3) e *E. coli* JM109(DE3). A produção do sensor de ComD não foi alcançada em nenhuma das condições testadas e não foram realizadas experiências adicionais com esta proteína. ComD poderia apresentar uma topologia semelhante à proteína AgrC, um recetor de membrana do tipo histidina cinase de *Staphylococcus aureus*, com o seu domínio sensor inserido na membrana através de 7 segmentos transmembranares conectados por pequenas regiões extra e intracelulares. Muitas das substituições de aminoácidos entre ComD1, ComD2 e ComD3 poderiam estar localizadas nos segmentos transmembranares. A presença de tantas hélices membranares pode ter sido a razão pela qual a produção do sensor de ComD não foi bem-sucedida.

O serogrupo 6 foi escolhido para avaliar a influência do ferótipo na recombinação entre os serotipos deste serogrupo, uma vez que estes apresentaram distintos ferótipos. Observou-se uma diminuição da variante 6B-2 na doença pneumocócica invasiva (IPD), mas não de 6B-1, após a introdução da vacina conjugada, sustentada por uma diminuição das estirpes do CC273. O serotipo 6C foi associado com a IPD em adultos e aumentou nesta faixa etária apresentando duas linhagens genéticas (CC315 e CC395), enquanto em crianças as mesmas linhagens expressaram outros serotipos do serogrupo 6. Em conjunto, estes resultados sugerem uma potencial proteção cruzada das PCVs contra o serotipo 6C entre as crianças vacinadas, mas não entre adultos. O serotipo 6A tornou-se o serotipo mais importante do serogrupo 6 em crianças, mas diminuiu na IPD em adultos. Não foram detetados outros serotipos do serogrupo 6, portanto, os ensaios fenotípicos disponíveis ou os ensaios genotípicos simples permanecem adequados para distinguir os serotipos dentro das estirpes do serogrupo 6. A influência do ferótipo na recombinação entre as estirpes do serogrupo 6 foi avaliada através da análise dos possíveis casos de troca capsular. No entanto, devido ao baixo número destes casos, o serogrupo 6 não proporcionou uma conclusão clara sobre este tema. Assim, foi utilizada uma coleção maior de estirpes isoladas de crianças com doença invasiva pneumocócica (n=903) e foram feitos testes de fluxo de genes e diferenciação genética entre populações com um ferótipo diferente usando os dados de MLST destas estirpes. No entanto, a influência do ferótipo na estrutura genética da bactéria *S. pneumoniae* não pôde ser elucidada com a informação obtida nesta tese.

S. pneumoniae apresenta um polimorfismo nos genes *comC* e *comD* e a interação dos seus produtos pode ser como um mecanismo de chave e fechadura. Foi visto que estes genes estão sob seleção positiva e, por isso, se ocorrerem mutações nesta região, estas poderiam diminuir consideravelmente a eficiência da indução de competência. A proporção dos ferótipos parece ser 70:28:2 (CSP1:CSP2:CSP3) e também parece que é mantida em populações provenientes de diversos locais e de datas diferentes. Porém, o mecanismo que mantém estas proporções continua por esclarecer, embora o facto das estirpes do ferótipo CSP1 apresentarem uma melhor capacidade do que as estirpes do ferótipo CSP2 para formar biofilmes e produzir transformantes possa explicar a maior proporção do alelo *comC1* na população pneumocócica.

Abstract

Keywords: *Streptococcus pneumoniae*, pherotype, competence, genetic transformation, genetic structure.

Streptococcus pneumoniae is a Gram-positive bacterium and is considered an opportunistic pathogen because it colonizes asymptomatically the human nasopharynx and colonization is a pre-requisite for local and systemic disease.

The widespread use of antimicrobials decreased dramatically the global burden of pneumococcal disease. However, resistance to antibiotics quickly arose after its initial use. Nowadays, *S. pneumoniae* has high frequency of resistance to a large group of antibiotics, including ß-lactams and quinolones. In order to prevent pneumococcal infections and with the advent of antimicrobial resistance, vaccines have been developed using capsular polysaccharides as immunogens. There are two types of vaccines, strictly polysaccharide and conjugate vaccines. The first is constituted only by polysaccharides whereas in the latter polysaccharides are linked to a carrier protein. Currently, in Portugal the two types of vaccines are licensed, the 23-valent polysaccharide vaccine (PPV23) and two pneumococcal conjugate vaccines (PCV), one with 10 serotypes (PCV10) and another with 13 serotypes (PCV13). PCV13 replaced the 7-valent conjugate vaccine (PCV7) increasing vaccine protection to 13 serotypes. PCV7 was the first conjugate vaccine introduced in Portugal and caused a change in the distribution of serotypes both in colonization and in invasive pneumococcal disease (IPD). Vaccine serotypes responsible for IPD in children decreased and, in a less pronounced way, also decreased in adults. The decrease in the incidence of serotypes included in PCV7 resulted in the expansion of non-vaccine serotypes (NVT) that increased in frequency in IPD. Genotypes that display serotypes included in the vaccines may also escape vaccine pressure by altering their capsule through the genetic recombination of their capsular genes, a process called capsular switching.

Bacterial genome plasticity is essential for rapid adaptation to changing environments and colonization of new niches. The molecular mechanisms responsible for genome plasticity are point mutations, genomic rearrangements, mobile genetic elements and horizontal gene transfer either by natural genetic transformation or transduction. Horizontal gene transfer and recombination are

important because they distribute and fix beneficial mutations and are also responsible for the acquisition of novel genetic information, having an essential role in the development of antibiotic resistance, capsular switching and the acquisition of virulence determinants. Among the mechanisms of horizontal gene transfer, natural genetic transformation has been the most studied in *S. pneumoniae* and it is crucial for genetic plasticity in pneumococci. Pneumococci undergo genetic transformation through natural competence which is tightly controlled by a quorum-sensing-like system. The product of the *comC* gene is processed and secreted by an ABC transporter (ComAB) resulting in the accumulation of a 17-aminoacid peptide pheromone, the competence-stimulating peptide (CSP), into the medium. A two-component system, consisting of a histidine kinase receptor (ComD) and its cognate response regulator (ComE), is responsible for sensing CSP concentration and triggering the competence response by activating a set of early genes including *comX*. The product of the *comX* gene is an alternative sigma factor that triggers the expression of late genes necessary for competence.

In pneumococci, several alleles of the *comC* gene have been identified producing at least 3 different mature CSPs. However, most strains present only one of two main alleles: CSP1 (≈70 % of the strains) or CSP2 (≈30 % of the strains). Each strain has only one copy of a *comC* allele and responds specifically to the CSP produced due to the specificity shown by the sensor domain of ComD. Consequently, pneumococci can be divided into subpopulations according to the CSP they respond to, defining the pherotype of an isolate (usually pherotype CSP1 or CSP2). It was termed pherotype because CSP behaves like a pheromone.

The main purpose of this thesis was to study pneumococcal pherotypes, the mechanism underlying the specific response of a given isolate to a CSP variant, and the impact of pherotype diversity on the genetic structure of the pneumococcal population. To achieve this goal, pherotype diversity and abundance were determined in pneumococcal isolates recovered from IPD in Portugal. To determine the structural determinants of pherotype specificity we attempted to study the 3D structure of the ComD sensor by X-ray crystallography and, ultimately, to crystallize it in complex with CSP. Specific interactions between ComD and CSP of each pherotype could be observed in detail using this technique. To perform these experiments, I visited Dr. Juan Hermoso's lab in Instituto Química Física Rocasolano, CSIC, Madrid. We also focused on serogroup 6, whose serotypes were

associated with different pherotypes, to see if recombination was restricted or not between the isolates. Study of serogroup 6 also had the aim of evaluating their clonal and serotype evolution during 14 years (1999-2012) and to check if the newly identified serogroup 6 serotypes were present in our pneumococcal collection.

It was performed a comprehensive study of the genetic diversity of the *comCDE* locus of a sample of invasive pneumococci recovered in Portugal. A total of 30 unique profiles of *comCDE* alleles were identified. Three CSP variants were identified corresponding to pherotypes CSP1, CSP2 and CSP3. However, the CSPs found in this study were already known, thus new variants of CSP were not discovered. The diversity of *comD* was associated with pherotype but that was not the case for *comE*. The majority of the genetic divergence was located in the region coding the mature CSP and the sensor domain of ComD. The gene *comD* was the most diverse while *comE* was more conserved presenting just 5 protein variants. It was the first time that the *comC*-3 allele and therefore the CSP3 pherotype were identified in pneumococcal isolates recovered in Portugal. *ComC*-1.1 (CSP1), *ComD*-1.1 and *ComE*-1 were the most frequent protein variants produced in these isolates.

We used an extensive pneumococcal invasive isolate collection (n=903) recovered from children (<18 years) during a period of 14 years to identify the abundance of each pherotype and evaluate their evolution, using a new PCR scheme to identify pherotype. Then, associations between pherotype and other characteristics were tested. The most abundant pherotype was CSP1 (n=681, 75.4 %), followed by CSP2 (n=192, 21.3 %) and then by CSP3 (n=20, 2.2 %). Pherotype proportions remained stable despite the changes that occurred in serotype distribution and pherotype CSP1 and CSP2 presented the same degree of diversity regarding serotype, ST and CC. This work confirmed that pherotype CSP3 was infrequent but it was distributed throughout several genetic backgrounds.

Another goal of this thesis was to study the specificity of ComD sensor domain towards CSP through the structural characterization by X-ray crystallography. Four constructs of parts of the ComD sensor were designed and produced with 6x His and LSL_{150} affinity tags and inserted into the *E. coli* strains BL21(DE3) and JM109(DE3). Production of the ComD sensor was not achieved in any of the conditions tested and further experiments were not performed with this protein. ComD could present a topology similar to AgrC, a histidine kinase membrane receptor from *Staphylococcus aureus*, presenting a membrane-

embedded sensor with 7 transmembrane segments connected by small extracellular regions, where many of the amino acid substitutions between ComD1, ComD2 and ComD3 could be located in transmembrane domains. The presence of so many membrane helices could be the reason why the production of ComD sensor was a not achieved.

We also chose serogroup 6 to see if pherotype was influencing recombination between serogroup 6 serotypes since they present distinct serotypes. It was seen a decrease of the 6B-2 variant among invasive pneumococcal disease (IPD), but not 6B-1, post conjugate vaccine introduction, underpinned by a decrease of CC273 isolates. Serotype 6C was associated with adult IPD and increased in this age group representing two lineages (CC315 and CC395), while the same lineages expressed other serogroup 6 serotypes in children. Taken together, these findings suggest a potential cross-protection of PCVs against serotype 6C IPD among vaccinated children but not among adults. Serotype 6A became the most important serogroup 6 serotype in children but it decreased in adult IPD. No other serogroup 6 serotypes were detected, so available phenotypic or simple genotypic assays remain adequate for distinguishing serotypes within serogroup 6 isolates. Influence of pherotype on the recombination between serogroup 6 isolates was evaluated by analyzing the cases of possible capsular switching events. However, due to the low number of these events, serogroup 6 did not provide a clear evidence to clarify this. Thus, a large collection of pneumococci causing IPD in children (n=903) was used and gene flow and genetic differentiation tests were performed using their MLST data. However, the influence of pherotype on the genetic structure of *S. pneumoniae* was not clear, even when considering all the information collected.

S. pneumoniae presents a polymorphism in the genes *comC* and *comD* and the interaction of their products might function as a lock and key mechanism. These genes are under positive selection and mutations in this region could decrease the efficiency of competence induction. Pherotype abundance in a 70:28:2 ratio (CSP1:CSP2:CSP3) seems to be maintained in populations from diverse geographic origins and isolation dates. However, the mechanism maintaining it remains to be elucidated, although the better capacity of CSP1 strains to form biofilms and yield more transformants than CSP2 strains may contribute to explain the higher frequency of the *comC1* allele in the pneumococcal population.

Abbreviations

AU: absorbance units

AW: Adjusted Wallace

CA: Cochran-Armitage test

CBD: choline binding domain

CC: clonal complex

CSF: cerebrospinal fluid

CSP: competence-stimulating peptide

DLV: double locus variant

DR: direct repeat

FDR: false discovery rate

FET: Fisher's exact test

IMAC: immobilized metal ion affinity chromatography

IPD: invasive pneumococcal disease

IPTG: isopropyl-ß-D-thiogalactopyranoside

IRL: inverted repeat left

IS: insertion sequence

LIC: ligation-independent cloning

MDR: multidrug resistance

MIC: minimum inhibitory concentration

MLST: multilocus sequence typing

MST: minimum spanning tree

NVT: non-vaccine serotype

PCR: polymerase chain reaction

PCV: pneumococcal conjugate vaccine

PNSP: penicillin non-susceptible pneumococci

RCF: relative centrifugal force

SID: Simpson's index of diversity

SLV: single locus variant

SNP: single nucleotide polymorphism

ST: sequence type

TEV: tobacco etch virus

Thesis Outline

The main purpose of this thesis was to study pneumococcal pherotypes, the mechanism underlying the specific response of a given isolate to a variant of the competence-stimulating peptide (CSP) and the impact of pherotype diversity in pneumococcal population genetic structure.

This thesis was divided into 4 parts:

- **Part I – Introduction:** it is an introductory part and consists of:
    - **Chapter 1 - Background:** in this chapter is presented a review of the literature providing the essential information to read this thesis.
    - **Chapter 2 – Aims of the thesis:** the questions pursued in this work to achieve the goals established are presented in this chapter.
    - **Chapter 3 – Pneumococcal isolates and typing methods:** the common materials and methods, transversal to all chapters, are presented here. Specific procedures of the studies performed are presented in their respective chapter.

- **Part II – Pherotype diversity and abundance:** this part presents an initial description of pherotype characteristics such as their genetic diversity and abundance among pneumococci populations. Two studies were performed and are presented in the following chapters:
    - **Chapter 4 – Diversity of the *comCDE* locus:** a genetic analysis of the pherotype-defining and competence-regulating genes was performed to check and compare the *comCDE* diversity of the pneumococci lineages circulating in Portugal with other studies.
    - **Chapter 5 – Pherotype abundance:** it was studied the proportion of pherotypes over a 14 years period when conjugate vaccines were introduced and substantial changes occurred in serotype distribution.

- **Part III – Pherotype specificity and influence on the genetic structure:** the work presented in this part addressed the specificity of pherotypes and its

consequences for the evolution of *S. pneumoniae*. This is the last section presenting results, which are divided into two chapters:

- **Chapter 6 – Structural study of pherotype specificity by X-ray crystallography:** production of the ComD sensor for crystallization experiments was attempted during my visit to Dr. Juan Hermoso lab in Instituto Química Física Rocasolano, CSIC, Madrid. The goal was to study the specific interaction between CSP and ComD through their 3D structure.
- **Chapter 7 – Serogroup 6 and influence of pherotype in genetic recombination:** this chapter presents the work performed to address the consequences of pherotype diversity on the genetic recombination among pneumococci. We focused on serogroup 6 to address this question and part of this work was published on the following article:

J. Diamantino-Miranda, S.I. Aguiar, J.A. Carriço, J. Melo-Cristino and M. Ramirez. (2017). *Clonal and serotype dynamics of serogroup 6 isolates causing invasive pneumococcal disease in Portugal: 1999-2012.* PLoS One, 12(2):e0170354. doi: 10.1371/journal.pone.0170354.

- **Part IV – Final remarks and future perspectives:** this is the final part of this thesis and a final reflection discussing the importance of the results of this work is performed.

I. Introduction

I. Introduction

1. Background

CHAPTER 1. BACKGROUND

1.1. Pneumococcus discovery and identification

Streptococcus pneumoniae, also known as pneumococcus, was first isolated simultaneous and independently by George Sternberg and Louis Pasteur in 1881 (Watson *et al.*, 1993). After its identification, it was confirmed to be an asymptomatically carried agent that could cause disease in humans and other mammalian hosts. The first name assigned to this bacterium was *Diplococcus pneumoniae* in 1920, which was later, changed to *Streptococcus pneumoniae* in 1974 (Grabenstein and Klugman, 2012). *S. pneumoniae* is a Gram-positive bacterium with a genome containing a high AT content. Morphologically, pneumococci typically present in pairs of cells with a lancet-shape. Phylogenetic studies placed *S. pneumoniae* in the Mitis group with *Streptococcus pseudopneumoniae* and *Streptococcus mitis* being the closest relatives of the pneumococcus (Kilian *et al.*, 2008).

The first approach to microbiological identification of *S. pneumoniae* consists in the detection of α-haemolytic colonies on blood agar plates, cell lysis by bile and susceptibility to optochin (Werno and Murdoch, 2008). Regarding cell lysis by bile, the major pneumococcal autolytic enzyme LytA is activated in the presence of the bile salt deoxycholate (Mosser and Tomasz, 1970). This feature distinguishes *S. pneumoniae* from other α-haemolytic streptococci. Another test to differentiate pneumococci from commensal streptococci is the susceptibility to optochin (also known as ethylhydrocupreine) because this antibiotic is able to inhibit the pneumococcal F_0F_1-ATPase (Angulo *et al.*, 2011). However, resistance to deoxycholate and to optochin was already reported (Aguiar *et al.*, 2006, Obregón *et al.*, 2002, Pikis *et al.*, 2001), making each of these phenotypic tests separately not fully conclusive for identification of pneumococci.

1.2. Colonization and disease

S. pneumoniae is considered an opportunistic pathogen because it colonizes asymptomatically the human nasopharynx and colonization is required before local and systemic disease. It is assumed that the human nasopharynx is the main

reservoir of pneumococci, especially in children where up to almost 70 % can be colonized (Sá-Leão *et al.*, 2009). Colonization starts with exposure to the bacteria and their attachment to epithelium. Then, bacteria replicate by obtaining nutrients and finally they persist by evading the host immune system. The cycle is complete with transmission to a new host (Siegel and Weiser, 2015). Colonization of the upper respiratory tract can not only evolve into disease but also drives the evolution of *S. pneumoniae* (Siegel and Weiser, 2015).

Mucus is an important physical barrier to colonization and infection. The pneumococcal polysaccharide capsule is the main virulence factor of *S. pneumoniae* and possibly evolved to allow survival in the mucus. Pneumococci undergo phase variation, a process that regulates capsule production. A thick capsule is produced to pass through the mucus by anionic interactions because most capsules are negatively charged. However, the presence of a thick capsule hampers attachment to the epithelium because the capsular polysaccharide masks the structures at the cell surface mediating adherence to host cells. For this reason, capsule production is downregulated to enable attachment to host surface carbohydrates. Pneumococci present several adhesion factors, for example NanA, a lectin-like protein, PspC also known as CbpA, and pili, the later divided into two major isoforms (Aguiar *et al.*, 2012, Horácio *et al.*, 2016). After adhesion to the epithelium, pneumococci form a biofilm, a highly-structured community of cells that produce an extracellular matrix and adhere to abiotic or biological surfaces (**Figure 1.1**). Forming biofilms during colonization may serve as protection against the host, by reducing production of inflammation inducing factors, and also promotes closeness between cells to facilitate the exchange of genetic material (Chao *et al.*, 2014).

Pneumococci interact with the host mucosal immune system and must evade it to persist. Capsule is a major player in this interaction because it limits opsonisation by complement and antibody. Pneumococci also possess an IgA1 protease which cleaves human immunoglobulin A1, the dominant IgA present in the human respiratory tract. During colonization, pneumococci face not only the host immune system but also need to compete with the indigenous host microbiota or other opportunistic pathogens. *Haemophilus influenzae* outcompetes *S. pneumoniae* by promoting its clearance (Lysenko *et al.*, 2005) while pneumococcal colonization protects against *Staphylococcus aureus* colonization (Bogaert *et al.*, 2004). Co-infection with Influenza virus is beneficial to pneumococci because it increases its replication in the nasopharynx (Siegel *et al.*, 2014). Pneumococci also

compete with other pneumococcal strains by producing bacteriocins that target them or strains of closely related species (Dawid *et al.*, 2007).

Colonizing pneumococci may progress to disease causing a range of infections. However, pneumococcal infections do not promote spread to new hosts and are therefore believed not to be a major selective pressure in the evolution of these bacteria. In contrast, the pressures to successfully colonize the host lead to the development of virulence factors that ultimately contribute to disease (Weiser, 2009). The range of infections may be from local and mild like acute otitis media, sinusitis and conjunctivitis, to systemic and severe like pneumonia, septicaemia or meningitis, with associated mortality rates around 5 %, 20 % and 30 %, respectively (Henriques-Normark and Tuomanen, 2013). The pneumococcal infections in which pneumococci invade a normally sterile body site such as blood and cerebrospinal fluid (CSF) are considered invasive pneumococcal disease (IPD). Annual IPD burden is around 14.5 million cases resulting in 800 000 deaths in children under the age of 5 and more than 20 % of the cases in the elderly result in death (Heron, 2012, Naucler *et al.*, 2013, O'Brien *et al.*, 2009). These data show that children and the elderly are the main groups at increased risk for pneumococcal infection followed by immunocomprised people.

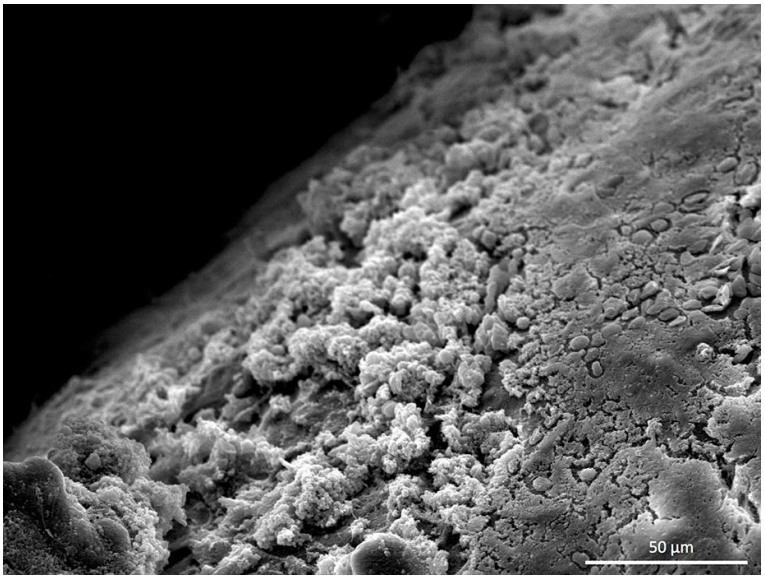


Figure 1.1 – Scanning electron microscopy image of *S. pneumoniae* biofilms formed on the nasal septum of a mouse. Mice were experimentally colonized 7 days prior. Biofilms are the non-contiguous aggregates on the left. Adapted from Gilley and Orihuela (Gilley and Orihuela, 2014).

1.3. Antimicrobial resistance

The widespread use of antimicrobials decreased dramatically the global burden of pneumococcal disease. However, resistance to antibiotics quickly arose after its initial use. Since the identification of the first *S. pneumoniae* strain with reduced susceptibility to penicillin (Hansman and Bullen, 1967), antibiotic resistance among *S. pneumoniae* has increased and it has become a global issue (Appelbaum, 1987). Moreover, at this point *S. pneumoniae* has inherent resistance to a large group of antibiotics, including ß-lactams, macrolides, aminoglycosides and first-generation quinolones (El Moujaber *et al.*, 2017). Below is presented a brief review of the molecular mechanisms underlying resistance to the antimicrobials for which susceptibility was tested in this thesis: penicillin and cefotaxime (ß-lactams), erythromycin (macrolide), clindamycin (lincosamide), tetracycline, chloramphenicol, co-trimoxazole (sulfonamide), levofloxacin (fluoroquinolone), vancomycin (glycopeptide) and linezolid (oxazolidinone).

1.3.1. β-lactams

β-lactam antibiotics inhibit cell-wall synthesis through the binding to specific enzymes called penicillin-binding proteins (PBPs) (Neu and Gootz, 1996). *S. pneumoniae* presents a total of six PBPs but only three (PBP1a, PBP2x and PBP2b) are associated with resistance (**Table 1.1**).

The main β-lactam resistance mechanism in *S. pneumoniae* is the acquisition of altered PBP genes, or of gene fragments which are recombined with the native genes, from related species such as *S. mitis* and *Streptococcus oralis* through genetic transformation (Jensen *et al.*, 2015). On the one hand, studies show that a high resistance level is achieved through the presence in the same strain of three altered PBPs: PBP1a, PBP2b and PBP2x. On the other hand, a low level of penicillin resistance could be based mainly on modifications of PBP2x and PBP2b, resulting in penicillin non-susceptible *S. pneumoniae* (PNSP) (El Moujaber *et al.*, 2017). MurM and MurN, encoded by the *murMN* operon, have also been associated with penicillin resistance (Filipe and Tomasz, 2000, Garcia-Bustos and Tomasz, 1990). These proteins are responsible for an irregularity in cell-wall synthesis because of the substitution of linearized muropeptides by atypical branched ones. However, MurM alone is insufficient to induce penicillin resistance, although it is

crucial to get the highest level of resistance to penicillin and cephalosporins, another class of β-lactams antimicrobials (Smith and Klugman, 2001).

Table 1.1 – Mutations associated with β-lactam resistance. Table adapted from (El Moujaber *et al.*, 2017).

	Mutation	Resistance
pbp2x	Thr550Ala	Cefotaxime
	T338A	Penicillin
pbp2b	Thr446Ala	Piperacillin
	T451A	Penicillin
pbp1a	Mosaic gene	Penicillin and cephalosporin
cpoA	Gly12Val	Piperacillin
ciaH	Thr230Pro	Cefotaxime
	Ala203Val	
murM	Mosaic gene	Penicillin and cephalosporin

1.3.2. Macrolide, lincosamide and streptogramin B

Resistance to macrolides in pneumococci can be due to modification of the ribosome and it was the first resistance mechanism described in *S. pneumoniae*. Ribosomal alteration is caused by the *erm* gene frequently located in the transposons Tn1545 (Courvalin and Carlier, 1986) and Tn917 (McDougal *et al.*, 1998). A methylase is encoded by *ermB* which is responsible for the dimethylation of the adenine A2058 in the domain V of 23S rRNA, reducing its affinity to macrolides (Arthur *et al.*, 1987). In addition, *ermA*, another methylase, was also implicated in macrolide resistance (Syrogiannopoulos *et al.*, 2001) and both *ermA* and *ermB* provide resistance to the macrolide-lincosamide-streptogramin B families (MLS_B phenotype) (McDougal *et al.*, 1998, Syrogiannopoulos *et al.*, 2001). This kind of resistance can be displayed in two distinct forms, either constitutive ($cMLS_B$) or inducible ($iMLS_B$). These resistance phenotypes are defined by the regulatory region and not by a specific gene type (Leclercq and Courvalin, 1991). Moreover, at the end of the 1990s it was identified a different resistance pattern. In this case, it was observed resistance to macrolides in the absence of *ermA* and *ermB* genes but vulnerability to clindamycin and streptogramin B. This phenotype, known as the M phenotype, was recognized in both *S. pneumoniae* and *Streptococcus pyogenes* (Sutcliffe *et al.*, 1996). Shortly after, *mefE* was identified as a key factor in conferring resistance with this phenotype (Tait-Kamradt *et al.*, 1997). The *mefE* gene encodes an efflux pump which expels the macrolide antibiotic from the cell (Sutcliffe *et al.*, 1996). Furthermore, it was identified a transposon,

Tn2010, which carried both *mefE* and *ermB*, making it a dual-macrolide resistance gene carrier transposon (Del Grosso *et al.*, 2007).

1.3.3. Tetracycline

Tetracycline resistance is mostly associated with the presence of *tet* genes that encode efflux proteins, like *tetA*, *tetB*, *tetC* and *tet31*, and proteins related with ribosomal protection, like *tetM, tetO, tetT* and *tetW*, being *tetM* the dominant tetracycline resistance gene found in *S. pneumoniae* (Chopra and Roberts, 2001). These genes were shown to be carried in a transposon together with another resistance gene conferring chloramphenicol resistance: *cat* (Shoemaker *et al.*, 1979). This genetic element was called Tn5253 (Ayoubi *et al.*, 1991) and found to be composed of two distinct conjugative transposons: the *tet*-carrier Tn5251 inserted within Tn5252, the carrier of the *cat* gene. Moreover, other transposons were also recognized as transporters of *tetM*, such as Tn1545 (Courvalin and Carlier, 1986), Tn2009 (Del Grosso *et al.*, 2004), Tn3872, Tn2010 and Tn2017 (Roberts and Mullany, 2009). Nevertheless, several works described tetracycline susceptibility among *tetM*-harbouring strains (Varaldo *et al.*, 2009). The reason for the low transcription level of *tetM* and consequently the susceptible profile was the occurrence of frame shift mutations originated by either a deletion or an insertion (Grohs *et al.*, 2012). Finally, a recent study revealed another tetracycline resistance mechanism in *S. pneumoniae* in the absence of *tet* genes (Lupien *et al.*, 2015). In this case, resistance was conferred by mutations in the *rpsJ* gene, which encodes the ribosomal protein S10 close to the tetracycline site of action, in parallel with mutations in the *patA* gene causing overexpression of this ABC transporter.

1.3.4. Chloramphenicol

The main mechanism of chloramphenicol resistance among bacteria, including *S. pneumoniae*, is the acquisition of chloramphenicol acetyltransferase (Cat) encoded by the *cat* gene. This protein is responsible for chloramphenicol acetylation generating *O*-acetoxy chloramphenicol products which decreases the binding of the antimicrobial to ribosomes (Dang-Van *et al.*, 1978, Yunis, 1988) . As referred before, in *S. pneumoniae* the *cat* gene is transported by the composite transposon Tn5253 (Ayoubi *et al.*, 1991).

1.3.5. Sulfonamide

Some studies have shown that the major increase in resistance to trimethoprim and trimethoprim-sulfamethoxazole (also known as co-trimoxazole) was caused by a single amino acid substitution (Ile100Leu) in dihydrofolate reductase (Adrian and Klugman, 1997, Schmitz *et al.*, 2001). This mutation can be accompanied by other mutations in this protein that result in increases of resistance. In the absence of the substitution in position 100, it was described another substitution, Asp92Ala, in more than 30 % of co-trimoxazole resistant pneumococci (Cornick *et al.*, 2014). Co-trimoxazole resistance also requires mutations in dihydropteroate synthase (Padayachee and Klugman, 1999).

1.3.6. Fluoroquinolone

Several mutations in the genes *gyrA*, *gyrB*, *parC* and *parE* have been identified as the main reason of quinolone resistance in *S. pneumoniae*. The level of fluoroquinolone resistance depends on the mutations that occurred. While mutations in *parC* alone confer low-level quinolone resistance, when they are accompanied by mutations in *gyrA* the resistance level is higher. Moreover, *gyrA* or *parC* mutations can also propitiate further substitutions increasing the resistance level (Gillespie *et al.*, 2003). Mutations in *gyrB* and *parE* were also observed but they conferred a lowest level of resistance than mutations in either *gyrA* or *parC* (Hooper, 2000).

It was thought that target modification was the only mechanism responsible for quinolone resistance. However, *pmrA*, a pneumococcal multidrug resistance gene, coding for an efflux pump, was associated with fluoroquinolone resistance and displayed more than 20 % similarity to the *Bacillus subtilis* multidrug resistance (*bmr*) and *S. aureus* norfloxacin efflux pump (*norA*) genes (El Moujaber *et al.*, 2017). Additionally, the overexpression of *patA* and *patB* genes was connected with fluoroquinolone resistance in clinical isolates of *S. pneumoniae* since they become susceptible to this antimicrobial after inactivation of these genes (Garvey *et al.*, 2011). Both genes are located in the same operon and they interact together as a heterodimer forming an ABC transporter. However, the regulatory mechanisms controlling expression of these genes are still unknown (Baylay *et al.*, 2015, Baylay and Piddock, 2015).

1.3.7. Glycopeptide

Vancomycin resistance has not yet been found in *S. pneumoniae*. However, a mutation in the histidine kinase *vncS* gene was reported in vancomycin-tolerant pneumococci (Henriques Normark *et al.*, 2001, Sung *et al.*, 2006). This mutation was regarded to be the main mechanism of vancomycin tolerance causing the suppression of the autolytic activity. It was also seen that a *vncS* mutant was capable of developing tolerance to other antibiotics such as quinolones and ß-lactams (Novak *et al.*, 1999).

1.3.8. Oxazolidinone

Oxazolidinones are a new class of synthetic antibiotics and linezolid was the first oxazolidinone to become available. Linezolid is effective against Gram-positive bacteria but has limited activity against Gram-negative bacteria (Diekema and Jones, 2001). Linezolid is one of the few truly new antibiotics that have been introduced in many years, as most of the antibiotics recently released are derivatives of existing drugs. FDA approved linezolid to be used against *S. pneumoniae* because of the increase of PNSP. Although the mode of action of oxazolidinones is not completely clear, they act as protein synthesis inhibitors by binding to the ribosomal peptidyl transferase center and stopping the growth of bacteria (Bozdogan and Appelbaum, 2004, Long and Vester, 2012). It was observed that oxazolidinones bind to the 50S ribosomal subunit but they have no affinity to the 30S subunit (Zhou *et al.*, 2002). Oxazolidinones compete with chloramphenicol and lincomycin for binding to the 50S subunit, which indicates that they have a close binding site (Lin *et al.*, 1997). The effect of oxazolidinones binding to 50S is inhibition of initiation complex formation and of 70S formation (Aoki *et al.*, 1997, Bobkova *et al.*, 2003, Swaney *et al.*, 1998). When 70S is already formed, binding of oxazolidinones inhibits translocation of the peptide chain from A site to P site during peptide bond formation. The mechanism of action of oxazolidinones is similar to those of other antibiotics binding to the ribosomal peptidyl transferase center.

One of the resistance mechanisms to linezolid is target modification and transferable resistance has not been described yet for oxazolidinones. The other nonribosomal resistance mechanism reported is due to mutations causing an

increased expression of ABC transporter genes in *S. pneumoniae* (Billal *et al.*, 2011, Feng *et al.*, 2009). Linezolid resistant *S. pneumoniae* strains selected in vitro carried mutations in the peptidyl transferase centre of 23S rRNA (Bozdogan and Appelbaum, 2004). Linezolid resistance in clinical isolates of *S. pneumoniae* was firstly reported for two *S. pneumoniae* isolates that presented a minimum inhibitory concentration (MIC) of 4 μg/mL (Farrell *et al.*, 2004). These isolates presented two separate mutations, A2059G in combination with G2057A in the 23S rRNA and a 6-bp deletion in the L4 riboprotein gene, 64PWRQ67 to 64P__Q67.

1.4. Vaccination

In order to prevent pneumococcal infections and with the advent of antimicrobial resistance, vaccines have been developed, using capsular polysaccharides as immunogenic agents. However, these pneumococcal vaccines only confer protection against the serotypes included in their composition.

The known serotypes are not equally prevalent among the isolates causing IPD and only a few serotypes are responsible for most of these infections. Their distribution also varies according to the age group considered and there are also differences in their prevalence in respect to colonization. Thus, some serotypes seem to have an enhanced invasive capacity, while others are more frequently found in colonization (Sá-Leao *et al.*, 2011). Serotype prevalence also varies over time and there are differences in their distribution between several countries (Hausdorff, 2002). For this reason, it is important to monitor strains in circulation in the population at the local level. At present, epidemiological surveillance of IPD makes it possible to evaluate the effect of vaccines on the distribution of serotypes in different countries and whether new formulations of vaccines are needed.

There are two types of vaccines, polysaccharide and conjugate vaccines. In the first only polysaccharides are included whereas in the latter polysaccharides are linked to a carrier protein. Currently, in Portugal, the two types of vaccines are licensed, the 23-valent polysaccharide vaccine (PPV23) and two pneumococcal conjugate vaccines (PCV), one with 10 serotypes (PCV10) and another with 13 serotypes (PCV13). PPV23 was introduced in Portugal in 1996, while the PCV10 and PCV13 vaccines were licensed by the European Medicines Agency in March and December 2009 (Aguiar *et al.*, 2010b), respectively, and were introduced in Portugal on the same date. PCV13 replaced the 7-valent conjugate vaccine (PCV7)

increasing vaccine protection to 13 serotypes. PCV7 was available for pediatric vaccination in Portugal through the private sector from June 2001 onwards and its uptake slowly increased over the years (Horácio *et al.*, 2016). In 2012, PCV13 received approval for use also in adults >50 years of age with an extension being made to all ages in 2013. Additionally, PCV13 entered the Portuguese National Immunization Program in June 2015 for children born from January 2015 onwards (Horácio *et al.*, 2016). The serotypes included in the vaccines (**Table 1.2**) were chosen because they represented the majority of cases of IPD at the time they were implemented, although the serotypes included in PCV7 were based mainly in the serotype distribution of USA while PCV10 and PCV13 also took into consideration other geographical areas. Some of the vaccine serotypes were associated with antimicrobial resistance.

Conjugate vaccines are recommended for children under 2 years of age and for children up to 6 years of age who have a medical condition that increases the risk of pneumococcal infection (Nuorti and Whitney, 2010). Polysaccharides that are conjugated to a protein induce a thymus-dependent response, so conjugate vaccines are effective in children younger than 2 years old because their immune system is able to mount this type of response efficiently. Thus, after three doses in the first year, these vaccines induce the activation of memory cells and their protection is prolonged. Conjugate vaccines also confer immunity in the mucosa (Nurkka *et al.*, 2001), preventing colonization of the nasopharynx by vaccine serotypes (Klugman, 2001).

PCV7 was the first conjugate vaccine introduced in Portugal and caused a change in the distribution of serotypes both in colonization (Sá-Leão *et al.*, 2009) and in IPD (Aguiar *et al.*, 2010b, Aguiar *et al.*, 2008). Vaccine serotypes responsible for IPD in children decreased and, in a less pronounced way, also decreased in adults (Aguiar *et al.*, 2008). Thus, this vaccine proved to be very effective in the prevention of pneumococcal infections by vaccine serotypes and in the USA there was a decline in the incidence of IPD in all age groups (Pilishvili *et al.*, 2010). However, in United Kingdom PCV7 introduction also reduced the incidence of IPD but its effect was quickly mitigated by serotype replacement which happened faster than in USA (Choi *et al.*, 2011, Miller *et al.*, 2011). The effect observed in adults, which is an unvaccinated group, was probably due to the interruption of the transmission of the vaccine serotypes by the vaccinated children. This protective effect exerted on the non-vaccinated groups is called herd effect.

The decrease in the incidence of serotypes included in PCV7 resulted in the expansion of non-vaccine serotypes (NVT) that increased their frequency in IPD. This replacement was due to the reduction or even elimination of nasopharyngeal colonization by vaccine serotypes, which allowed the expansion of non-vaccine serotypes in this ecological niche (Dagan, 2009b). Among NVT, serotype 19A showed the greatest expansion following PCV7 use, both in the USA (Pilishvili *et al.*, 2010) and in Europe (Ardanuy *et al.*, 2009, Miller *et al.*, 2011). This serotype also increased in areas where PCV7 was not implemented, therefore, PCV7 only strengthened or accelerated the expansion of serotype 19A (Aguiar *et al.*, 2010c). Genotypes that display serotypes included in the vaccine may also escape vaccine pressure by altering their capsule through genetic recombination of the capsular locus, a process called capsular switching. It has been documented in the USA that a genotype associated with serotype 4 obtained serotype 19A capsular genes, evading PCV7 (Brueggemann *et al.*, 2007). PCV7 includes serotypes associated with antimicrobial resistance and, shortly after its introduction, the frequency of resistant strains decreased. However, resistance to antibiotics has increased again, as some of the NVT that have expanded are associated with resistance, as was the case of serotype 19A (Dagan, 2009a).

The use of PCV13 in the vaccination of adults aged 50 years and over has recently been authorized by the European Medicines Agency. PCV13 has been shown to be more immunogenic than PPV23 for most of the serotypes shared by both vaccines and is well tolerated in healthy adults (Scott *et al.*, 2007). The use of this vaccine in adults has the advantage of conferring direct protection against vaccine serotypes.

Table 1.2 – Serotype composition of pneumococcal vaccines.

Vaccine	Included Serotypes																							
PCV7				4			6B			9V				14			18C		19F			23F		
PCV10[a]	1			4	5		6B	7F		9V				14			18C		19F			23F		
PCV13[b]	1		3	4	5	6A	6B	7F		9V				14			18C	19A	19F			23F		
PPV23[c]	1	2	3	4	5		6B	7F	8	9N	9V	10A	11A	12F	14	15B	17F	18C	19A	19F	20	22F	23F	33F

[a]PCV10 has all serotypes of PCV7 plus serotypes 1, 5 and 7F.

[b]PCV13 has all PCV10 serotypes plus serotypes 3, 6A and 19A.

[c]PPV23 has all serotypes of PCV13, except serotype 6A, plus another eleven serotypes.

1.5. Genetic transformation

Bacterial genome plasticity is essential for rapid adaptation to changing environments and colonization of new niches. The molecular mechanisms responsible for genome plasticity are point mutations, genome rearrangements, mobile genetic elements and horizontal gene transfer either by natural genetic transformation or transduction (Straume *et al.*, 2015).

Horizontal gene transfer and recombination are important because they distribute and fix beneficial mutations and are also responsible for the gain of new genetic information. Between the mechanisms of horizontal gene transfer, natural genetic transformation has been the most studied in *S. pneumoniae* and is crucial for genetic plasticity in pneumococci (Chewapreecha *et al.*, 2014, Croucher *et al.*, 2011). Although it is recognized that pneumococcus is a highly transformable bacterium, there is a great variation in transformation frequencies among the pneumococcal population and it does not correlate with genetic relatedness (Evans and Rozen, 2013).

Natural genetic transformation is a phenomenon where cells are in a state of competence being able to recognize and bind to double-stranded DNA fragments present in the external environment, independent of their origin. These fragments are then transported through the membrane as single-stranded DNA and can be integrated into the genome through genetic recombination often leading to permanent modification of the cell genotype (Claverys and Havarstein, 2002). This mechanism was discovered in *S. pneumoniae* when, in 1928, Griffith found that non-virulent and non-capsulated strains were able to acquire a polysaccharide capsule and cause deadly infections in mice if they were administered concomitantly with encapsulated and virulent strains previously killed by heat (Griffith, 1928). It is now known that *S. pneumoniae* is not the only naturally competent species, with more than 60 other species found to be naturally transformable (Johnsborg *et al.*, 2007). Competence in *S. pneumoniae* is temporary, reaching a maximum about twenty minutes after the beginning of the process and occurs usually during the exponential phase of growth (Luo and Morrison, 2003).

1.5.1. Competence regulation

The mechanism of competence regulation is a form of intercellular communication that is called quorum-sensing because it allows monitoring the population density through the concentration of an extracellular peptide, the competence-stimulating peptide (CSP). A large number of bacterial species present this type of communication, which reveals the importance of quorum-sensing in the adaptation and survival of these microorganisms. Coordination of competence was among the first examples of intercellular communication between bacteria (Claverys *et al.*, 2006). The process of competence regulation is schematized in **Figure 1.2** and each step will be further detailed in the next sections.

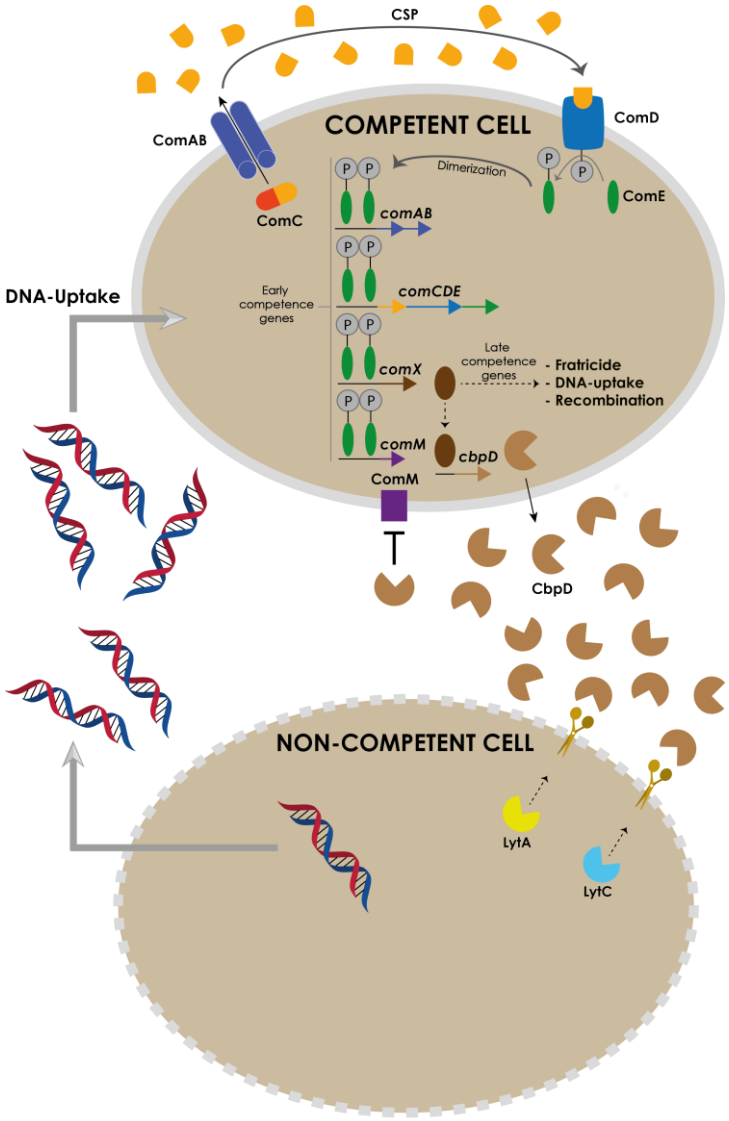


Figure 1.2 – Pneumococcal quorum-sensing-like system of competence regulation for natural genetic transformation. The products of the operons *comAB* and *comCDE* are the main constituents of the two-component signaling system regulating competence in *S. pneumoniae*. Secretion and processing of the product of *comC* by the ABC transporter ComAB results in the accumulation of the mature CSP in the medium until it reaches a critical concentration to trigger competence. This peptide is sensed by the transmembrane histidine kinase ComD, which autophosphorylates upon CSP binding, transferring the phosphoryl group to the response regulator ComE. Then this protein becomes activated and, after dimerization, induces the expression of the early competence genes, including *comAB*, *comCDE*, *comX* and *comM*, resulting in an amplification of the production of CSP and driving the cells into the competent state. ComX is an alternative sigma factor that upon association with the RNA polymerase activates the expression of the late competence genes involved in fratricide, DNA-uptake and recombination of homologous DNA. ComM protein provides immunity against CbpD induced lysis. Thus, CbpD, a secreted murein hydrolase, induces the lysis of non-competent pneumococci by enhancing the activity of the peptidoglycan hydrolases LytA and LytC in the process termed fratricide. The DNA released from the lysed bacteria can be taken up by competent cells and integrated into their genomes.

1.5.2. Pherotypes determined by competence-stimulating peptide (CSP)

In the 1960s it was discovered that competence was induced in a coordinated way when the cultures reached a defined cell density and that the sterile filtered medium of competent cultures contained a proteinaceous compound that induced competence in non-competent cultures (Johnsborg and Havarstein, 2009). However, this protein was purified and identified only in 1995 and was found to be a small unaltered peptide consisting of 17 amino acid residues, which was named CSP (Havarstein *et al.*, 1995). The amino acid sequence of this peptide was used to identify the gene that produces it. This gene, termed *comC*, encodes a precursor of CSP with 41 amino acids, of which 24 belong to the N-terminal leader peptide (**Figure 1.3**). At the processing site this peptide contains two glycine residues which are characteristic of double-glycine leader peptides. Several alleles of the *comC* gene have been identified producing at least 3 different mature CSPs (Pozzi *et al.*, 1996, Ramirez *et al.*, 1997, Whatmore *et al.*, 1999). However, most strains present only one of the two main alleles: CSP1 (≈70 % of the strains) or CSP2 (≈30 % of the strains) (Carrolo *et al.*, 2009). The third variant of CSP is not very well studied because of its rarity and its nomenclature varies among the studies that report it (Ramirez *et al.*, 1997, Whatmore *et al.*, 1999). Here in this thesis, this third variant will be called CSP3. Each strain has only one copy of a *comC* allele and responds specifically to the CSP produced due to the specificity shown by the sensor domain of ComD (see next section). Consequently, pneumococci can be divided into subpopulations regarding the CSP they respond to, defining the pherotype of an isolate (usually pherotype CSP1 or CSP2). It was termed pherotype because CSP behaves like a pheromone.

Prior to CSP identification, two genes (*comA* and *comB*, **Figure 1.2**) were identified which were essential for the presence of this peptide in the supernatant of *S. pneumoniae* cultures (Chandler and Morrison, 1988). ComA was later found to belong to a family of ABC (ATP-binding cassette) transporters specific for double-glycine leader peptides (Claverys and Havarstein, 2002, Hui and Morrison, 1991, Hui *et al.*, 1995) and which, together with ComB, specifically exported CSP. The CSP precursor is cleaved after the two glycine residues and its continuous export leads to its accumulation in the medium (**Figure 1.2**).

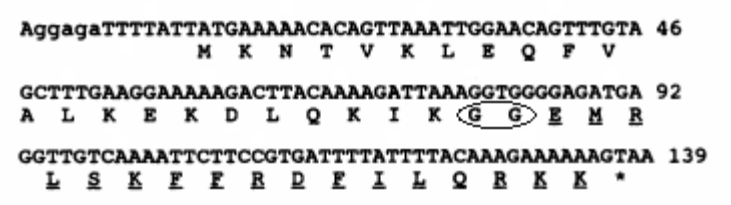


Figure 1.3 – *comC1* gene sequence and respective encoded amino acid residues. Residues that form CSP1 are underlined and the processing site formed by two glycine residues is pointed out. Adapted from Håvarstein *et al.*, 1995 (Havarstein *et al.*, 1995).

1.5.3. ComD and ComE, a histidine kinase receptor and a response regulator protein

The *comC* gene belongs to an operon containing two other genes, *comD* and *comE* (**Figure 1.2**). These encode two proteins, ComD and ComE, which form a two-component sensor system that monitors the external concentration of CSP and controls the expression of specific genes. ComD is a histidine kinase and has been identified as the CSP-specific receptor (Havarstein *et al.*, 1996). Like most histidine kinases, ComD has a sensor domain at the N-terminus associated with the membrane and a cytoplasmic C-terminal kinase domain containing the histidine residue that is phosphorylated. This protein has some diversity in its binding domain (Havarstein *et al.*, 1997) which, together with the variation of a hydrophobic domain in each CSP variant (Johnsborg *et al.*, 2006), is thought to be responsible for the specific interaction between these two proteins. Thus, each strain synthesizes a receptor specific for the CSP variant it produces, thereby defining different pherotypes.

The binding of CSP to ComD is believed to induce a conformational change in its transmembrane domain which activates its cytoplasmic kinase domain. This activation leads to the phosphorylation of the respective response regulator, the ComE protein. In its phosphorylated state this protein induces the transcription of several genes by binding to a specific sequence in their promoters (Martin *et al.*, 2010, Ween *et al.*, 1999). Among these genes are the *comCDE* and *comAB* operons and the *comX* gene. With the induction of the *comCDE* and *comAB* operons, extracellular CSP and intracellular levels of phosphorylated ComE rapidly increase in an autocatalytic process allowing the coordinated induction of competence across the population. However, genes encoding proteins required for genetic transformation, such as *dprA*, *sssB* and *recA* (Claverys *et al.*, 2009), are not directly

induced by ComE, and the *comX* gene has been shown to be responsible for this induction (Lee and Morrison, 1999). Thus, we can divide the genes of the cascade of competence into two classes, early (*comCDE*, *comAB* and *comX*) and late (*dprA*, *sssB* and *recA*) (**Figure 1.2**), whose difference of expression time was confirmed by DNA microarray studies (Peterson *et al.*, 2004).

1.5.4. ComX, the alternative sigma factor

Genes belonging to the late class, which include all those encoding the proteins required for the acquisition and recombination of exogenous DNA, share in the -10 region a consensus sequence (TACGAATA) named com-box and a thymine-rich site in the -25 region (Campbell *et al.*, 1998). The com-box is not present in *ComE*-activated genes and is different from the standard Pribnow box, suggesting that an alternative sigma factor is responsible for the regulation of the genes that contain this motif (Claverys and Havarstein, 2002). In fact, it was demonstrated that the product of the *comX* gene was the regulating factor of these genes. ComX is a sigma factor (σ^X) that associates with the RNA polymerase during the period of competence (**Figure 1.2**) (Lee and Morrison, 1999, Luo *et al.*, 2003). It has also been found that there are two copies of the *comX* gene (designated *comX1* and *comX2*) in the *S. pneumoniae* genome and that both are functional.

Although ComX is the only link between early competence genes and the genes required for genetic transformation, it was later found that another factor was necessary for the efficient development of competence and accumulation of ComX. The expression of this factor was induced by ComE and, therefore, it belongs to the early class (Luo *et al.*, 2003). This competence-positive regulator is encoded by the *comW* gene and acts post-transcriptionally on ComX (Luo *et al.*, 2004). The ClpEP complex is the major protease responsible for the degradation of ComX when competence declines and ComW was shown to protect σ^X from proteolysis (Piotrowski *et al.*, 2009, Sung and Morrison, 2005). Thus, the presence of ComW is essential for *S. pneumoniae* to develop and maintain the state of competence. Furthermore, regulation of ComX by ComW ensures that competence is not induced at a wrong time due to basal transcription of *comX* when the RNA polymerase overcomes the terminator of transcription of the genes upstream of *comX* or when there are mutations affecting the expression or function of ComX (Sung and Morrison, 2005).

1.5.5. Fratricide

The number of genes induced by CSP is high, with 105 genes being identified in one study (Dagkessamanskaia *et al.*, 2004) and 124 in another study (Peterson *et al.*, 2004). However, only 7 early and 14 late class genes are required for the genetic transformation process to occur (Eldholm *et al.*, 2009), indicating that CSP can regulate other physiological processes. In fact, it was discovered a molecular mechanism induced during competence that causes lysis of non-competent cells with the consequent active release of DNA into the medium (Moscoso and Claverys, 2004, Steinmoen *et al.*, 2002, Steinmoen *et al.*, 2003). As this mechanism kills cells genetically similar to the attacking cells it was termed fratricide (murder of brother or sister) (**Figure 1.2**) and it was shown that cell lysis occurs by contact between cells and not by diffusion in the medium of the lytic factors (Steinmoen *et al.*, 2003).

Proteins possessing choline binding domains (CBDs) have the ability to bind non-covalently to the phosphorylcholine residues present in the teichoic and lipoteichoic acids on the cell surface of *S. pneumoniae* (Kausmally *et al.*, 2005). Of the various choline-binding proteins encoded by *S. pneumoniae*, three of them have been described as important for competence-induced lysis and all are murein hydrolases: LytA, LytC and CbpD (Eldholm *et al.*, 2009, Guiral *et al.*, 2005, Kausmally *et al.*, 2005, Moscoso and Claverys, 2004, Steinmoen *et al.*, 2002).

The *lytA* gene encodes the major autolysin of *S. pneumoniae* and the *lytC* gene synthesizes the LytC lysozyme. The *lytA* gene, as the *lytC,* is constitutively transcribed but, unlike *lytC*, during competence its expression increases because it belongs to the CSP-induced *cinA-recA-dinF* operon (Mortier-Barriere *et al.*, 1998, Rimini *et al.*, 2000). The CbpD protein is only produced during competence and its gene belongs to the late class (Kausmally *et al.*, 2005, Peterson *et al.*, 2004).

The action of CbpD is essential in fratricide because in strains without this protein the lysis of non-competent cells is absent (Eldholm *et al.*, 2009, Kausmally *et al.*, 2005). However, strains with a double mutation in the *lytA* and *lytC* genes are practically unable to lyse cells during the competence, i.e., CbpD alone does not have sufficiently effective hydrolytic activity (Eldholm *et al.*, 2009, Kausmally *et al.*, 2005, Moscoso and Claverys, 2004) and the presence of LytA or LytC significantly increases cell lysis. Thus, when CbpD is present, LytA and LytC assume an important role in the lytic process (Eldholm *et al.*, 2009).

CbpD is only produced by the competent cells but LytA and LytC are synthesized by both the attacking and the target cells. It was discovered that the proteins produced by the non-competent cells resulted in greater lysis than the same proteins produced by the attacking cells (Eldholm *et al.*, 2009). LytC can be found in the supernatant medium while LytA is not detected in the medium, therefore LytC is secreted and LytA is an intracellular protein thought to be associated with the membrane (Eldholm *et al.*, 2009). Cleavage of the peptidoglycan by CbpD facilitates the hydrolysis of these chains by LytC (Perez-Dorado *et al.*, 2010). Thus, during the fratricide, the attacking cells through the production of CbpD are able to activate LytC and, after membrane degradation, LytA is released from the non-competent cells and consequently also becomes active. The more cells are lysed, the greater the concentration of active LytC and LytA, and this amplification effect allows fratricide to be efficient at a reduced cost to the attacking cells.

Attacking cells must have a mechanism of protection against their own lysins. The protein conferring protection to competent cells has already been identified and is named ComM (Havarstein *et al.*, 2006). Strains that do not produce ComM are susceptible to the action of CbpD, LytA and LytC even though they are in the competent state. The *comM* gene belongs to the early class, being expressed before the *cbpD* gene, and its product is a transmembrane protein. Although the competent cell immunity factor is already known, the mechanism through which it provides protection has not yet been elucidated.

I. Introduction

1. Background

2. Aims of the thesis

CHAPTER 2. AIMS OF THE THESIS

The aim of this thesis was to strengthen the current knowledge on horizontal gene transfer and its importance for the evolution of *S. pneumoniae,* focusing in the pherotypes variants that exist in this population. The specific goals of each study presented in this thesis were the following:

Part II – Pherotype diversity and abundance: the general aim of this part was to characterize the isolates recovered in IPD in Portugal regarding their pherotype diversity and abundance.

- **Chapter 4 – Diversity of the *comCDE* locus:** the aim of this study was to evaluate the genetic diversity of the pherotype-defining region *comCDE* of a sample of isolates recovered from IPD in Portugal during 1999-2001, a period before introduction of PCV7. Thus, we wanted to report the diversity present in Portuguese isolates and compare our results with other studies of isolates recovered in other geographical locations. Another aim was to detect possible new variants of CSP.
- **Chapter 5 – Pherotype abundance:** the proportion of pherotypes is well described in the literature and CSP1 is always the dominant pherotype over CSP2 in a proportion close to 70:30. A few isolates were reported to present a different pherotype, here named CSP3, suggesting that this pherotype is very uncommon. It was previously noted that there was an association between serotype and pherotype. From this scenario, the main goal of this study was to evaluate the evolution of pherotype abundance over a 14 years period where conjugate vaccines were introduced and dramatic changes occurred in serotype distribution. Another aim was to reliably identify the abundance of CSP3 in natural populations of *S. pneumoniae* by using a large collection (n=903).

Part III – Pherotype specificity and influence on the genetic structure: the aim of the work performed in this part was to describe the structural determinants of the specificity of pherotypes and if there are any consequences of this specificity affecting the evolution of *S. pneumoniae* by influencing recombination among pneumococci.

- **Chapter 6 – Structural study of pherotype specificity by X-ray crystallography:** our strategy to determine the structural determinants of pherotype specificity was to study the 3D structure of the ComD sensor by X-ray crystallography and, ultimately, to crystallize it in complex with CSP. Specific interactions between ComD and CSP of each pherotype could be observed in detail using this technique. To perform this experiment, I visited Dr. Juan Hermoso lab in Instituto Química Física Rocasolano, CSIC, Madrid. His team has an extensive experience in studying protein 3D structures by X-ray crystallography.
- **Chapter 7 – Serogroup 6 and influence of pherotype in genetic recombination:** phenotypic differences among pherotypes were observed regarding their capacity to form biofilms and their recombination efficiency (Carrolo *et al.*, 2014). Thus, the main goal of this work was to evaluate possible effects of pherotype diversity on the evolution of *S. pneumoniae*. We wanted to elucidate if pherotype has any influence on the recombination between isolates that share or not the same pherotype. To achieve this goal, we focused on serogroup 6, whose serotypes present different pherotypes, to see if recombination was restricted or not between isolates. Study of serogroup 6 also had the aim to evaluate their clonal and serotype evolution during 14 years (1999-2012) and to check if newly identified serogroup 6 serotypes were present in our pneumococcal collection.

I. Introduction

1. Background

2. Aims of the thesis

3. Pneumococcal isolates and typing methods

CHAPTER 3. PNEUMOCOCCAL ISOLATES AND TYPING METHODS

This chapter describes the general procedures that are common to all studies presented in this thesis.

3.1. Pneumococcal isolates

Since 1999 the isolates of *S. pneumoniae* responsible for infections in Portugal have been monitored by the Portuguese Group for the Study of Streptococcal Infections. This surveillance is carried out by more than 30 hospital laboratories located in all regions of Portugal, including the Autonomous Regions of Azores and Madeira. These laboratories collect isolates from patients with pneumococcal infections which are then sent to the Institute of Microbiology of the Faculty of Medicine of the University of Lisbon.

The largest collection studied in this study was composed of isolates that caused IPD in children (<18 years) during the years 1999 and 2012 in Portugal. Although the laboratories were contacted periodically, no audit was carried out to ascertain whether all the collected isolates were sent. In cases where there was more than one invasive isolate from the same patient only one was included in the study, giving preference to isolates collected from the CSF.

3.2. Storing, identification and growing of isolates

The isolates, on arrival at the Institute of Microbiology, were inoculated onto blood agar plates (Tryptone Soy Agar, Oxoid, Hampshire, England, supplemented with 5 % (v/v) sheep blood, Probiológica, Belas, Portugal) and incubated overnight at 35 °C with an atmosphere enriched with 5 % CO_2. Laboratory identification of *S. pneumoniae* consisted of an analysis of the colony morphology, hemolysis type after growth in solid medium enriched with blood, susceptibility to optochin and solubility in bile salts.

Colonies of *S. pneumoniae* are round, non-pigmented, and exhibit a characteristic central depression of this species resulting from self-induced cell lysis. Some isolates have a mucous aspect which is due to the increased production

of the polysaccharide capsule. In blood-containing media, *S. pneumoniae* colonies are α-hemolytic with a green pigmentation around the growth due to partial erythrocyte degradation and consequent haemoglobin degradation. However, other viridans streptococci are also α-hemolytic, making it difficult to distinguish from *S. pneumoniae*. In the laboratory, this distinction is made by the susceptibility test to optochin, in which viridans streptococci are resistant and *S. pneumoniae* is susceptible.

The optochin susceptibility test was performed by placing a filter paper disc with 5 µg of optochin (Oxoid, Hampshire, England) on the blood agar plate. Inhibition halos were measured after incubation. The isolates were identified as sensitive when their halo was equal to or greater than 14 mm. Isolates showing halos less than 14 mm or that did not even have inhibition halos were tested for solubility in bile salts because isolates of *S. pneumoniae* resistant to optochin have already been described (Aguiar *et al.*, 2006). The bile solubility test was done by suspending the bacterial cultures in 0.5 mL of 0.85 % (w/v) NaCl with an approximate turbidity of McFarland 1 and by adding an equal volume of 2 % (w/v) sodium deoxycholate (DOC), which is a salt present in bile, or 0.85 % (w/v) NaCl as control. The isolates were incubated at 35 °C for 15 min and the turbidity of the DOC suspension was compared to the turbidity of the control. *S. pneumoniae* cells lyse in the presence of bile salts due to the activation of the autolysin LytA, while the cells of other species remain intact. Thus, if the DOC suspension was clear or substantially less turbid than the control it was indicative of *S. pneumoniae*. In the case where no differences were detected, the suspensions were further incubated at 35 °C for up to 2 h. If lysis did not occur after this incubation time, the isolate was not considered *S. pneumoniae*.

After confirmation of the species, the isolates were stored at -70 °C in Tryptone Soy Broth medium (Oxoid, Hampshire, England) with 15 % (v/v) glycerol. Every time the isolates were thawed they were plated onto blood agar plates and incubated overnight at 35 °C with an atmosphere enriched with 5 % CO_2 for further analysis.

3.3. Serotyping

The serotype of the isolates was determined by the capsular reaction test (or Quellung reaction) in which the capsular polysaccharides react with specific antibodies. When there is reaction, agglutination is either observed macroscopically, upon observation under the optical microscope (the bacteria are aggregated) or by the capsule becoming visible because its refractive index is altered by the immunoprecipitation with specific serum (Quellung reaction) (Sorensen, 1993). This test was performed with a kit of sera (Statens Serum Institut, Copenhagen, Denmark) that is divided into pools (identifying several serogroups or serotypes), sera that identify serogroups and sera that identify serotypes. Testing of each serum was performed on a microscopy slide by mixing about 1 µL of serum with about 3 µL of a suspension of the test strain. Agglutination was a sign of a positive reaction. Due to the high number of existing serotypes, a "chessboard" system (**Table 3.1**) was used to optimize serotyping (Sorensen, 1993).

Sera from the new pool set (serum P to T) containing the serotypes included in vaccines were first sequentially tested. Upon a positive reaction, sera from the existing pool set (sera A-F and H) were tested to determine the serotype or serogroup in common with the serum that tested positive. In the absence of agglutination with sera from P to T, sera containing the serotypes or serogroups not included in the vaccines (C-I sera) were tested. The identified serotype (or serogroup) was confirmed with the corresponding specific serum. When a serogroup was found, the serotype was determined by testing specific sera according to a table provided by Statens Serum Institut at http://www.ssi.dk.

Table 3.1 – Chessboard system used for serotyping *S. pneumoniae* isolates. Adapted from (Sorensen, 1993).

Pool	Serotype or serogroup with a new pool[c]					NVT serotypes or serogroups[a]
	P	Q	R	S	T	
A	**1**	**18***	**4**	**5**	**2**	
B	**19***	**6***	**3**	**8**		
C	**7***				**20**	24*, 31, 40
D			**9***		**11***	16*, 36, 37
E			**12***	**10***	**33***	21, 39
F				**17***	**22***	27, 32*, 41*
H	**14**	**23***		**15***		13, 28*
G[b]						29, 34, 35*, 42, 47*
I[b]						25*, 38, 43, 44, 45, 46, 48

[a]The serogroups are marked with an asterisk and the serotypes or serogroups included in vaccines are in bold.
[b]Sera G and I do not react with any vaccine serotype and therefore are not included in the chessboard system.

3.4. Multilocus Sequence Typing (MLST)

The internal fragments of seven conserved polymorphic genes were amplified by PCR and sequenced. Each isolate was inoculated into Brain Heart Infusion (BHI) (Becton Dickinson, Maryland, EUA) and incubated in a water bath at 37 °C until reaching an OD_{600} of at least 0.80. To prepare the DNA, 9 µL of culture was boiled in 441 µL of water for 2 min and then immediately placed on ice for 5 min. For each gene the following amplification mixture was prepared: 20 µL of boiled DNA; 0.02 U/µL GoTaq® DNA polymerase (Promega, Wisconsin, USA); 1x buffer suitable for the enzyme and supplied by the manufacturer; 2 mM magnesium chloride (Promega, Wisconsin, USA); 0.2 mM deoxyribonucleotide triphosphates (dNTPs) (Fermentas, Vilnius, Lithuania); 0.4 pmol/µL of each primer (up and dn) of the gene to be amplified and 10.8 µL of water purified by the Milli-Q system (Millipore, Massachusetts, USA), completing a final volume of 50 µL.

The sequences of the primers used, as well as the size of the generated fragment of each gene, are shown in **Table 3.2**. Each primer contains a tail upstream of the gene-specific sequence. This tail is a universal primer, in this case M13 phage, which makes it possible to sequence several genes in a single plate, allowing optimization of sequencing. The fragments were amplified in the thermocycler MyCycler® (Bio-Rad, California, USA). To amplify all fragments, except for the *recP* gene, the following program was used: denaturation at 95 °C (5 min); 4 cycles of denaturation at 95 °C (1 min), hybridization at 48 °C (45 s) and elongation at 72 °C (1 min); 29 cycles of denaturation at 95 °C (1 min), hybridization at 65 °C (45 s) and elongation at 72 °C (1 min); elongation at 72 °C (5 min). The primers used have a lower DNA binding specificity due to the presence of the tail, so the lower annealing temperature of the initial 4 cycles allowed hybridization with less specificity, facilitating the initiation of amplification of the target sequence. The *recP* gene fragments were amplified with the following program: denaturation at 95 °C (4 min); 30 cycles of denaturation at 95 °C (30 s), hybridization at 58 °C (30 s) and elongation at 72 °C (30 s); elongation at 72 °C (10 min). Amplification products were visualized on a 1 % (w/v) agarose gel in 0.5x TBE buffer with 0.5 µg/mL ethidium bromide and purified using the commercial kit High Pure PCR Product Purification Kit (Roche, Mannheim, Germany) according to the manufacturer's instructions. After purification, the amplification products were sequenced by GATC Biotech (Constance, Germany) using the

universal primer M13F (**Table 3.2**). When new alleles were identified and when there were doubts in the visual inspection of the chromatogram, the results were confirmed by sequencing the other strand with the universal primer M13R-pUC (**Table 3.2**).

The sequences obtained were analyzed through BioNumerics®, where an allele number was assigned to each of the genes and the sequence type (ST) to each allelic profile by comparison with the database available at https://pubmlst.org/spneumoniae/. The newly identified alleles and STs were submitted to the database curator for assignment of a new number. The analysis of the genetic relationship between the STs found was made through the software PHYLOViZ that uses the algorithm goeBURST (Francisco *et al.*, 2009). To perform clustering of the isolates into clonal complexes (CCs), the complete database (all STs currently identified) was used. This clustering in CCs was done at the single locus variant (SLV) level, i.e. all STs within a clonal complex are linked to at least one of their SLV. Otherwise, the conditions to define CCs in PHYLOViZ are described in the appropriate section if the method performed differs from this procedure.

Table 3.2 – Primers used in MLST.

Primer[a]	Sequence (5'→ 3')[b]	Size (bp)[c,d]	Fragment size (bp)[d]
*aroe*_up	M13F – CGT TTA GCT GCA GTT GTT GC	38	405
*aroe*_dn	M13R-pUC – CCC ACA CTG GTG GCA TTA AC	38	
*ddl*_up	M13F – TTG CCA TGG ATA AAA TCA CGA C	40	441
*ddl*_dn	M13R-pUC – CGC GCT TGT CAA AAC TTT CC	38	
*gdh*_up	M13F – GTG CTG AAA AGA TTA AGG TCT	39	460
*gdh*_dn	M13R-pUC – TGC TTC CAG CTT TAT AGT CAT C	40	
*gki*_up	M13F – GGC ATT GGA ATG GGA TCA CC	38	483
*gki*_dn	M13R-pUC – TCT CCC GCA GCT GAC AC	35	
*recP*_up	M13F – GCC AAC TCA GGT CAT CCA GG	38	450
*recP*_dn	M13R-pUC – GCT TCC AAG TCT GTT CCA TTT TC	41	
*spi*_up	M13F – CGC TTA GAA AGG TAA GTT ATG	39	474
*spi*_dn	M13R-pUC – AGG CTG AGA TTG GTG ATT CTC	39	
*xpt*_up	M13F – GGA GGT CTT ATG AAA TTA TTA G	40	486
*xpt*_dn	M13R-pUC – AGA TCT GCC TCC TTA AAT AC	38	
M13F	TGT AAA ACG ACG GCC AGT	18	
M13R-pUC	CAG GAA ACA GCT ATG ACC	18	

[a]The universal primers of phage M13 are shaded.

[b]All the up and dn genes have the sequence of M13F and M13R-pUC at the beginning, respectively.

[c]The size of the primers is the sum of the nucleotides of the tail with those of the specific sequence for the gene.

[d]bp - base pair.

3.5. Antimicrobial susceptibility

The choice of antimicrobials to be tested was based on their clinical use, their importance as resistance markers and the fact that it is possible to compare them with previous studies. Thus, susceptibility profiles to penicillin and cefotaxime were determined by Etest® (BioMérieux, Marcy l'Etoile, France) according to the manufacturer's instructions (Melo-Cristino *et al.*, 2003). The susceptibility profiles for erythromycin, clindamycin, tetracycline, levofloxacin, co-trimoxazole, chloramphenicol, vancomycin and linezolid were determined using the Kirby-Bauer method (Bauer *et al.*, 1966), each with a defined concentration, according to the standards of the Clinical and Laboratory Standards Institute (CLSI) (C.L.S.I., 2009). Briefly, to determine the susceptibility to the antimicrobials chosen, suspensions of the isolates were prepared in 0.85 % (w/v) NaCl with a turbidity of McFarland 0.5. This standardization of the inoculum is important for the reproducibility of the results. Thus, Mueller-Hinton plates (Oxoid, Hampshire, England) supplemented with 5 % (v/v) sheep blood (Probiológica, Belas, Portugal) were uniformly inoculated with these suspensions and discs or strips of Etest® of each antimicrobial were placed onto the plate. The plates were incubated at 35 °C with 5 % CO_2 for 24 h. After incubation, the inhibition halos were measured around the discs and MIC values were read in the Etest® strips. In order to classify the susceptibility of the isolates, the diameter of the halos and MIC values were interpreted according to CLSI (C.L.S.I., 2014). In 2008 a change was introduced in the CLSI breakpoint for resistance to penicillin, establishing different values for cases of meningitis and non-meningitis. To facilitate comparison with previous epidemiological studies, the 2007 CLSI indications (C.L.S.I., 2007) were used to define levels of susceptibility to penicillin. Thus, MIC values of >0.06 to <2 µg/mL defined intermediate resistance and values ≥2 µg/mL defined high-level resistance. Multidrug resistance (MDR) was defined by non-susceptibility to at least three classes of antibiotics.

To validate the results obtained, the reference strain *S. pneumoniae* ATTC 49619 was used as a control and susceptibility testing for this strain was regularly performed with the antimicrobials used in this study. The results were validated by making sure the values obtained fell within the ranges indicated in the CLSI guidelines (C.L.S.I., 2014).

II. Pherotype diversity and abundance

II. PHEROTYPE DIVERSITY AND ABUNDANCE

In this section we will describe the diversity of pherotypes found in invasive isolates from Portugal and also a report of their abundance. Pherotype diversity was determined by a genetic study of the *comCDE* locus of a collection of pneumococcal invasive isolates recovered before conjugate vaccines were introduced (Chapter 4), while pherotype abundance was evaluated by identifying the pherotype of a large collection of invasive isolates recovered from children over a 14 year period including conjugate vaccination (Chapter 5).

II. Pherotype diversity and abundance

4. Diversity of the *comCDE* locus

CHAPTER 4. DIVERSITY OF THE *COMCDE* LOCUS

This chapter presents a comprehensive study of the genetic diversity of the *comCDE* locus of an invasive pneumococcal sample recovered in Portugal. There are two major published studies about the genetic diversity of the *comC* gene (Ramirez *et al.*, 1997, Whatmore *et al.*, 1999) and one of them also included the *comD* gene (Whatmore *et al.*, 1999). However, in those studies the complete sequence of *comD* and also of the *comE* gene was not analyzed. Furthermore, in contrast to those studies, the sample studied in this work had more isolates and these were recovered in the same period and geographical site, allowing us to evaluate diversity in the same population.

The aim of this study was to sequence the *comCDE* locus of a collection of invasive pneumococcal isolates and identify the genetic profiles of this locus present in Portugal. Then, these profiles were compared with published *comCDE* sequences and related with other isolate characteristics, such as serotype and genetic lineage. This study also enabled the identification of potential new variants of CSP.

4.1. Materials and methods

4.1.1. Bacterial isolates

To evaluate the genetic diversity of the *comCDE* locus, the same invasive pneumococcal sample studied previously by our laboratory was chosen (Carrolo *et al.*, 2009). This sample was composed of isolates recovered from invasive infections from all age groups and isolated between 1999 and 2001, representing the diversity of genetic lineages circulating in Portugal in a period when the effects of conjugate vaccination had not been felt yet. The initially chosen sample included a total of 90 pneumococcal invasive isolates but one of them was not recovered when isolates were recovered from storage, making a total of 89 isolates available for this study. The serotype and MLST data of these isolates were described previously by our laboratory (Serrano *et al.*, 2005, Serrano *et al.*, 2004).

4.1.2. Sequencing of comCDE locus

An analysis of the public sequences of *comCDE* locus was required to identify conserved regions as targets for designing primers for sequencing. Thus, a BLAST search of the nucleotide database was performed in February 2013 using as query the sequence of *comD* gene of strain R6 (accession no. AE007317), restricting the results to the species *Streptococcus pneumoniae*. After downloading the resulting complete original[1] sequences, an analysis of the *comCDE* locus was restricted only to complete sequences of this operon, resulting in a total of 31 *comCDE* sequences (**Table 4.1**). These sequences were aligned with the exception of the isolates gamPNI0373 and AP200 because they presented a deletion in the gene *comE* and an insertion in the gene *comC*, respectively. Then, the primers *comC*-fw, *comD*-fw, *comD*-rv and *comE*-rv (**Table 4.2**) were designed in conserved regions in the alignment (**Figure 4.1**) as universal primers to be used in our screening.

While all primers were used to sequence the *comCDE* locus, only the primer pair *comC*-fw and *comE*-rv were used in the PCR reaction to yield the product of interest in a first step. However, some isolates did not yield a PCR product with sufficient quality for sequencing, making it necessary the use of the inner primers

[1]The complete original sequences were the entire sequence of the matching accession number, for example, it was entire genomes in the case where the match was from a complete or draft genome.

(**Figure 4.1**). The DNA of the isolates was extracted using CTAB (Wilson, 2001) because extraction from boiled cells did not yield good results. PCR products were sequenced by GATC Biotech (Constance, Germany) using the four primers indicated.

The PCR reaction (V=50 µl) for the amplification of *comCDE* locus was composed of:

- 0.05 U/µl GoTaq® DNA polymerase (Promega, Wisconsin, USA);
- 1x enzyme's buffer provided by manufacturer;
- 2 mM magnesium chloride;
- 0.4 mM dNTPs;
- 0.4 pmol/µl of each primer (*comC*-fw and *comE*-rv);
- 4 ng/µl of template DNA.

With the following PCR program:

- 5 min at 95 ºC;
- 30x (30 s at 95 ºC, 30 s at 50 ºC and 3 min at 72 ºC)
- 10 min at 72 ºC.

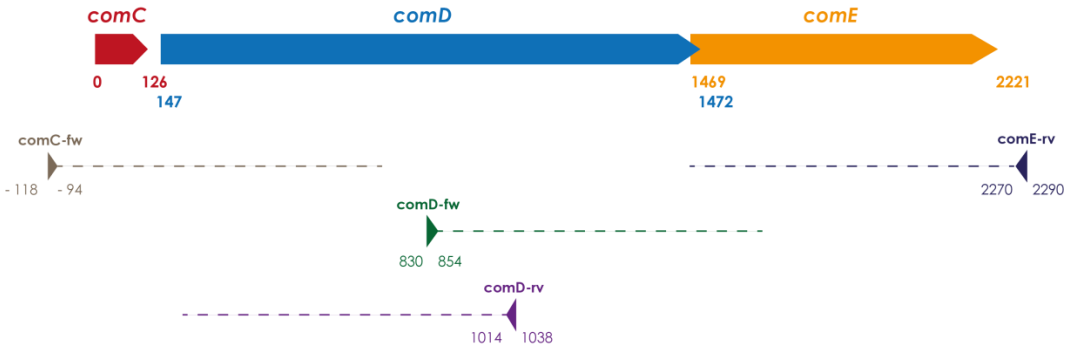


Figure 4.1 – Primers used for sequencing the *comCDE* locus. The first base of *comC* was used as reference to assign positions numbers. Small triangles represent primers and their direction. Dashed lines represent the expected size of the region sequenced from each primer (800 bp). Gene's length is at scale.

Table 4.1 – Strains with complete *comCDE* locus sequence in the public database and comparison with the alleles found in Portugal. The alleles of these strains were not included in the allele classification described in the section *4.2. Results*.

Accession no.	Strain	Pherotype	*comC*	*comD*	*comE*
AE005672	TIGR4	CSP2	*comC*-2.1	*comD*-2.8	*comE*-10
AE007317	R6	CSP1	*comC*-1.1	*comD*-1.1	*comE*-15
AF067659	A66	CSP2	*comC*-2.1	*comD*-2.1	*comE*-5
CM001835	PCS8235	CSP1	*comC*-1.1	*comD*-1.1	*comE*-2
CP000410	D39	CSP1	*comC*-1.1	*comD*-1.1	*comE*-15
CP000918	70585	CSP1	*comC*-1.2	*comD*-1.5	No match[b]
CP000919	JJA	CSP2	*comC*-2.2	No match[b]	*comE*-7
CP000920	P1031	CSP1	*comC*-1.1	No match[b]	*comE*-15
CP000921	Taiwan19F-14	CSP1	*comC*-1.1	*comD*-1.1	*comE*-2
CP000936	Hungary19A-6	CSP1	*comC*-1.1	*comD*-1.1	No match[b]
CP001015	G54	CSP1	*comC*-1.1	No match[b]	*comE*-1
CP001033	CGSP14	CSP1	*comC*-1.1	*comD*-1.1	*comE*-2
CP001845	gamPNI0373	CSP1	*comC*-1.1	No match[b]	*Deletion of 555 bp* No match[b]
CP001993	TCH8431/19A	CSP1	No match[b]	*comD*-1.1	*comE*-2
CP002121	AP200	ND[a]	*Insertion of 1248 bp* No match[b]	No match[b]	*comE*-1
CP002176	670-6B	CSP1	*comC*-1.1	*comD*-1.1	*comE*-4
CP003357	ST556	CSP1	*comC*-1.1	*comD*-1.1	*comE*-2
CP006844	A026	CSP1	No match[b]	*comD*-1.1	*comE*-2
FM211187	ATCC 700669	CSP2	*comC*-2.2	*comD*-2.2	*comE*-7
FQ312027	OXC141	CSP1	*comC*-1.2	*comD*-1.3	*comE*-8
FQ312029	INV200	CSP1	*comC*-1.1	*comD*-1.1	*comE*-2
FQ312030	INV104	CSP2	*comC*-2.1	*Deletion of 9 bp* No match[b]	*comE*-10
FQ312039	SPN032672	CSP1	*comC*-1.1	*comD*-1.1	*comE*-2
FQ312041	SPN994038	CSP1	*comC*-1.2	*comD*-1.5	*comE*-8
FQ312042	SPN033038	CSP1	*comC*-1.1	*comD*-1.1	*comE*-2
FQ312043	SPN034183	CSP1	*comC*-1.2	*comD*-1.5	*comE*-8
FQ312044	SPN994039	CSP1	*comC*-1.2	*comD*-1.5	*comE*-8
FQ312045	SPN034156	CSP1	*comC*-1.2	No match[b]	*comE*-8
HE983624	SPNA45	CSP1	*comC*-1.2	No match[b]	No match[b]
U33315	CP1200	CSP1	*comC*-1.1	*comD*-1.1	*comE*-15
U76218	R6x	CSP1	*comC*-1.1	No match[b]	No match[b]

[a]ND – Not determined because an insertion was found in *comC* of the strain AP200.
[b]No match with the alleles found in this study.

Table 4.2 – Sequence of the primers used for sequencing the *comCDE* locus.

Primer	Sequence (5'→ 3')	Size (bp)
comC-fw	GTACACTTTGGGAGAAAAAAATGA	24
comD-fw	TTTACAGAATTACACAGATGAAATT	25
comD-rv	CTATGTTGTTCAAATCAAAGTAAGT	25
comE-rv	CCCCTTGACCAACGGACCTTC	21

4.1.3. Analysis of comCDE locus

The analysis of the sequences was performed in Geneious Pro (version 5.3.6). For each isolate, the *comCDE* sequence was obtained by assembling the reads from the four sequencing primers using as reference the *comCDE* locus of R6 strain. Then, the genes *comC*, *comD* and *comE* were identified and translated to amino acid sequence. DNA and protein sequences were aligned using MUSCLE algorithm with default settings and then alleles and *comCDE* profiles were determined.

Alignments of the alleles were performed using MUSCLE algorithm with default settings in Geneious and were edited in BioEdit (version 7.2.5). The matrix BLOSUM62 was used to define similar amino acids. The alignments were used to build phylogenetic trees by neighbor-joining method and resampling was done by bootstrap for 1000 times.

4.2. Results

4.2.1. Pneumococcal invasive isolates and allele classification

The sequence of the *comCDE* locus was obtained for n=89 invasive isolates recovered in 1999 (n=26), 2000 (n=33) and 2001 (n=30) from all age groups (<18 years, n=21; ≥18 years, n=68). These isolates presented a total of 27 serotypes and 57 STs (**Figure 4.2**) and were recovered from blood (n=76), CSF (n=12) and pleural fluid (n=1). The alleles of *comC*, *comD* and *comE* of these isolates were identified, translated into protein and classified. The alleles of the strains with complete *comCDE* locus sequence in the public database were not taken into account in this classification. However, a comparison was performed with the alleles found in this study (**Table 4.1**).

The following criteria were used for allele classification. In the case of the genes *comC* and *comD*, numbers were attributed to them including the corresponding pherotype followed by a number identifying their frequency rank in the current study. For example, the *comD*-1.3 and *comD*-2.2 alleles presented pherotype CSP1 and CSP2 alleles, respectively, and were the third and second most frequent allele of their pherotype, respectively. However, for the *comE* gene, classification was done only considering their frequency because a correlation with pherotype was not obvious. Protein alleles inherited the name of their respective DNA allele and when more than one allele yielded the same protein sequence, the name of the most frequent was chosen.

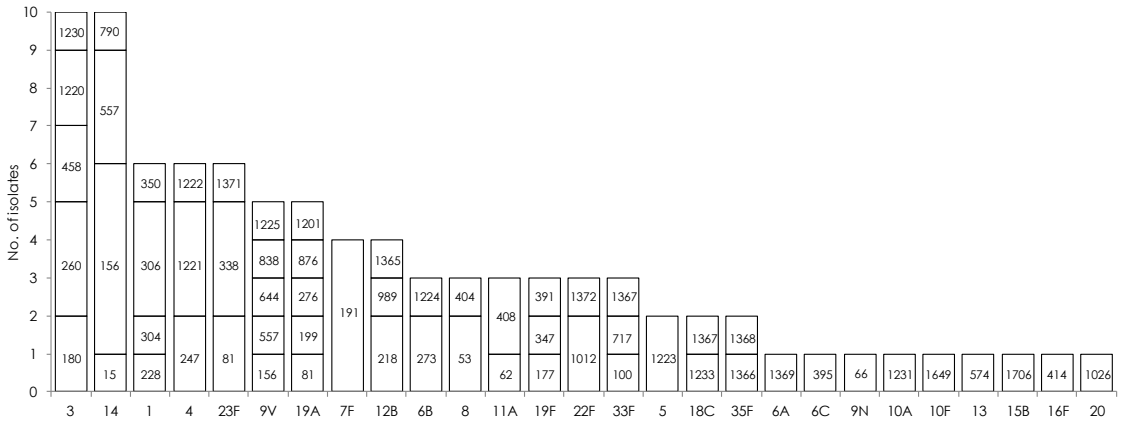


Figure 4.2 – Serotype and ST of the invasive isolates used to study the diversity of *comCDE* locus.

4.2.2. Gene comC

A total of 6 alleles of the gene *comC* were identified among the isolates (**Table 4.3** and **Figure 4.3**). These alleles could be divided into three groups representing pherotypes CSP1, CSP2 and CSP3. Pherotype CSP1 (n=66, 74.2 %) was the most abundant among the isolates, followed by CSP2 (n=21, 23.6 %) and CSP3 (n=2, 2.2 %). Translation of these alleles yielded a total of 5 variants of the precursor of CSP (**Figure 4.4**) because the *comC*-1.2 allele presented the silent mutation T82C (position 97 in **Figure 4.3**) in relation to *comC*-1.1, producing the same protein variant. However, looking only at the mature exported CSPs, which are represented by the sequence after the double glycine cleavage site, only 3 distinct protein sequences were identified, each representing a distinct pherotype. Thus, one of each of the ComC variants of the pherotypes CSP1 and CSP2 presented an amino acid substitution in the signal peptide, while the mature CSP was conserved.

Regarding the *comC*-3 allele, the main feature observed was its higher genetic diversity and larger size in comparison with *comC*-1 and *comC*-2 alleles (141 bp vs. 126 bp, respectively). Translation of *comC*-3 allele showed a precursor of CSP3 with a signal peptide of the same size, albeit with one amino acid alteration relative to the other signal peptides, but the mature CSP3 presented 22 amino acids while CSP1 and CSP2 mature peptides were composed of 17 amino acids. As far as we know, it was the first time the *comC*-3 allele and therefore the CSP3 pherotype were identified in pneumococcal isolates recovered in Portugal.

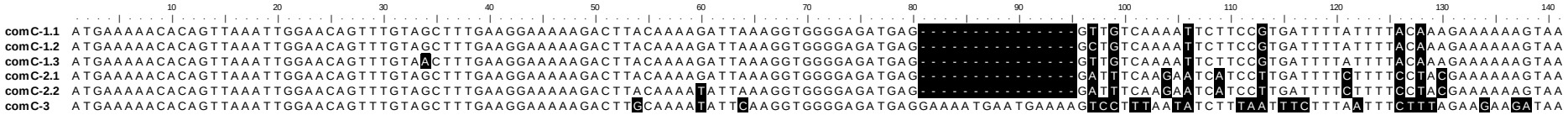


Figure 4.3 – DNA sequence of *comC* alleles. Black boxes highlight SNPs.

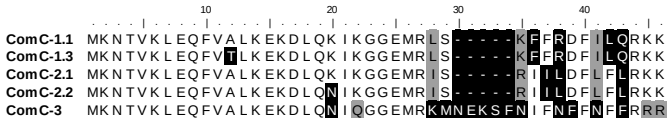


Figure 4.4 – Protein sequence of CSP precursors. Black boxes highlight SNPs and grey boxes highlight similar amino acids.

Table 4.3 – Alleles of the genes *comC*, *comD* and *comE* and respective serotypes and STs.

Allele	n	Pherotype	Serotype	ST
comC-1.1	56	CSP1	14 (n=10); 1 (n=6); 9V (n=5); 7F, 23F (n=4 each); 4, 11A, 12B, 33F (n=3 each); 6B, 8, 18C, 22F (n=2 each); 6A, 13, 15B, 16F, 19F, 20, 35F (n=1 each)	156 (n=6); 191, 557 (n=4 each); 306, 338, 1221 (n=3 each); 53, 218, 408, 1012, 1367 (n=2 each); 15, 62, 100, 228, 273, 304, 347, 350, 414, 574, 644, 717, 790, 838, 1026, 1224, 1225, 1233, 1365, 1368, 1369, 1371, 1706 (n=1 each)
comC-1.2	9	CSP1	3 (n=5); 5, 19A (n=2 each)	180, 458, 1223 (n=2 each); 199, 876, 1230 (n=1 each)
comC-1.3	1	CSP1	22F (n=1)	1372 (n=1)
comC-2.1	16	CSP2	3 (n=5); 4 (n=3); 19F (n=2); 6B, 6C, 8, 10A, 10F, 12B (n=1 each)	260 (n=3); 247, 1220 (n=2 each); 177, 273, 391, 395, 404, 989, 1222, 1231, 1649 (n=1 each)
comC-2.2	5	CSP2	19A, 23F (n=2 each); 9N (n=1)	81 (n=3); 66, 276 (n=1 each)
comC-3	2	CSP3	19A, 35F (n=1 each)	1201, 1366 (n=1 each)
comD-1.1	51	CSP1	14 (n=10); 1 (n=6); 9V (n=5); 23F (n=4); 4, 11A, 12B, 22F, 33F (n=3 each); 6B, 18C (n=2 each); 6A, 8, 13, 15B, 16F, 20, 35F (n=1 each)	156 (n=6); 557 (n=4); 306, 338, 1221 (n=3 each); 218, 408, 1012, 1367 (n=2 each); 15, 53, 62, 100, 228, 273, 304, 350, 414, 574, 644, 717, 790, 838, 1026, 1224, 1225, 1233, 1365, 1368, 1369, 1371, 1372, 1706 (n=1 each)
comD-1.2	4	CSP1	7F (n=4)	191 (n=4)
comD-1.3	3	CSP1	3 (n=3)	180 (n=2); 1230 (n=1)
comD-1.4	2	CSP1	19A (n=2)	199, 876 (n=1 each)
comD-1.5	2	CSP1	5 (n=2)	1233 (n=2)
comD-1.6	1	CSP1	19F (n=1)	347 (n=1)
comD-1.7	1	CSP1	8 (n=1)	53 (n=1)
comD-1.8	2	CSP1	3 (n=2)	458 (n=2)
comD-2.1	5	CSP2	3 (n=4); 4 (n=1)	260, 1220 (n=2 each); 1222 (n=1)
comD-2.2	4	CSP2	23F (n=2); 9N, 19A (n=1 each)	81 (n=3); 66 (n=1)
comD-2.3	3	CSP2	19F (n=2); 10F (n=1)	177, 391, 1649 (n=1 each)
comD-2.4	1	CSP2	8 (n=1)	404 (n=1)
comD-2.5	1	CSP2	6C (n=1)	395 (n=1)
comD-2.6	1	CSP2	12B (n=1)	989 (n=1)
comD-2.7	1	CSP2	6B (n=1)	273 (n=1)
comD-2.8	1	CSP2	10A (n=1)	1231 (n=1)
comD-2.9	1	CSP2	19A (n=1)	276 (n=1)
comD-2.10	2	CSP2	4 (n=2)	247 (n=2)
comD-2.11	1	CSP2	3 (n=1)	260 (n=1)
comD-3.1	1	CSP3	19A (n=1)	1201 (n=1)
comD-3.2	1	CSP3	35F (n=1)	1366 (n=1)
comE-1	18	CSP1, CSP2	11A, 33F (n=3 each); 8, 18C, 19F, 22F (n=2 each); 6C, 10F, 13, 20 (n=1 each)	53, 408, 1012, 1367 (n=2 each); 62, 100, 177, 391, 395, 574, 717, 1026, 1233, 1649 (n=1 each)
comE-2	16	CSP1	1 (n=6); 7F (n=4); 4 (n=3); 14, 16F, 19F (n=1 each)	191 (n=4); 306, 1221 (n=3 each); 15, 228, 304, 347, 350, 414 (n=1 each)
comE-3	15	CSP1	14 (n=9); 9V (n=5); 12B (n=1)	156 (n=6); 557 (n=4); 218, 644, 790, 838, 1225 (n=1 each)
comE-4	7	CSP1, CSP2	5, 6B (n=2 each); 4, 6A, 12B (n=1 each)	1223 (n=2); 273, 989, 1222, 1224, 1369 (n=1 each)
comE-5	5	CSP2	3 (n=5)	260 (n=3); 1220 (n=2)
comE-6	5	CSP1	23F (n=4); 15B (n=1)	338 (n=3); 1371, 1706 (n=1 each)
comE-7	4	CSP2	19A, 23F (n=2 each)	81 (n=3); 276 (n=1)
comE-8	3	CSP1	3 (n=3)	180 (n=2); 1230 (n=1)
comE-9	3	CSP1, CSP2	19A (n=2); 8 (n=1)	199, 404, 876 (n=1 each)
comE-10	3	CSP2	4 (n=2); 6B (n=1)	247 (n=2); 273 (n=1)
comE-11	2	CSP1	12B (n=2)	218, 1365 (n=1 each)
comE-12	2	CSP1	3 (n=2)	458 (n=2)
comE-13	2	CSP3	19A, 35F (n=1 each)	1201, 1366 (n=1 each)
comE-14	1	CSP2	9N (n=1)	66 (n=1)
comE-15	1	CSP1	22F (n=1)	1372 (n=1)
Insertion of 1274 bp	1	CSP1	35F (n=1)	1368 (n=1)
Insertion of 1246 bp	1	CSP2	10A (n=1)	1231 (n=1)

4.2.3. Gene comD

The genetic diversity observed in the gene *comD* was higher with a total of 21 alleles identified, resulting in 17 ComD variants (**Table 4.3, Figure 4.5 and 4.6**). The majority of the genetic diversity was observed in the N-terminus corresponding to the sensor domain and, similarly to the gene *comC*, *comD* alleles could also be

divided into three major groups, each representing a pherotype: *comD*-1, *comD*-2 and *comD*-3. This relation with pherotype was still observed when alleles were translated into protein (**Figure 4.7**). The division into three groups is due to the sensor domain because the kinase domain located in the C-terminus was more conserved and an association with pherotype could not be observed for this domain.

A total of 8 alleles were identified in *comD*-1 group and all produced different ComD1 proteins. The allele *comD*-1.8 presented a deletion of 24 bp which translated in a protein with a deletion of 8 amino acids (**Figure 4.5 and 4.6**). Analyzing the phylogenetic tree built with the sequences of ComD proteins, it could be observed that ComD1 variants clearly formed two subgroups: canonical (*ComD*-1.1, *ComD*-1.2, *ComD*-1.6 and *ComD*-1.7) and non-canonical (*ComD*-1.3, *ComD*-1.4, *ComD*-1.5 and *ComD*-1.8) (**Figure 4.7**).

Regarding *comD*-2 group, 11 alleles were observed resulting in 7 protein variants of ComD2. The *comD*-2.10 and *comD*-2.11 alleles presented an insertion and a deletion of 1 bp, respectively (**Figure 4.5**), resulting in truncated proteins of 29 bp and 135 bp length, respectively, and for this reason they were not included in the protein variants analysis. The *comD*-2.6, *comD*-2.7 and *comD*-2.8 alleles produced the same protein.

The two alleles of *comD*-3 and their respective protein variants presented the highest genetic variation in relation to *comD*-1 and *comD*-2 groups, resembling the results observed in *comC*.

Analyzing the protein sequence, ComD1 variants shared His21, Phe47, Leu51 and Ile53 as exclusive amino acids residues whereas ComD2 variants shared Arg27, Tyr34, Glu48, Ser50, Lys58 and Ile59 and ComD3 presented 13 exclusive amino acids in the first 200 bases. ComD3 also shared Ala62 with all the variants from the non-canonical ComD1 subgroup. The canonical ComD1 variants presented Phe22 as an exclusive amino acid. These positions could be candidates to be the determinants of specificity between ComD receptors and CSPs.

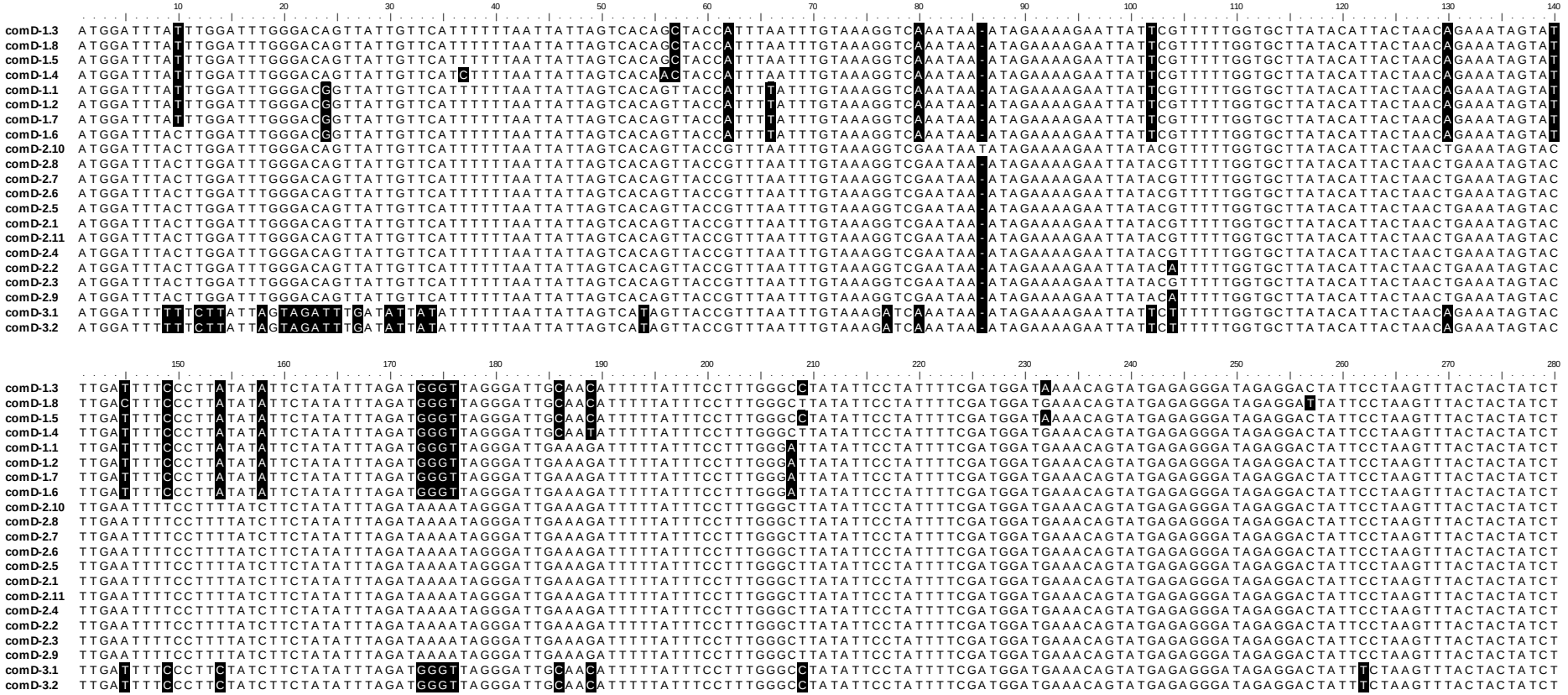


Figure 4.5 – DNA sequence of *comD* alleles. Black boxes highlight SNPs.

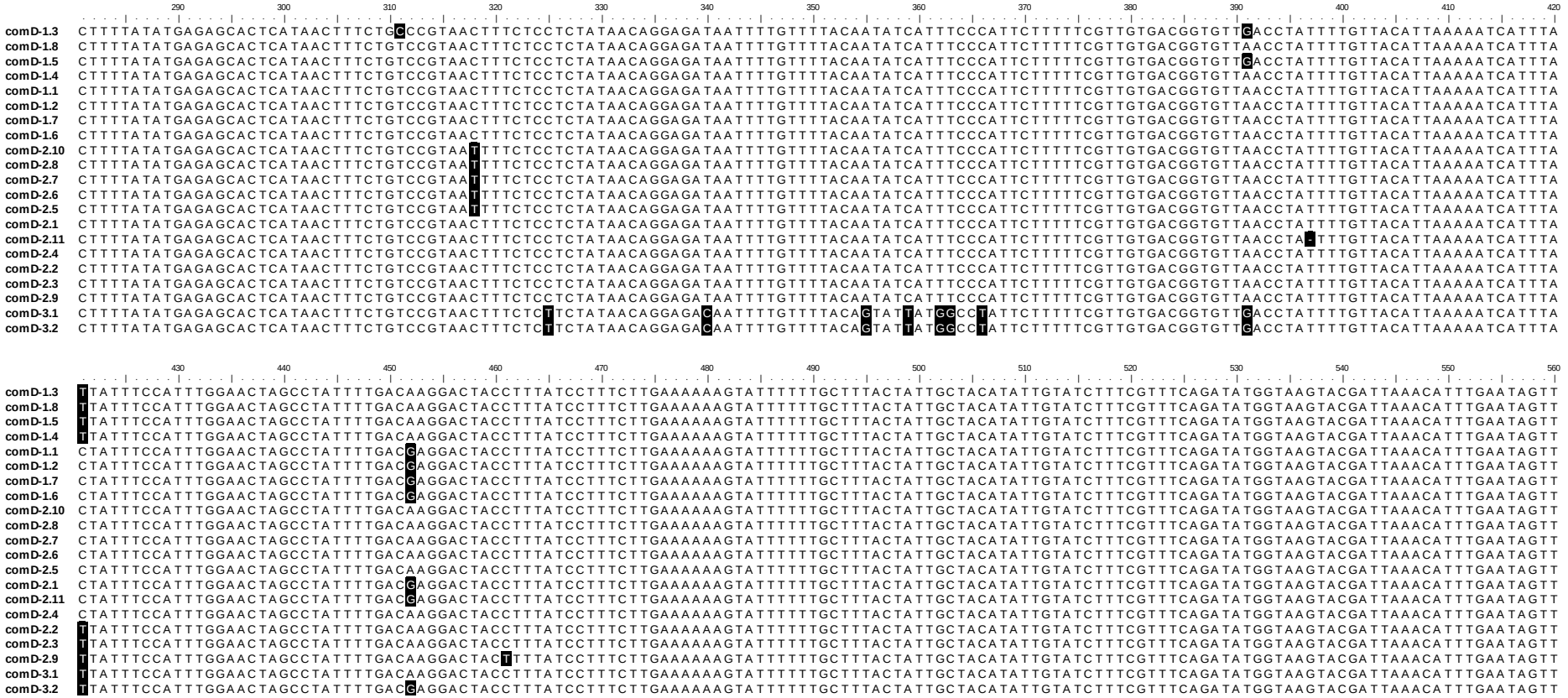


Figure 4.5 – DNA sequence of *comD* alleles *(continued).*

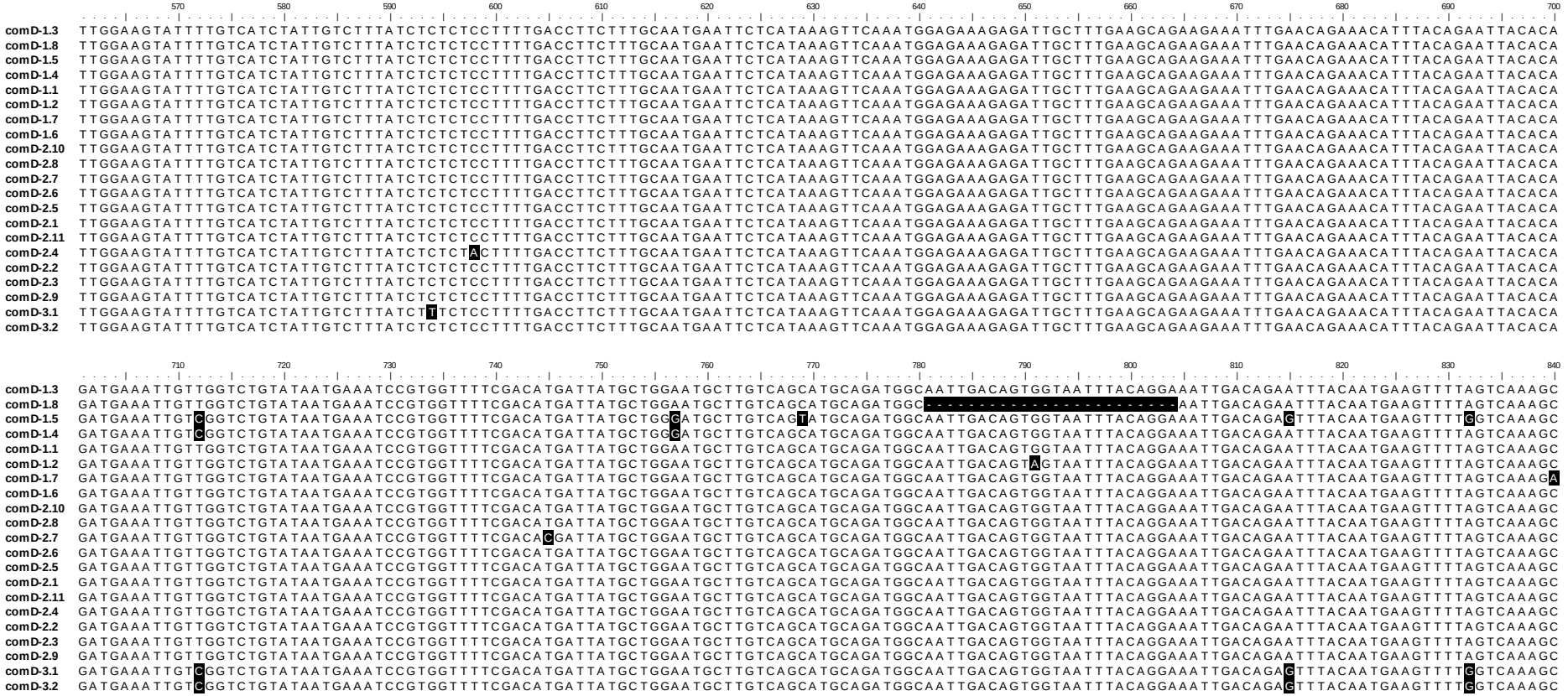


Figure 4.5 – DNA sequence of *comD* alleles *(continued).*

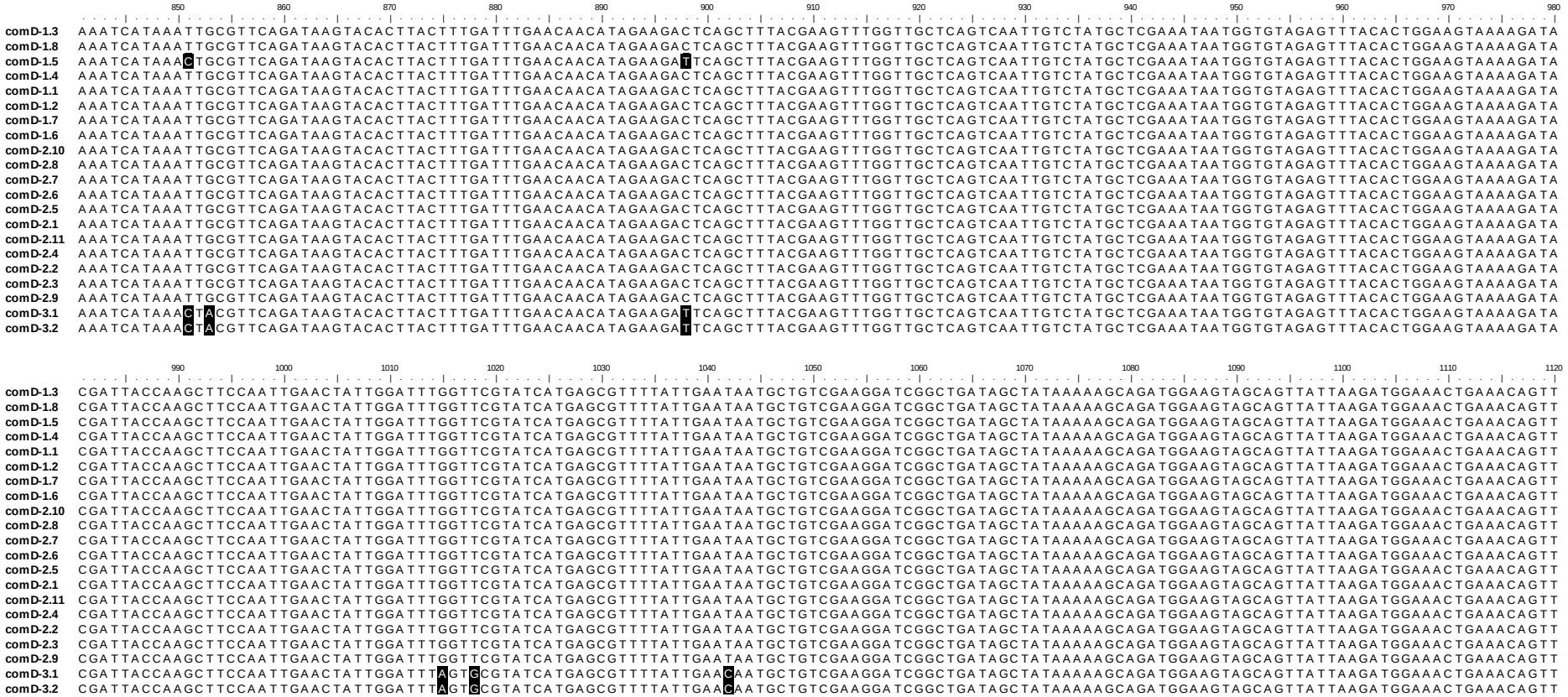


Figure 4.5 – DNA sequence of *comD* alleles *(continued).*

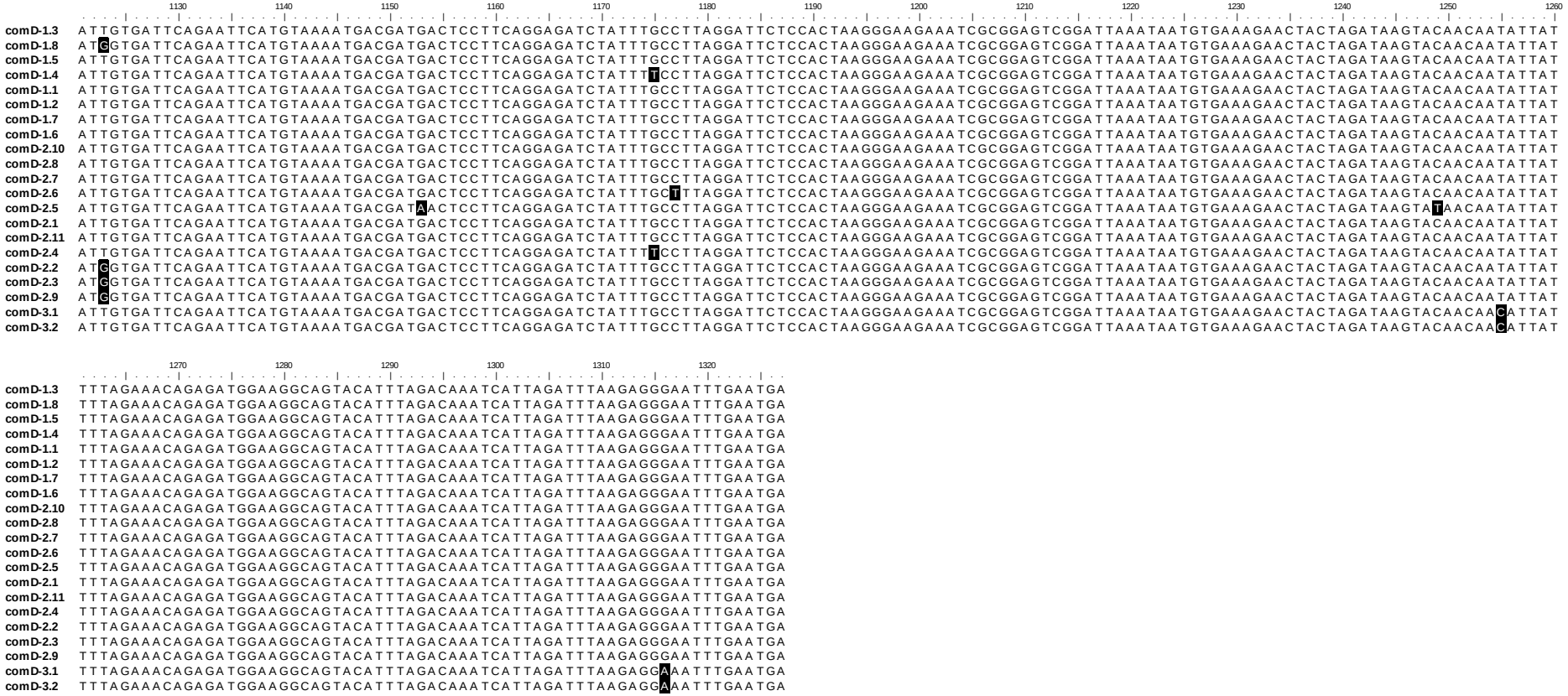


Figure 4.5 – DNA sequence of *comD* alleles *(continued).*

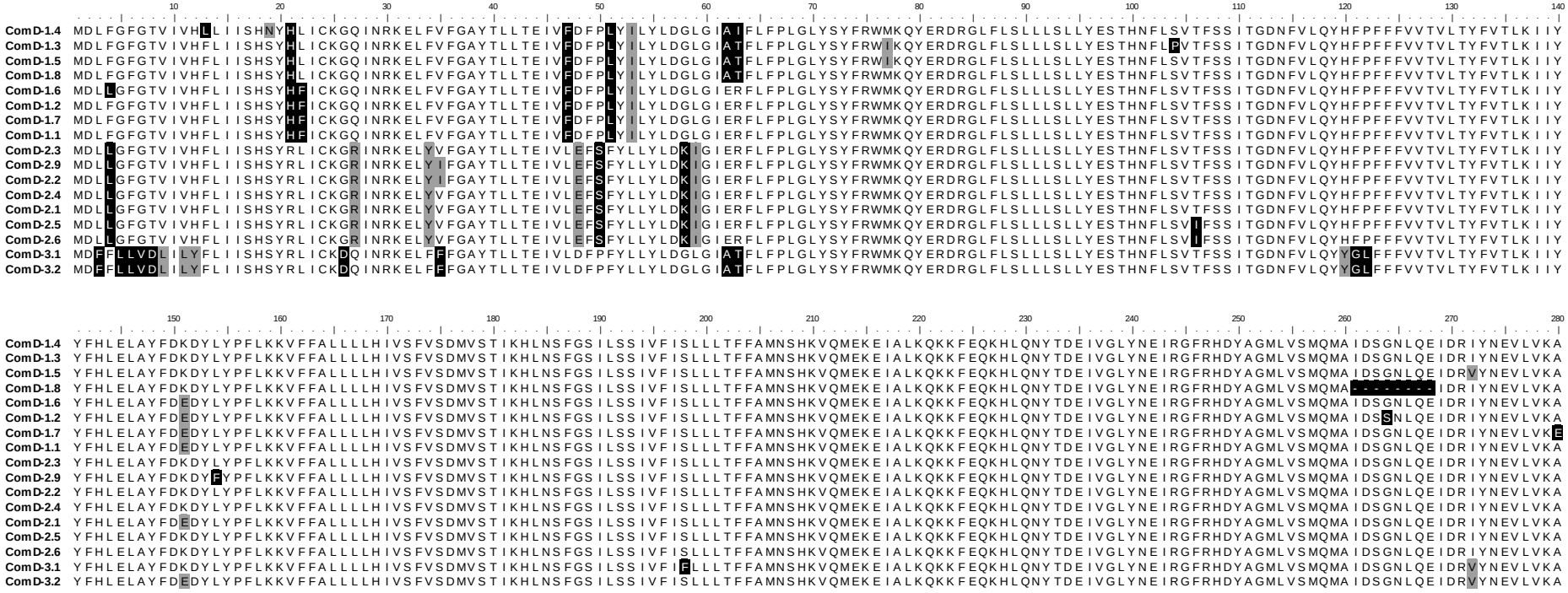


Figure 4.6 – Protein sequence of ComD variants. Black boxes highlight SNPs and grey boxes highlight similar amino acids.

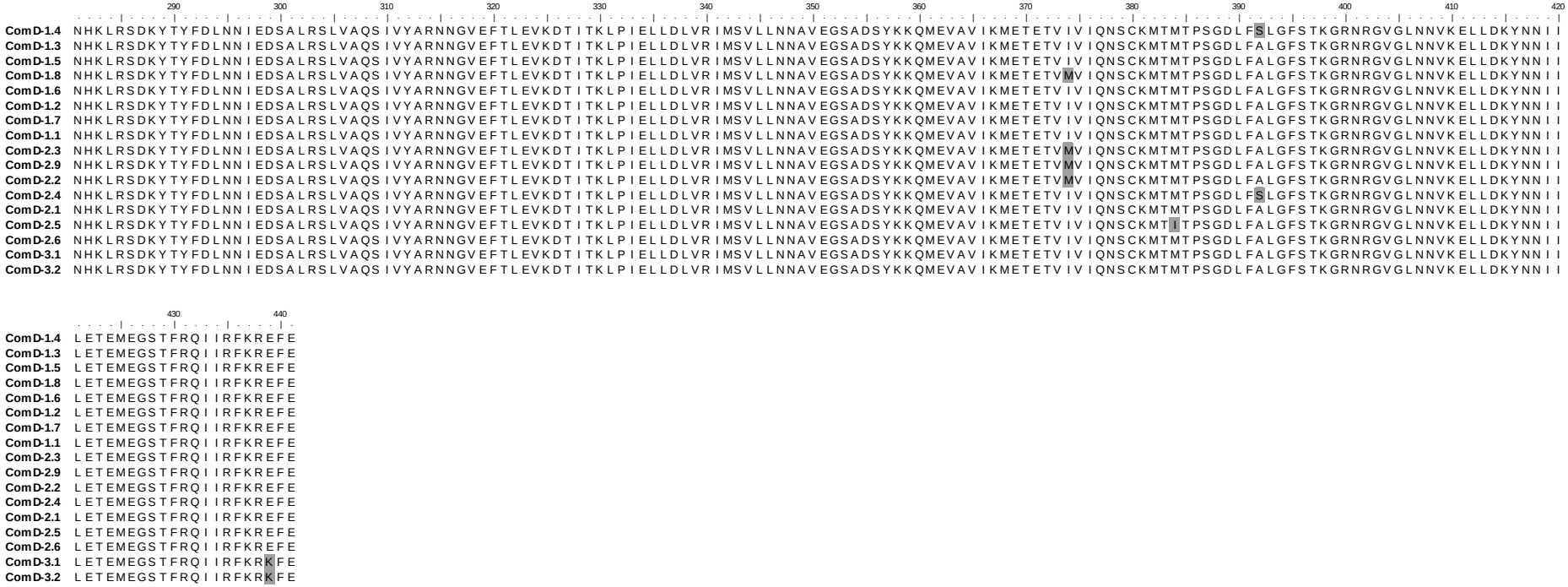


Figure 4.6 – Protein sequence of ComD variants *(continued).*

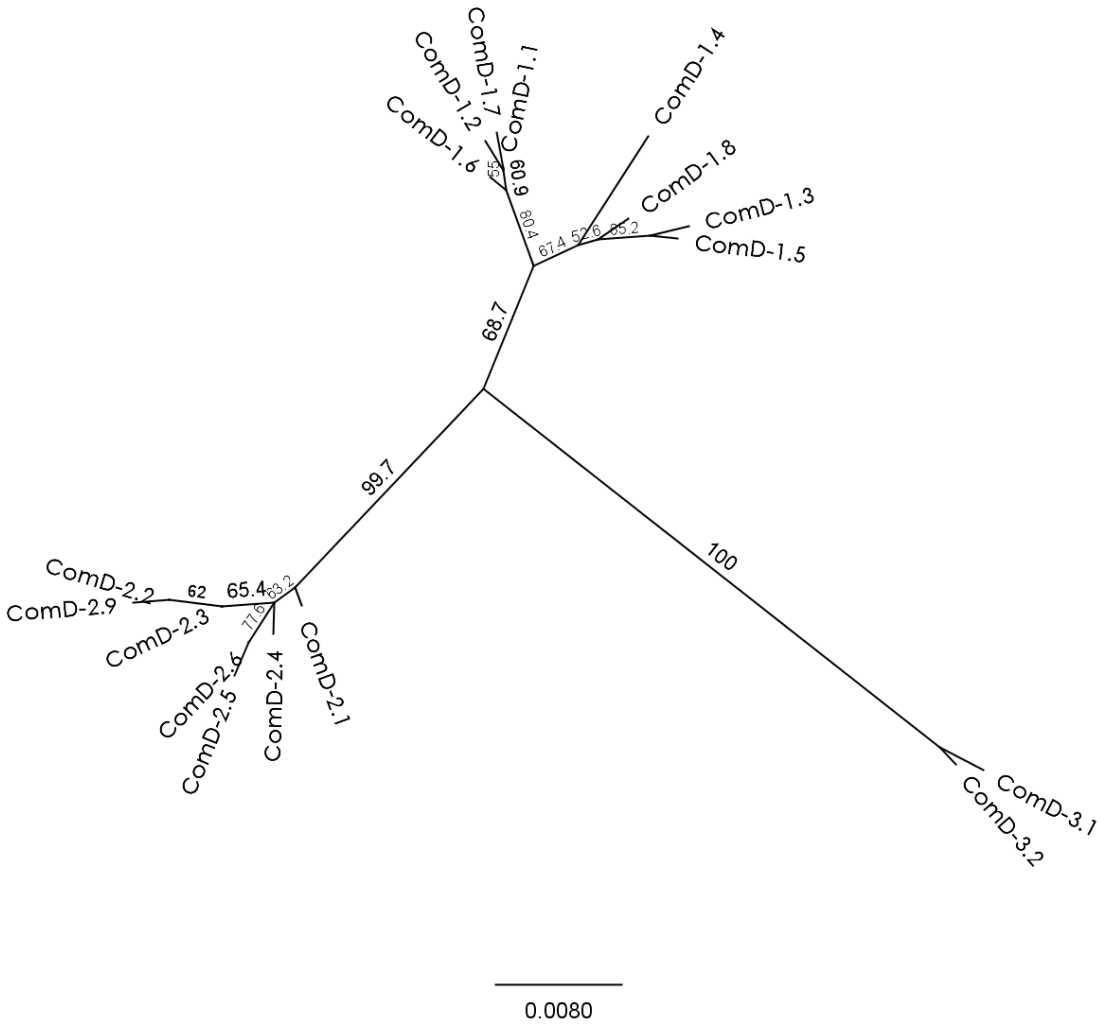


Figure 4.7 – Phylogenetic tree of ComD variants. Tree built by neighbor-joining method and resampling was done by bootstrap for 1000 times.

4.2.4. Gene comE

The *comE* gene presented a total of 15 alleles, however, two isolates presented an insertion sequence of >1200 bp in this gene and they were excluded from this analysis (**Table 4.3**, **Figure 4.8**). The insertion sequence was ISSpn8 (IS1239) from the family IS30 according to ISfinder database available at https://www-is.biotoul.fr (Siguier *et al.*, 2006). This IS was inserted in the positions 721 and 749 of these alleles, i.e., close to the end of the *comE* gene, but in a different orientation.

The alleles could not be grouped and related with pherotype because in some cases the same allele was identified in different pherotypes (**Table 4.3**). The majority of the SNPs must have originated by point mutations because many were silent mutations resulting in only 5 protein variants of ComE (**Figure 4.9**). The allele *comE*-1 produced the same protein as all the other alleles except *comE*-5, *comE*-9, *comE*-12 and *comE*-13 which produced a different ComE variant. Nevertheless, the ComE variants were conserved presenting only 6 positions with amino acid substitutions.

The *comE*-13 and *comE*-14 alleles presented the highest genetic diversity. This could be expected for *comE*-13, because it is the allele presented by CSP3 isolates, but the case of *comE*-14 was special because it presented 16 silent mutations in relation to the allele *comE*-1, producing the same protein. Thus, the allele *comE*-14 could be defined as a case of stabilizing selection.

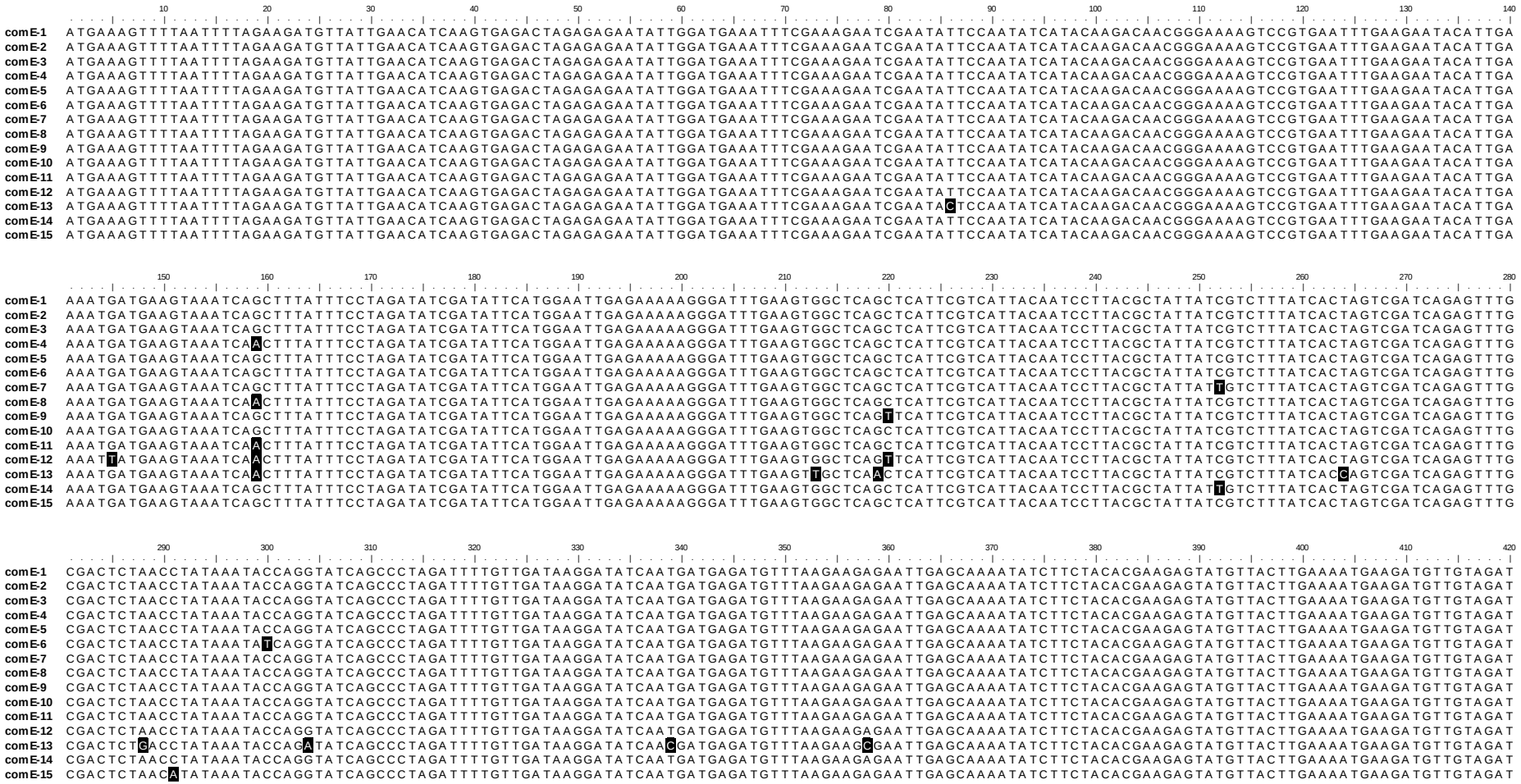


Figure 4.8 – DNA sequence of *comE* alleles. Black boxes highlight SNPs.

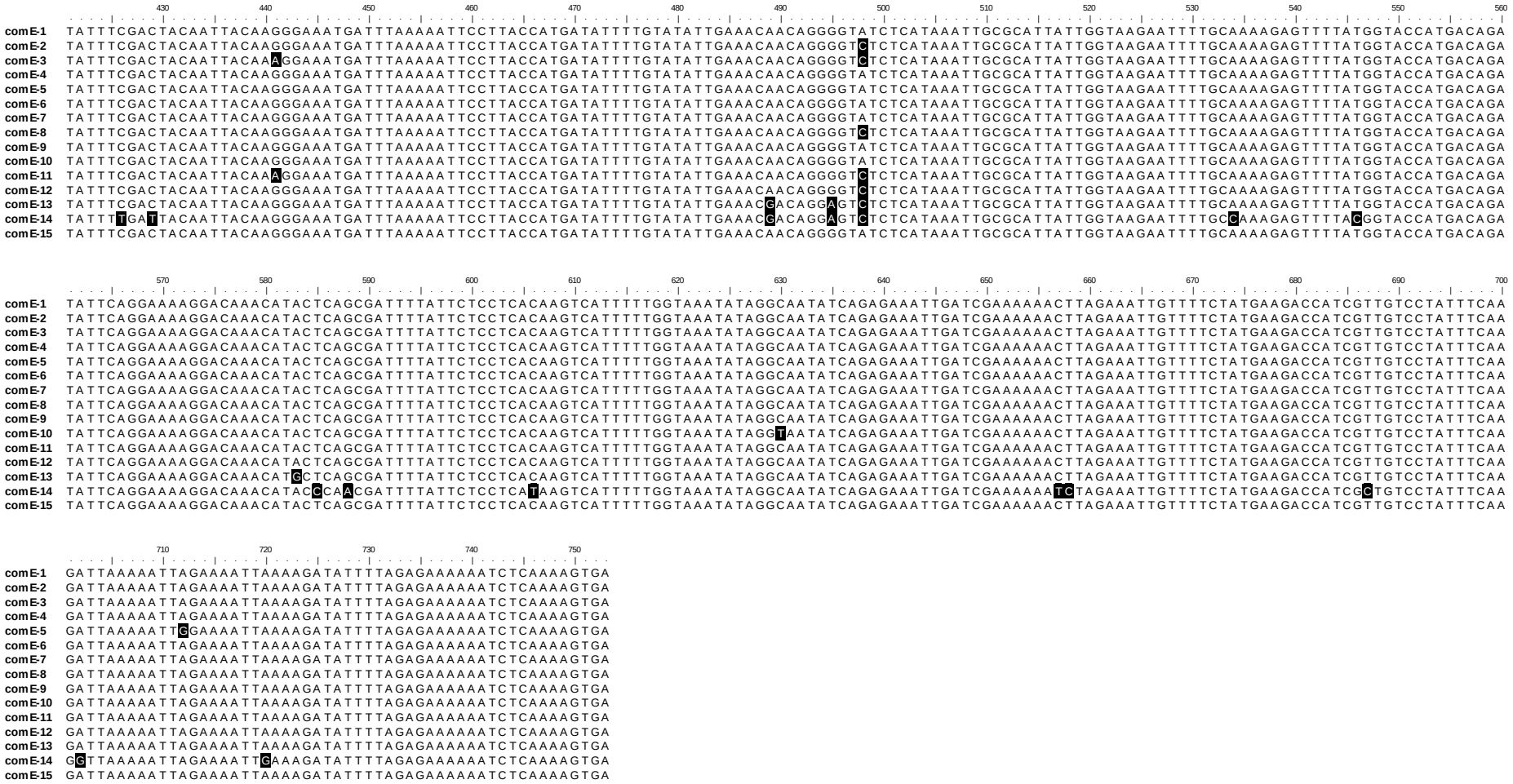


Figure 4.8 – DNA sequence of *comE* alleles *(continued).*

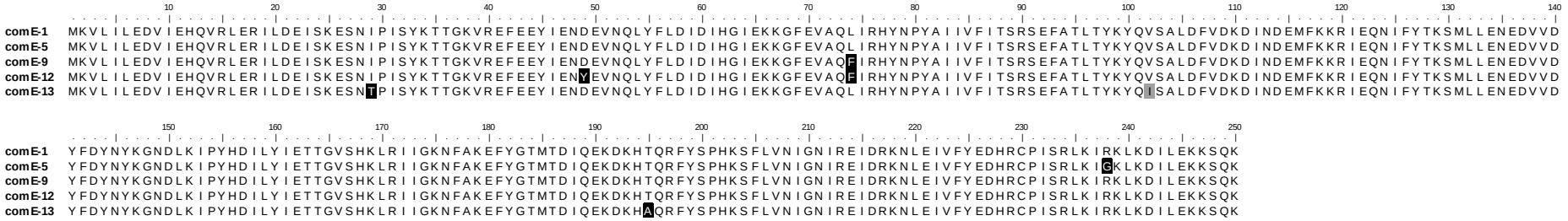


Figure 4.9 – Protein sequence of ComE variants. Black boxes highlight SNPs and grey boxes highlight similar amino acids.

4.2.5. Profiles of comCDE

After concatenating the sequences of the three genes, a total of 30 unique profiles of *comCDE* alleles were identified, including the isolates with the two *comE* alleles with ISSpn8 (IS1239) (**Table 4.4**). The canonical subgroup of *comD1* alleles was correlated with *comC*-1.1 and *comC*-1.3 alleles, while *comC*-1.2 allele presented only non-canonical *comD1* alleles. The *comC*-2.1 and *comC*-2.2 alleles also presented their own *comD* alleles, i.e., they did not share *comD* alleles between them.

The profiles with *comC*-1.1, *comD*-1.1 and four diverse *comE* alleles (*comE*-1, *comE*-2, *comE*-3 and *comE*-6) were the most frequently identified. Although various *comE* alleles were associated with the canonical *comD1* alleles, they produced the same ComE protein variant. Thus, *ComC*-1.1 (CSP1), *ComD*-1.1 and *ComE*-1 were the most frequent protein variants produced in these isolates.

Table 4.4 – Profiles of *comCDE* sequences and respective serotypes and STs.

Pherotype	n	*comC*	*comD*	*comE*	Serotype	Sequence Type
CSP1	15	*comC*-1.1	*comD*-1.1	*comE*-3	14 (n=9); 9V (n=5); 12B (n=1)	156 (n=6); 557 (n=4); 218, 644, 790, 838, 1225 (n=1 each)
CSP1	13	*comC*-1.1	*comD*-1.1	*comE*-1	11A, 33F (n=3 each); 18C, 22F (n=2 each); 8, 13, 20 (n=1 each)	408, 1012, 1367 (n=2 each); 53, 62, 100, 574, 717, 1026, 1233 (n=1 each)
CSP1	11	*comC*-1.1	*comD*-1.1	*comE*-2	1 (n=6); 4 (n=3); 14, 16F (n=1 each)	306, 1221 (n=3 each); 15, 228, 304, 350, 414 (n=1 each)
CSP1	5	*comC*-1.1	*comD*-1.1	*comE*-6	23F (n=4); 15B (n=1)	338 (n=3); 1371, 1706 (n=1 each)
CSP1	4	*comC*-1.1	*comD*-1.2	*comE*-2	7F (n=4)	191 (n=4)
CSP1	3	*comC*-1.1	*comD*-1.1	*comE*-4	6B (n=2); 6A (n=1)	273, 1224, 1369 (n=1 each)
CSP1	3	*comC*-1.2	*comD*-1.3	*comE*-8	3 (n=3)	180 (n=2); 1230 (n=1)
CSP1	2	*comC*-1.1	*comD*-1.1	*comE*-11	12B (n=2)	218, 1365 (n=1 each)
CSP1	2	*comC*-1.2	*comD*-1.4	*comE*-9	19A (n=2)	199, 876 (n=1 each)
CSP1	2	*comC*-1.2	*comD*-1.5	*comE*-4	5 (n=2)	1223 (n=2)
CSP1	2	*comC*-1.2	*comD*-1.8	*comE*-12	3 (n=2)	458 (n=2)
CSP1	1	*comC*-1.1	*comD*-1.6	*comE*-2	19F (n=1)	347 (n=1)
CSP1	1	*comC*-1.1	*comD*-1.7	*comE*-1	8 (n=1)	53 (n=1)
CSP1	1	*comC*-1.1	*comD*-1.1	Insertion of 1274 bp	35F (n=1)	1368 (n=1)
CSP1	1	*comC*-1.3	*comD*-1.1	*comE*-15	22F (n=1)	1372 (n=1)
CSP2	4	*comC*-2.1	*comD*-2.1	*comE*-5	3 (n=4)	260, 1220 (n=2 each)
CSP2	3	*comC*-2.1	*comD*-2.3	*comE*-1	19F (n=2); 10F (n=1)	177, 391, 1649 (n=1 each)
CSP2	3	*comC*-2.2	*comD*-2.2	*comE*-7	23F (n=2); 19A (n=1)	81 (n=3)
CSP2	2	*comC*-2.1	*comD*-2.10	*comE*-10	4 (n=2)	247 (n=2)
CSP2	1	*comC*-2.1	*comD*-2.1	*comE*-4	4 (n=1)	1222 (n=1)
CSP2	1	*comC*-2.1	*comD*-2.5	*comE*-1	6C (n=1)	395 (n=1)
CSP2	1	*comC*-2.1	*comD*-2.6	*comE*-4	12B (n=1)	989 (n=1)
CSP2	1	*comC*-2.1	*comD*-2.7	*comE*-10	6B (n=1)	273 (n=1)
CSP2	1	*comC*-2.1	*comD*-2.8	Insertion of 1246 bp	10A (n=1)	1231 (n=1)
CSP2	1	*comC*-2.1	*comD*-2.11	*comE*-5	3 (n=1)	260 (n=1)
CSP2	1	*comC*-2.2	*comD*-2.2	*comE*-14	9N (n=1)	66 (n=1)
CSP2	1	*comC*-2.2	*comD*-2.4	*comE*-9	8 (n=1)	404 (n=1)
CSP2	1	*comC*-2.2	*comD*-2.9	*comE*-7	19A (n=1)	276 (n=1)
CSP3	1	*comC*-3	*comD*-3.1	*comE*-13	19A (n=1)	1201 (n=1)
CSP3	1	*comC*-3	*comD*-3.2	*comE*-13	35F (n=1)	1366 (n=1)

4.3. Discussion

In this chapter the genetic diversity of the *comCDE* locus was presented. Three CSP variants were identified corresponding to pherotypes CSP1, CSP2 and CSP3. The diversity of *comD* could be related with pherotype but that was not the case for *comE*. The majority of the genetic divergence was located in the region coding the mature CSP and the sensor domain of ComD. The gene *comD* was the most diverse while *comE* was more conserved presenting just 5 protein variants.

Regarding pherotype, which is identified by the type of CSP produced, most isolates presented CSP1 (n=66, 74.2 %) while the remaining isolates were CSP2 (n=21, 23.6 %) and CSP3 (n=2, 2.2 %). Proportion of CSP1 and CSP2 was similar to previous work in our lab (Carrolo *et al.*, 2009) but the PCR used by this study was not able to detect CSP3 strains. Until now CSP3 abundance has not been properly evaluated because only a few isolates of this pherotype were identified (Evans and Rozen, 2013, Ramirez *et al.*, 1997, Whatmore *et al.*, 1999). The proportion of CSP3 reported here is not trustworthy because only 2 CSP3 isolates were found in the 89 isolates tested, making it necessary to screen a larger number of isolates. The CSPs found in this study were already known, thus new variants of CSP were not discovered.

Regarding *comC* alleles, *comC*-1.3 was found in a single isolate and resulted from a point mutation in the region coding the signal-peptide. This allele was not identified in the published studies of the diversity of *comC* (Evans and Rozen, 2013, Pozzi *et al.*, 1996, Ramirez *et al.*, 1997, Vestrheim *et al.*, 2011, Whatmore *et al.*, 1999) but *comC*-1.1, *comC*-1.2, *comC*-2.2, *comC*-2.2 and *comC*-3 alleles were described in those studies.

The CSP3 identified here is the same as CSP-4 described by Whatmore *et al.* (Whatmore *et al.*, 1999). Furthermore, Whatmore *et al.* and Ramirez *et al.* (Ramirez *et al.*, 1997) also identified a similar CSP (CSP-3 and CSPγ, respectively) that presented an additional NFF repetition before the terminal triple arginine (**Figure 4.4**). However, this variant was not found in this study. The primers designed for sequencing the *comCDE* locus made it possible to detect pherotype CSP3 among our isolates and it was the first time CSP3 strains were identified in Portugal.

The gene presenting more genetic diversity was *comD*. However, this diversity was not distributed evenly across the encoded protein. Most of the genetic variation was seen in the N-terminus thought to correspond to the ComD sensor

domain, which interacts with CSP, while the kinase domain located at the C-terminus, which interacts with ComE, was more conserved. Thus, because of the diversity in the sensor domain, *comD* also forms 3 groups that correlate with pherotypes. However, when analyzing just the kinase domain, the relation with pherotype is lost, as also happens with *comE*. This strongly suggests that pherotype specificity is due to the interaction between CSP and the ComD sensor domain located in the N-terminus by an unknown mechanism.

Interestingly, ComD1 variants could be divided into two groups, canonical and non-canonical (**Figure 4.7**), with the former corresponding to ComD1 in Iannelli *et al*. (Iannelli *et al.*, 2005) and the later to ComD3 and ComD4 in the same study. The results of Iannelli *et al*. suggest a low affinity binding of non-canonical ComD1 to CSP1, requiring a higher dose of CSP1 to reach the same transformation efficiency as canonical ComD1 (**Figure 4.10**). They also observed that a non-canonical ComD1 (termed ComD3 in their study) was cross-induced by high doses of CSP2 and the same was also observed when high doses of CSP1 were given to ComD2-carrying strains. Their results also suggest that ComD2 has a lower affinity binding to CSP2 than canonical ComD1 to CSP1, because at a dose of 30 ng/mL of synthetic CSP, ComD2 is not induced while canonical ComD1 is fully induced (**Figure 4.10**). However, Evans and Rozen (Evans and Rozen, 2013) did not see any difference in transformation efficiency between canonical and non-canonical ComD1 variants, but they observed significantly more admixture in isolates carrying non-canonical ComD1. Their study included a pherotype CSP3 strain which was not able to transform in the presence of CSP1 synthetic peptide. Ramirez *et al*. reported a low response of a CSP3 isolate to CSP1 and also observed cross-induction of CSP2 isolates by adding CSP1 (Ramirez *et al.*, 1997). Taking together the results of these studies, isolates carrying non-canonical ComD1 variants respond to CSP1 and, although to a lesser extent, could also respond to CSP2. This could be the reason why they present more admixture than isolates carrying canonical ComD1 variants because, hypothetically, they would be able to more easily exchange DNA with both pherotype CSP1 and CSP2 populations. Induction of non-canonical ComD1 by CSP2 cannot be explained by looking at the amino acid sequence because non-canonical ComD1 do not share unique amino acids with ComD2 protein variants. Thus, a study comparing the transformation efficiencies of isolates carrying diverse ComD variants (ComD2, ComD3 and canonical and non-canonical ComD1) in the presence of synthetic CSPs (CSP1,

CSP2 and CSP3) complemented by a population structure analysis could shed light on this hypothesis. ComD3 variants seem a hybrid between both ComD1 subgroups and ComD2, sharing many of the positions characteristic of each of these variants. Response of ComD3 to CSP2 or CSP3 has not been reported yet.

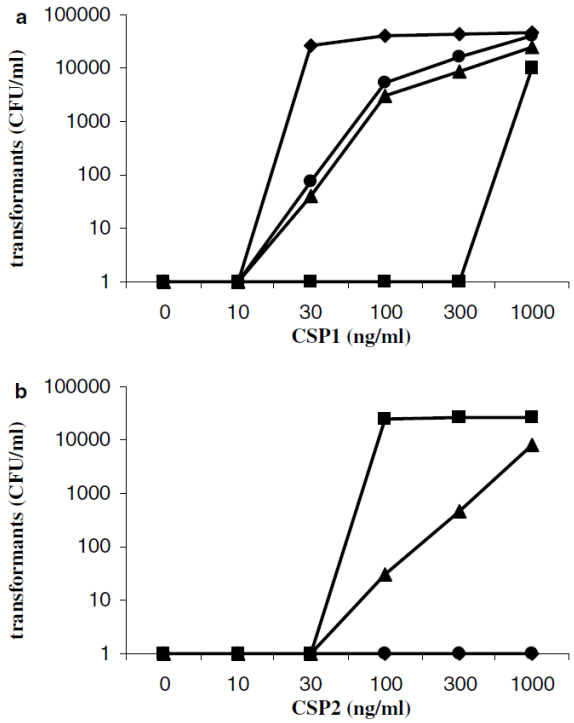


Figure 4.10 - Competence pheromone specificity in *S. pneumoniae*. Figure from Iannelli *et al.* (Iannelli *et al.*, 2005). Four isogenic strains, each expressing a ComD variant in the same competence-negative genetic background, were induced to competence with synthetic CSP1 (a) and CSP2 (b) peptides. CSP doses range from 10 to 1000 ng/mL. Transformation efficiency is expressed as number of CFU transformed with a streptomycin resistance marker in a 106 CFU/mL culture. Strains expressing the ComD1/ComD3/ComD4 receptors were induced to competence by CSP1, whereas strain expressing the ComD2 receptor was induced by CSP2. Cross-induction of competence was also observed in strains expressing ComD2 or ComD3. In the absence of CSPs, no transformants were obtained. The ComD1-carrying strain is indicated as a diamond, the ComD2-carrying strain as a square, the ComD3-carrying strain as a triangle, and the ComD4-carrying strain as a circle.

The protein ComE is highly conserved among the pneumococcal population with only 5 variants identified in this study. It seems to be a gene under strong purifying selection as the majority of the SNPs observed were silent mutations. ComE does not take part in the recognition of CSP, thus the relation between pherotype and alleles which was seen for *comC* and *comD* is lost. ComE is conserved probably because of its gene activator function and the requirement to both interact with ComD for phosphorylation and the target DNA sequences.

Two isolates presented the insertion sequence ISSpn8 (IS1239) belonging to the family IS30 (Siguier *et al.*, 2006) (ISfinder database available at https://www-is.biotoul.fr) at the C-terminal end of *comE*. The resulting ComE proteins of these isolates are 244 aa and 296 aa in length while the canonical ComE is 250 aa. These proteins could be functional if the C-terminal end of ComE does not have a functional role. However, the 296 aa protein could be too large to be functional. Moreover, an insertion sequence was also reported in the *comC* gene of the strain AP200 (**Table 4.1**) and an insertion (*comD*-2.10) and deletion (*comD*-2.11) was observed here in *comD*, indicating that there could be alterations in all the genes of the *comCDE* locus leading to a loss of function. It was seen that mutations in ComD severely attenuate virulence in mouse models of pneumonia and bacteraemia infections (Lau *et al.*, 2001). Thus, although pneumococci can survive with alterations in their competence induction genes, these may be ultimately detrimental to them or at least to their virulence potential.

This chapter analyzed a relatively small sample chosen to represent the diversity of the pneumococcal population before conjugate vaccine introduction. Even though, *ComC*-1.1 (CSP1), *ComD*-1.1 and *ComE*-1 proteins were much more frequent than any other protein variants. It is not clear if these are the most efficient protein combination in controlling competence. Our results are similar to what was previously described in the literature. The dominance of CSP1 pherotype is well known, although it is unclear why CSP1 dominates over other pherotypes. The most immediate explanation could be that CSP1 and canonical ComD1 presents the most effective peptide-receptor interaction translating in a best response to CSP. Maybe Phe22 of canonical ComD1 plays an important role because it is specific of this group. However, this is a hypothesis that will require formal testing. A previous study by our laboratory with mutants lacking the *comC* gene showed that CSP1 strains had a higher capacity to form *in vitro* biofilms and yielded more transformants than CSP2 strains (Carrolo *et al.*, 2014). It was suggested that both effects could help explain the higher prevalence of CSP1 strains in the pneumococcal population by enabling them to be more transmissible and more efficient at persisting in carriage.

II. Pherotype diversity and abundance

4. Diversity of the *comCDE* locus

5. Pherotype abundance

CHAPTER 5. PHEROTYPE ABUNDANCE

In the last chapter, pherotype CSP3 was identified for the first time in Portugal. The abundance of pherotypes CSP1 and CSP2 is well described in literature (Carrolo *et al.*, 2009, Evans and Rozen, 2013, Ramirez *et al.*, 1997, Valente *et al.*, 2012, Vestrheim *et al.*, 2011, Whatmore *et al.*, 1999). Regarding CSP3, it is assumed to be a rare pherotype because those studies identify no strain or very few strains of this pherotype. Our laboratory previously described the abundance of pherotypes CSP1 and CSP2 of an invasive pneumococcal collection recovered between 1999 and 2002 (Carrolo *et al.*, 2009). However, the PCR scheme designed in that study was unable to distinguish pherotype CSP1 from CSP3.

The discovery of CSP3 isolates in Portugal prompted us to explore the abundance of this pherotype in natural populations of pneumococci. We used an extensive pneumococcal invasive isolate collection recovered during a period of 14 years to identify the abundance of each pherotype and evaluate their evolution, using a new PCR scheme which would be able to identify the three pherotype variants. Then, associations between pherotype and other clonal characteristics were evaluated.

5.1. Materials and methods

5.1.1. Bacterial isolates

To determine the abundance of each pherotype, all invasive pneumococcal isolates recovered from children (<18 years) during 1999 and 2012 in Portugal and sent to the Instituto de Microbiologia of Faculdade de Medicina of Universidade de Lisboa (n=903) were selected. They represent a geographically well-defined population and this period included the introduction of conjugate vaccines, when changes in serotype distribution were described previously (Aguiar *et al.*, 2010a, Aguiar *et al.*, 2014, Aguiar *et al.*, 2008). Although this collection does not include isolates recovered from adults, it is larger than any other study where pherotype was determined. A high number of isolates was needed in order to have a reliable estimation of the proportion of CSP3 isolates. The serotype, MLST characterization and antibiotic susceptibility of these isolates were previously described by our laboratory (Aguiar *et al.*, 2010b, Aguiar *et al.*, 2014, Aguiar *et al.*, 2010d, Aguiar *et al.*, 2008, Serrano *et al.*, 2005, Serrano *et al.*, 2004).

5.1.2. Pherotype identification

In the previous work performed by our laboratory (Carrolo *et al.*, 2009), a PCR was developed to identify the pherotype of pneumococcal isolates. This PCR used three primers (**Table 5.1**): CSP-fw, able to recognize all isolates, and CSP1-rv and CSP2-rv, specific for CSP1 and CSP2 isolates, respectively (**Figure 5.1**). However, this PCR was unable to distinguish CSP1 from CSP3, since CSP1-rv was able to recognize both pherotypes. Therefore, a new primer was designed: CSP3-rv (**Table 5.1**, **Figure 5.1**). The identification of pherotypes was performed by evaluating the size and number of the products amplified (**Figure 5.2**).

The PCR reaction for identification of pherotype was composed of:

- 0.05 U/µl GoTaq® DNA polymerase (Promega, Wisconsin, USA);
- 1x enzyme's buffer provided by manufacturer;
- 2 mM magnesium chloride;
- 0.4 mM dNTPs;
- 0.4 pmol/µl of each primer (CSP-fw, CSP1-rv, CSP2-rv and CSP3-rv);
- 20 µl of boiled bacterial cells.

With the following PCR program:

- 5 min at 95 ºC;
- 30x (1 min at 95 ºC, 30 s at 55 ºC and 1 min at 72 ºC)
- 10 min at 72 ºC.

Table 5.1 – Sequence of the primers used for pherotype determination.

Primer	Sequence (5'→ 3')	Size (bp)
CSP-fw	TGAAAAACACAGTTAAATTGGAAC	24
CSP1-rv	TCAAGAAAGGATAAAGGTAGTCCTC	25
CSP2-rv	TAAAAATCTTTCAATCCCTATTTT	24
CSP3-rv	AAGATATTAAAGGACTTTTCATTCA	25

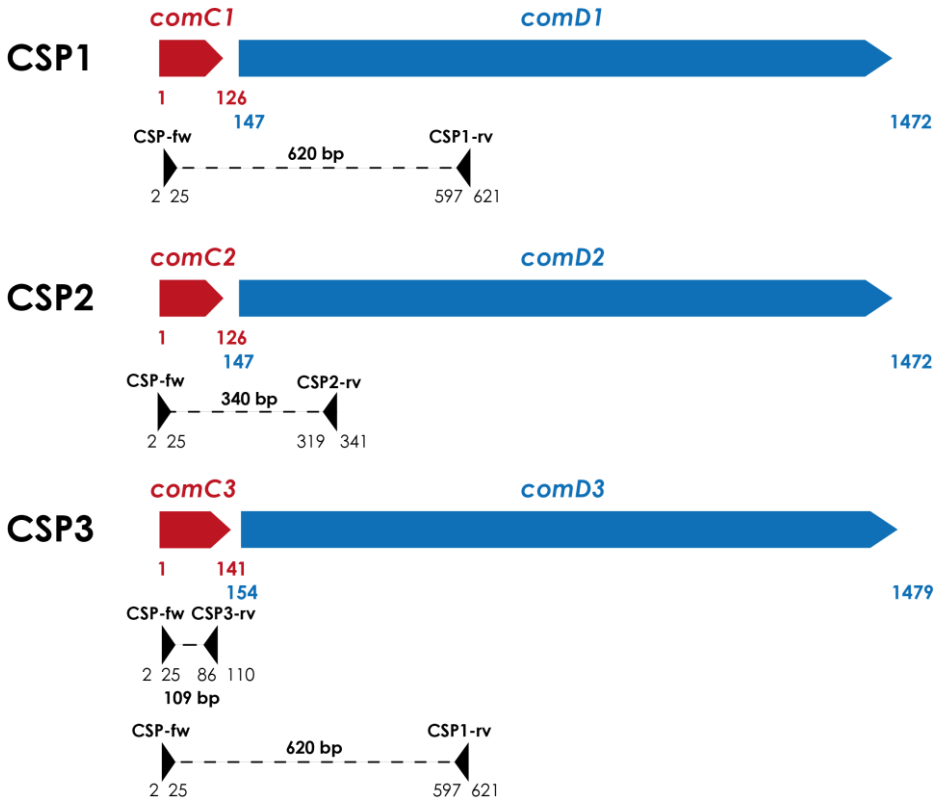


Figure 5.1 – Primers used for pherotype identification. Primer CSP1-rv also binds to *comD3* in CSP3 strains, whereas *comD2* in CSP2 strains only have a single difference in the site of CSP1-rv binding, so this primer can also bind to CSP2 strains but with less specificity.

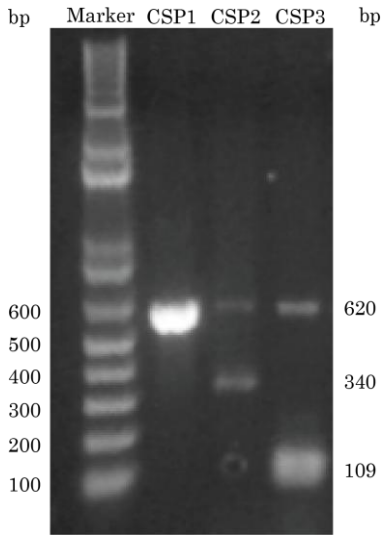


Figure 5.2 – Example gel of pherotype determination. Primer CSP1-rv also binds to *comD3* in CSP3 strains, whereas *comD2* in CSP2 strains only have a single difference in the site of CSP1-rv binding, so this primer can also bind to CSP2 strains but with less specificity.

5.1.3. Statistical analysis

The Cochran-Armitage test (CA) was used to evaluate the temporal trends of pherotype, serotype, ST, CC and antimicrobial resistance. Fisher's exact test (FET) was used to test association of pherotype with serotype, ST, CC and antimicrobial resistance. The false discovery rate (FDR) correction for multiple testing (Benjamini and Hochberg, 1995) was used in both tests. A $p<0.05$ was considered significant for all tests.

Genetic diversity was evaluated using Simpson's index of diversity (SID) and respective 95 % confidence intervals ($CI_{95\,\%}$) (Carriço *et al.*, 2006). Adjusted Wallace (AW) was used to evaluate the congruence between the typing methods and to quantify the ability of each trait to predict pherotype (Carriço *et al.*, 2006). SID and AW were calculated using the online tool available at www.comparingpartitions.info.

5.2. Results

5.2.1. Pherotype abundance and evolution

A total of n=903 isolates were recovered from children (<18 years) during a period of 14 years (1999-2012) (**Table 5.2**). These isolates were recovered from blood (n=768), CSF (n=94), pleural (n=35), synovial (n=4) and ascitic (n=2) fluids. The total number of isolates recovered per year increased over time, stabilizing around 2007. After 2010, a decrease of children IPD incidence was observed (**Table 5.2**, (Aguiar *et al.*, 2014)).

The most abundant pherotype was CSP1 (n=681, 75.4 %), followed by CSP2 (n=192, 21.3 %) and then by CSP3 (n=20, 2.2 %) (**Table 5.2**). The pherotype of 10 isolates was not determined because either a PCR product was not amplified (serotype 25A/38, n=9) or it had a higher size indicating the presence of an insertion sequence (serotype 6C, n=1). The case of the isolates expressing serotype 25A/38, which were all identified as ST393 (*see 5.2.4. Sequence type and pherotype*), will be discussed in the next section (*5.2.2. ST393 serotype 25A/38 isolates*).

Analyzing pherotype proportions over the years, a decrease from 85.4 % in the period 1999-2002 to 67.1 % in 2012 was observed for pherotype CSP1, although this was only supported before FDR correction (CA p=0.0314) (**Figure 5.3**). However, no significant changes were seen for pherotypes CSP2 and CSP3. Therefore, pherotype proportions remained relatively stable despite the changes that occurred in serotype distribution (see *5.2.3. Serotype and pherotype*).

In this study four age groups were defined: 0-11 months (n=261), 12-23 months (n=178), 2-4 years (n=207) and 5-17 years (n=257). Between the period 1999-2002 and 2012, a decrease from 82.8 % to 44.4 % of pherotype CSP1 proportion and an increase from 13.8 % to 50.0 % of pherotype CSP2 proportion were observed in the age group 0-11 months, although, as before, these changes were only supported before FDR correction (CA p=0.0385 and CA p=0.0316, respectively) (**Figure 5.4**). In the other age groups the pherotype proportions remained constant. Comparing the pherotype proportion between age groups, CSP1 proportion increased from 67.8 % to 85.2 % in the younger to older age groups, while CSP2 proportion decreased from 27.6 % to 12.5 % between the same age groups (CA p<0.001 for both pherotypes, significant after FDR) (**Figure 5.5**). In the section *5.2.3. Serotype and pherotype* is presented the contribution of serotypes 1 and 19A for this association between pherotype and age.

Table 5.2 – Pherotype distribution of pneumococcal invasive isolates recovered in children in Portugal during 1999-2012.

Pherotype	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	**Total**
CSP1	20	15	25	22	22	29	35	60	98	80	116	62	48	49	**681**
CSP2	0	3	5	5	6	11	12	22	21	26	32	12	17	20	**192**
CSP3	0	0	1	0	1	1	1	2	3	1	2	4	2	2	**20**
ND[a]	0	0	0	0	0	0	0	2	3	0	0	0	3	2	**10**
Total	**20**	**18**	**31**	**27**	**29**	**41**	**48**	**86**	**125**	**107**	**150**	**78**	**70**	**73**	**903**

[a]Not determined: 25A/38 (n=9) and 6C (n=1, insertion in *comC*).

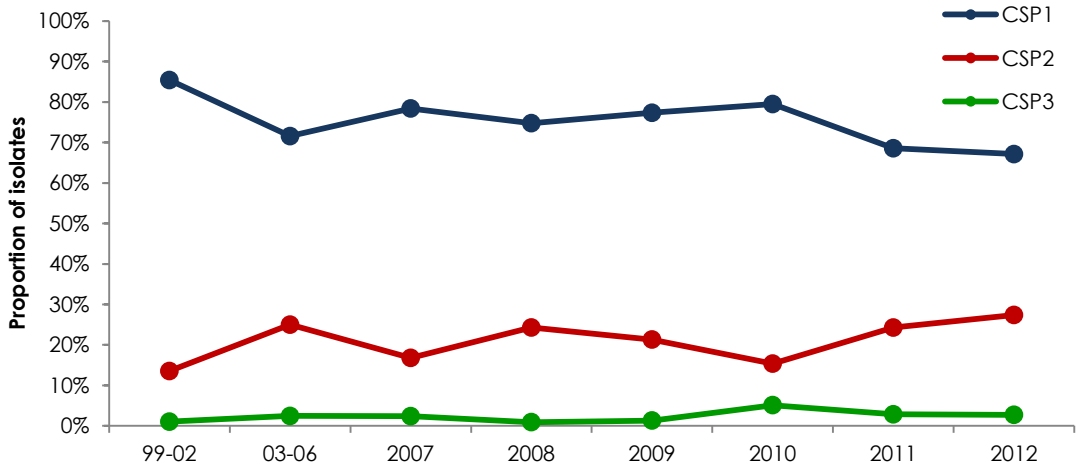


Figure 5.3 – Evolution of pherotype proportion in invasive isolates recovered from children between 1999 and 2012. PCV7 and PCV13 were available in children vaccination since 2001 and 2010, respectively.

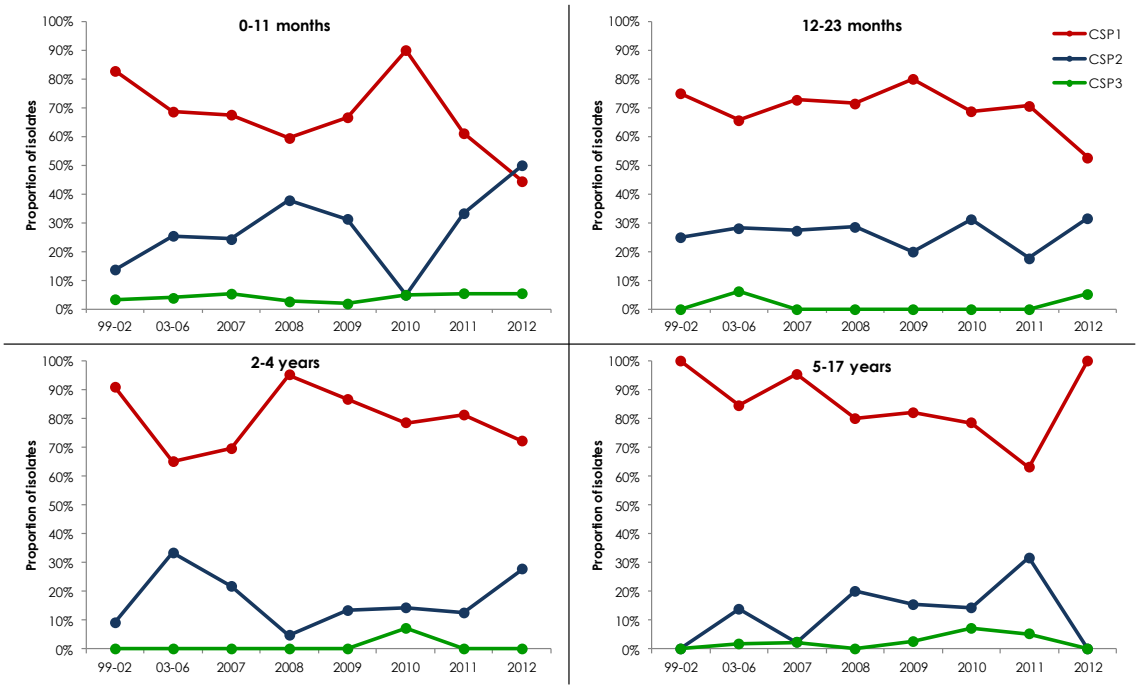


Figure 5.4 – Evolution of pherotype proportion in invasive isolates by age group. PCV7 and PCV13 were available in children vaccination since 2001 and 2010, respectively.

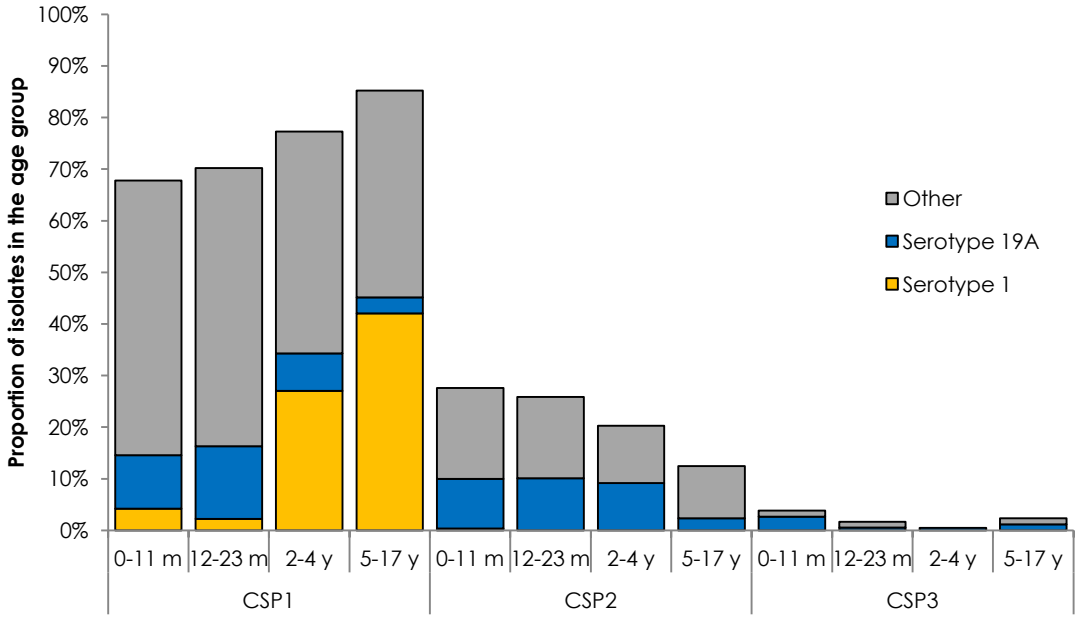


Figure 5.5 – Pherotype, serotype 1 and serotype 19A proportion by age group. Other: other serotypes.

5.2.2. ST393 serotype 25A/38 isolates

All serotype 25A/38 isolates were identified as ST393 and a PCR product was not obtained in the pherotype determination PCR. This was intriguing because these isolates could present a new type of CSP undetected by the PCR setup used. Thus, we focused in sequencing the *comCDE* locus of these strains. However, despite trying several primer pairs aimed at amplification of this locus, a PCR product with the complete operon was not obtained.

To clarify this, we obtained preliminary data of the high throughput sequencing of an ST393 isolate. The contigs obtained from the assembly of this genome were queried with the *comC* and *comD* alleles from strain R6. Both genes were found in the same contig but they had a sequence of ≈ 70 000 bp between them, resulting from an inversion rearrangement of the genome (**Figure 5.6**). The *comC* gene was complete and its sequence was equal to *comC*-2.1 allele (*see Chapter 4*, **Figure 4.3**). However, downstream of *comC*, the first 48 bp of *comD* were identified but then this gene was truncated, appearing the rest of it upstream of the insertion sequence ISSpn8 and the *fcsR* gene (**Figure 5.6**). ISSpn8 was identified in the last chapter in two *comE* alleles (*see 4.2.4 Gene comE*). The primer CSP2-rv (**Figure 5.1**) was located in the *comD* region that was separated from *comC*, explaining why pherotype determination by PCR was unsuccessful. The inverted region (≈ 70 000 bp) between *comC* and *fcsR* in this ST393 isolate was similar to other pneumococcal isolates, maintaining the gene synteny from *comD* to *adcA*. The observed differences that are worth mentioning were: i) the presence of a direct repeat (DR) sequence in both *comD* extremities in the truncated zones; ii) the presence of a sequence of 26 bp with just a single SNP regarding the inverted repeat left (IRL) of ISSpn8, corresponding to the start of this IS; and iii) the presence of ISSpn8 between *comD* and *fcsR* (**Figure 5.6**). This suggests that the first event was the insertion of ISSpn8 within *comD* forming the DR sequences and then a recombination event promoted by this IS originated the inversion rearrangement. However, these preliminary data need to be confirmed by sequencing additional serotype 25A/38-ST393 isolates, including long read technologies which could provide further support for the detected rearrangement.

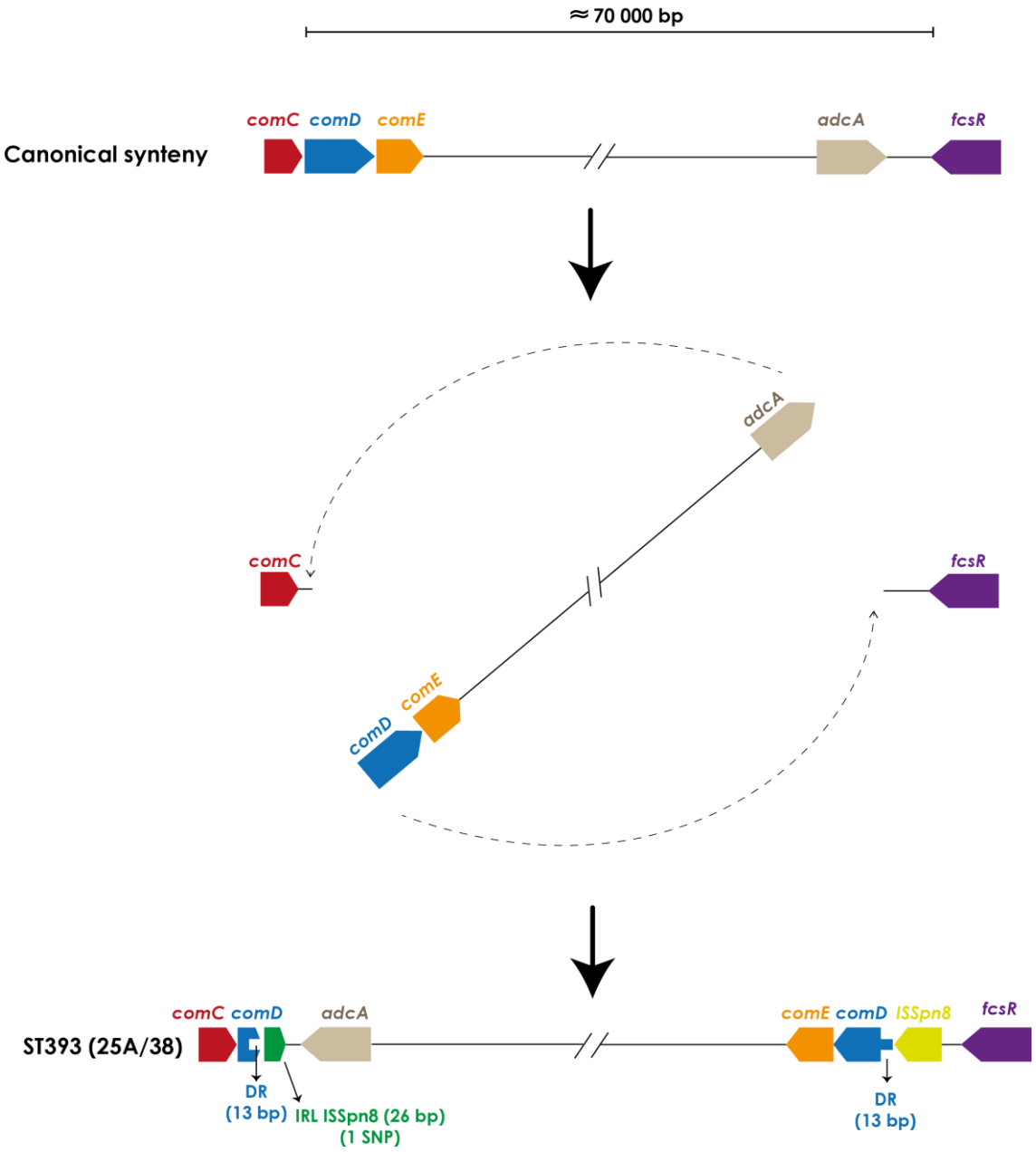


Figure 5.6 – Inversion of a 70 000 bp region between *comC* and *comD* genes in a serotype 25A/38-ST393 isolate. DR – direct repeat; IRL – inverted repeat left; SNP – single nucleotide polymorphism.

5.2.3. Serotype and pherotype

A total of 46 serotypes were identified among the isolates. However, serotypes 1, 19A, 7F, 14 and 23F accounted for 61.4 % (n=554) of the population (**Figure 5.7**). Serotypes 1, 7F, 14 and 23F were associated with pherotype CSP1 (FET p<0.0001: 1, 7F and 14; p=0.0015: 23F; all significant after FDR), serotypes 6A, 9N, 10A, 19A, 19F, 24A, and 24F with pherotype CSP2 (FET p<0.0001: 6A, 9N, 10A, 19A, 19F

and 24F; p=0.0020: 24A; all significant after FDR) and serotype 19A with pherotype CSP3 (FET p<0.0001, significant after FDR). Regarding vaccine serotypes, PCV7 serotypes (4, 6B, 9V, 14, 18C, 19F and 23F, n=229) were associated with pherotype CSP1 [n=193 (84.3 %), FET p=0.0003, significant after FDR] and NVT (n=182) with CSP2 [n=57 (31.3 %), FET p=0.0004, significant after FDR]. However, an association with a pherotype was not found for PCV13 exclusive serotypes (1, 3, 5, 6A, 7F and 19A, n=492) possibly because of the inclusion of 19A, associated to both CSP2 and CSP3.

The pherotype proportion of each serotype remained stable over time, despite the changes that occurred in serotype distribution. PCV7 serotypes decreased from 56.3 % to 19.2 %, while NVT increased from 10.4 % to 46.6 % between 1999-2002 and 2012 (CA p<0.0001, significant after FDR for both cases) (**Figure 5.8a**). PCV13 exclusive serotypes increased from 33.3 % to 60.7 % between 1999-2002 and 2009 (CA p<0.0001, significant after FDR) but then decreased to 34.2 % between 2009 and 2012 (CA p=0.0001, significant after FDR). The PCV7 serotypes 9V, 14 and 23F presented the most significant decreases between 1999 and 2012 (CA p<0.0001: 9V, 23F; p=0.0023: 14; all significant after FDR) (**Figure 5.8b**). The PCV13 serotypes 1, 7F and 19A increased between 1999 and 2009 (CA p=0.0410, p=0.0075 and p=0.0085, respectively, only supported before FDR correction) but between 2009 and 2012 serotypes 1 and 19A decreased (CA p=0.0027 and p=0.0313, only supported before FDR correction), while serotype 7F remained constant (**Figure 5.8c**). Regarding NVT, serotypes 23B and 24F increased between 1999 and 2012 (CA p=0.0170 and p=0.0238, respectively, only supported before FDR correction) while the increase of serotype 10A was noted only between 2009 and 2012 (CA p=0.0011, significant after FDR) (**Figure 5.8d**).

In the section *5.2.1. Pherotype abundance and evolution* it was seen that pherotype CSP1 proportion increased with age while pherotype CSP2 decreased. Analyzing the relation of serotype with age group, serotype 19A decreased from 22.6 % to 6.6 % between 0-11 months and 5-17 years age groups (CA p<0.0001, significant after FDR) and serotype 23F decreased from 8.4 % to 2.7 % between the same age groups (CA p=0.0022, significant after FDR) (**Figure 5.5**, serotype 23F data not shown). However, serotype 1 increased from 4.6 % in 0-11 months to 42.0 % in 5-17 years (CA p<0.0001, significant after FDR) (**Figure 5.5)**. Previously it was seen that serotypes 1 and 23F were associated with pherotype CSP1 while serotype 19A was associated with pherotypes CSP2 and CSP3. When the data of

serotypes 1 and 19A was removed from the analysis, the proportions of pherotypes CSP1 and CSP2 remained constant between the age groups, suggesting that the association of serotypes 1 and 19A with older and younger age groups, respectively, was the determining factor of the association observed between pherotype and age.

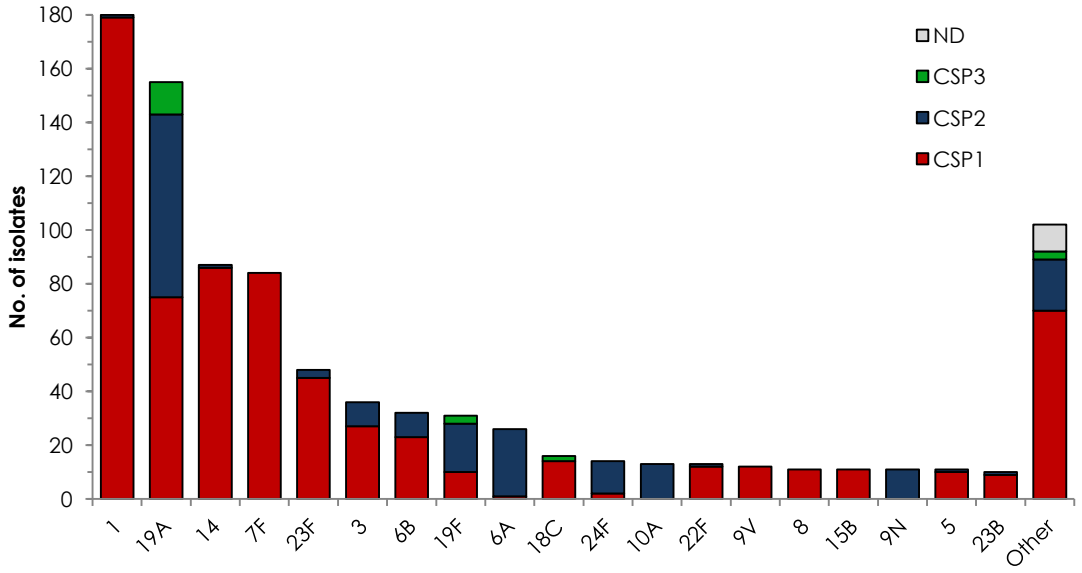


Figure 5.7 –Pherotype proportion within serotype. ND – not determined. Other – serotypes with <10 isolates (CSP1, n=9: 15A; n=8: 33A; n=7: 11A, 15C; n=6: 20, 21, 33F; n=3: 4, 16F; n=2: 13, 29, 35F, NT; n=1: 6C, 12B, 17F, 18B, 23A, 28A, 35C/42; CSP2, n=4: 24A; n=3: 35F; n=2: 6C, 12B, 20, 34; n=1: 29, 31, 37, 45; CSP3, n=2: 16F; n=1: 7C; ND: n=9: 25A/38; n=1: 6C).

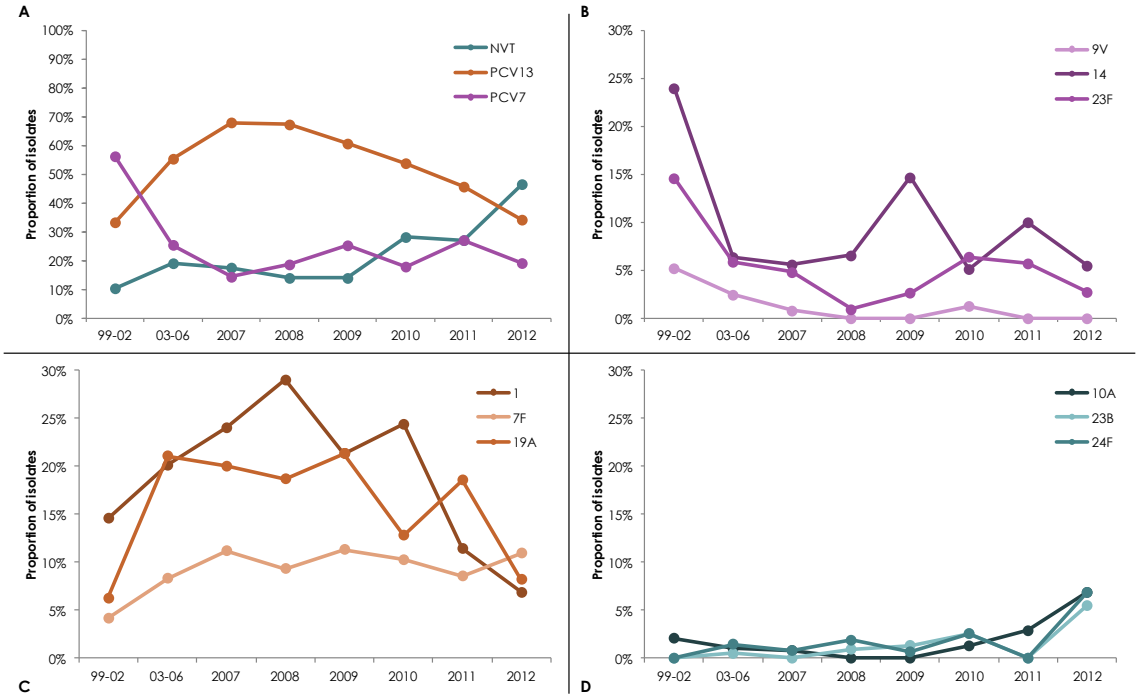


Figure 5.8 – Evolution of (A) vaccine serotypes, (B) selected PCV7 serotypes, (C) selected PCV13 serotypes and (D) selected non-vaccine serotypes (NVT) in invasive isolates recovered from children between 1999 and 2012. It is only shown the serotypes where significant changes were observed. PCV13 data included only serotypes 1, 3, 5, 6A, 7F and 19A. PCV7 and PCV13 were available in children vaccination since 2001 and 2010, respectively.

5.2.4. Sequence type and pherotype

All isolates were characterized by MLST and PHYLOViZ was used to define CCs at the SLV level taking in account all STs deposited in the pneumococcal MLST database (downloaded on 17th October 2016). A total of 210 STs were identified and distributed in 59 CCs and 12 singletons (**Table 5.3**, **Figure 5.9**). However, five genetic lineages accounted for more than half of the isolates (n=529, 58.6 %): CC306 (n=179, 19.8 %), CC156 (n=150, 16.6 %), CC191 (n=86, 9.5 %), CC230 (n=66, 7.3 %) and CC177 (n=48, 5.3 %).

Analyzing the association of pherotype and CC, it was observed that CC306, CC156, CC191, CC199, CC180 and CC1151 were associated with pherotype CSP1 (FET p<0.0001: CC306, CC156, CC191; p=0.0003: CC199; p=0.0009: CC180; p=0.0014: CC1151; all significant after FDR), while CC230, CC81, CC460, CC994, CC315, CC473, CC260 and CC395 were associated with pherotype CSP2 (FET p<0.0001: CC230, CC81, CC460, CC994, CC473, CC260, CC395; p=0.0016: CC315; all significant after FDR) and CC15, CC994 and CC102 were associated with

pherotype CSP3 (FET p<0.0001: CC15, CC994; p=0.0005: CC102; all significant after FDR). Most STs presented a single pherotype but there were eight exceptions: ST306 (n=150 CSP1, n=1 CSP2), ST994 (n=12 CSP2, n=6 CSP3), ST143 (n=9 CSP1, n=1 CSP2), ST439 (n=7 CSP1, n=1 CSP2), ST273 (n=5 CSP1, n=1 CSP2), ST30 (n=4 CSP1, n=1 CSP2), ST4577 (n=1 CSP1, n=1 CSP2) and ST1012 (n=1 CSP1, n=1 CSP2). However, it was common that a clonal complex presented two pherotypes, although a dominant pherotype could generally be identified, with the exception of CC177 (**Table 5.3**, **Figure 5.9**).

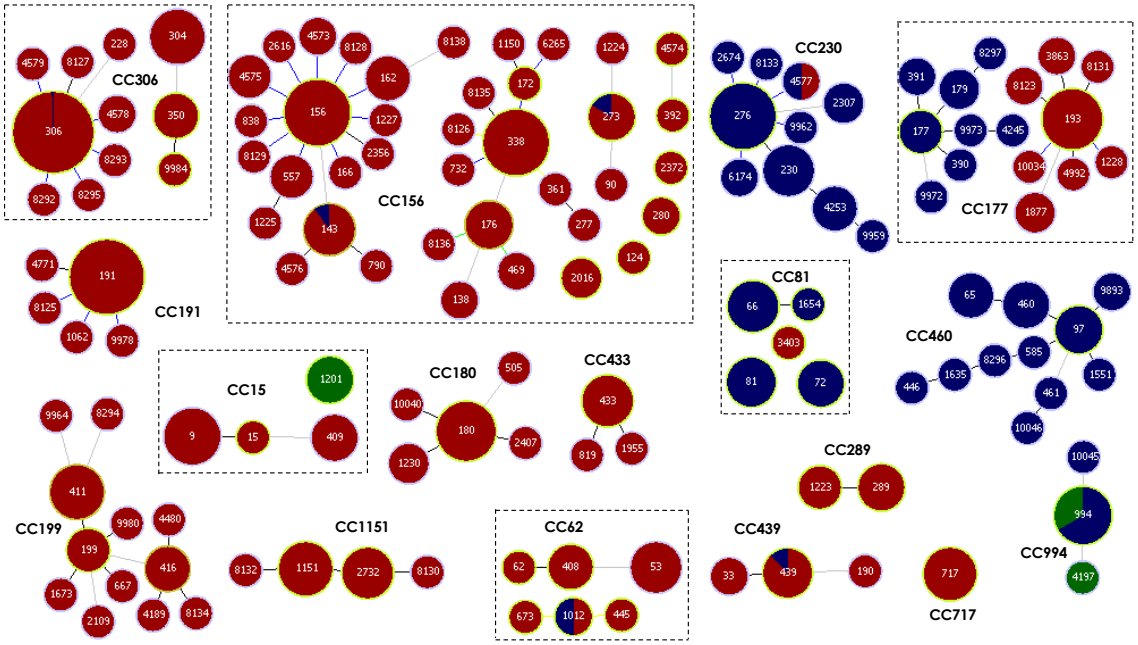
Similarly to serotypes, changes in the distribution of clonal complexes occurred over time. Between 1999-2002 and 2012, CC156 decreased from 42.7 % to 16.4 % (CA p<0.0001, significant after FDR), while CC191 increased from 4.2 % to 11.0 % in the same period (CA p=0.0175, only supported before FDR) (**Figure 5.10**). The CC306 and CC230 increased from 14.6 % and 1.0 % in 1999-2002 to 21.3 % and 7.3 % in 2009, respectively (CA p=0.0410 and p=0.0067, respectively, only supported before FDR) but then CC306 decreased from 21.3 % to 5.5 % between 2009 and 2012 (CA p=0.0012, only supported before FDR), while CC230 remained stable after 2009. A decrease of the CC15 from 5.3 % to 0.0 % was noted only in the period between 2009 and 2012 (CA p=0.0360, only supported before FDR).

Table 5.3 – Clonal complexes and sequence types identified in the invasive isolates recovered from children between 1999 and 2012.

CC	ST	Pherotype	Serotype	PMEN clone
CC306 (n=179)	306 (n=151); 304 (n=14); 350 (n=5); 228, 4578, 4579, 8127, 8292, 8293, 829, 9984 (n=1 each)	CSP1 (n=178) CSP2 (n=1)	1 (n=179)	Sweden^{1}-ST304 Sweden^{1}-ST306
CC156 (n=150)	156, 338 (n=39 each); 143 (n=10); 176 (n=7); 273 (n=6); 162 (n=5); 557, 4573 (n=4 each); 2016 (n=3); 138, 280, 469, 4573 (n=2 each); 90, 124, 166, 172, 277, 361, 392, 732, 790, 838, 1150, 1224, 1225, 1227, 2356, 2372, 2616, 4576, 6265, 8126, 8128, 8129, 8135, 8136, 8138 (n=1 each)	CSP1 (n=148) CSP2 (n=2)	14 (n=62); 23F (n=41); 6B (n=23); 9V (n=12); 19A, 19F, 23B, 24F (n=2 each); 6C, 15B, 17F, 21 (n=1 each)	S. Africa^{19A}-ST75 Spain^{6B}-ST90 Netherlands^{14}-ST124 Spain^{9V}-ST156 Poland^{23F}-ST173 S. Africa^{6B}-ST185 Hungary^{19A}-ST268 Finland^{6B}-ST270 Greece^{6B}-ST273 Colombia^{23F}-ST338 Maryland^{6B}-ST384
CC191 (n=86)	191 (n=82); 1062, 4771, 8125, 9978 (n=1 each)	CSP1 (n=86)	7F (n=84); 8, NT (n=1 each)	Netherlands^{7F}-ST191
CC230 (n=66)	276 (n=40); 230 (n=10); 4253 (n=5); 2307 (n=3); 4577, 6174 (n=2 each); 2674, 8133, 9959, 9962 (n=1 each)	CSP1 (n=1) CSP2 (n=65)	19A (n=51); 24F (n=8); 24A (n=4); 19F (n=3)	Denmark^{14}-ST230
CC177 (n=48)	193 (n=23); 177 (n=4); 179, 1877 (n=3 each); 391, 3863, 8123 (n=2 each); 390, 1228, 4245, 4992, 8131, 8297, 9972, 9973, 10034 (n=1 each)	CSP1 (n=34) CSP2 (n=14)	19A (n=26); 19F (n=16); 21 (n=3); 15A, 15B, 15C (n=1 each)	Portugal^{19F}-ST177 Greece^{21}-ST193
CC15 (n=30)	9 (n=17); 409, 1201 (n=6 each); 15 (n=1)	CSP1 (n=24) CSP3 (n=6)	14 (n=24); 19A (n=5), 7C (n=1)	England^{14}-ST9 Spain^{14}-ST18 CSR^{14}-ST20
CC199 (n=29)	411 (n=13); 416 (n=5); 199 (n=4); 667, 1673, 2109, 4189, 4480, 8134, 9980 (n=1 each)	CSP1 (n=29)	19A (n=16); 15B (n=8); 15C (n=4); 18C (n=1)	Netherlands^{15B}-ST199
CC81 (n=27)	66 (n=10); 81 (n=8); 72 (n=6); 1654, 3403, 4574 (n=1 each)	CSP1 (n=2) CSP2 (n=25)	9N (n=11); 19A (n=5); 24F (n=4); 23F (n=3); 6A, 13, 19F, 20 (n=1 each)	Tennessee^{14}-ST67 Spain^{23F}-ST81
CC180 (n=27)	180 (n=21); 1230 (n=3); 505, 2407, 10040 (n=1 each)	CSP1 (n=27)	3 (n=26); 14 (n=1)	Netherlands^{3}-ST180
CC460 (n=27)	97 (n=7); 460 (n=6); 65 (n=5); 9893 (n=2); 446, 460, 461, 585, 1551, 1635, 8296, 9893, 10046 (n=1 each)	CSP2 (n=27)	10A (n=13); 6A (n=10); 35F (n=3); 6B (n=1)	
CC1151 (n=25)	1151 (n=13); 2732 (n=10); 8130, 8132 (n=1 each)	CSP1 (n=25)	19A (n=23); 19F (n=2)	
CC62 (n=20)	53 (n=10); 408 (n=5); 1012 (n=2); 62, 445, 673 (n=1 each)	CSP1 (n=19) CSP2 (n=1)	8 (n=10); 11A (n=6); 22F (n=2); 33F, NT (n=1 each)	Netherlands^{8}-ST53
CC994 (n=19)	994 (n=18); 4197 (n=1)	CSP2 (n=12) CSP3 (n=7)	19A (n=19)	
CC433 (n=12)	433 (n=10); 819, 1955 (n=1 each)	CSP1 (n=12)	22F (n=11); 35C/42 (n=1)	
CC717 (n=12)	717 (n=12)	CSP1 (n=12)	33A (n=8); 33F (n=4)	
CC439 (n=11)	439 (n=8); 33 (n=2); 190 (n=1)	CSP1 (n=10) CSP2 (n=1)	23B (n=8); 23F (n=2); 23A (n=1)	Tennessee^{23F}-ST37
CC289 (n=10)	289 (n=6); 1223 (n=4)	CSP1 (n=10)	5 (n=10)	Columbia^{5}-ST289
CC393 (n=9)	393 (n=9)	ND (n=9)	25A/38 (n=9)	
CC63 (n=8)	63 (n=6); 1621, 4268 (n=1 each)	CSP1 (n=8)	15A (n=6); 15C, 19A (n=1 each)	Sweden^{15A}-ST63
CC315 (n=8)	386 (n=3); 315, 887 (n=2 each); 9985 (n=1)	CSP1 (n=1) CSP2 (n=6) ND (n=1)	6B (n=6); 6C (n=2)	Poland^{6B}-ST315
CC113 (n=7)	113 (n=3); 1766 (n=2); 123, 1073 (n=1 each)	CSP1 (n=5) CSP2 (n=2)	18C (n=4); 19A, 20, 31 (n=1 each)	Netherlands^{18C}-ST113
CC473 (n=7)	1876 (n=4); 473, 5679, 9988 (n=1 each)	CSP2 (n=7)	6A (n=7)	
CC1381 (n=7)	1233 (n=6); 1381 (n=1)	CSP1 (n=7)	18C (n=6); 18B (n=1)	
CC260 (n=6)	260 (n=4); 1220, 8124 (n=1 each)	CSP2 (n=6)	3 (n=6)	
CC395 (n=6)	327 (n=4); 395 (n=2)	CSP2 (n=6)	6A (n=5); 6C (n=1)	Portugal^{6A}-ST327
CC30 (n=5)	30 (n=5)	CSP1 (n=4) CSP2 (n=1)	16F (n=3); 5, 21 (n=1 each)	
CC320 (n=5)	320 (n=3); 4251, 8139 (n=1 each)	CSP1 (n=5)	19A (n=3); 19F (n=2)	Taiwan^{19F}-ST236

Table 5.3 – *(Continued).*

CC	ST	Pherotype	Serotype	PMEN clone
CC235 (n=3)	235 (n=2); 10047 (n=1)	CSP1 (n=3)	20 (n=3)	
CC1026 (n=3)	1026 (n=3)	CSP1 (n=3)	20 (n=3)	
CC1221 (n=3)	1221 (n=3)	CSP1 (n=3)	4 (n=3)	
CC102 (n=2)	102 (n=2)	CSP3 (n=2)	18C (n=2)	
CC198 (n=2)	198 (n=2)	CSP1 (n=2)	29 (n=2)	
CC378 (n=2)	232 (n=2)	CSP2 (n=2)	3 (n=2)	
CC458 (n=2)	458, 2633 (n=1 each)	CSP1 (n=2)	3, 33F (n=1 each)	
CC1368 (n=2)	1368 (n=2)	CSP1 (n=2)	35F (n=2)	
CC1650 (n=2)	1650 (n=2)	CSP1 (n=2)	18C (n=2)	
CC2669 (n=2)	2669 (n=2)	CSP1 (n=2)	19A (n=2)	
CC99 (n=1)	99 (n=1)	CSP1 (n=1)	11A (n=1)	
CC217 (n=1)	3081 (n=1)	CSP1 (n=1)	1 (n=1)	Sweden^{1}-ST217
CC218 (n=1)	218 (n=1)	CSP1 (n=1)	12B (n=1)	Denmark^{12F}-ST218
CC251 (n=1)	251 (n=1)	CSP3 (n=1)	19F (n=1)	
CC342 (n=1)	1371 (n=1)	CSP1 (n=1)	23F (n=1)	
CC346 (n=1)	1706 (n=1)	CSP1 (n=1)	15B (n=1)	
CC347 (n=1)	347 (n=1)	CSP1 (n=1)	19F (n=1)	
CC432 (n=1)	432 (n=1)	CSP1 (n=1)	21 (n=1)	
CC447 (n=1)	447 (n=1)	CSP2 (n=1)	37 (n=1)	
CC476 (n=1)	476 (n=1)	CSP3 (n=1)	19F (n=1)	
CC546 (n=1)	494 (n=1)	CSP1 (n=1)	28A (n=1)	
CC558 (n=1)	558 (n=1)	CSP2 (n=1)	29 (n=1)	Utah^{35B}-ST377
CC912 (n=1)	4248 (n=1)	CSP2 (n=1)	6A (n=1)	
CC989 (n=1)	989 (n=1)	CSP2 (n=1)	12B (n=1)	
CC1046 (n=1)	1046 (n=1)	CSP2 (n=1)	34 (n=1)	
CC1262 (n=1)	9975 (n=1)	CSP1 (n=1)	15C (n=1)	
CC1778 (n=1)	1778 (n=1)	CSP2 (n=1)	34 (n=1)	
CC2186 (n=1)	8322 (n=1)	CSP1 (n=1)	15A (n=1)	
CC2402 (n=1)	2402 (n=1)	CSP3 (n=1)	16F (n=1)	
CC2831 (n=1)	2831 (n=1)	CSP2 (n=1)	45 (n=1)	
CC3324 (n=1)	3324 (n=1)	CSP2 (n=1)	6A (n=1)	
CC5067 (n=1)	28 (n=1)	CSP1 (n=1)	18C (n=1)	
Singleton	923 (n=1)	CSP1 (n=1)	13 (n=1)	
Singleton	1646 (n=1)	CSP2 (n=1)	3 (n=1)	
Singleton	1648 (n=1)	CSP2 (n=1)	6A (n=1)	
Singleton	1662 (n=1)	CSP2 (n=1)	6B (n=1)	
Singleton	2228 (n=1)	CSP3 (n=1)	19F (n=1)	
Singleton	8060 (n=1)	CSP2 (n=1)	12B (n=1)	
Singleton	8122 (n=1)	CSP1 (n=1)	23F (n=1)	
Singleton	8137 (n=1)	CSP2 (n=1)	6B (n=1)	
Singleton	8294 (n=1)	CSP1 (n=1)	19F (n=1)	
Singleton	9964 (n=1)	CSP1 (n=1)	15A (n=1)	
Singleton	9976 (n=1)	CSP3 (n=1)	16F (n=1)	
Singleton	10045 (n=1)	CSP2 (n=1)	19A (n=1)	



Red: CSP1, **Blue**: CSP2 and **Green**: CSP3.

Figure 5.9 – Distribution of pherotype among the main clonal complexes (≥10 isolates) of the pneumococcal isolates recovered from children between 1999 and 2012. Each circle represents an ST and the diameter represents its frequency in a logarithmic scale. Grey lines connect STs that are double locus variants (DLVs), while lines of other colors connect STs that are single locus variants (SLVs) according to the PHYLOViZ tie-break rule reached. Unconnected STs using only our isolates that belong to the same CC when using all STs deposited in the public database are presented in the same box.

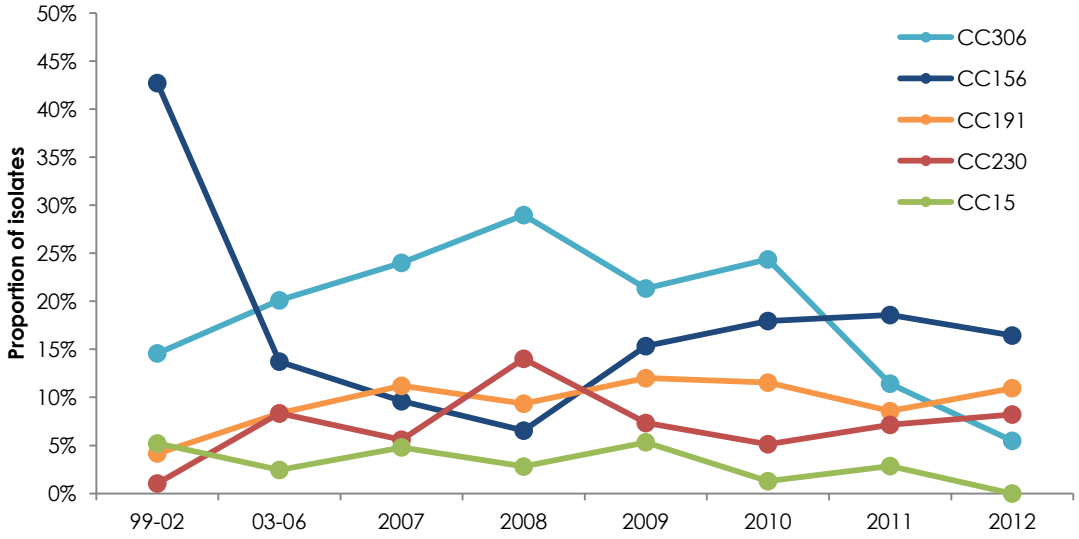


Figure 5.10 – Evolution of clonal complexes in invasive isolates recovered from children between 1999 and 2012. PCV7 and PCV13 were available in children vaccination since 2001 and 2010, respectively.

5.2.5. Antibiotic resistance and pherotype

The antibiotic susceptibility profile was determined for all isolates. Tetracycline presented the highest number of non-susceptible isolates (n=317, 35.1 %), followed by penicillin (n=237, 26.2 %), erythromycin (n=208, 23.0 %), co-trimoxazole (n=196, 21.7 %), clindamycin (n=171, 18.9 %), chloramphenicol (n=39, 4.3 %) and cefotaxime (n=29, 3.2 %) (**Figure 5.11**). All isolates were susceptible to levofloxacin, linezolid and vancomycin. Although the number of non-susceptible isolates of pherotype CSP1 was higher than other pherotypes, as expected given its higher abundance, it was observed that pherotype CSP2 was associated with non-susceptibility to most antibiotics (FET $p<0.0001$: penicillin, erythromycin, clindamycin, tetracycline and co-trimoxazole; $p=0.0002$: cefotaxime, all significant after FDR) with the exception of chloramphenicol (FET $p=0.1598$).

Resistance to antibiotics remained stable over time with the exception of the decrease in non-susceptibility to penicillin (from 41.7 % in 1999-2002 to 20.5 % in 2012, CA $p=0.0056$, significant after FDR), co-trimoxazole and chloramphenicol (from 33.3 % to 26.0 % and from 6.3 % to 0.0 %, CA $p=0.0290$ and $p=0.0363$, respectively, only supported before FDR) (**Figure 5.12**). The decrease in non-susceptibility to penicillin and co-trimoxazole was due to a decrease of CSP1 non-susceptible isolates to these antibiotics (CA, $p=0.0002$ and $p=0.0010$, respectively, both significant after FDR), while the decrease in resistance to chloramphenicol was due to a decrease of CSP2 non-susceptible isolates (CA, $p=0.0006$, significant after FDR).

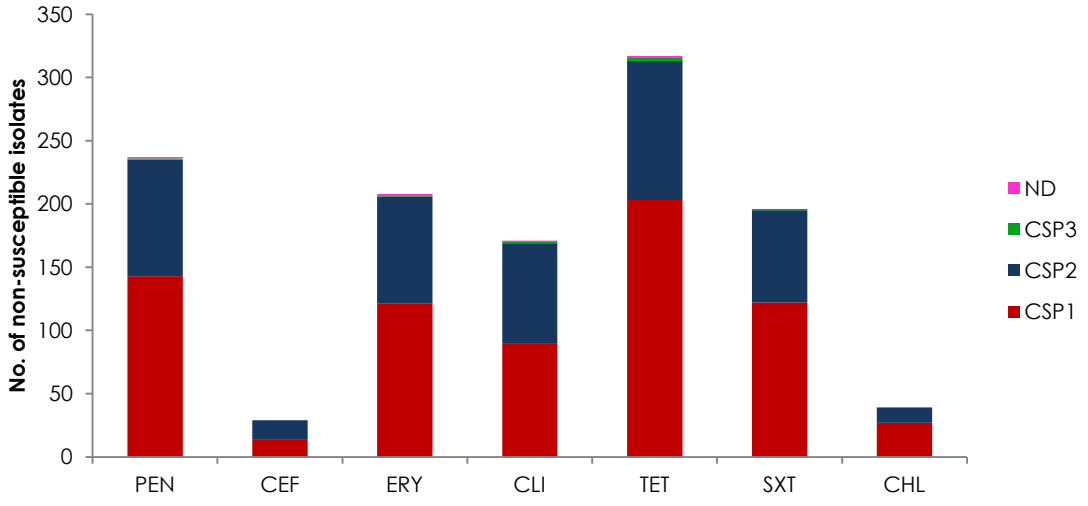


Figure 5.11 – Number of isolates of each pherotype among non-susceptible isolates. PEN: penicillin, CEF: cefotaxime, ERY: erythromycin, CLI: clindamycin, TET: tetracycline, SXT: co-trimoxazole, CHL: chloramphenicol. All isolates were susceptible to levofloxacin, linezolid and vancomycin.

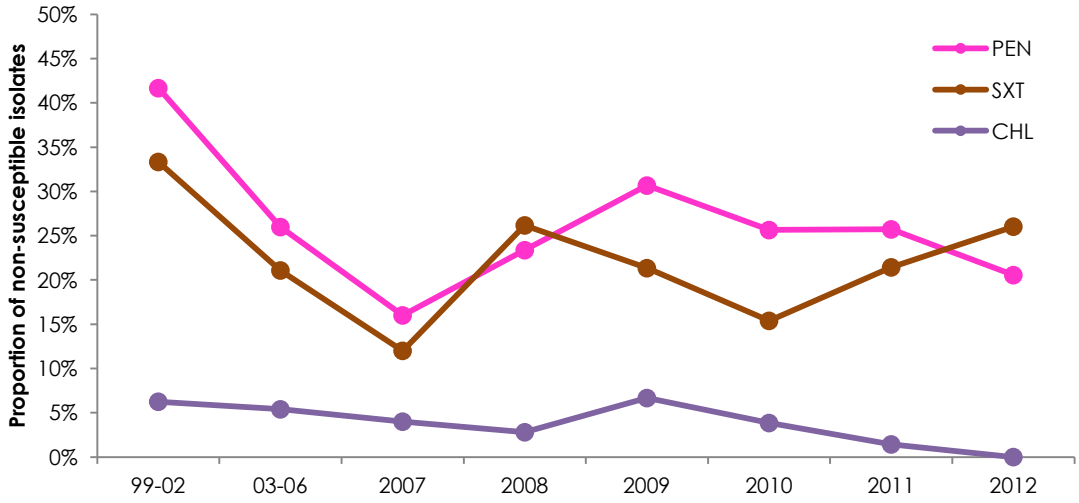


Figure 5.12 – Evolution of non-susceptibility to penicillin (PEN), co-trimoxazole (SXT) and chloramphenicol (CHL) in invasive isolates recovered from children between 1999 and 2012. PCV7 and PCV13 were available in children vaccination since 2001 and 2010, respectively.

5.2.6. Comparing partitions

The Simpson's index of diversity (SID) was calculated with the typing data of serotype, ST and CC for all the isolates and then for CSP1, CSP2 and CSP3 subpopulations (**Table 5.4**). ST presented the highest discriminatory power while serotype and CC presented similar SID values when analyzing all isolates.

Comparing CSP1 with CSP2 isolates, their SID of serotype, CC and ST were similar, however, the difference between serotype diversity of each pherotype was close to be statistically supported (p=0.0504). CSP3 isolates presented lower SID values but very large confidence intervals due to their small number.

Adjusted Wallace (AW) was used to evaluate the congruence between the typing methods and to assess the ability of each trait to predict pherotype (**Table 5.5**). ST and CC were the best typing methods to predict the type of CSP produced, reflecting the clonal property of pherotype. ST was also a good predictor of serotype and antibiotic susceptibility, with the exception of non-susceptibility to tetracycline and cefotaxime. The case of tetracycline could be because the transmission of resistance to this antibiotic is through transposons which can more easily transfer between isolates. Serotype presented lower AW values than ST, revealing the heterogeneity of some serotypes regarding their genetic background and antibiotic susceptibility.

Table 5.4 – Simpson's index diversity (SID) regarding serotype, ST and CC of the overall pneumococcal invasive isolates and of the CSP1, CSP2 and CSP3 isolates.

	Serotype	SID (95 % CI)[a]	ST	SID (95 % CI)[a]	CC	SID (95 % CI)[a]
Overall (n=903)	46	0.904 (0.894-0.914)	210	0.954 (0.946-0.962)	71	0.909 (0.899-0.919)
CSP1 (n=681)	37	0.879 (0.864-0.893)	140	0.925 (0.912-0.938)	44	0.859 (0.843-0.875)
CSP2 (n=192)	24	0.835 (0.794-0.877)	68	0.941 (0.919-0.964)	30	0.839 (0.800-0.877)
CSP3 (n=20)	5	0.626 (0.382-0.871)	9	0.837 (0.717-0.956)	8	0.805 (0.675-0.935)

[a]Jackknife pseudo-values 95 % confidence intervals.

Table 5.5 – Adjusted Wallace values to assess congruence between pherotype, serotype, sequence type (ST), clonal complex (CC) and susceptibility to antimicrobials. Values >0.70 are presented in bold.

	CSP	Serotype	ST	CC			
Serotype	0.467 (0.427-0.507)	-	0.432 (0.392-0.473)	0.628 (0.609-0.647)			
ST	**0.965 (0.924-1.000)**	**0.960 (0.943-0.977)**	-	**1.000 (1.000-1.000)**			
CC	**0.906 (0.865-0.946)**	0.670 (0.650-0.689)	0.480 (0.437-0.523)	-			
	PEN	CEF	ERY	CLI	TET	SXT	CHL
Serotype	0.363 (0.318-0.408)	0.000 (0.000-0.406)	0.275 (0.205-0.346)	0.213 (0.125-0.301)	0.162 (0.086-0.238)	0.237 (0.157-0.316)	0.000 (0.000-0.333)
ST	**0.878 (0.826-0.930)**	0.564 (0.419-0.708)	**0.899 (0.833-0.964)**	**0.878 (0.802-0.954)**	0.351 (0.242-0.460)	**0.724 (0.624-0.824)**	**0.799 (0.600-0.997)**
CC	0.533 (0.470-0.597)	0.000 (0.000-0.401)	0.508 (0.411-0.606)	0.512 (0.392-0.631)	0.277 (0.189-0.365)	0.308 (0.243-0.372)	0.551 (0.367-0.735)

5.3. Discussion

Pherotype abundance was determined in this chapter and three pherotypes were identified: CSP1 (n=681, 75.4 %), CSP2 (n=192, 21.3 %) and CSP3 (n=20, 2.2 %) (**Table 5.2**). Pherotype association with serotype, ST and antibiotic resistance was evaluated. PCV7 serotypes were associated with CSP1 while NVTs were associated with CSP2. However, despite the decrease of PCV7 serotypes and the increase of NVTs, pherotype abundance remained stable over 14 years. Most STs presented a single pherotype but it was common that a CC presented two pherotypes. Pherotype CSP2 was associated with non-susceptibility to most antibiotics and, together with the fact that NVTs were also associated with this pherotype, CSP2 abundance could potentially increase in the pneumococcal population. Pherotypes CSP1 and CSP2 are equally diverse and distributed through several genetic lineages. Pherotype CSP3 is rare but it was definitively identified as a third pneumococcal pherotype and it was present in a variety of serotypes and genetic backgrounds.

Over the 14 years covered by this study, when pneumococcal conjugate vaccines were introduced and serotype distribution was widely changed, pherotype proportions remained stable, although a small decrease was seen for pherotype CSP1 which was not statistically supported. The pherotype proportion within each serotype also remained stable over time. In all the studies that identified the pherotype of pneumococcal isolates, CSP1 was always the dominant pherotype over CSP2 and pherotype CSP3 was rarely seen, if at all (**Table 5.6**). More importantly, the abundance of pherotypes determined in all those collections was similar, with the exception of three of them (Evans and Rozen, 2013, Ramirez *et al.*, 1997, Whatmore *et al.*, 1999). It is possible that the process of selection of the isolates of those studies introduced some bias which resulted in different pherotype abundance. The studies included in **Table 5.6** represented temporal and geographically diverse isolates. Excluding the three studies mentioned above, CSP1 abundance ranged from 62.7 % to 76.3 %. The fact that pherotype abundance remained stable during the study period and that it was similar to other collections from other geographical sites and isolation dates, suggests that there may be a selective force which was responsible for maintaining CSP1 as the dominant pherotype. However, it could also be expected that CSP2 abundance would increase

with the pressure of vaccination and antimicrobial use because PCV7 serotypes were associated with pherotype CSP1, whereas NVT serotypes were associated with CSP2, and because CSP2 strains were associated with resistance to most antibiotics. The association of CSP2 strains with antimicrobial resistance was also observed in another Portuguese study with pneumococci isolated from nasopharyngeal samples of children attending day-care centers (Valente *et al.*, 2012). Nevertheless, it was seen that CSP1 strains formed biofilms that were thicker and denser than those formed by CSP2 strains and also that they yielded more transformants than CSP2 strains in both planktonic and biofilm growth (Carrolo *et al.*, 2014). Maybe the more robust biofilm formation and higher transformation efficiency of pherotype CSP1 strains in relation to CSP2 strains were the forces ensuring the dominance of CSP1 over CSP2, despite the potential contrary forces of antimicrobial consumption and vaccine use. The nature of the selective forces maintaining the 7:3 ratio of CSP1:CSP2 remain elusive but it is possible that some form of balancing selection is involved (Fijarczyk and Babik, 2015). It would be important to continue monitoring pherotype abundance now that PCV13 was included in the National Vaccination Plan to evaluate if pherotype abundance remains stable or there is an increase of CSP2 strains.

The association of CSP1 with older age groups was concluded to be mainly due to the association of serotypes 1 and 19A with IPD in older and younger children, respectively. However, the possibility that pherotype CSP1 strains have a higher capacity than other pherotypes to cause invasive infections in individuals with a robust immunologic system cannot be completely discarded. In fact, CSP1 strain abundance was observed to be slightly higher in IPD than in carriage (**Table 5.6**), supporting the hypothesis that CSP1 strains could have a higher invading capacity than CSP2.

Table 5.6 – Comparison of pherotype abundance between several studies.

Study	Collection	Age	n	CSP1 (%)	CSP2 (%)	CSP3 (%)
This study[a]	Invasive	Children	893[a]	681 (76.3)	192 (21.5)	20 (2.2)
(Carrolo *et al.*, 2009)	Invasive	All	483	341 (70.6)	142 (29.4)	Not found
(Valente *et al.*, 2012)	Colonization	Children	366	233 (63.7)	133 (36.3)	Not found
(Vestrheim *et al.*, 2011)[b]	Colonization	Children	150	94 (62.7)	56 (37.3)	Not found[b]
(Cornejo *et al.*, 2010)	Clinical	Adults	88	65 (73.9)	23 (26.1)	Not found
(Whatmore *et al.*, 1999)	Not described	Not described	58	29 (50.0)	27 (46.6)	2 (3.4)
(Evans and Rozen, 2013)	Colonization	3-36 months	54	53 (98.1)	Not found	1 (1.9)
(Ramirez *et al.*, 1997)	Clinical/colonization	Not described	50	24 (48.0)	25 (50.0)	1 (2.0)
(Pozzi *et al.*, 1996)	Clinical	All	43	30 (69.8)	13 (30.2)	Not found
(Iannelli *et al.*, 2005)	Clinical	All	17	11 (64.7)	6 (35.3)	Not found

[a]The 10 isolates with unidentified pherotype were excluded for this comparison.

[b]Two strains of this study did not amplify a PCR product and their pherotype was not determined. However, one of them was 19F-ST476 and in our collection we found a CSP3 isolate presenting the same serotype and ST, so maybe at least this isolate could be CSP3.

The genetic diversity of each pherotype was similar, as suggested by their SID values regarding ST, CC and serotype (**Table 5.4**). This means pherotypes were not only distributed through several genetic backgrounds, but they also presented the same diversity of serotypes, consistent with an ancient origin of each of the pherotypes in the pneumococcal population. Another characteristic of pherotype was that it corresponded to a clonal property which reflects the observation that ST and CC were the best typing methods to predict the type of CSP produced. Most STs presented a single pherotype, with eight exceptions which could have been the result of horizontal gene transfer including the *comCDE* locus. Regarding CC, a dominant pherotype could frequently be identified, with the exception of CC177, but it was quite common to find some isolates of a different pherotype grouped in the same CC suggesting that occasional horizontal gene transfer of the *comCDE* locus can generate stable lineages. In the previous chapter it was seen that indels and insertion sequences were present at the *comCDE* locus of some strains, so recombination involving this region could be an effective way to repair the competence regulating genes when these lose their functionality.

Pherotype CSP3 strains presented a total of 9 STs (ST102, ST251, ST476, ST994, ST1201, ST2228, ST2402, ST4197 and ST9976) distributed through 6 CCs (CC15, CC102, CC251, CC476, CC994 and CC2402) and 2 singletons (ST2228 and ST9976) and expressed a total of 5 serotypes (7C, 16F, 18C, 19A and 19F) (**Table 5.3**). This indicates that this pherotype was genetically diverse, despite its rarity in the pneumococcal population, suggesting it has not emerged recently in this population. As far as we know, this study provided the most accurate data concerning the abundance of pherotype CSP3 among pneumococci by identifying a total of n=20 isolates (2.2 %) of this pherotype in a collection of n=903 invasive pneumococcal isolates recovered from a defined geographic region. Thus, CSP3 was confirmed to be an established third pneumococcal pherotype in the population but found in a very low proportion, as was suggested by other studies that identified CSP3 strains (Evans and Rozen, 2013, Ramirez *et al.*, 1997, Whatmore *et al.*, 1999). The rarity of the CSP3 pherotype could be due to the fact that CSP3 competence regulating alleles are less efficient than the CSP1 and CSP2 alleles at triggering the competence response. Preliminary studies were performed to determine the sequence of the *comCDE* locus of most of the CSP3 strains identified here and also some of these strains were subject to transformation assays in the presence of

synthetic CSP3 peptide. These preliminary data identified at least 3 variants of the CSP3 peptide differing in the number of NFF amino acid repeats before the terminal triple arginine (1, 2 or 3 repeats, see **Figure 4.4**). The variants with 2 and 3 repeats were already identified in the previous chapter and in other studies (Ramirez *et al.*, 1997, Whatmore *et al.*, 1999) but these preliminary data could be the first report of a CSP3 peptide variant with just 1 repeat of the NFF amino acids. The transformation ability of some CSP3 strains tested in the presence of synthetic CSP3 showed the specific response of at least two strains to this peptide, with most of the strains tested being unable to yield transformants. Although not strong, this suggests that CSP3 is a bona fide competence inducer. However, it is possible that the CSP3 mature peptide varies in length, as opposed to CSP1 and CSP2 peptides that were restricted to exactly 17 amino acids. The variation of NFF amino acid repeats in CSP3 peptide variants could be a factor affecting the transformation efficiency of CSP3 strains and could explain in part the absence of response of many of the strains tested. However, formal testing of this hypothesis is required with the synthesis of peptides with variable numbers of the NFF repeat.

Regarding the case of serotype 25A/38-ST393 isolates, preliminary data suggests that an inversion rearrangement occurred in the genome of these isolates, resulting in the separation of *comC* from the *comDE* genes by an intervening sequence of approximately 70 000 bp. Genetic recombination events involve typically shorter sequences but recombination of sequences in the order of 70 000 bp in length had been also reported in the literature in *S. pneumoniae* (Croucher *et al.*, 2011, Vijayakumar *et al.*, 1986). This inversion could have been caused by the insertion sequence ISSpn8 but a comprehensive analysis including more isolates should be done to confirm this hypothesis. It is unknown if these strains can enter the competence state and therefore yield transformants. Preliminary studies of transformation, including the addition of excess synthetic CSP peptide, with these strains were performed and no transformants were obtained, suggesting that a loss of function of the competence-regulating genes could have occurred concomitantly with the inversion event. One possibility is that the expression of the ComD and ComE proteins, essential for sensing extracellular CSP, was abrogated.

Finally, comparing the results of this and the last chapters, the pherotype abundance identified in this work was very similar to the abundance observed in the pneumococcal sample studied in Chapter 4: CSP1 n=681, 75.4 %; CSP2 n=192, 21.3 %; and CSP3 n=20, 2.2 % vs. CSP1 n=66, 74.2 %; CSP2 n=21, 23.6 %; and CSP3 n=2, 2.2 %, respectively. This clearly showed that the sample used was representative of the Portuguese pneumococcal invasive population circulating in Portugal and it was another evidence that pherotype abundance remained stable because that sample was selected from a period when conjugate vaccines were still not introduced in children immunization.

III. Pherotype specificity and influence on the genetic structure

III. PHEROTYPE SPECIFICITY AND INFLUENCE ON THE GENETIC STRUCTURE

After studying the genetic diversity of the *comCDE* locus and determining the abundance of pherotypes, our next question was if pherotype has a role in the evolution of *S. pneumoniae*, particularly if it has an influence in shaping the genetic structure of this pathogen. To answer this question, we attempted to use X-ray crystallography to study the interaction between CSP and ComD by determining the 3D structure of this receptor and also of the receptor in complex with its ligand. Afterwards, we studied the clonal and serotype dynamics of serogroup 6 isolates causing IPD in Portugal during 1999-2012 and we used this work to also evaluate the influence of pherotype on the genetic recombination between serogroup 6 isolates.

III. Pherotype specificity and influence on the genetic structure

6. Structural study of pherotype specificity by X-ray crystallography

CHAPTER 6. STRUCTURAL STUDY OF PHEROTYPE SPECIFICITY BY X-RAY CRYSTALLOGRAPHY

This chapter presents the work performed during my visit to Dr. Juan Hermoso lab (Instituto de Química-Física "Rocasolano", CSIC, Madrid), a laboratory expert in X-ray crystallography.

The pneumococcal pherotypes exhibit different phenotypes. A previous study by our laboratory with mutants lacking the *comC* gene indicated that CSP1 strains were able to respond much better to their cognate peptide than CSP2 strains, resulting in a better capacity to form biofilms and a higher transformation efficiency of CSP1 in relation to CSP2 strains (Carrolo *et al.*, 2014). These results suggested either a stronger binding of CSP1 to its receptor or an amplification of the response in the CSP1 genetic background and could explain the higher prevalence of *comC1* allele in the pneumococcal population.

To determine the reasons for these differences, we wanted to study the pherotype specificity by X-ray crystallography. The goal was to solve the 3D structure of ComD variants in the absence of CSP and in complex with their cognate CSP. This approach would allow us to identify with atomic resolution which regions of the receptor interact with the pheromone peptide and are responsible for its specificity.

Determination of the 3D structure of a protein by X-ray crystallography requires several steps (**Figure 6.1**). The first bottle-neck in the process is the production of pure protein in sufficient quantity to perform protein crystallization. The strategies attempted to produce ComD protein will be described in this chapter, although due to its complexity, this goal was not achieved.

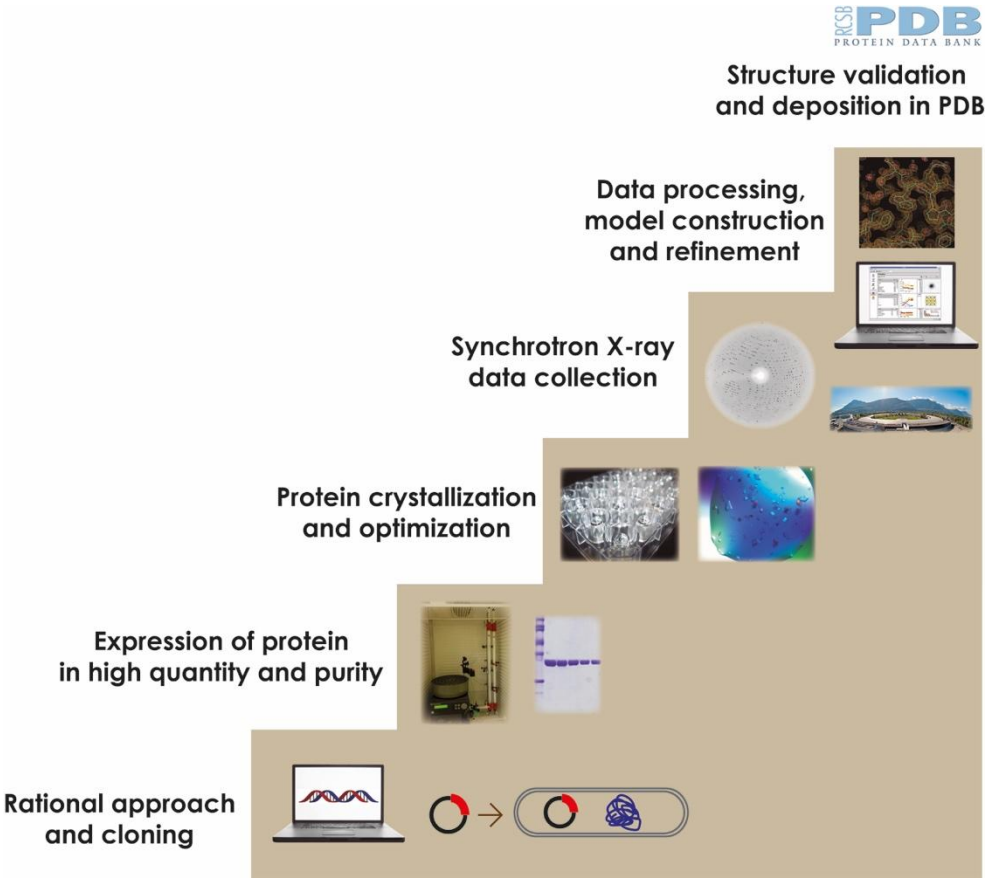


Figure 6.1 – Crystallography pipeline work depicted as a ladder. Each step represents a possible bottleneck in the process to solve a protein structure by X-ray crystallography.

6.1. Designing of ComD constructs

ComD is an integral membrane histidine kinase protein composed of two domains: an extracellular sensor and an intracellular kinase. Our aim was the crystallization of the soluble extracellular sensor domain instead of the full-length protein due to the high complexity of an integral membrane protein production. The reference strain R6 ComD sensor domain, which presents pherotype CSP1 and its sequence is equal to the *ComD*-1.1 variant (*see Chapter 4*, **Table 4.1**, **Figures 4.5** and **4.6**), was chosen to be produced. Although the sensor domain is reported to be placed in the N-terminus of ComD, very little is known about its extension and topology. Therefore, we ran an *in silico* simulation to predict the topology of ComD using the algorithm TopPred (Claros and von Heijne 1994; von Heijne 1992). Seven membrane helices were identified but only four were classified as certain (**Table 6.1**, **Figure 6.2**). Primers were designed (**Table 6.2**), considering these predictions, to generate several constructs summarized in **Table 6.3**. We decided to attempt the

production of constructs with and without the first predicted helix (from Met1 or His21, respectively) and until the third putative helix (Arg84). After the initial set of expression experiments, constructs until the fourth helix (Asn114) were also produced. The four constructs were named 1-84, 21-84, 1-114 and 21-114 (**Table 6.3**).

Table 6.1 – Amino acid residues and classification of membrane helices identified for ComD by the algorithm TopPred (Claros and von Heijne 1994; von Heijne 1992).

Helix	Begin-End	Score	Certainty[a]
1	1-21	1.510	Certain
2	32-52	0.803	Putative
3	85-105	0.645	Putative
4	119-139	1.856	Certain
5	160-180	1.422	Certain
6	186-206	2.296	Certain
7	379-399	0.603	Putative

[a]Helices were considered certain or putative for scores >1.00 and >0.60, respectively.

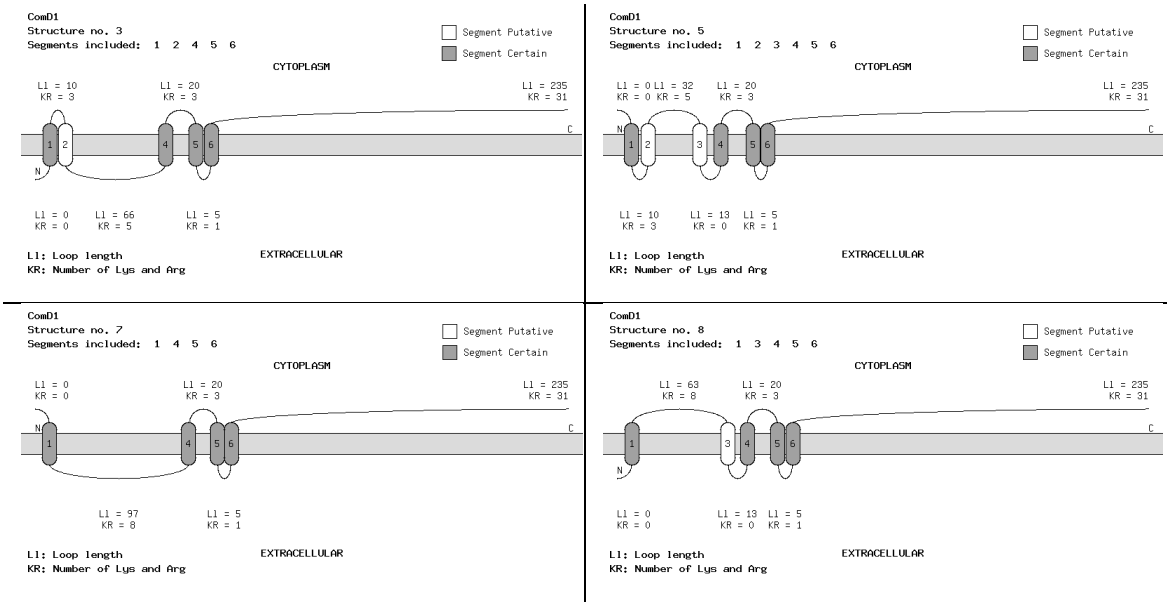


Figure 6.2 – Possible topologies identified with the algorithm TopPred (Claros and von Heijne 1994; von Heijne 1992). The topologies with the putative helix 7 are not shown because they showed an extracellular C-terminus.

Table 6.2 – Primers used to produce constructs of the ComD sensor domain.

Primer	Sequence (5'→ 3')[a]	Size (bp)
ComD FM1	GGTGAAAACCTGTATTTCCAGGGC**ATG**GATTTATTTGGATTTGGGACG	48
ComD FM21	GGTGAAAACCTGTATTTCCAGGGC**ATG**CATTTTATTTGTAAAGGTCAA	48
ComD RM84	AAGCTTAGTTAGCTATTATGCGTA**TCA**TCTATCCCTCTCATACTGTTTCATC	52
EcoRI Fw1	CCGAATTC**ATG**GATTTATTTGGATTTGGGACG	32
EcoRI Fw21	CCGAATTCCATTTTATTTGTAAAGGTCAA	29
HindIII Rv84	GGCCAAGCTT**TTA**TCTATCCCTCTCATACTGTTTCATC	38
HindIII Rv114	GGCCAAGCTT**TTA**ATTATCTCCTGTTATAGAGGA	34

[a]Underlined: sequence pairing the vector pURI3-TEV (ComD FM1, FM21 and RM84) and restriction site (EcoRI Fw1 and Fw21 and HindIII Rv84 and Rv114. **Bold:** start/stop codon.

Table 6.3 – Constructs of the ComD sensor domain produced during this work.

Construct	Begin-End	Vector	Primer Forward[a]	Primer Reverse[a]	M.W. (Da)[b]
1-84	Met1-Arg84	pURI3-TEV	ComD FM1	ComD RM84	12208.12[c]
21-84	His21-Arg84	pURI3-TEV	ComD FM21	ComD RM84	10060.62[c]
1-114	Met1-Asn114	pKLSLt	EcoRI Fw1	HindIII Rv114	32021.43[d]
21-114	His21-Asn114	pKLSLt	EcoRI Fw21	HindIII Rv114	29742.73[d]

[a]The sequence of the primers is presented in the Table 6.2.

[b]Molecular weight calculated with ProtParam (Gasteiger *et al.*, 2005) (accessed in http://web.expasy.org/protparam/protpar-ref.html)

[c]Calculated with His-tag and TEV recognition site. ComD 1-84 and ComD 21-84 predicted molecular weights were 10143.94 Da and 7865.24 Da, respectively.

[d]Calculated with LSL_{150} and TEV recognition site. ComD 1-114, ComD 21-114 and LSL_{150} predicted molecular weights were 13415.65 Da, 11136.95 Da and 17163.27, respectively.

6.2. Cloning of ComD sensor in vector pURI3-TEV

The plasmid pURI3-TEV (Curiel *et al.*, 2011) was chosen as the vector for the cloning of ComD sensor. This plasmid, harbouring the ampicillin resistance gene as a selective marker, can produce a recombinant protein with a His-tag in the N-terminus which can be later removed by using tobacco etch virus (TEV) protease. We decided to clone the constructs 1-84 and 21-84 (**Table 6.3**) in this vector using the primers ComD FM1, ComD FM21 and ComD RM84 (**Table 6.2**).

Due to its ease of use and high efficiency, the ligation independent cloning (LIC) methodology was chosen to introduce all targets into the vector. The LIC approach consisted in two consecutive PCRs: the first one amplifies the gene of interest introducing specific complementary overhangs adaptors for the desired plasmid. The second PCR takes advantage of the adaptors to perform the annealing between the first PCR product and the vector.

The conditions of the first PCR to amplify the insert were:

- 0.02 U/μL KOD Hot Start DNA polymerase (Toyobo, Osaka, Japan)
- 1x KOD buffer
- 0.2 mM dNTPs
- 1.5 mM $MgSO_4$
- 1 M Betaine
- 0.4 pmol/μL of each primer (**Table 6.2**)
- 4 ng/μL of R6 genomic DNA extracted through CTAB method (Wilson, 2001)

With the following PCR program:

- 5 min at 94 ºC
- 30x (1 min at 94 ºC, 45 s at 55 ºC and 2 min at 68 ºC)
- 5 min at 68 ºC

Agarose gel electrophoresis was used to confirm the amplification of the fragment of interest. The DNA fragment was purified by using the QuickClean II PCR Extraction Kit® (GenScript, Piscataway, USA) following the manufacturer's instructions. Cleaned DNA was used in the second PCR to generate vectors with the inserted gene using the following conditions:

- 0.02 U/μL KOD Hot Start DNA polymerase (Toyobo, Osaka, Japan)
- 1x KOD buffer,
- 0.2 mM dNTPs
- 1.5 mM $MgSO_4$
- 0.3 M Betaine
- 4 ng/μL vector pURI3-TEV
- 19 μL of the purified gene fragment from first PCR

With the following PCR program:

- 4 min at 95 ºC
- 25x (20 s at 98 ºC, 30 s at 55 ºC, 4 min at 68 ºC)

After the second PCR an additional digestion with 20 U of the enzyme *Dpn*I (New England Biolabs, Ipswich, USA) was performed in the buffer provided by the manufacter at 37 ºC overnight. This digestion eliminates the parent pURI3-TEV

plasmids because *Dpn*I recognizes and digests only methylated DNA. This was followed by a second digestion using 10 U of the enzyme *Not*I (New England Biolabs, Ipswich, USA) in the buffer provided by the manufacturer at 37 ºC during 6 h because the pURI3-TEV plasmid was engineered to be digested by *NotI* in the absence of an inserted fragment.

The product obtained after these two digestions was used to transform *Escherichia coli* DH5α competent cells created by the heat-shock method (Froger and Hall, 2007). PCR was used to identify the presence of the insert in the several colonies obtained after transformation. Briefly, transformed *E. coli* DH5α cells were boiled for 20 min at 98 ºC before addition to a PCR reaction which was performed under the following conditions:

- 0.1 U/µL *Taq* DNA polymerase (New England BioLabs, Ipswich, USA)
- 1x *Taq* buffer with $MgSO_4$
- 0.2 mM dNTPs
- 0.4 pmol/µL of each primer (**Table 6.2**)
- 5 µL of boiled cells

With the following PCR program:

- 4 min at 95 ºC
- 30x (30 s at 95 ºC, 1 min at 55 ºC and 3 min at 72 ºC)
- 10 min at 72 ºC

The plasmids from the positive colonies were extracted using QuickClean II Plasmid Miniprep Kit® (GenScript, Piscataway, USA) following the manufacturer's instructions and were sent for sequencing in Secugen (Madrid, Spain) to verify if the fragment sequence was correct and in frame.

In the end, we obtained the constructs 1-84 and 21-84 inserted into the vector pURI3-TEV (**Table 6.3**).

6.3. Expression tests of ComD sensor with a His-tag

The pURI3-TEV constructs 1-84 and 21-84 were inserted into the bacterial expression hosts *E. coli* BL21(DE3) and *E. coli* JM109(DE3), made competent by the heat shock method, for recombinant protein expression. However, only four

colonies were obtained for each of these expression hosts. This section describes the expression tests performed with these strains, which are summarized in **Table 6.4**.

Table 6.4 – Summary of the expression tests performed using the vector pURI3-TEV.

Construct	Expression host	Medium	Volume	Purification
1-84	*E. coli* BL21(DE3)	LB	200 mL	HisTrap HP® Ni-NTA/Superdex 75
1-84	*E. coli* BL21(DE3)	LB	2 L	HisTrap HP® Ni-NTA
1-84	*E. coli* JM109(DE3)	2xTY	1 L	HisTrap HP® Ni-NTA
21-84	*E. coli* JM109(DE3)	2xTY	1 L	HisTrap HP® Ni-NTA

6.3.1. Expression in E. coli BL21(DE3)

The construct 1-84 was used for the first small-scale expression test in *E. coli* BL21(DE3). The cells were inoculated in 20 mL of LB medium supplemented with ampicillin (100 µg/mL) followed by incubation at 37 ºC overnight with agitation. This pre-inoculum was diluted 1:100 in 200 mL of LB medium supplemented with ampicillin (100 µg/mL) and grown at 37 ºC with agitation until an OD_{600nm}=0.6. Protein production was induced at this point by the addition of 1 mM isopropyl-ß-D-thiogalactopyranoside (IPTG) to the culture followed by overnight incubation at 16 ºC with agitation. Cells were harvested by centrifugation and suspended in 10 mM Tris pH 7.5, 100 mM NaCl.

The step of protein purification started with the lysis of the cells by sonication (10 cycles of 30 s at maximum intensity with 10 s of interval, model SONOPULS HD 2200, Bandelin, Berlin, Germany). The soluble and insoluble fractions of the lysate were separated by centrifugation at 38000 RCF during 40 min at 4 ºC. Since the protein was engineered to include a six-histidine purification tag (His-tag), enrichment of the soluble fraction was performed by immobilized metal ion affinity chromatography (IMAC) using a 1 mL HisTrap HP® Ni-NTA column (GE HealthCare, Chicago, USA) and following the manufacturer's instructions. Elution was performed in 10 mM Tris pH 7.5, 100 mM NaCl with an isocratic gradient ranging from 0 to 1 M imidazole on a low pressure liquid chromatography system (BioLogic LP System®, Bio-Rad, Hercules, USA). The chromatogram of this purification presented a peak of 0.51 AU in the elution phase (**Figure 6.3a**). The result of this purification was analyzed by SDS-PAGE. Two weak protein bands with molecular weights of approximately 31 kDa and between 6.5 and 14.4 kDa were seen in the eluted fractions, although no overexpressed protein was observed in the lysate nor in the insoluble fraction (data not shown).

The predicted molecular weight of the construct 1-84 was 12.2 kDa (**Table 6.3**), so we thought the protein band with a molecular weight between 6.5 and 14.4 kDa could be the ComD sensor protein. For this reason, a gel-filtration chromatography was performed to further purify the putative protein. Sample volume was reduced to 5ml by using 10 kDa cutoff centrifugal concentration devices (Millipore) and injected into a superdex 75 column (GE HealthCare, Chicago, USA) attached to a fast protein liquid chromatography system (BioLogic DuoFlow®, Bio-Rad, Hercules, USA). Afterwards, a sample of molecular weight markers (Gel Filtration Standard®, Bio-Rad, Hercules, USA) was also passed through the column with the same conditions to allow the estimation of the molecular weights of the purified proteins. The chromatogram showed three protein species, two with more than 44 kDa and the third with less than 17 kDa but the quantity of this protein was much smaller than the other two (**Figure 6.4**).

A scale-up of the previous experiment was performed by repeating the same procedures. In this experiment, 2 L of cell culture were induced for protein overexpression. After centrifugation separation, Lysate supernatants were enriched by using a 5 mL HisTrap HP® Ni-NTA column (GE HealthCare, Chicago, USA). The resulting chromatogram was similar to that of the previous experiment but showed a peak of 1.1 AU in the elution phase. The result of this purification step was analyzed by SDS-PAGE but this time a band between 6.5 and 14.4 kDa was not seen in the fractions of the elution phase. As in the previous experiment, an overexpressed protein band was also not seen in the lysate or in the insoluble fraction. Therefore, the purification was terminated and the gel-filtration step was not performed.

6.3.2. Expression in E. coli JM109(DE3)

After performing the expression tests of construct 1-84 in *E. coli* BL21(DE3), the production of the ComD sensor was attempted in *E. coli* JM109(DE3) following the same general procedures. One liter of culture was used with both constructs, 1-84 and 21-84 for protein overexpression tests. The purification through the Ni-NTA column of constructs 1-84 and 21-84 yielded peaks of 0.94 and 0.54 AU (**Figure 6.3b**), respectively. The result of this purification step was evaluated by SDS-PAGE but it was similar to the result obtained with *E. coli* BL21(DE3). A protein band between 6.5 and 14.4 kDa molecular weight was not observed in the eluted

sample of both constructs and an overexpressed protein band was not detected in any fraction (**Figure 6.5**).

Production of ComD sensor with an N-terminus His-tag was not achieved and a new strategy needed to be devised.

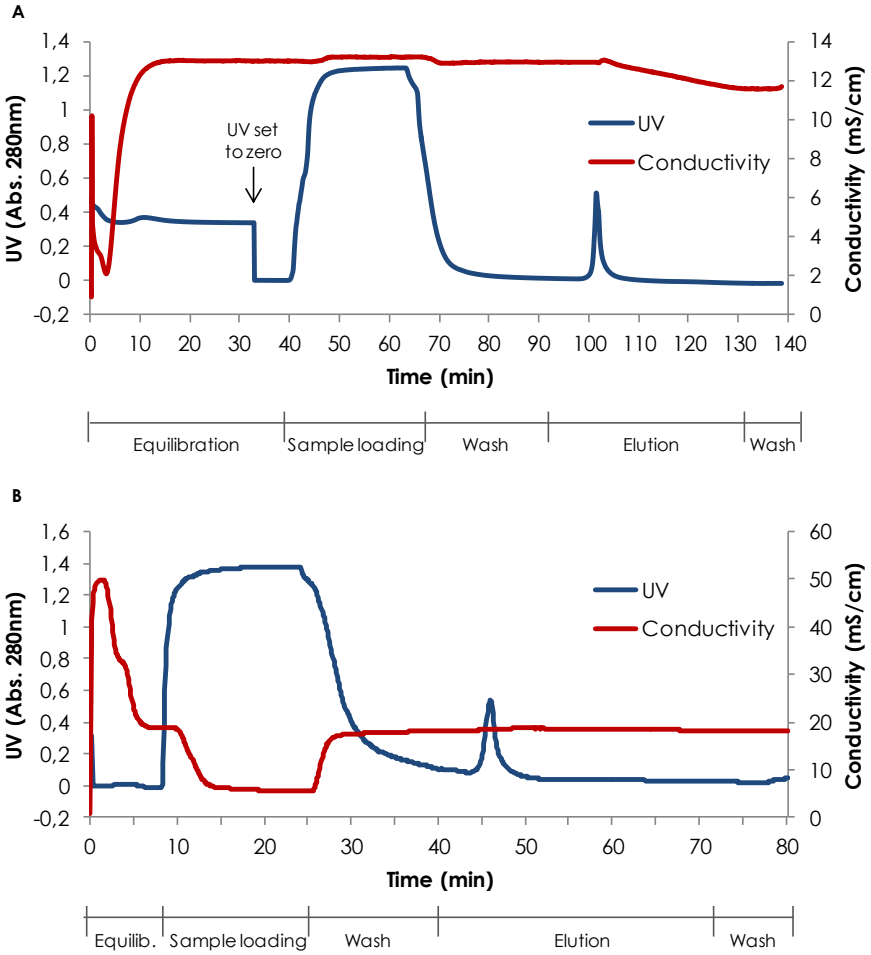


Figure 6.3 – Chromatograms of the purifications through a Ni-NTA column of (A) the construct 1-84 expressed in *E. coli* BL21(DE3) and (B) the construct 21-84 expressed in *E. coli* JM109(DE3).

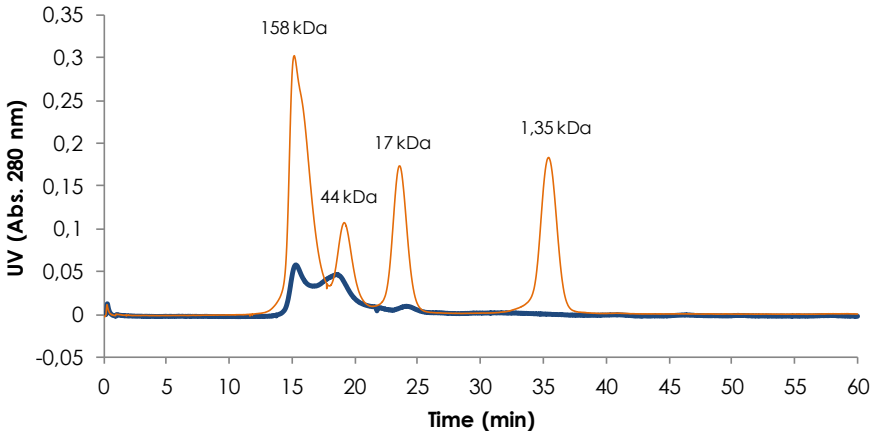


Figure 6.4 – Chromatogram of the gel-filtration of the protein sample (blue line) purified through a Ni-NTA column from the construct 1-84 expressed in *E. coli* BL21(DE3). The molecular weights presented were the weights of the proteins from the standard markers (orange line).

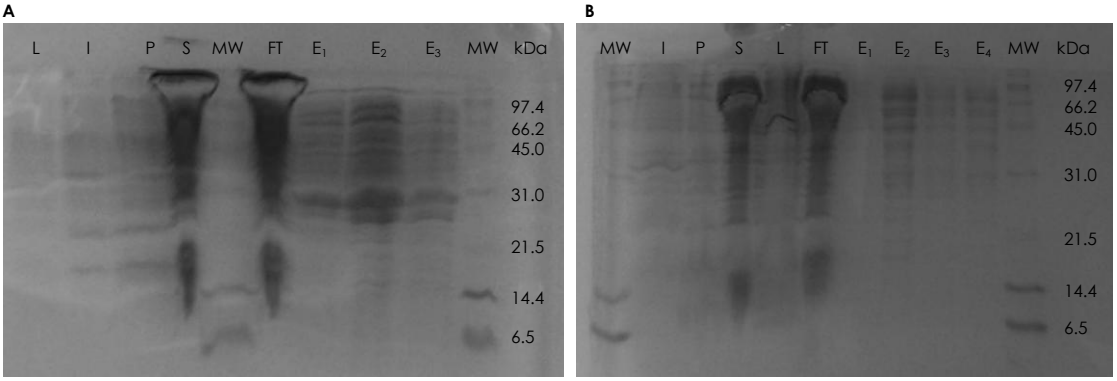


Figure 6.5 - SDS-PAGE 17.5 % of the purification through a Ni-NTA column of (A) the construct 1-84 and (B) the construct 21-84, both expressed in *E. coli* JM109(DE3). L – lysate, I – insoluble fraction, P – precipitate of DNA, S – soluble fraction, FT – flow-through after passing the protein sample through a Ni-NTA column, E_1 to E_4 – elution fractions of the peak, MW – molecular weight markers (Molecular Weight Standards Broad Range, Bio-Rad, Hercules, USA).

6.4. Cloning of the fusion protein LSL_{150}-ComD in vector pKLSLt

After the experiments to produce his-tagged recombinant ComD sensor, which were not satisfactory, another strategy was followed. We decided to try the production of ComD sensor fused to an affinity tag protein. Tag proteins often promote greater expression and solubility of recombinant proteins because they fold into a stable and highly soluble protein upon translation.

LSLa is a hemolytic lectin produced by the mushroom *Laetiporus sulphureus* and its N-terminus domain, LSL_{150}, was shown to be both a good affinity tag

protein and a solubility enhancer of other proteins (Angulo *et al.*, 2011). It was seen that LSL_{150} is an autonomous folding unit highly translated in *E. coli* BL21(DE3) that increase the solubility and the ratio of the target fused protein. Moreover, LSL_{150} behaved as an excellent affinity tag because of its binding specificity to lactose which enabled purification in a single-step using Sepharose 4B® columns (Angulo *et al.*, 2011).

The vector pKLSLt, which was kindly provided by Dr. José Miguel Mancheño and Dr. Iván Acebrón, produces the LSL_{150} domain followed by a linker sequence and a TEV cleavage site for tag removal (**Figure 6.6**). Constructs were inserted after the TEV cleavage site by using the restriction enzymes *Eco*RI and *Hind*III, whose recognition sequences were not present in the target proteins. At this point, the construct design was revisited and two more constructs terminating in the predicted fourth helix, 1-114 and 21-114, were decided to be produced in addition to constructs 1-84 and 21-84 (**Tables 6.1 and 6.3 and Figure 6.2**).

The four target proteins were amplified by PCR, following the same conditions used for the amplification of the inserts for the vector pURI3-TEV, with the exception that the primers used had the *Eco*RI and *Hind*III recognition sequences (**Table 6.2**), inserting these sequences in the N- and C-terminus of the fragments, respectively. The inserts and the vector pKLSLt were digested with *Eco*RI and *Hind*III (New England BioLabs, Ipswich, USA) at 37 ºC during 6 h. Digested DNA was further purified by using 0.75 % agarose gel and recovered by using QIAquick Gel Extraction Kit® (Qiagen, Venlo, Netherlands) following the manufacturer's instructions. Ligation of the digested products was performed using T4 ligase (New England BioLabs, Ipswich, USA) during 10 h at 16 ºC. The product of ligation was used to transform heat-shock competent *E. coli* DH5α cells. After various attempts, only 3 and 5 transformant colonies of constructs 1-114 and 21-114 were obtained, respectively. The presence of the inserts was confirmed by PCR following the same conditions used in the cloning of pURI3-TEV. The plasmids of a single positive colony of each construct were extracted using QuickClean II Plasmid Miniprep Kit® (GenScript, Piscataway, USA) following the manufacturer's instructions and sent to sequencing in Secugen (Madrid, Spain) to verify if the insert sequence was correct and in frame. Although the sequencing results of the construct 1-114 showed bad quality in the beginning of ComD, we performed the initial expression tests experiments using the plasmids from this colony.

In summary, we successfully inserted the constructs 1-114 and 21-114 into the vector pKLSLt (**Table 6.3**). However, the constructs 1-84 and 21-84 were not obtained in this vector.

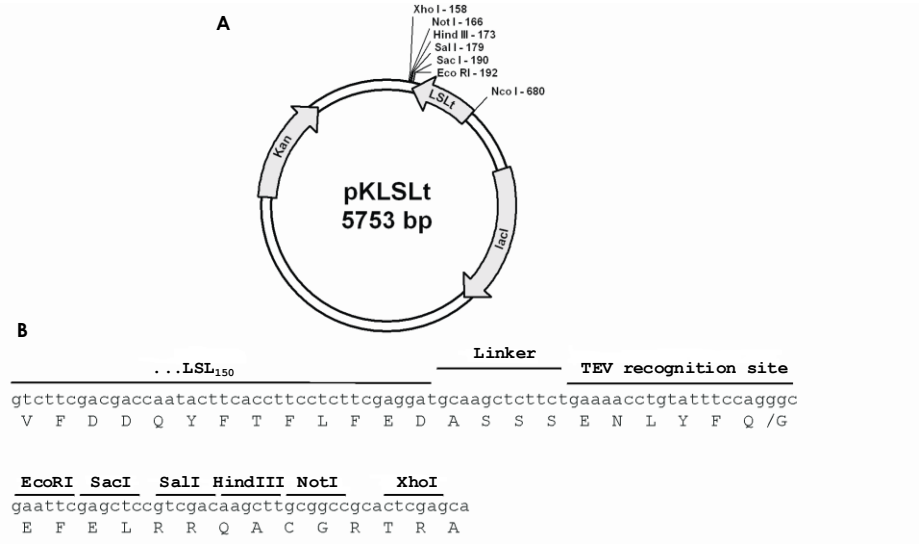


Figure 6.6 – (A) Schematic representation of the vector pKLSLt and (B) sequence of this vector containing the end of LSL_{150}, the linker, the TEV and restriction enzymes recognition sites. Figure kindly provided by Dr. Iván Acebrón and adapted to this work.

6.5. Expression tests of the fusion protein LSL_{150}-ComD

The production of a fusion protein with the LSL_{150} domain in the N-terminus and the ComD sensor domain in the C-terminus was attempted by inserting the vector pKLSLt with constructs 1-114 and 21-114 in the strains *E. coli* BL21(DE3) and *E. coli* JM109(DE3). All the expression tests described in this section are summarized in **Table 6.5**.

Table 6.5 – Summary of the expression tests performed using the vector pKLSLt.

Construct	Expression host	Medium	Volume	Purification
1-114[a]	*E. coli* BL21(DE3)	LB	500 mL	Sepharose 4B®
1-114[a]	*E. coli* JM109(DE3)	LB	500 mL	Sepharose 4B®
1-114[a]	*E. coli* JM109(DE3)	LB	2 L	Sepharose 4B®
21-114	*E. coli* JM109(DE3)	LB	2 L	Sepharose 4B®
21-114	*E. coli* BL21(DE3)	LB	500 mL	Sepharose 4B®
1-114	*E. coli* BL21(DE3)	LB	500 mL	Sepharose 4B®

[a]ComD sequence presented a nucleotide deletion.

Four independent 500 mL cultures were prepared to test expression of constructs 1-114 and 21-114 in both *E. coli* BL21(DE3) and *E. coli* JM109(DE3) strains. Protein overexpression of these cultures was induced at OD_{600nm}=0.6 with 1 mM IPTG followed by incubation at 16 ºC overnight with agitation. Interestingly, despite all our efforts, the construct 21-114 in *E. coli* JM109(DE3) did not grow for reasons that we could not identify. After induction, the cells were harvested in 20 mM Tris pH 8.0, 100 mM NaCl and kept at -20 ºC until use. From the three pellets obtained, the construct 1-114 expressed in *E. coli* BL21(DE3) and in *E. coli* JM109(DE3) was chosen to be purified first to allow the comparison of the expression between both *E. coli* strains. Cells were lysed by sonication (10 cycles of 30 s at maximum intensity with 10 s of interval, model SONOPULS HD 2200, Bandelin, Berlin, Germany) and soluble and insoluble fractions were separated by centrifugation at 38000 RCF during 50 min. The soluble fraction was applied to a column with 3 mL of Sepharose 4B® resin (GE HealthCare, Chicago, USA) which was previously equilibrated with the buffer 20 mM Tris pH 8.0, 100 mM NaCl. Elution was performed in 3 mL of 20 mM Tris pH 8.0, 100 mM NaCl, 200 mM lactose. The result of both purifications was assessed by SDS-PAGE. Both *E. coli* BL21(DE3) and *E. coli* JM109(DE3) strains presented a protein in the elution fractions with a molecular weight close to 21.5 kDa, although in *E. coli* BL21(DE3) the SDS denaturing gels showed that this protein was heavily contaminated with other unspecific bands (data not shown). The predicted molecular weight of the fusion protein of LSL with construct 1-114 was 32.0 kDa (**Table 6.3**), so we concluded that the protein was not being properly expressed under this condition.

Another expression test was performed preparing 2 L of cultures of both constructs in *E. coli* JM109(DE3) because this strain produced purer protein than *E. coli* BL21(DE3) in the previous experiment. The protocol of the previous expression experiment was followed, although, to improve cell lysis, a French Press (Thermo Fisher Scientific, Waltham, USA) was used instead of the ultra-wave sonicator and the volume of the Sepharose 4B® resin was increased to 5 mL to cope with the increased cell lysate volume. The column was connected to a low pressure liquid chromatography system (BioLogic LP System®, Bio-Rad, Hercules, USA) and elution of the protein was performed in the same buffer but with a 30ml gradient of 0-200 mM lactose. The result of the purification was observed by SDS-PAGE. The 21.5 kDa protein observed in the previous experiment was also present in the elution fractions of construct 1-114 (**Figure 6.7a**) but was absent in construct 21-

114 purification. In fact, the lack of expression of the fusion protein by construct 21-114 was further confirmed by purification of the same construct expressed in *E. coli* BL21(DE3) using a similar protocol (**Figure 6.7b**).

After confirming that the construct 21-114/LSL fusion protein was not being produced, the focus was turned to the construct 1-114. The protein purified from this construct was shorter than what was expected and since the sequence of the plasmid of this construct showed a zone with bad quality in the beginning of ComD, sequencing was repeated for the 3 transformant colonies obtained (*see 6.4. Cloning of the fusion protein LSL$_{150}$-ComD in vector pKLSLt*). The plasmid that was originally used in the previous experiments showed a deletion in the zone that presented bad quality in the first sequencing. The deleted nucleotide was T18 affecting the coding of the amino acid residue Phe6 and onwards of ComD. This mutation led to the production of LSL fused with a 13 amino acid peptide with a predicted molecular weight of 20.1 kDa, which correspond to the weight observed by SDS-PAGE for the protein purified from the construct 1-114, explaining the results described above. The other two colonies presented the expected sequence and the plasmid of one of them was transformed into *E. coli* BL21(DE3) and *E. coli* JM109(DE3) strains.

A new expression test was performed with the new construct using *E. coli* BL21(DE3) and *E. coli* JM109(DE3) hosts following the protocol of previous experiments. However, the culture of *E. coli* JM109(DE3) did not grow, similarly to what had happened with construct 21-114, and induction of expression was not performed. Thus, only the expression of the culture of 500 mL of LB of *E. coli* BL21(DE3) was tested. After purification and observing the result by SDS-PAGE, expression of the fusion protein was not observed, confirming that construct 1-114 was not produced as a fusion protein in any of the host strains used with the affinity tag LSL$_{150}$, as was the case of construct 21-114.

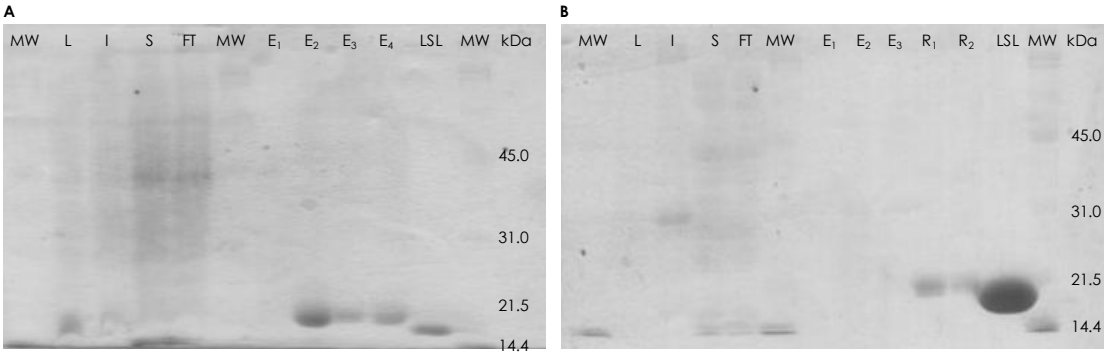


Figure 6.7 - SDS-PAGE 12 % of the purification through a Sepharose 4B® resin of (A) the construct 1-114 expressed in *E. coli* JM109(DE3) and (B) the construct 21-114 expressed in *E. coli* BL21(DE3). MW – molecular weight markers (Molecular Weight Standards Broad Range, Bio-Rad, Hercules, USA), L – lysate, I – insoluble fraction, S – soluble fraction, FT – flow-through after passing protein sample through a Sepharose 4B® column, E_1 to E_4 – elution fractions, R_1 and R_2 – references: purified protein of construct 1-114 expressed in *E. coli* BL21(DE3) and *E. coli* JM109(DE3) respectively, LSL – purified LSL_{150}.

6.6. Discussion

The goal of this work was to study the specificity of ComD sensor domain towards CSP through the structural characterization by X-ray crystallography. Four constructs of parts of the ComD sensor were designed and produced with His and LSL_{150} affinity tags and inserted into the *E. coli* strains BL21(DE3) and JM109(DE3). Production of the ComD sensor was not achieved in any of the conditions tested and further experiments were not performed with this protein.

The construct design was based on the results of the prediction of ComD topology performed with the algorithm TopPred, but the online tool that was used is no longer available. A review of currently available tools identified CCTOP (Dobson *et al.*, 2015) as a robust method since it combines the results of ten algorithms of protein topology prediction. Prediction of ComD topology by CCTOP identified a total of 7 membrane helices in ComD between the positions 4-23, 34-53, 58-74, 85-104, 121-140, 161-180 and 187-206. These positions are very similar to the positions of the helices predicted by TopPred with the exceptions of 58-74, helix not identified by TopPred, and helix 7 (379-399), not identified by CCTOP (**Table 6.1**). Regarding the pneumococcal strain TIGR4, which presents pherotype CSP2, the sequence of its *comD* is equal to allele *comD*-2.8 (*see Chapter 4*, **Figure 4.5**). Performing the topology prediction of TIGR4 ComD with CCTOP tool, a total of 5 helices were identified: 4-22, 34-53, 121-139, 161-179 and 187-206. Comparing with the prediction of R6 ComD, two helices were not detected (58-74 and 85-104), indicating that ComD could present a topology with 5 or 7 transmembrane segments.

AgrC is a histidine kinase membrane receptor from *S. aureus* that was classified in the same family as ComD, the HPK10 family (Grebe and Stock, 1999). AgrC topology was shown to present 7 transmembrane segments, a periplasmatic N-terminus and a cytoplasmatic C-terminus and it was suggested that this protein has a membrane-embedded sensor (Wang et al. 2014). ComD could present a similar topology presenting a membrane-embedded sensor with 7 transmembrane segments connected by small extracellular regions, where many of the amino substitutions between ComD1, ComD2 and ComD3 could be located in transmembrane domains as predicted by CCTOP (**Figure 6.8**). The presence of so

many membrane helices could be the reason why the production of ComD sensor was a failure. We chose to limit the number of putative transmembrane domains in the hope of facilitating expression but including all membrane segments could also diminish toxicity and improve expression or recovery of the protein. Cloning of constructs containing all the predicted transmembrane domains (residues 1-207) was started but due to the end of the visit period this was not finished.

Little is known about the mechanism of recognition of ligands by membrane-embedded sensors. The lack of structural studies of this type of sensors is due to the difficulty of producing and analyzing proteins with so many transmembrane segments, as was probably the case of ComD. Although the interaction of ComD kinase domain with ComE was already structurally characterized (Sanchez *et al.*, 2015), the atomic structural characterization of ComD sensors and their interaction with CSPs would not only elucidate the mechanism of recognition and specificity of these peptides but would also be helpful to design possible inhibitors with a therapeutic function. Abolishment of competence activation was shown to result in an attenuated virulence in *in vivo* models (Lau et al. 2001; Oggioni et al. 2004; Zhu and Lau 2011). Several studies about the structure and specificity of CSPs and also of their analogues were performed to evaluate their ability to activate or inhibit competence (Duan et al. 2012; Johnsborg et al. 2006; Yang et al. 2017; Zhu and Lau 2011). These studies showed that the first three amino acids of CSP were crucial for receptor binding and activation, while the central region of CSP was responsible for specificity towards the type of ComD, possibly through hydrophobic interactions. The degree of α-helix conformation of the central region was correlated with the biological activity of the CSP analogues (Yang *et al.*, 2017). The analogue CSP1-E1A was found to competitively inhibit CSP1-mediated competence development in a concentration dependent-manner (Zhu and Lau, 2011), while the analogue CSP1-K6A was found to be a pan-group activator of competence development independent of ComD type (Yang *et al.*, 2017). However, the actual interaction between ComD and the CSPs was not elucidated.

Regarding the expression experiments to produce the ComD sensor with a His-tag, some proteins were being eluted from the Ni-NTA column (**Figures 6.3** and **6.5**). However, these proteins were probably host proteins with a small affinity to the Ni-NTA matrix. A small concentration of imidazole could have been used in

the equilibration buffer to prevent these unspecific interactions, although the success of such an approach would also not be guaranteed and could result on a lower yield of the protein of interest.

The construct 1-114 presenting a premature stop codon originated an overexpressed protein with the affinity tag LSL_{150}. However, when this was recognized and expression was tried with the correct ComD sequence, a fusion protein was not obtained. This suggested that the experimental setup used was working and that it was the production of the ComD sensor that prevented the overexpression of the fusion protein LSL_{150}-ComD. However, observing **Figure 6.7b**, a protein band around 31 kDa was seen in the insoluble fraction. This molecular weight was compatible with the molecular weight expected for the fusion protein of LSL_{150} with construct 21-114 (**Table 6.3**). Although it is probably improperly folded protein, ComD could be a sensor inserted into the membrane, so taking into account the literature published in the meantime, it cannot be excluded the possibility that the recombinant ComD was located in the membrane and was therefore in the insoluble fraction. Indeed, during the experimental work, we did not consider this possibility because we were assuming that the sensor domain would be exposed in the outer face of the membrane and so we were trying to produce a predicted soluble sensor domain. We did not perform any further manipulations to try to extract any possibly expressed protein located in the *E. coli* membrane.

In the future, structural studies of ComD should focus not only in the ComD sensor containing the 7 predicted transmembrane segments but also on the full-length protein because nowadays increasingly efficient protocols for purification and solubilization of membrane proteins for crystallization are being established (Newby et al. 2009). In fact, several structures of membrane proteins are being solved (Moraes *et al.*, 2014), although only a few families of membrane proteins have been successfully crystallized and the crystallographic structure of a full-length histidine kinase has not been described yet. In addition to ComD1, production of ComD2 should also be attempted because its sensor seems to be less hydrophobic with just 5 membrane helices identified by CCTOP, raising the possibility that the two sensors do not share the same structure.

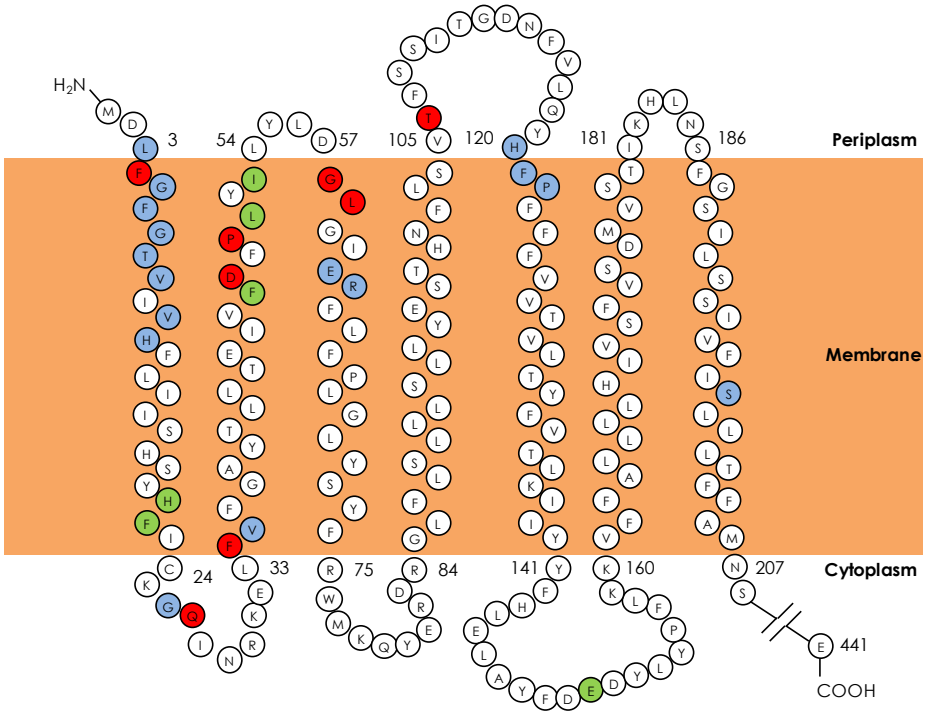


Figure 6.8 – Possible topology of ComD1 based in the prediction by CCTOP. The cytoplasmatic kinase domain was omitted. Colored amino acid residues are positions presenting substitutions in ComD2.8 and/or ComD3.1: red – substitution in ComD2.8; blue – substitution in ComD3.1; green – substitution in both ComD2.8 and ComD3.1.

III. Pherotype specificity and influence on the genetic structure

6. Structural study of pherotype specificity by X-ray crystallography

7. Serogroup 6 and influence of pherotype in genetic recombination

CHAPTER 7. SEROGROUP 6 AND INFLUENCE OF PHEROTYPE IN GENETIC RECOMBINATION

The polysaccharide capsule of *S. pneumoniae* is a major virulence factor and is the target of the current available pneumococcal conjugate vaccines. The type of capsule produced (serotype) is defined by the *cps* region which contains several capsular genes. Currently, almost one hundred serotypes are identified and many of them are grouped into serogroups due to their similar capsular composition. However, new serotypes are constantly being identified. Although historically only two serotypes were recognized within serogroup 6 (6A, 6B), five more were recently discovered (6C, 6D, 6F, 6G and 6H) and shown to be the result of alterations in the *wciN* and *wciP* genes (Bratcher *et al.*, 2011, Burton *et al.*, 2016). The distinction between serotype 6A and 6B was seen to correspond to a single amino acid in the gene *wciP* (Mavroidi *et al.*, 2004). Serotype 6A pneumococci have a serine in the amino acid position 195 of this gene, while serotype 6B pneumococci have an asparagine, resulting in a 1→3 or 1→4 linkage, respectively, between the rhamnose and the ribitol of their capsule repeating unit. However, in 2007 two alleles of the gene *wciN* (*wciN*$_\alpha$ and *wciN*$_\beta$) were identified leading to the identification of serotypes 6C and 6D, which have the *wciN*$_\beta$ allele, among serotypes 6A and 6B pneumococci, respectively (Bratcher *et al.*, 2009, Jin *et al.*, 2009, Park *et al.*, 2007a, Park *et al.*, 2007b). The capsule repeating unit of isolates with *wciN*$_\alpha$ (6A and 6B) is composed of galactose-(1→3)-glucose before rhamnose and ribitol, while isolates with *wciN*$_\beta$ have another glucose instead of the galactose (Park *et al.*, 2007b).

An additional serotype "6E" was proposed (Elberse *et al.*, 2011) and its prevalence has been investigated (Baek *et al.*, 2014a, Baek *et al.*, 2014b, Kawaguchiya *et al.*, 2015, Marimon *et al.*, 2016, Shi *et al.*, 2014, van Tonder *et al.*, 2015). The proposal of this new serotype was based on the divergent capsular genetic sequence regarding typical serogroup 6 capsular loci (≈5 % on average, including SNPs and indels), and the differences observed among, most frequently, some phenotypically serotype 6B isolates. The capsular sequences of these isolates were defined as class 2 sequences (Mavroidi *et al.*, 2004) and it was proposed that they could produce a different capsule (Elberse *et al.*, 2011). However, the capsule of an isolate containing class 2 capsular sequences was recently confirmed to have the same biochemical composition as the capsule of a serotype 6B isolate with class

1 capsular sequences (Burton *et al.*, 2016). Thus, isolates with class 2 capsular sequences are henceforth designated serotype 6B-2 isolates, whereas isolates with canonical serotype 6B loci are designated serotype 6B-1 isolates in this work.

PCV7, which includes serotype 6B, was introduced in Portugal in 2001. However, PCV7 was not included in the National Immunization Plan so its uptake increased slowly. Previous work from our laboratory identified a period when no PCV7 attributable changes in serotypes causing IPD occurred (Pre-PCV7, 1999-2002), distinguishable from early-PCV7 (2003-2006) and late-PCV7 (2007-2009) periods (Aguiar *et al.*, 2010b, Aguiar *et al.*, 2014, Aguiar *et al.*, 2008, Horácio *et al.*, 2012, Horácio *et al.*, 2013, Serrano *et al.*, 2004) when significant serotype changes were seen. By mid-2009 and start of 2010, PCV10 and PCV13 were introduced replacing PCV7. While PCV10 was used only briefly and does not include additional serogroup 6 serotypes, PCV13 includes serotype 6A.

A previous study by our lab showed that two genetically distinct subpopulations could be identified associated with CSP1 and CSP2 pherotypes (Carrolo *et al.*, 2009). It was suggested that "assortative mating" mediated by different pherotypes and ongoing genetic drift could be driving an incipient speciation process within pneumococci, supporting theoretical predictions that the existence of barriers to recombination allow the accumulation of significant genetic drift, even within highly recombinogenic bacterial species. However, another study about this topic was unable to find a restriction in recombination imposed by pherotype (Cornejo *et al.*, 2010). We chose to use serogroup 6 as a case-study to evaluate pherotype influence on genetic recombination because serogroup 6 serotypes presented diverse pherotypes. Serotype 6A was associated with pherotype CSP2 (*see section 5.2.3. Serotype and pherotype*) and serotype 6B presented mostly CSP1 (**Figure 5.7** in Chapter 5). Capsular switching between 6A and 6B can be just a point mutation, but between 6A and 6C the recombination of the *wciN* gene is necessary and between 6B and 6C the recombination must include the *wciN* and *wciP* genes. Thus, capsular switching events were analyzed among serogroup 6 isolates to evaluate if pherotype could be influencing recombination between serogroup 6 isolates.

In summary, on the one hand this study aimed to evaluate which serogroup 6 serotypes were causing IPD in Portugal, the genetic lineages associated and their antimicrobial resistance as well as any changes occurring during a period when

PCVs were being introduced. While on the other hand, serogroup 6 was used to evaluate the pherotype influence on genetic recombination.

7.1. Materials and methods

7.1.1. Bacterial isolates, serotyping, multilocus sequence typing (MLST) and pherotype identification

In 1999-2012, n=4812 isolates causing IPD (n=985 in children <18 years and n=3847 in adults ≥18 years) were identified and characterized regarding serotype and antimicrobial susceptibility (Aguiar *et al.*, 2010b, Aguiar *et al.*, 2014, Aguiar *et al.*, 2008, Horácio *et al.*, 2012, Horácio *et al.*, 2013, Serrano *et al.*, 2004). Isolates were serotyped by the standard capsular reaction test (Statens Serum Institut, Copenhagen, Denmark). Only isolates expressing serogroup 6 (n=242, 5 %) were retained for further study.

To confirm the phenotypically determined serotypes and to identify the most recent serotypes and locus class, three PCR reactions were used (**Table 7.1, Figures 7.1 and 7.2**), complemented by sequencing of *wciN* and *wciP*. Identification of the alleles of the *wciP* gene was done as described in other studies (Bratcher *et al.*, 2011, Mavroidi *et al.*, 2004, Yun *et al.*, 2014) and a similar procedure was followed for the *wciN* gene.

MLST was performed as described in Chapter 3 [*see 3.4 - Multilocus Sequence Typing (MLST)]* with the exception of the definition of CCs by PHYLOViZ. In this work, CCs were defined at the double-locus-variant (DLV) level using only the STs found in this study and not the entire MLST database.

Pherotype identification was performed as described in Chapter 5 (*see 5.1.2 – Pherotype identification*).

7.1.2. Statistical analysis

Genetic diversity was evaluated using Simpson's index of diversity (SID) and respective 95 % confidence intervals ($CI_{95\ \%}$) (Carriço *et al.*, 2006). The Cochran-Armitage test (CA) was used to evaluate the temporal trends of serotypes, STs and CCs. Only STs and CCs with ≥5 isolates were considered for the CA analysis. Fisher's exact test (FET) was used to test association of serotypes, STs and CCs with age group. The false discovery rate (FDR) correction for multiple testing (Benjamini and Hochberg, 1995) was used in both tests. A $p<0.05$ was considered significant for all tests.

Table 7.1 – PCRs used for serotype and class identification.

Purpose of the PCR	Target gene	Primer	Sequence (5'-3')	Target serotype	Product size	Reference
Multiplex PCR for identification of 6A, 6B, 6C and 6D serotypes	*wzy*	wzy-f	CGACGTAACAAAGAACTAGGTGCTGAAAC	Serogroup 6	220 bp	(Brito *et al.*, 2003)
		wzy-r	AAGTATATAACCACGCTGTAAAACTCTGAC			
	wciP	wciP-f1	ATATGTAGAAGAACTGGCTCAGGGTAG	6A, 6C and 6F	128 bp	This study
		wciP-r1	GATGACTAGATGGTACATTATGTCCAT			
	wciN	wciN-f1	CATTTTAGTGAAGTTGGCGGTGGAGTT	6C and 6D	727 bp	This study
		wciN-r1	AGCTTCGAAGCCCATACTCTTCAATTA			
Class identification	*wzh*	wzh-f	TGATATTCATTCGCACATTGTC	Class 2 sequences	578 bp	(Kawaguchiya *et al.*, 2015)
		wzh-r	TATGAACCAAATCACGCTCCAAG			
	wze	wze-f	CTCACAGGCAAAATTGGATTC	Class 1 sequences	217 bp	(Kawaguchiya *et al.*, 2015)
		wze-r	AACAGAATTGCGAATATCTC			
wciN sequencing	*wciN*	wciN-f2	TGGAAAGATATTGAAATTTT	Serogroup 6	1.4 kb (6A/6B-1/6F/6G) 1.2 kb (6C/6D) 1.7 kb (6B-2)	This study
		wciN-r2	GTT TTTCTTTCAATATCTTTA			
wciP sequencing	*wciP*	wciP-f2	CGATTAATTTTTTATTAATG	Serogroup 6	1.0 kb	This study
		wciP-r2	ATATGAATAAGAAATTTAAAAG			

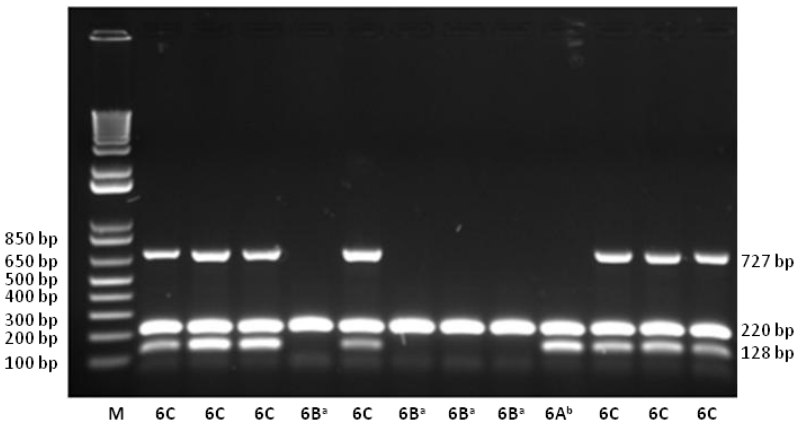


Figure 7.1 – Multiplex PCR for identification of 6A, 6B, 6C and 6D serotypes. M – size marker. [a]Could also be 6G or 6H and it cannot distinguish the class (6B-1 or 6B-2). [b]Can also be 6F.

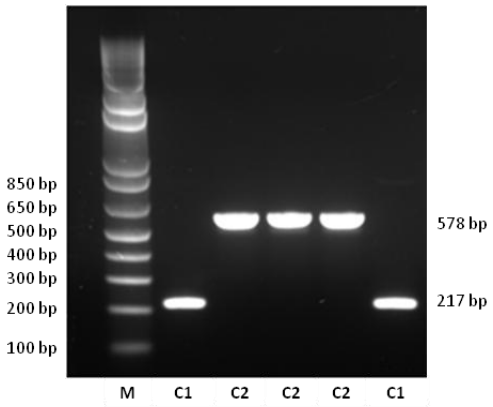


Figure 7.2 – PCR for class identification. M – size marker; C1 – class 1 sequence; C2 – class 2 sequence.

7.1.3. Gene flow and genetic differentiation analyses

To estimate the level of gene flow and the genetic differentiation between subpopulations the classical F_{ST} parameter (Hudson *et al.*, 1992b) and the statistics K^*_{ST} (Hudson *et al.*, 1992a) and S_{nn} (Hudson, 2000) were used. These analyses were performed with the sequence of the MLST genes (*aroE*, *ddl*, *gdh*, *gki*, *recP*, *spi* and *xpt*) and also with the sequence obtained from the concatenation of these MLST genes, with the exception of *ddl* since it was previously shown that this gene showed a hitchhiking effect with *pbp2b* involved in penicillin resistance (Enright and Spratt, 1999). The DnaSP v5.10.01 program was used to perform the distribution of the sequences in populations regarding their pherotype and to perform the analyses of gene flow and genetic differentiation. K^*_{ST} and S_{nn} statistics were used in combination with permutation tests performing a total of 1000 permutations, considering $p<0.05$ statistically significant.

7.2. Results

7.2.2. Evolution of serogroup 6 serotypes in children and in adults

The serotypes identified were 6A (n=80), 6B (n=79) and 6C (n=83). All class 2 loci were identified among serotype 6B isolates (6B-2, n=52; 6B-1, n=27). No representatives of serotypes 6D, 6F, 6G and 6H were found, confirming their rarity in Europe (Marimón *et al.*, 2016, van der Linden *et al.*, 2013b). The proportion of serogroup 6 isolates was higher in children than in adults (6.4 % and 4.7 %, respectively, FET p=0.032) and differences in serotype distribution were also observed (**Figure 7.3a and 7.3b** and **Tables 7.2** and **7.3**). Most 6C pneumococci were recovered from adults (n=79/83) resulting in an association with this age group (FET p<0.001, significant after FDR correction). There was no association between serotype and isolate source. Despite no overall change in frequency of serogroup 6 pneumococci following the introduction of PCVs, there were changes in proportion of serotypes (**Figure 7.3a and 7.3b**). In the pre-PCV7 period, serotype 6B (6B-2) was the most frequent in contrast to neighboring Spain where 6B-1 predominated (Marimón *et al.*, 2016), but it subsequently decreased in both adults and children, similarly to Spain (supported only in children: CA p<0.001, significant after FDR; CA p=0.160, in adults). Class 6B-1 accounted for a small proportion of IPD in both age groups before vaccine introduction and, although there was an increase in the proportion of 6B-1 isolates in both age groups, this was supported only in adults and before FDR correction (CA p=0.041). So although both 6B-1 and 6B-2 express the same polysaccharide, they showed different dynamics following vaccination. It was hypothesized that the two classes could present differences in capsule expression (Burton *et al.*, 2016) potentially explaining their contrasting variations seen here. Serotype 6A increased to become the most frequent serotype in children in late-PCV7 and PCV13 periods. In adults, although 6A was an important serotype up to 2009, a decrease was seen afterwards (CA p=0.006, significant after FDR), coinciding with PCV13 introduction in children (**Figure 7.3b**). While the proportion of isolates expressing serotype 6C remained low and stable in children, in adults an increase in serotype 6C was noted after 2008, although not statistically supported (CA p=0.392), establishing it as the most frequent serotype of serogroup 6 in adults.

Table 7.2 – No. of isolates of serogroup 6 responsible for invasive infections in children (<18 years) in Portugal between 1999 and 2012.

	Pre-vaccine				PCV7							PCV13			
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	Total
Serogroup 6	2	3	3	2	3	1	2	5	7	7	10	5	7	5	62
6A	0	1	0	1	2	0	1	0	4	4	6	2	4	2	27
6B-1	0	0	0	1	0	1	0	1	0	2	2	1	2	2	12
6B-2	2	2	3	0	1	0	1	2	3	1	2	1	1	0	19
6C	0	0	0	0	0	0	0	2	0	0	0	1	0	1	4
All invasive pneumococci	20	19	31	27	29	42	51	93	145	115	160	87	72	74	965

Table 7.3 – No. of isolates of serogroup 6 responsible for invasive infections in adults (≥18 years) in Portugal between 1999 and 2012.

	Pre-vaccine				PCV7							PCV13			
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	Total
Serogroup 6	4	1	6	6	13	10	17	17	16	11	26	18	19	16	180
6A	0	0	3	2	6	4	6	8	6	6	7	2	1	2	53
6B-1	0	0	0	0	1	0	0	2	2	0	3	0	2	5	15
6B-2	0	0	3	3	2	3	2	3	3	1	3	3	7	0	33
6C	4	1	0	1	4	3	9	4	5	4	13	13	9	9	79
All invasive pneumococci	67	91	113	105	167	214	312	285	406	409	448	404	413	413	3847

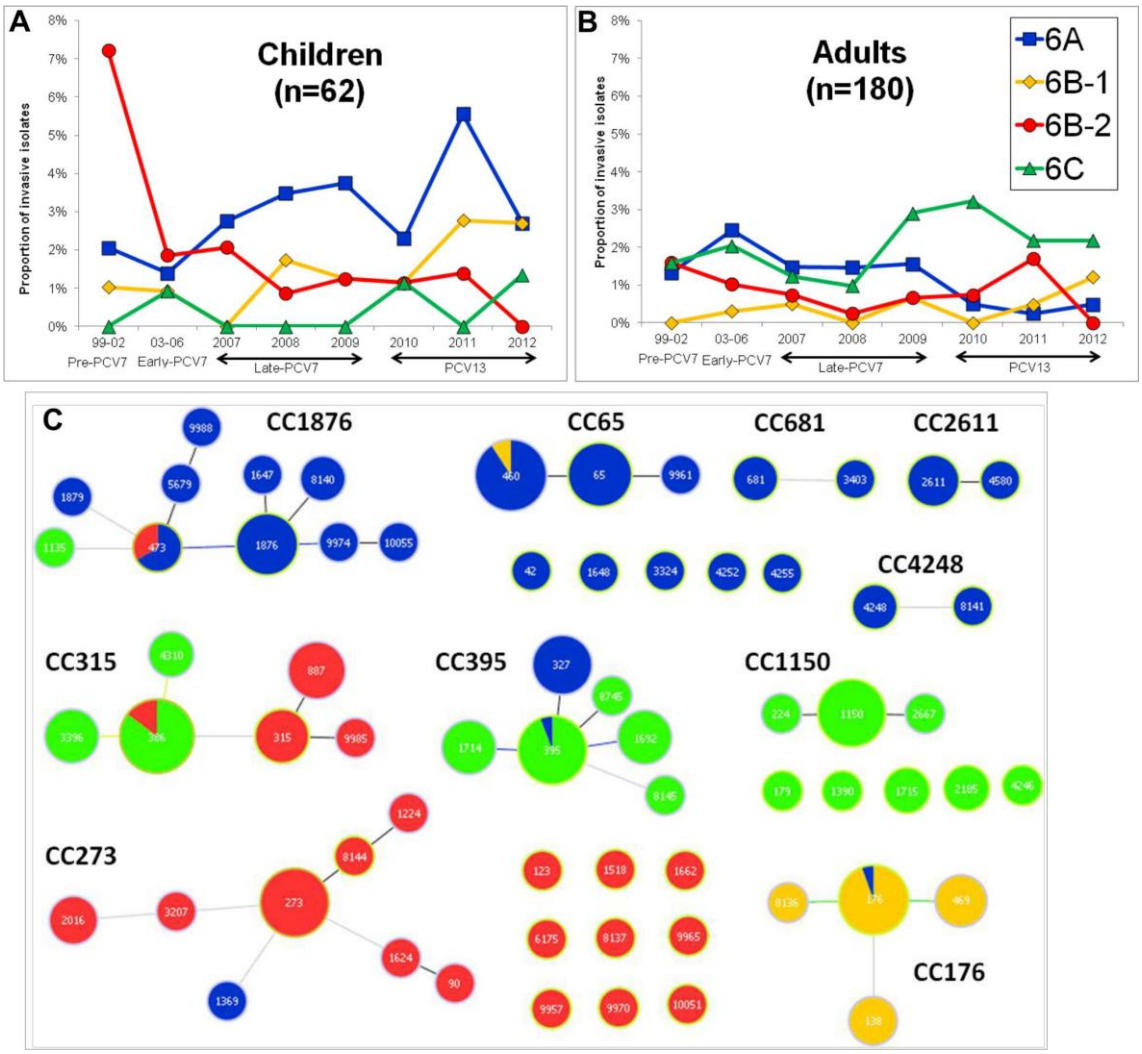


Figure 7.3 - Clonal composition and changes in serogroup 6 serotypes among invasive pneumococci recovered in Portugal (1999-2012). (A) Shows the variations of the serotypes and serotype classes in children and (B) in adults. The years before PCV7 introduction (1999-2002) and subsequent periods were defined as described in the text. (C) Shows STs and CCs identified colored by serotype. Each circle represents an ST and the diameter represents its frequency in a logarithmic scale. Grey lines connect STs that are double-locus variants, while lines of other colors connect STs that are single-locus variants according to the PHYLOViZ tie-break rule reached. STs that are linked belong to the same CC. This data set can be explored online at http://bit.do/PHYLOViZ_sero6.

7.2.2. Serogroup 6 genetic diversity

All isolates were characterized by multilocus sequence typing (MLST) to complement the information already available (Horácio *et al.*, 2016) and PHYLOViZ was used to define CCs (**Figure 7.3c**). Serotype 6A was the most diverse (27 STs and 13 CCs) followed by 6B-2 (21 STs and 12 CCs), 6C (17 STs and 9 CCs) and 6B-1 (5 STs and 2 CCs). Still, the majority of serogroup 6 isolates

(n=210, 87 %) were distributed into only 7 genetic lineages, three of which included PMEN clones (McGee *et al.*, 2001): CC315 (Poland^{6B}315, n=47), CC395 (Portugal^{6A}327, n=37), CC65 (n=35), CC176 (n=26), CC273 (Greece^{6B}273 and Spain^{6B}90, n=26), CC1876 (n=22) and CC1150 (n=17) (**Tables 7.4** and **7.5**). Mostly, each ST presented only one serotype, with the exception of STs ST176, ST460, ST473, ST386 and ST395, possibly reflecting capsular switching events (**Figure 7.3c**). The 6B-2 ST90 lineage dominant in Asia (Baek *et al.*, 2014b, Kawaguchiya *et al.*, 2015) was represented by a single isolate.

The genetic diversity of serogroup 6 isolates recovered from both children and adults was similar when considering both STs and CCs, with all seven major lineages present in both age groups (**Tables 7.4** and **7.5**). Although CC176 was found more frequently in children (FET p=0.017, not supported after FDR), no association of particular STs or CCs with age group was observed. While in adults the majority of CC315 and CC395 isolates expressed serotype 6C (n=28/39 and n=27/31, respectively), in children these lineages expressed mostly other serotypes: CC315/6B-2 (n=6/8) and CC395/6A (n=5/6). Serotype 6C was found in 3.0 % of carriage isolates in children (Nunes *et al.*, 2009) but in only 0.4 % of IPD (**Table 7.2**). Taken together these observations suggest that children may be particularly protected against serotype 6C IPD, potentially through cross-protection due to the presence of polysaccharides 6A and 6B in PCV13. CC315 increased in adults (CA p=0.012, significant after FDR). This CC included both 6C and 6B-2 isolates but only the 6C drove the increase (CA p=0.001). In contrast, and in agreement with the decrease of 6B-2, there was a decrease in the major lineage expressing this class (CC273) in both children and adults (CA p<0.001 and p=0.002, respectively, both significant after FDR) (**Figure 7.4** and **Tables 7.4** and **7.5**). Increase of the prevalence of serotype 6C pneumococci was reported in several regions (Loman *et al.*, 2013, Rolo *et al.*, 2011b, van der Linden *et al.*, 2013b). However, the genetic lineage that increased was not always CC315. For example, in Southampton, England, the increase of serotype 6C pneumococci in carriage was due to the clonal expansion of CC395 (Loman *et al.*, 2013). In Spain, the authors associated the increase of serotype 6C isolates with spread of CC1150 (Rolo *et al.*, 2011b), although isolates of CC315 emerged in 2007, coinciding with the data from IPD and carriage in Portugal.

Table 7.4 – No. of isolates of STs and CCs of serogroup 6 responsible for invasive infections in children (<18 years) in Portugal between 1999 and 2012.

	Pre-vaccine				PCV7							PCV13			
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	Total
CC176	-	-	-	1	-	1	-	1	-	2	2	1	2	2	12
ST176	-	-	-	1	-	1	-	-	-	1	2	1	-	1	7
ST138	-	-	-	-	-	-	-	1	-	-	-	-	1	-	2
ST469	-	-	-	-	-	-	-	-	-	-	-	-	1	1	2
ST8136	-	-	-	-	-	-	-	-	-	1	-	-	-	-	1
CC65	-	-	-	1	-	-	-	-	2	2	2	2	1	1	11
ST460	-	-	-	1	-	-	-	-	1	-	2	1	-	1	6
ST65	-	-	-	-	-	-	-	-	1	2	-	1	1	-	5
CC273	2	1	3	-	-	-	-	1	2	-	2	-	-	-	11
ST273	1	-	3	-	-	-	-	-	1	-	1	-	-	-	6
ST2016	-	-	-	-	-	-	-	1	1	-	1	-	-	-	3
ST90	1	-	-	-	-	-	-	-	-	-	-	-	-	-	1
ST1224	-	1	-	-	-	-	-	-	-	-	-	-	-	-	1
CC315	-	1	-	-	1	-	1	1	1	-	-	1	1	1	8
ST386	-	-	-	-	-	-	-	1	1	-	-	-	-	1	3
ST315	-	-	-	-	1	-	1	-	-	-	-	-	-	-	2
ST887	-	1	-	-	-	-	-	-	-	-	-	-	1	-	2
ST9985	-	-	-	-	-	-	-	-	-	-	-	1	-	-	1
CC1876	-	-	-	-	-	-	-	-	2	2	-	-	3	-	7
ST1876	-	-	-	-	-	-	-	-	2	2	-	-	-	-	4
ST473	-	-	-	-	-	-	-	-	-	-	-	-	1	-	1
ST5679	-	-	-	-	-	-	-	-	-	-	-	-	1	-	1
ST9988	-	-	-	-	-	-	-	-	-	-	-	-	1	-	1
CC395	-	-	-	-	1	-	1	1	-	-	3	-	-	-	6
ST327	-	-	-	-	1	-	-	-	-	-	3	-	-	-	4
ST395	-	-	-	-	-	-	1	1	-	-	-	-	-	-	2
CC681	-	-	-	-	-	-	-	-	-	-	1	-	-	-	1
ST3403	-	-	-	-	-	-	-	-	-	-	1	-	-	-	1
CC1150	-	-	-	-	-	-	-	-	-	-	-	1	-	-	1
ST1150	-	-	-	-	-	-	-	-	-	-	-	1	-	-	1
CC4248	-	1	-	-	-	-	-	-	-	-	-	-	-	-	1
ST4248	-	1	-	-	-	-	-	-	-	-	-	-	-	-	1
ST1648	-	-	-	-	1	-	-	-	-	-	-	-	-	-	1
ST1662	-	-	-	-	-	-	-	1	-	-	-	-	-	-	1
ST3324	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1
ST8137	-	-	-	-	-	-	-	-	-	1	-	-	-	-	1

Table 7.5 – No. of isolates of STs and CCs of serogroup 6 responsible for invasive infections in adults (≥18 years) in Portugal between 1999 and 2012.

	Pre-vaccine				PCV7							PCV13			
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	Total
CC315	-	-	1	1	-	2	3	2	1	1	8	7	9	4	39
ST386	-	-	-	-	-	2	1	1	1	1	4	6	5	3	24
ST887	-	-	1	1	-	-	1	-	-	-	-	1	1	-	5
ST3396	-	-	-	-	-	-	-	-	-	-	3	-	1	1	5
ST315	-	-	-	-	-	-	1	1	-	-	-	-	1	-	3
ST4310	-	-	-	-	-	-	-	-	-	-	1	-	1	-	2
CC395	1	-	-	3	2	3	5	2	3	-	3	5	3	1	31
ST395	1	-	-	1	2	2	2	-	-	-	2	2	3	-	15
ST1692	-	-	-	-	-	-	-	2	-	-	1	2	-	-	5
ST1714	-	-	-	-	-	-	2	-	2	-	-	-	-	1	5
ST327	-	-	-	2	-	1	-	-	-	-	-	1	-	-	4
ST8145	-	-	-	-	-	-	1	-	-	-	-	-	-	-	1
ST8745	-	-	-	-	-	-	-	-	1	-	-	-	-	-	1
CC65	-	-	-	-	2	2	3	3	4	3	2	1	1	3	24
ST460	-	-	-	-	2	2	3	1	3	2	-	-	-	3	16
ST65	-	-	-	-	-	-	-	2	1	-	2	1	1	-	7
ST9961	-	-	-	-	-	-	-	-	-	1	-	-	-	-	1
CC1150	1	-	-	-	1	-	2	1	1	2	2	2	1	3	16
ST1150	1	-	-	-	1	-	2	1	1	1	2	2	1	2	14
ST224	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1
ST2667	-	-	-	-	-	-	-	-	-	1	-	-	-	-	1
CC273	-	-	3	1	2	2	-	2	3	-	1	-	1	-	15
ST273	-	-	2	1	1	2	-	1	3	-	-	-	1	-	11
ST1369	-	-	1	-	-	-	-	-	-	-	-	-	-	-	1
ST1624	-	-	-	-	1	-	-	-	-	-	-	-	-	-	1
ST3207	-	-	-	-	-	-	-	-	-	-	1	-	-	-	1
ST8144	-	-	-	-	-	-	-	1	-	-	-	-	-	-	1
CC1876	-	-	1	-	3	-	1	3	-	2	3	-	-	2	15
ST1876	-	-	1	-	1	-	-	2	-	1	1	-	-	-	6
ST473	-	-	-	-	-	-	-	-	-	1	1	-	-	-	2
ST8140	-	-	-	-	1	-	1	-	-	-	-	-	-	-	2
ST1135	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1
ST1647	-	-	-	-	1	-	-	-	-	-	-	-	-	-	1
ST1879	-	-	-	-	-	-	-	1	-	-	-	-	-	-	1
ST9974	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1
ST10055	-	-	-	-	-	-	-	-	-	-	1	-	-	-	1
CC176	-	-	-	-	1	-	-	2	2	-	4	-	2	3	14
ST176	-	-	-	-	1	-	-	2	1	-	3	-	2	2	11
ST469	-	-	-	-	-	-	-	-	1	-	-	-	-	1	2
ST138	-	-	-	-	-	-	-	-	-	-	1	-	-	-	1
CC2611	-	-	-	-	1	1	1	1	-	-	1	-	-	-	5
ST2611	-	-	-	-	1	1	1	-	-	-	1	-	-	-	4
ST4580	-	-	-	-	-	-	-	1	-	-	-	-	-	-	1
CC681	-	-	-	-	-	-	-	1	-	-	1	-	-	-	2
ST681	-	-	-	-	-	-	-	1	-	-	1	-	-	-	2
CC4248	-	-	1	-	-	-	1	-	-	-	-	-	-	-	2
ST4248	-	-	1	-	-	-	-	-	-	-	-	-	-	-	1
ST8141	-	-	-	-	-	-	1	-	-	-	-	-	-	-	1
ST1715	-	-	-	-	-	-	1	-	-	-	-	1	-	-	2
ST2185	1	-	-	-	1	-	-	-	-	-	-	-	-	-	2
ST42	-	-	-	-	-	-	-	-	-	1	-	-	-	-	1
ST123	-	-	-	1	-	-	-	-	-	-	-	-	-	-	1
ST179	-	1	-	-	-	-	-	-	-	-	-	-	-	-	1
ST1390	-	-	-	-	-	-	-	-	-	1	-	-	-	-	1
ST1518	-	-	-	-	-	-	-	-	-	-	-	-	1	-	1
ST4246	1	-	-	-	-	-	-	-	-	-	-	-	-	-	1
ST4252	-	-	-	-	-	-	-	-	1	-	-	-	-	-	1
ST4255	-	-	-	-	-	-	-	-	1	-	-	-	-	-	1
ST6175	-	-	-	-	-	-	-	-	-	1	-	-	-	-	1
ST9957	-	-	-	-	-	-	-	-	-	-	-	1	-	-	1
ST9965	-	-	-	-	-	-	-	-	-	-	-	-	1	-	1
ST9970	-	-	-	-	-	-	-	-	-	-	-	1	-	-	1
ST10051	-	-	-	-	-	-	-	-	-	-	1	-	-	-	1

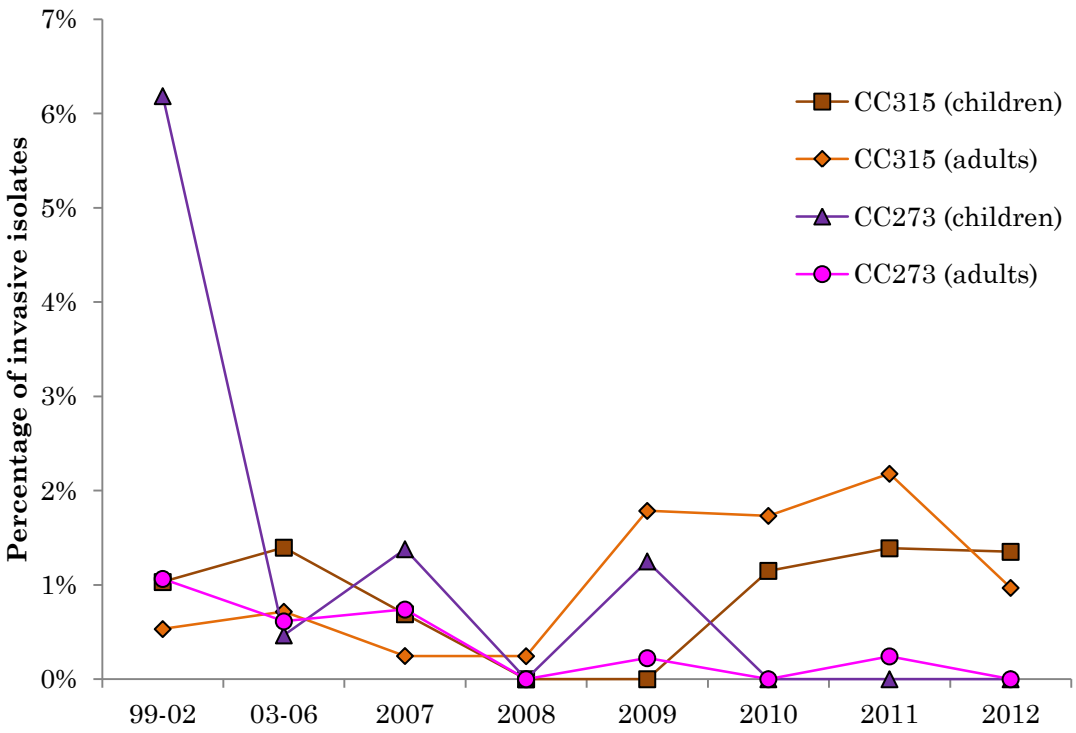


Figure 7.4 – Evolution among all invasive pneumococci of the proportion of CC273 and CC315 before vaccine introduction (1999-2002), in the early PCV7 use (2003-2006) period and following years.

7.2.3. Diversity of the wciP and wciN genes

Sequencing of the capsular genes *wciN* and *wciP* identified 5 and 14 alleles, respectively (**Table 7.6**). There was a strong correspondence between the alleles at these loci and serotype and locus type but two isolates of ST460 presented unusual capsular loci. In one isolate (presenting *wciP*-24 allele) a point mutation seems to have occurred in an allele characteristic of serotype 6A switching it to 6B (6B-1). The other possibly resulted from horizontal DNA transfer, presenting a hybrid locus including class 1 sequences associated with 6A and 6B-1 (*wciN*-1) and class 2 sequences (*wciP*-8). These observations confirm that serotype switching within serogroup 6 may occur in the wild by both point mutation and recombination.

Table 7.6 – Allelic profiles of genes *wciN* and *wciP* and respective serotype and CC/ST.

Allelic profile		Serotype				Clonal complex								
wciN[a]	*wciP*[b]	6A	6B-1	6B-2	6C	CC315	CC395	CC65	CC273	CC176	CC1876	CC1150	CC2611	Other[c]
1	1	33	-	-	-	-	-	32	-	-	-	-	-	1
4	1	1	-	-	-	-	-	1	-	-	-	-	-	-
1	2	24	-	-	-	-	-	-	-	1	15	-	-	8
3	2	14	-	-	-	-	8	-	-	-	-	-	5	1
5	2	1	-	-	-	-	1	-	-	-	-	-	-	-
1	4	-	25	-	-	-	-	-	-	25	-	-	-	-
1	8	-	1	-	-	-	-	1	-	-	-	-	-	-
2	8	-	-	51	-	17	-	-	25	-	1	-	-	8
wciN$_{\beta}$	9	-	-	-	60	28	27	-	-	-	1	-	-	4
wciN$_{\beta}$	13	-	-	-	18	-	-	-	-	-	-	16	-	2
1	14	5	-	-	-	-	-	-	1	-	4	-	-	-
1	22	1	-	-	-	-	-	-	-	-	1	-	-	-
1	23	1	-	-	-	-	-	-	-	-	-	-	-	1
1	24	-	1	-	-	-	-	1	-	-	-	-	-	-
wciN$_{\beta}$	25	-	-	-	1	-	1	-	-	-	-	-	-	-
wciN$_{\beta}$	26	-	-	-	1	1	-	-	-	-	-	-	-	-
wciN$_{\beta}$	27	-	-	-	1	-	-	-	-	-	-	-	-	1
2	28	-	-	1	-	-	-	-	-	-	-	-	-	1

[a]Since the changes leading to the expression of serotypes 6F e 6G were identified in *wciNa* variants, only these were sequenced for the identification of allelic variants. *wciN$_{\beta}$* variants detected by PCR (all found in serotype 6C isolates) are indicated as such in the table. The sequence of each allele is available at https://dx.doi.org/10.6084/m9.figshare.3437429.v1.

[b]In two isolates serotype 6C it was not possible to determine the allele of the *wciP* gene. The sequence of each allele is available at https://dx.doi.org/10.6084/m9.figshare.3437429.v1.

[c]Other CCs or STs not included in those discriminated. These were (*wciN-wciP*) – (1-1): ST4255, n=1; (1-2): CC681, n=3; CC4248, n=3; ST3324, n=1; ST4252, n=1; (3-2): ST42, n=1; (2-8): ST123, n=1; ST1518, n=1; ST1662, n=1; ST8137, n=1; ST9957, n=1; ST9965, n=1; ST9970, n=1; ST10051, n=1; (wciN$_{\beta}$-9): ST1715, n=2; ST179, n=1; ST4246, n=1; (wciN$_{\beta}$-13): ST2185, n=2; (1-23): ST1648, n=1; (wciN$_{\beta}$-27): ST1390, n=1; (2-28): ST6175, n=1.

7.2.4. Antimicrobial susceptibility

The antimicrobial susceptibility of the isolates is indicated in **Table 7.7**. Serotype 6B-2 presented the highest proportion of multidrug resistant isolates (67 %) followed by 6C (36 %) (MDR, defined as non-susceptibility to at least 3 antimicrobial classes). The majority of isolates non-susceptible to penicillin (PNSP) (C.L.S.I., 2007) expressed low level resistance (MIC=0.12-1 µg/mL), with the exception of a single 6B-2 isolate that expressed high level resistance (MIC=2 µg/mL). None of the PNSP isolates was 6B-1. Considering the current CLSI guidelines (C.L.S.I., 2014), 6 cerebrospinal fluid isolates (6B-2, n=3; 6C, n=2; 6A, n=1) would be considered resistant and all remaining isolates would be considered fully susceptible to penicillin using the non-meningitis breakpoints. Erythromycin resistance (ERP) was identified in 77 isolates and 53 isolates were simultaneously PNSP and ERP (serotype 6C, n=29; serotype 6B-2, n=20; and serotype 6A, n=4). CC315 and CC273 presented the highest proportions of non-susceptible isolates (**Table 7.7**). The proportion of isolates representing these CCs changed over time and this was the basis of changes in resistance within serotypes 6C and 6B-2.

Among 6C, the proportions of PNSP (p<0.001), ERP (p=0.001), clindamycin (p=0.001), tetracycline (p<0.001) and MDR (p=0.001) increased (CA, all significant after FDR), changes driven mostly by increases in CC315. In fact, the remaining two CCs representing almost all serotype 6C isolates were CC1150, including mostly PNSP, and CC395 including mostly susceptible isolates. In 6B-2, decreases of ERP (p=0.014), clindamycin (p=0.005), tetracycline (p=0.031) and MDR (p=0.024) resistance were observed (CA, all significant after FDR correction), reflecting the decrease in CC273.

Table 7.7 – Antimicrobial resistance of the serogroup 6 isolates responsible for invasive infections in Portugal (1999-2012).

	No. of non-susceptible isolates (%)											
	Serotype				Clonal complex							
Antimicrobial[a]	6A	6B-1	6B-2	6C	CC315	CC395	CC65	CC273	CC176	CC1876	CC1150	Other[b]
MDR	4 (5.0)	3 (11.1)	35 (67.3)	30 (36.1)	47 (100.0)	0 (0)	1 (2.9)	17 (65.4)	3 (12.0)	1 (4.5)	0 (0)	3 (9.4)
PEN	10 (12.5)	0 (0)	26 (50.0)	42 (50.6)	45 (95.7)	1 (2.7)	2 (5.7)	9 (34.6)	0 (0)	4 (18.2)	13 (76.5)	4 (12.5)
MIC_{50}	0.023	0.016	0.047	0.064	0.125	0.016	0.023	0.023	0.016	0.016	0.094	0.023
MIC_{90}	0.064	0.032	0.19	0.19	0.19	0.032	0.047	0.38	0.023	0.19	0.125	0.064
ERY	4 (5.0)	9 (33.3)	34 (65.4)	30 (36.1)	47 (100.0)	0 (0)	0 (0)	16 (61.5)	9 (34.6)	3 (13.6)	0 (0)	2 (6.3)
CLI	1 (1.3)	4 (14.8)	33 (63.5)	30 (36.1)	47 (100.0)	0 (0)	0 (0)	16 (61.5)	4 (15.4)	0 (0)	0 (0)	1 (3.1)
LEV	0 (0)	0 (0)	0 (0)	1 (1.2)	1 (2.1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
SXT	10 (12.5)	6 (22.2)	27 (51.9)	3 (3.6)	6 (12.8)	3 (8.1)	4 (11.4)	16 (61.5)	5 (19.2)	1 (4.5)	2 (11.8)	9 (26.1)
TET	5 (6.3)	3 (11.1)	34 (65.4)	32 (38.6)	44 (93.6)	0 (0)	1 (2.9)	21 (80.8)	3 (11.5)	0 (0)	1 (5.9)	4 (12.5)
CHL	1 (1.3)	0 (0)	7 (13.5)	2 (2.4)	2 (4.3)	0 (0)	1 (2.9)	6 (23.1)	0 (0)	0 (0)	0 (0)	1 (3.1)

[a]All isolates were susceptible to cefotaxime, vancomycin and linezolid. MDR: multidrug resistance, PEN: penicillin, MIC: minimum inhibitory concentration, ERY: erythromycin, CLI: clindamycin, LEV: levofloxacin, SXT: trimethoprim-sulfamethoxazole, TET: tetracycline, CHL: chloramphenicol. Isolates presenting PEN MIC≥0.12µg/ml were considered resistant and isolates presenting PEN MIC<0.12µg/ml were considered susceptible.

[b]Other CC/ST – MDR: CC4248 (n=1), ST1518 (n=1), ST1662 (n=1); PEN: CC4248 (n=1), ST1518 (n=1), ST1662 (n=1), ST4255 (n=1); ERY: ST1518 (n=1), ST1662 (n=1); CLI: ST1662 (n=1); SXT: CC681 (n=1), CC4248 (n=1), ST42 (n=1), ST1518 (n=1), ST1662 (n=1), ST3324 (n=1), ST9957 (n=1), ST9965 (n=1), ST9970 (n=1); TET: CC4248 (n=2), ST1715 (n=1), ST3324 (n=1); CHL: ST1662 (n=1).

7.2.5. Pherotype and influence on genetic recombination

The pherotype of serogroup 6 isolates was identified (**Table 7.8**). Taking into consideration only the serogroup 6 isolates, serotype 6A and 6C isolates were associated with CSP2 (FET, p<0.001 and p=0.007, respectively, significant after FDR correction) and serotype 6B-1 and 6B-2 with CSP1 (FET, p<0.001, significant after FDR correction for both). These associations led us to choose serogroup 6 as a case-study of the influence of pherotype on genetic recombination.

Although pherotype was observed to be a clonal property, ST273 (n=17) presented a CSP2 isolate (n=1) whereas the remaining isolates were CSP1 (n=16) (all serotype 6B-2) and ST473 (n=3) presented two CSP2 (serotype 6A) and one CSP1 (serotype 6B-2) isolate.

A total of 5 STs presented more than a single serogroup 6 serotype (**Figure 7.3c**) reflecting possible capsular switching events:

- ST176 (n=18): all the isolates were CSP1 (**Table 7.9**) and it could be assumed that the capsular switch was from 6B-1 (n=17) to 6A (n=1), namely by the exchange of a *wciP*-4 for *wciP*-2 allele. These alleles have 3 nucleotide differences, suggesting that the capsular switch was not driven by point mutations. The allele *wciP*-2 (n=39) was identified both in CSP1 (n=6) and CSP2 isolates (n=33), whereas *wciP*-4 (n=25) was identified only in isolates with CSP1 pherotype. Thus, although the donor strain could have been a CSP2 strain, the hypothesis that the exchange occurred between two CSP1 isolates cannot be ruled out.

- ST386 (n=26): this ST presented isolates of serotype 6C (n=22) and serotype 6B-2 (n=4) which means the capsular switch was between class 1 and class 2 capsular sequences. All these isolates presented pherotype CSP2 (**Table 7.9**) and, in this case, it is not completely clear which serotype was that of the recipient strain. However, it was reported in this thesis that CC315 increased driven by an increase of serotype 6C isolates (*see 7.2.2. Serogroup 6 genetic diversity*). It is possible that this genetic lineage was 6B-2 to start with and then switched to 6C, enabling it to escape vaccine conferred immunity. If that was the case, recombination could have occurred between two CSP2 strains because the *wciP*-9 allele, the allele found in ST386 6C isolates, was also found in other 6C isolates presenting the CSP2 pherotype, namely CC395, which could have acted as donors (**Tables 7.6** and **7.9**).

- ST395 (n=17): the isolates of this ST were all CSP2 (**Table 7.9**) and a capsular switch could have occurred between 6C (n=16) and 6A (n=1). The serotype 6A isolate presented the *wciN*-3 and *wciP*-2 alleles that were also identified in ST327, a SLV of ST395, and in CC2611 (**Table 7.6**). These genetic lineages also presented the pherotype CSP2 (**Table 7.9**), indicating that the recombination event could have occurred between two CSP2 strains.

- ST460 (n=22): all the isolates were CSP2 (**Table 7.9**) and the majority of the isolates presented serotype 6A (n=20), whereas the serotype 6B of the two

remaining isolates was originated by a point mutation and by recombination with a *wciP* allele typical of class 2 capsular sequences (6B-2), as described above (*see 7.2.3. Diversity of the wciP and wciN genes*). Although serotype 6B-2 was associated with CSP1, an important fraction of the isolates of this serotype was CSP2 and the horizontal gene transfer could have occurred between CSP2 strains.

- ST473 (n=3): two isolates presented pherotype CSP2 and serotype 6A with the *wciP*-14 allele, an allele present mostly in isolates of CC1876 (**Table 7.6**). The other isolate was serotype 6B-2 and pherotype CSP1. Given the small number of isolates it is difficult to say which is the most frequent combination and horizontal gene transfer in these isolates could have occurred both with capsular and competence genes.

In summary, the capsular switching events identified among serogroup 6 isolates were:

- 6A and 6B-1: 1 recombination event (ST176);
- 6A and 6B-2: 2 recombination events (ST460, ST463);
- 6A and 6C: 1 recombination event (ST395);
- 6B-2 and 6C: 1 recombination event (ST386).

Table 7.8 – Pherotype of serogroup 6 isolates.

Serotype	CSP1	CSP2	CSP3	Total
6A	7 (8.75 %)	73 (91.25 %)	-	80
6B	56 (70.89 %)	21 (26.58 %)	2 (2.53 %)	79
6B-1	*25 (92.59 %)*	*2 (7.41 %)*	*-*	*27*
6B-2	*31 (59.62 %)*	*19 (36.54 %)*	*2 (3.85 %)*	*52*
6C[a]	19 (23.17 %)	63 (76.83 %)	-	82

[a]One isolate presented a PCR product with a size higher than expected indicating the presence of an insertion sequence somewhere between the *comC* and *comD* genes. This isolate was not taken into account in pherotype analysis because its pherotype was not identified.

Table 7.9 – Pherotype of clonal complexes and STs.

CC/ST	CSP1	CSP2	CSP3
CC315	1	45	-
ST386	-	*26*	-
ST887	-	*7*	-
ST315	-	*5*	-
ST3396	-	*5*	-
ST4310	-	*2*	-
ST9985	*1*	-	-
CC395	-	37	-
ST395	-	*17*	-
ST327	-	*8*	-
ST1692	-	*5*	-
ST1714	-	*5*	-
ST8145	-	*1*	-
ST8745	-	*1*	-
CC65	-	35	-
ST460	-	*22*	-
ST65	-	*12*	-
ST9961	-	*1*	-
CC176	26	-	-
ST176	*18*	-	-
ST469	*4*	-	-
ST138	*3*	-	-
ST8136	*1*	-	-
CC273	25	1	-
ST273	*16*	*1*	-
ST2016	*3*	-	-
ST90	*1*	-	-
ST1224	*1*	-	-
ST1369	*1*	-	-
ST1624	*1*	-	-
ST3207	*1*	-	-
ST8144	*1*	-	-
CC1876	1	21	-
ST1876	-	*10*	-
ST473	*1*	*2*	-
ST8140	-	*2*	-
ST1135	-	*1*	-
ST1647	-	*1*	-
ST1879	-	*1*	-
ST5679	-	*1*	-
ST9974	-	*1*	-
ST9988	-	*1*	-
ST10055	-	*1*	-
CC1150	17	-	-
ST1150	*15*	-	-
ST224	*1*	-	-
ST2667	*1*	-	-
CC2611	-	5	-
ST2611	-	*4*	-
ST4580	-	*1*	-
CC681	3	-	-
ST681	*2*	-	-
ST3403	*1*	-	-
CC4248	-	3	-
ST4248	-	*2*	-
ST8141	-	*1*	-
ST1715	2	-	-
ST2185	-	2	-
ST42	1	-	-
ST123	1	-	-
ST179	-	1	-
ST1390	-	1	-
ST1518	1	-	-
ST1648	-	1	-
ST1662	-	1	-
ST3324	-	1	-
ST4246	-	1	-
ST4252	1	-	-
ST4255	-	1	-
ST6175	-	-	1
ST8137	-	1	-
ST9957	-	-	1
ST9970	1	-	-
ST9965	1	-	-
ST10051	1	-	-

The capsular switching events identified on our isolates were not enough to extract any conclusions about a possible pherotype influence on recombination. For that reason, we used the isolate data deposited on the MLST database (downloaded on 16th December 2015) to check how many STs presented two or more serogroup 6 serotypes. The isolate data were filtered by serotype to match exactly 6A, 6B or 6C. Serotype 6D was not taken into account because it was not found in our study and an association with pherotype could not be established. A total of 3955 isolates and 1952 STs were obtained. Then, the STs presenting two or more serogroup 6 serotypes were selected resulting in a total of 95 STs reflecting possible capsular switching events (**Table 7.10**). STs presenting a capsular switch between serotypes 6A and 6B were the most frequent (n=45), followed by the capsular switch between serotypes 6A and 6C (n=32), whereas STs sharing serotypes 6B and 6C were less common (n=7). This may suggest that horizontal gene transfer could occur frequently between CSP1 and CSP2 isolates since serotype 6B isolates were associated with the former pherotype while serotype 6A with the latter. However, it would be important to distinguish between 6B-1 and 6B-2 isolates because the latter presented several CSP2 isolates and maybe the majority of the recombination events could have been between isolates of the same pherotype.

The serogroup 6 isolates characterized in this thesis were divided into subpopulations considering both their serotype and pherotype. An estimation of the gene flow between these subpopulations was performed using the classical F_{ST} parameter (Hudson *et al.*, 1992b) calculated with the concatenated sequence of the MLST genes excluding *ddl*. Low values of F_{ST} represent higher gene flow between the populations, i.e., higher Nm values, where N is the number of individuals in each subpopulation and m is the fraction of migrants in each subpopulation in each generation. The subpopulations 6B-1 CSP2 and 6B-2 CSP3 were not included because they presented just 2 isolates (**Table 7.8**). The results of the estimations showed that the range of the level of gene flow between subpopulations of the same pherotype (CSP1: 0.20-1.71, CSP2: 0.53-1.80) is similar to the level estimated for CSP1 vs. CSP2 subpopulations (0.22-1.88) (**Table 7.11**).

Table 7.10 – No. of STs presenting two or more serogroup 6 serotypes on the MLST isolate database.

Serotype	No. of STs
6A, 6B	45
6A, 6C	32
6A, 6B, 6C	11
6B, 6C	7

Table 7.11 – Gene flow between serogroup 6 subpopulations.

Population 1	Population 2	F_{ST}	*Nm*
CSP1	*CSP1*		
6A CSP1 (n=7)	6B-2 CSP1 (n=31)	0.22601	1.71
6A CSP1 (n=7)	6B-1 CSP1 (n=25)	0.49898	0.50
6A CSP1 (n=7)	6C CSP1 (n=19)	0.54933	0.41
6B-1 CSP1 (n=25)	6B-2 CSP1 (n=31)	0.68763	0.23
6B-2 CSP1 (n=31)	6C CSP1 (n=19)	0.69869	0.22
6B-1 CSP1 (n=25)	6C CSP1 (n=19)	0.71524	0.20
CSP2	*CSP2*		
6B-2 CSP2 (n=19)	6C CSP2 (n=63)	0.21784	1.80
6A CSP2 (n=73)	6C CSP2 (n=63)	0.23343	1.64
6A CSP2 (n=73)	6B-2 CSP2 (n=19)	0.48585	0.53
CSP1	*CSP2*		
6A CSP1 (n=7)	6C CSP2 (n=63)	0.21026	1.88
6B-2 CSP1 (n=31)	6C CSP2 (n=63)	0.28067	1.28
6A CSP1 (n=7)	6B-2 CSP2 (n=19)	0.29402	1.20
6A CSP1 (n=7)	6A CSP2 (n=73)	0.31865	1.07
6B-2 CSP1 (n=31)	6A CSP2 (n=73)	0.39235	0.77
6B-2 CSP1 (n=31)	6B-2 CSP2 (n=19)	0.42847	0.67
6B-1 CSP1 (n=25)	6C CSP2 (n=63)	0.50661	0.49
6C CSP1 (n=19)	6C CSP2 (n=63)	0.52256	0.46
6B-1 CSP1 (n=25)	6A CSP2 (n=73)	0.54957	0.41
6C CSP1 (n=19)	6A CSP2 (n=73)	0.61838	0.31
6B-1 CSP1 (n=25)	6B-2 CSP2 (n=19)	0.63801	0.28
6C CSP1 (n=19)	6B-2 CSP2 (n=19)	0.69149	0.22
Serogroup 6 CSP1	Serogroup 6 CSP2	0.15065	2.82

Since poor evidence of pherotype influence on the genetic structure was obtained from serogroup 6 isolates, a further analysis was performed using the MLST data of the children invasive collection (n=903) studied in detail in Chapter 5. For this analysis, the serotype 25A/38-ST393 isolates and the 6C isolate with an insertion in the *comC* gene (*see 5.2.1. Pherotype abundance and evolution*) were excluded because their pherotype was not determined, giving a total of n=893 sequences for this analysis.

A similar approach as described in Carrolo *et al.* (Carrolo *et al.*, 2009) was followed using the F_{ST} parameter (Hudson *et al.*, 1992b) to estimate the level of gene flow and the K^*_{ST} (Hudson *et al.*, 1992a) and S_{nn} (Hudson, 2000) statistics to evaluate genetic differentiation between pherotype subpopulations. However, in addition to the analysis using all strains (All dataset, n=893, **Table 7.12**), another two analyses were performed with smaller datasets. One of these datasets was formed by non-duplicate sequences within each pherotype subpopulation (Unique dataset, n=199, **Table 7.12**). This corresponds almost to using a single sequence

per ST, but a single sequence was kept in the cases when two STs differed only in the *ddl* gene, since this gene was not included in this analysis. It also means that unique sequences presenting different pherotypes were duplicated because they were included in different subpopulations. The use of non-duplicated sequences reduced a potential bias of sampling which could have altered the frequency of each ST in the sample relative to the one actually found in the pneumococcal population. The other dataset used contained just the sequences of the founder STs of the CCs presented in Chapter 5 defined by the software PHYLOViZ (CC founders dataset, n=71, **Table 7.12**) (see **Table 5.3** and **Figure 5.9**). If a founder ST presented two pherotypes, its sequence was included in the dominant pherotype of its CC (a pherotype was considered dominant if expressed by >50 % of the isolates of that CC). With this approach, I attempted to reduce the effect of the clonal structure of the pneumococcal population but this came with the cost of dismissing all the information provided by the evolution within the CCs.

Regarding the genetic analyses including all isolates (All dataset) and all pherotype populations (CSP1, CSP2 and CSP3) of each of the MLST genes separately and the sequence that resulted from their concatenation, it could be observed that statistically significant K^*_{ST} and S_{nn} values were obtained in all the tests performed, indicating that pherotype populations were genetically distinct (**Table 7.13**). A rather low gene flow between the pherotype populations was also seen when considering the F_{ST} and *Nm* parameters and, as expected due to its hitchhiking effect with *pbp2b* and penicillin resistance, *ddl* presented the highest gene flow (**Table 7.13**). The analysis between each population (CSP1 vs. CSP2, CSP1 vs. CSP3 and CSP2 vs. CSP3) with all the sequences also produced significant results regarding genetic differentiation (**Table 7.14**, All dataset). The highest differentiation was obtained when comparing CSP1 and CSP3 populations. When I minimized the effect of the frequency of each ST, the pherotype populations were still considered genetically different, with the exception of CSP2 and CSP3 populations regarding the K^*_{ST} statistic, but the gene flow became higher (**Table 7.14**, Unique dataset). However, by removing the influence of the intrinsic pneumococcal clonal structure in these tests the populations were no longer considered to be genetically different (**Table 7.14**, CC founders dataset). The smaller number of sequences included in the analyses of Unique and CC founders datasets could have reduced the power of the tests, although, at least for the statistic S_{nn}, it was proven by previous literature to retain its power even when

using sequences from few individuals (Hudson, 2000), so it is unlikely that this is the only explanation.

The phenotype distribution among the pneumococcal genetic lineages was evaluated by constructing a minimum spanning tree (MST) using PHYLOViZ online tool (Ribeiro-Gonçalves *et al.*, 2016) (**Figure 7.5**). Phenotype was evenly distributed through the MST and genetic subpopulations could not be identified regarding phenotype.

Table 7.12 – Number of sequences per population in the datasets used in gene flow and genetic differentiation analyses.

Dataset	CSP1	CSP2	CSP3	Total
All	681	192	20	893
Unique	127	63	9	199
CC founders	40	24	7	71

Table 7.13 – Gene flow and genetic differentiation of MLST genes and the concatenated sequence of them, with the exception of *ddl,* using all the isolates and all phenotype populations.

Dataset: All	Populations: CSP1, CSP2, CSP3						
Allele	π	F_{ST}	*Nm*	K^*_{ST}	p (K^*_{ST})	S_{nn}	p (S_{nn})
aroE	0.00466	0.21766	1.80	0.07097	**<0.001**	0.74358	**<0.001**
ddl	0.00922	0.06180	7.59	0.03778	**<0.001**	0.75812	**<0.001**
gdh	0.00803	0.16990	2.44	0.05818	**<0.001**	0.83261	**<0.001**
gki	0.01880	0.12372	3.54	0.04209	**<0.001**	0.75104	**<0.001**
recP	0.00497	0.09893	4.55	0.04652	**<0.001**	0.78980	**<0.001**
spi	0.00798	0.07396	6.26	0.03160	**<0.001**	0.71080	**<0.001**
xpt	0.00821	0.10782	4.14	0.05144	**<0.001**	0.79408	**<0.001**
concatenated	0.00894	0.12523	3.49	0.04043	**<0.001**	0.94655	**<0.001**

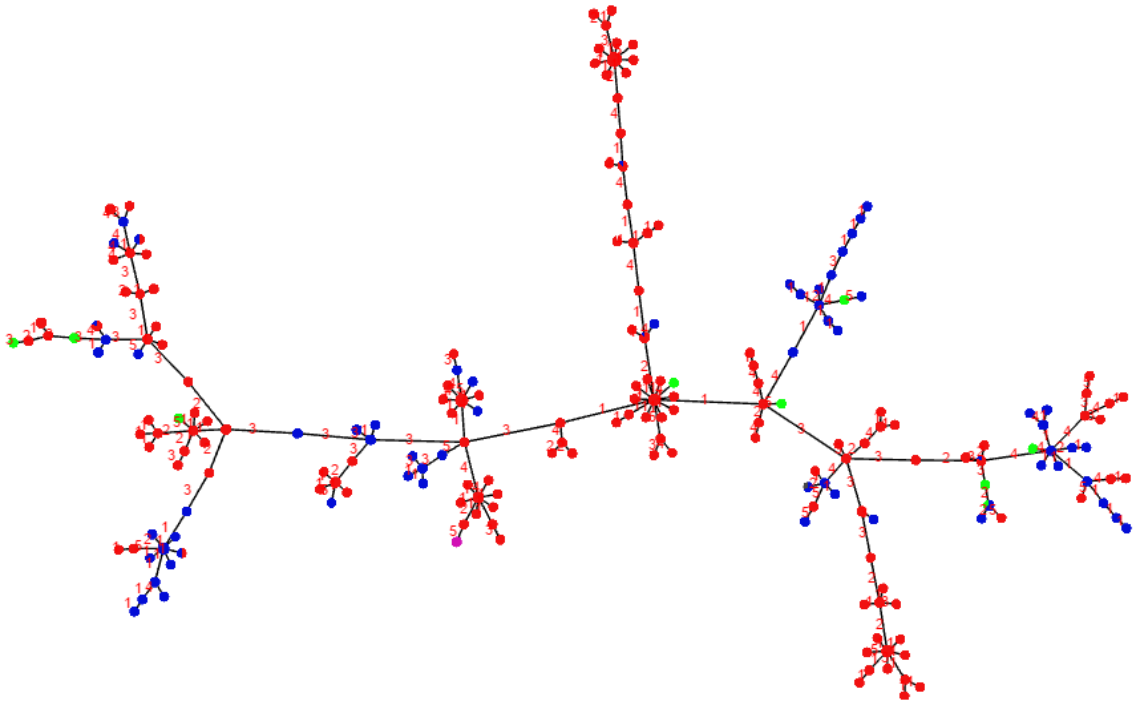
Table 7.14 – Gene flow and genetic differentiation between pherotype populations using the concatenated sequence of the MLST genes, with exception of *ddl,* and the datasets All, Unique and CC founders.

Dataset	Populations	n	π	F_{ST}	*Nm*	K^*_{ST}	p (K^*_{ST})	S_{nn}	p (S_{nn})
All[a]	CSP1, CSP2, CSP3	893	0,00894	0,12523	3,49	0,04043	**<0.001**	0,94655	**<0.001**
	CSP1, CSP2	873	0,00894	0,12383	3,54	0,03629	**<0.001**	0,96400	**<0.001**
	CSP1, CSP3	701	0,00869	0,14248	3,01	0,00787	**<0.001**	0,99061	**<0.001**
	CSP2, CSP3	212	0,00817	0,10816	4,12	0,02062	**<0.001**	0,92396	**<0.001**
Unique[b]	CSP1, CSP2, CSP3	199	0,01007	0,03047	15,91	0,00943	**<0.001**	0,72822	**<0.001**
	CSP1, CSP2	190	0,01004	0,04900	9,70	0,00867	**<0.001**	0,77588	**<0.001**
	CSP1, CSP3	136	0,01013	0,04399	10,87	0,00276	**0,0180**	0,93382	**<0.001**
	CSP2, CSP3	72	0,00947	<0	<0	0,00200	0,1880	0,84722	**0,0460**
CC founders[c]	CSP1, CSP2, CSP3	71	0,00990	<0	<0	0,00151	0,2340	0,44601	0,4450
	CSP1, CSP2	64	0,00979	0,00356	139,87	0,00071	0,2600	0,51823	0,5500
	CSP1, CSP3	43	0,00976	0,00870	56,98	0,00231	0,1750	0,78723	0,2740
	CSP2, CSP3	31	0,01030	<0	<0	-0,00039	0,5040	0,67742	0,3300

[a]All the sequences.

[b]Non-duplicated sequences within each pherotype population.

[c]Sequences of the founders of CCs.



Red: CSP1, **Blue**: CSP2, **Green**: CSP3 and **Pink**: Unidentified pherotype.

Figure 7.5 – Pherotype distribution among the STs connected in a minimum spanning tree (MST) constructed from the MLST data of a collection of pneumococci causing IPD in children (n=903). The MST was constructed by using PHYLOViZ online (Ribeiro-Gonçalves *et al.*, 2016). Each circle represents a ST colored by the pherotype identified in its isolates. These data are available at http://bit.do/children_invasive.

7.3. Discussion

The recent identification of new serotypes (6D, 6F, 6G and 6H) in serogroup 6 prompted us to re-evaluate our collection of invasive isolates. However, these new serotypes were not present among our collection and their rarity indicates that currently available phenotypic or simple genotypic assays remain adequate for distinguishing serotypes within serogroup 6 isolates. The data presented here also revealed that 66 % (52/79) of the 6B isolates presented class 2 capsular sequences and the observed decline was due to these isolates. In fact, the decrease of serotype 6B was the result of the reduction of one major genetic lineage, the CC273, both in children and in adults.

Serotype 6B-2 isolates were confirmed to produce the same capsule as serotype 6B-1 pneumococci (Burton *et al.*, 2016), so it is reasonable to think that PCV7 confers the same protection against them. In fact, killing of serotype 6B-2 pneumococci mediated by vaccine-induced serotype 6B antibodies was shown *in vitro* (van Tonder *et al.*, 2015). However, the data reported here only indicate a decrease of serotype 6B-2 isolates after PCV7 introduction in Portugal, especially from the CC273, while serotype 6B-1 isolates did not present a significant change. In South Korea a similar situation was reported but in this country it was the frequency of serotype 6B-1 pneumococci that decreased after introduction of PCV7, while the observed increase of serotype 6B-2 isolates was not statistically supported (Baek *et al.*, 2014b). Conjugate vaccine protection against both types of serotype 6B could be similar and the fact that this protection against serotype 6B-1 and serotype 6B-2 pneumococci was not seen in Portugal and South Korea, respectively, could be because the prevalence of these serotypes was already low before introduction of PCV7 in the respective countries. The universal uptake of PCV13 by children, expected with the introduction of this vaccine in the National Immunization Plan in 2015 in Portugal, could reduce the prevalence of serotype 6B-1 pneumococci in the next years. Regarding data of pneumococcal carriage in children in Portugal, a decrease of not only isolates belonging to CC273 but also of CC315 was observed after introduction of PCV7 (Simões *et al.*, 2011). In the study reported here, CC315 presented mostly serotype 6B-2 isolates in children, so the observed decrease in carriage of this clonal complex could have been due to the reduction of serotype 6B-2 and not serotype 6C pneumococci. A decrease of serotype

6B-2 pneumococci was also observed in San Sebastian, Spain, and CC273 and CC315 lineages were identified in this region (Marimon *et al.*, 2016). A decrease of isolates of CC273 was also reported in Barcelona, Spain, during 1994-2008 (Rolo *et al.*, 2011a), confirming the reduction of this genetic lineage in Spain. In a whole genome genetic screen of serogroup 6 isolates from diverse regions of the world, the majority of serotype 6B pneumococci were revealed to present class 2 capsular sequences and they were worldwide distributed (van Tonder *et al.*, 2015). Thus, a decrease of serotype 6B-2 pneumococci is expected to have occurred in the countries that implemented conjugate vaccination. The antibiotic resistance profile of serotype 6B-1 and 6B-2 pneumococci was distinct, with serotype 6B-2 isolates presenting a higher degree of MDR than serotype 6B-1 pneumococci. This could be one of the reasons for the selection and expansion of serotype 6B-2 prior to the availability of conjugate vaccines. However, further studies are necessary to clarify potential differences between 6B-1 and 6B-2 and their response to vaccination.

Following the reduction of serotype 6B invasive infections, serotype 6A isolates became the most frequently found among serogroup 6 pneumococci after the introduction of PCV7 in Portugal suggesting that a cross-protection of PCV7 against these pneumococci did not occur. Thus, the introduction of PCV13, which includes serotype 6A polysaccharides in its formulation, was expected to reduce the proportion of invasive cases caused by serotype 6A pneumococci. However, this reduction was observed only in adults and not in children who were directly vaccinated. The stabilization of the number of isolates received by our laboratory from 2007 onwards made it possible to calculate the incidence of IPD in children and a reduction of the total invasive cases was observed after the introduction of PCV13 (Aguiar *et al.*, 2014). Thus, analysis of the proportions of infections in children in the post-PCV13 period (2010-2012) needs to be carefully done since it may be misleading. For example, the increase of the proportion of serotype 6A isolates seen in 2011 in children does not correspond to an increase of the number of cases caused by isolates expressing this serotype comparing with the period 2007-2009 (**Table 7.2**). In fact, the number of infections caused by serotype 6A pneumococci decreased from 14 to 8 comparing this period with the period after PCV13 introduction. Thus, a protection of PCV13 against serotype 6A pneumococci cannot be discarded.

The prevalence of serogroup 6 isolates in Portugal was higher in children than in adults. However, serotype 6C pneumococci were responsible for a small proportion of IPD cases in children. Several countries also reported infrequent cases of invasive infections by serotype 6C pneumococci among children (du Plessis *et al.*, 2008, Millar *et al.*, 2010, Rolo *et al.*, 2011b, van der Linden *et al.*, 2013a). The proportion of serotype 6C pneumococci among isolates recovered from carriage in children during the years 1999-2007 in Portugal was 3.0 % (Nunes *et al.*, 2009), contrasting with the prevalence of just 0.4 % (2/457, **Table 7.2**) in IPD in children of these pneumococci during the same period. This could mean that serotype 6C isolates could have a decreased ability to cause invasive disease in this age group, although some clones may be more successful than others and its prevalence could rise in children IPD. In fact, although not seen in this study, a rise in IPD cases of serotype 6C pneumococci was observed in 2009 in Spain (Rolo *et al.*, 2011b). On the other side, the proportion of serotype 6C pneumococci infections in adults was completely different. These pneumococci became the most prevalent of the invasive serogroup 6 isolates recovered from adults and were mainly distributed between three major genetic lineages (CC315, CC395 and CC1150). Although an increase of serotype 6C isolates in adults was not statistically supported, expansion of CC315 due to serotype 6C pneumococci occurred in this age group starting around 2008. In children's carriage in Portugal, isolates of ST3396 (CC315) were recovered only in the last years of the study (2006-2007) and isolates of CC395 and CC1150 were already present before introduction of conjugate vaccine (Nunes *et al.*, 2009). Increase of the prevalence of serotype 6C pneumococci was reported in several regions (Loman *et al.*, 2013, Millar *et al.*, 2010, Nahm *et al.*, 2009, Rolo *et al.*, 2011b, van der Linden *et al.*, 2013a). However, the genetic lineage that increased was not always the CC315. For example, in Southampton, England, the increase of serotype 6C pneumococci in carriage was due to the clonal expansion of CC395 (Loman *et al.*, 2013). In Spain, the authors associated the increase of serotype 6C isolates with spread of CC1150 (Rolo *et al.*, 2011b), although isolates of CC315 emerged in 2007, coinciding with the data from IPD and carriage in Portugal. Expansion of just CC315 from the three major serotype 6C genetic lineages present in Portugal could have occurred due to selection by antibiotics consumption, since CC315 presented MDR and CC1150 and CC395 were more susceptible to antimicrobials. Thus, expansion of resistant 6C lineages in adults, but not in children where the same lineages express other serogroup 6 serotypes, may be a

hallmark of the post-PCV period and suggest that vaccination and antimicrobial use are the two major forces currently shaping pneumococcal populations.

Capsular switching events are relatively common in serogroup 6 (Mavroidi *et al.*, 2004). In this study, five STs suggested that a capsular switch has occurred in some of their isolates. In serogroup 6, capsular switching may occur by horizontal transfer (whole capsular loci or part of it) or by a spontaneous mutation in a single nucleotide, namely in the amino acid position 195 of WciP. Although capsular switch by horizontal transfer appeared to be the most common mechanism, one isolate in this study was confirmed to switch its capsule by a point mutation like the case described by Mavroidi and colleagues (Mavroidi *et al.*, 2004). Another characteristic of serogroup 6 is the occurrence of hybrid capsular sequences, particularly between class 1 and class 2 sequences (Mavroidi *et al.*, 2004). This was recently highlighted in a whole genome screen of serogroup 6 isolates that reported a hybrid between serotype 6C and serotype 6B-2 capsular loci (van Tonder *et al.*, 2015). In this study, a ST460 isolate also presented a hybrid sequence between classes 1 and 2 sequences. It remains unclear if some cases of hybrid capsular loci in serogroup 6 represent new serotypes but the hybrid described here should be serotype 6B since the genetic sequences are from two classes of serotype 6B capsular sequences.

A decline of serotype 6B-2 pneumococci, particularly of CC273, and not of serotype 6B-1 pneumococci was observed in Portugal. Along with this decline, expansion of CC315 due to serotype 6C isolates was also observed. This suggested that introduction of PCV7 protected against serotype 6B-2 pneumococci infections, while this protection should not be discarded in serotype 6B-1 pneumococci and, with introduction of PCV13, in serotype 6A isolates.

Expansion of CC315 was not responsible for the increase of serotype 6C pneumococci in some regions, so genetic surveillance either by MLST or whole genome sequencing is important to perform along with serotyping and determination of the antimicrobial resistance profiles.

The pherotype influence on the pneumococcal genetic structure was also explored in this chapter. The serogroup 6 collection was used as a case-study in a first approach to this subject under the premise that serogroup 6 serotypes were associated to different pherotypes and the capsular switching events occurring between them would be representative of intra- and inter-pherotype genetic recombination. A total of 5 cases of probable capsular switching events were identified among serogroup 6 isolates, a low number that did not allow any solid conclusions. In most of these cases, a dominant serotype was identified, and the capsular switch was initially assumed to have been from the dominant serotype to the least commonly found serotype in a given ST. By taking ST386 isolates (n=26) as an example, where most of them were serotype 6C (n=22) while serotype 6B-2 was identified in fewer isolates (n=4), the capsular switch could have been from 6C to 6B-2. However, a careful interpretation should be done because the capsular switch could have been exactly in the opposite direction. In fact, it may have been from 6B-2 to 6C, followed by an increase of 6C ST386 isolates, an interpretation supported by the observed increase of CC395 due to 6C isolates documented in this thesis. The MLST isolate database was also used to further explore if pherotype influence on recombination could be inferred from serogroup 6 capsular switching events. Most of the capsular switching events identified using this database were between serotype 6A and 6B, which, under the premise of this study, would mean recombination between pherotype CSP2 and CSP1. However, in addition to the fact that a significant proportion of 6B-2 isolates presented CSP2 pherotype, the difficulty to quantify the expected capsular switching rate between serogroup 6 serotypes under different scenarios have not allowed any conclusion to be drawn from this analysis. Thus, pherotype influence on capsular switching events and hence in genetic recombination could not be properly evaluated using serogroup 6 isolates.

Pherotype influence on the genetic recombination was further explored by using a collection of pneumococcal isolates causing IPD in children between 1999 and 2012, which was studied in Chapter 5. By using this large collection (n=903), it was expected to elucidate if pherotype was really defining genetically different subpopulations as was found in previous work from our lab (Carrolo *et al.*, 2009) or if it did not have an effect on recombination between populations of different pherotypes as was suggested by Cornejo *et al.* (Cornejo *et al.*, 2010). These studies

used rather small collections so the analyses performed here with a larger collection could shed light on the subject. The results obtained here showed that pherotype populations were indeed genetically different even when the frequency of each ST was not taken into account. However, this result could be expected because of the clonal structure of the pneumococcal population. For example, if two subpopulations were chosen from the CSP2 population and if the STs belonging to the same CC were kept together in the same subpopulation, the genetic differentiation tests would also report those subpopulations as genetically different, despite all being CSP2 strains. By using the CC founder dataset, we tried to eliminate the effect of the clonal structure on the results obtained from the genetic differentiation tests using the All and Unique datasets. In this case, a genetic differentiation between CSP1 and CSP2 populations could no longer be seen. Moreover, the analysis of the pherotype distribution through the genetic backgrounds connected between them in a minimum spanning tree failed to identify subpopulations segregated by pherotype. However, these results do not unequivocally prove that pherotype does not restrict recombination between pherotypes and further analyses should be performed. Recently, a study using whole genome data from a total of 4089 isolates was unable to find an effect of pherotype on genetic structure and recombination (Miller *et al.*, 2017). Their results showed that pherotype does not restrict recombination between strains presenting a different pherotype and they even suggested that recombination between those strains could be slightly facilitated. Definitively, their results should be taken into account in the scope of the work performed in this chapter. Thus, the influence of the pherotype on the genetic structure of *S. pneumoniae* was not clear considering all the information collected in this thesis.

IV. Final remarks and future perspectives

IV. FINAL REMARKS AND FUTURE PERSPECTIVES

The work performed under the scope of this thesis aimed to strengthen the knowledge on horizontal gene transfer and its importance for the evolution of *S. pneumoniae*. To achieve this goal it was proposed to:

- Study the genetic diversity of the pherotype-defining region *comCDE*, which was done by sequencing this region from a collection of n=89 pneumococcal isolates causing IPD in Portugal during 1999-2001;
- Evaluate pherotype abundance and its evolution and the pherotypes that could be identified in the pneumococcal population. This goal was accomplished by studying a large collection of invasive pneumococci recovered from children IPD during a period of 14 years when conjugate vaccines were introduced.
- Explore pherotype specificity by identifying the structural determinants responsible for this putative lock-and-key mechanism. Here the strategy followed was to determine the 3D structure of the ComD sensor in complex with the respective CSP. Although several experiments were performed, this goal was not achieved.
- Further study the influence of pherotype on genetic structure, which was explored by using a serogroup 6 pneumococci collection and also a large collection of isolates causing IPD in children.

The main findings of this thesis were:

- Pherotype abundance remained stable in a period when extensive changes occurred in serotype distribution and in frequency of some pneumococcal clones;
- Recognition of CSP3 as a pneumococcal pherotype by identification of CSP3 strains circulating in Portugal and quantification of the CSP3 pherotype abundance;
- ComD could be a membrane-embedded sensor making it difficult the production of this protein;
- Influence of the pherotype on the genetic structure of *S. pneumoniae* is not clear.

S. pneumoniae presents polymorphism in the *comCD* genes and the interaction between their products could function like a lock-and-key mechanism. These genes were observed to be under positive selection (Ichihara *et al.*, 2006), indicating that there is a high probability that mutations in this region cause a decrease of the efficiency of competence induction and hence of fitness. Pneumococci enter in the competent state in response to cell density and environmental conditions such as pH and antibiotic stress (Moreno-Gámez *et al.*, 2017, Weyder *et al.*, 2018). In this state, pneumococci are not only able to undergo genetic recombination but they also express other genes. Fratricide was identified as a molecular mechanism induced during competence and the lysis of non-competent cells was proposed to increase the homologous DNA available for genetic transformation. The main benefit of pneumococcal transformation was suggested to be gene repair instead of the acquisition of new traits (Ambur *et al.*, 2016). Naturally, genetic recombination proved to be extremely beneficial in some cases where strong selective pressures existed, as for example in recombination events involving the acquisition of antimicrobial resistance genes or capsular switching events enabling the evasion of vaccine protection. Inter-pherotype recombination could theoretically be more advantageous because pneumococci would have a bigger reservoir of genetic diversity than if they were restricted to exchange information with just strains of the same pherotype. However, it can be also the case that the specificity between some ComD variants and their cognate CSPs may be not so strict, allowing responses to heterologous CSPs further confusing the issue of the potential genetic segregation mediated by the pherotypes. Moreover, it was also suggested that the main reservoir of genetic variability of *S. pneumoniae* resides in other streptococcal species, namely *Streptococcus mitis* (Donati et al., 2010) which do not share the same CSP alleles of *S. pneumoniae*.

Pherotype proportions of approximately 70:28:2 (CSP1:CSP2:CSP3) seem to be maintained in pneumococcal populations from diverse geographic sites and isolation dates, suggesting that a mechanism enforcing these proportions could exist, although it is still not clear which are the forces behind it. The possibility of a yet undiscovered mechanism that could contribute to stabilize pherotype abundance could exist because up to 124 genes are induced during competence and only a small set of them were identified to be directly involved in genetic transformation. Thus, like the discovery of fratricide, there could be still unrevealed phenotypes induced during competence which could also be under

selection. However, a bigger temporal scale may be needed to evaluate if pherotype abundance is indeed being actively stabilized by selection or if it is slowly changing at a pace not measurable by the temporal scale of 14 years used in this work, although the introduction of the vaccine would be expected to have been a sufficiently strong perturbation to cause changes in the absence of other stabilizing selective forces. The pneumococcal pherotypes exhibit different phenotypes as observed by a previous study performed by our laboratory with mutants lacking the *comC* gene which indicated that CSP1 strains were able to respond much better to their cognate peptide than CSP2 strains, resulting in a better capacity to form biofilms and a higher transformation efficiency of CSP1 in relation to CSP2 strains (Carrolo *et al.*, 2014). These results suggested either a stronger binding of CSP1 to its receptor or an amplification of the response in the CSP1 genetic background and could explain the higher prevalence of the *comC1* allele in the pneumococcal population.

It is difficult to state if there was an original pneumococcal pherotype, but it is likely that it was either CSP1 or CSP2, or maybe an ancestor of both pherotypes. CSP3 seems to have been acquired from a *Streptococcus mitis* strain because the *comC3* allele and the intergenic space between *comC* and *comD* presented sequence lengths different from CSP1 and CSP2 strains whereas that intergenic space was of the same length as that seen in a sequence of the *comCDE* operon from an *S. mitis* strain (Whatmore *et al.*, 1999).

Future work should focus on the evaluation of pherotype abundance in the period after introduction of PCV13 to ascertain if pherotype abundance really remained stable or an increase of CSP2 has occurred. The pherotype of non-invasive isolates and of adult isolates could also be identified to check possible differences in comparison to invasive isolates causing IPD in children. The structural determination of the specificity between ComD and CSP should consider that ComD may be a membrane-embedded sensor and production of this protein should be attempted using systems mimicking cellular membranes.

V. References

V. REFERENCES

Adrian, P. V. and K. P. Klugman (1997). *Mutations in the dihydrofolate reductase gene of trimethoprim-resistant isolates of Streptococcus pneumoniae*. Antimicrob Agents Chemother, 41(11), 2406-13.

Aguiar, S. I., M. J. Brito, J. Goncalo-Marques, J. Melo-Cristino and M. Ramirez (2010a). *Serotypes 1, 7F and 19A became the leading causes of pediatric invasive pneumococcal infections in Portugal after 7 years of heptavalent conjugate vaccine use*. Vaccine, 28(32), 5167-73.

Aguiar, S. I., M. J. Brito, J. Gonçalo-Marques, J. Melo-Cristino and M. Ramirez (2010b). *Serotypes 1, 7F and 19A became the leading causes of pediatric invasive pneumococcal infections in Portugal after 7 years of heptavalent conjugate vaccine use*. Vaccine, 28(32), 5167-73.

Aguiar, S. I., M. J. Brito, A. N. Horácio, J. P. Lopes, M. Ramirez and J. Melo-Cristino (2014). *Decreasing incidence and changes in serotype distribution of invasive pneumococcal disease in persons aged under 18 years since introduction of 10-valent and 13-valent conjugate vaccines in Portugal, July 2008 to June 2012*. Euro Surveill, 19(12), 20750.

Aguiar, S. I., M. J. Frias, L. Santos, J. Melo-Cristino and M. Ramirez (2006). *Emergence of optochin resistance among Streptococcus pneumoniae in Portugal*. Microb Drug Resist, 12(4), 239-45.

Aguiar, S. I., J. Melo-Cristino and M. Ramirez (2012). *Use of the 13-valent conjugate vaccine has the potential to eliminate pilus carrying isolates as causes of invasive pneumococcal disease*. Vaccine, 30(37), 5487-90.

Aguiar, S. I., F. R. Pinto, S. Nunes, I. Serrano, J. Melo-Cristino, R. Sa-Leao, M. Ramirez and H. de Lencastre (2010c). *Denmark14-230 clone as an increasing cause of pneumococcal infection in Portugal within a background of diverse serotype 19A lineages*. J Clin Microbiol, 48(1), 101-8.

Aguiar, S. I., F. R. Pinto, S. Nunes, I. Serrano, J. Melo-Cristino, R. Sá-Leão, M. Ramirez and H. de Lencastre (2010d). *Denmark14-230 clone as an increasing cause of pneumococcal infection in Portugal within a background of diverse serotype 19A lineages*. J Clin Microbiol, 48(1), 101-8.

Aguiar, S. I., I. Serrano, F. R. Pinto, J. Melo-Cristino and M. Ramirez (2008). *Changes in Streptococcus pneumoniae serotypes causing invasive disease with non-universal vaccination coverage of the seven-valent conjugate vaccine*. Clin Microbiol Infect, 14(9), 835-43.

Ambur, O. H., J. Engelstadter, P. J. Johnsen, E. L. Miller and D. E. Rozen (2016). *Steady at the wheel: conservative sex and the benefits of bacterial transformation*. Philos Trans R Soc Lond B Biol Sci, 371(1706).

Angulo, I., I. Acebrón, B. de las Rivas, R. Muñoz, I. Rodríguez-Crespo, M. Menéndez, P. García, H. Tateno, I. J. Goldstein, B. Pérez-Agote*, et al.* (2011). *High-resolution structural insights on the sugar-recognition and fusion tag properties of a versatile beta-trefoil lectin domain from the mushroom Laetiporus sulphureus*. Glycobiology, 21(10), 1349-61.

Aoki, H., K. Dekany, S. L. Adams and M. C. Ganoza (1997). *The gene encoding the elongation factor P protein is essential for viability and is required for protein synthesis*. J Biol Chem, 272(51), 32254-9.

Appelbaum, P. C. (1987). *World-wide development of antibiotic resistance in pneumococci*. Eur J Clin Microbiol, 6(4), 367-77.

Ardanuy, C., F. Tubau, R. Pallares, L. Calatayud, M. A. Dominguez, D. Rolo, I. Grau, R. Martin and J. Linares (2009). *Epidemiology of invasive pneumococcal disease among adult patients in barcelona before and after pediatric 7-valent pneumococcal conjugate vaccine introduction, 1997-2007*. Clin Infect Dis, 48(1), 57-64.

Arthur, M., A. Andremont and P. Courvalin (1987). *Distribution of erythromycin esterase and rRNA methylase genes in members of the family Enterobacteriaceae highly resistant to erythromycin*. Antimicrob Agents Chemother, 31(3), 404-9.

Ayoubi, P., A. O. Kilic and M. N. Vijayakumar (1991). *Tn5253, the pneumococcal omega (cat tet) BM6001 element, is a composite structure of two conjugative transposons, Tn5251 and Tn5252*. J Bacteriol, 173(5), 1617-22.

Baek, J. Y., I. H. Park, T. M. So, M. K. Lalitha, N. Shimono, R. M. Yasin, C. C. Carlos, J. Perera, V. Thamlikitkul, P. R. Hsueh*, et al.* (2014a). *Prevalence and characteristics of Streptococcus pneumoniae "putative serotype 6E" isolates from Asian countries*. Diagn Microbiol Infect Dis, 80(4), 334-7.

Baek, J. Y., I. H. Park, J. H. Song and K. S. Ko (2014b). *Prevalence of isolates of Streptococcus pneumoniae putative serotype 6E in South Korea*. J Clin Microbiol, 52(6), 2096-9.

Bauer, A. W., W. M. Kirby, J. C. Sherris and M. Turck (1966). *Antibiotic susceptibility testing by a standardized single disk method*. Am J Clin Pathol, 45(4), 493-6.

Baylay, A. J., A. Ivens and L. J. Piddock (2015). *A novel gene amplification causes upregulation of the PatAB ABC transporter and fluoroquinolone resistance in Streptococcus pneumoniae*. Antimicrob Agents Chemother, 59(6), 3098-108.

Baylay, A. J. and L. J. Piddock (2015). *Clinically relevant fluoroquinolone resistance due to constitutive overexpression of the PatAB ABC transporter in Streptococcus pneumoniae is conferred by disruption of a transcriptional attenuator*. J Antimicrob Chemother, 70(3), 670-9.

Benjamini, Y. and Y. Hochberg (1995). *Controlling the false discovery rate: a pratical and powerful approach to multiple testing*. Journal of the Royal Statistical Society. Series B (Methodological), 57(1), 289-300.

Billal, D. S., J. Feng, P. Leprohon, D. Legare and M. Ouellette (2011). *Whole genome analysis of linezolid resistance in Streptococcus pneumoniae reveals resistance and compensatory mutations*. BMC Genomics, 12, 512.

Bobkova, E. V., Y. P. Yan, D. B. Jordan, M. G. Kurilla and D. L. Pompliano (2003). *Catalytic properties of mutant 23 S ribosomes resistant to oxazolidinones*. J Biol Chem, 278(11), 9802-7.

Bogaert, D., A. van Belkum, M. Sluijter, A. Luijendijk, R. de Groot, H. C. Rumke, H. A. Verbrugh and P. W. Hermans (2004). *Colonisation by Streptococcus pneumoniae and Staphylococcus aureus in healthy children*. Lancet, 363(9424), 1871-2.

Bozdogan, B. and P. C. Appelbaum (2004). *Oxazolidinones: activity, mode of action, and mechanism of resistance*. Int J Antimicrob Agents, 23(2), 113-9.

Bratcher, P. E., I. H. Park, S. K. Hollingshead and M. H. Nahm (2009). *Production of a unique pneumococcal capsule serotype belonging to serogroup 6*. Microbiology, 155(Pt 2), 576-83.

Bratcher, P. E., I. H. Park, M. B. Oliver, M. Hortal, R. Camilli, S. K. Hollingshead, T. Camou and M. H. Nahm (2011). *Evolution of the capsular gene locus of Streptococcus pneumoniae serogroup 6*. Microbiology, 157(Pt 1), 189-98.

Brito, D. A., M. Ramirez and H. de Lencastre (2003). *Serotyping Streptococcus pneumoniae by multiplex PCR*. J Clin Microbiol, 41(6), 2378-84.

Brueggemann, A. B., R. Pai, D. W. Crook and B. Beall (2007). *Vaccine escape recombinants emerge after pneumococcal vaccination in the United States*. PLoS Pathog, 3(11), e168.

Burton, R. L., K. A. Geno, J. S. Saad and M. H. Nahm (2016). *Pneumococcus with the "6E" cps Locus Produces Serotype 6B Capsular Polysaccharide*. J Clin Microbiol, 54(4), 967-71.

C.L.S.I. (2007). *Performance standards for antimicrobial susceptibility testing - seventeenth informational supplement*. CLSI document M100-S17, Clinical and Laboratory Standards Institute, Wayne, PA.

C.L.S.I. (2009). *Methods for dilution antimicrobial susceptibility tests for bacteria that grow aerobically; approved standard - eight edition*. CLSI document M07-A8, Clinical and Laboratory Standards Institute, Wayne, PA.

C.L.S.I. (2014). *Performance standards for antimicrobial susceptibility testing - twenty-fourth informational supplement*. CLSI document M100-S24, Clinical and Laboratory Standards Institute, Wayne, PA.

Campbell, E. A., S. Y. Choi and H. R. Masure (1998). *A competence regulon in Streptococcus pneumoniae revealed by genomic analysis*. Mol Microbiol, 27(5), 929-39.

Carriço, J. A., C. Silva-Costa, J. Melo-Cristino, F. R. Pinto, H. de Lencastre, J. S. Almeida and M. Ramirez (2006). *Illustration of a common framework for relating multiple typing methods by application to macrolide-resistant Streptococcus pyogenes*. J Clin Microbiol, 44(7), 2524-32.

Carrolo, M., F. R. Pinto, J. Melo-Cristino and M. Ramirez (2009). *Pherotypes are driving genetic differentiation within Streptococcus pneumoniae*. BMC Microbiol, 9, 191.

Carrolo, M., F. R. Pinto, J. Melo-Cristino and M. Ramirez (2014). *Pherotype influences biofilm growth and recombination in Streptococcus pneumoniae*. PLoS One, 9(3), e92138.

Chandler, M. S. and D. A. Morrison (1988). *Identification of two proteins encoded by com, a competence control locus of Streptococcus pneumoniae*. J Bacteriol, 170(7), 3136-41.

Chao, Y., L. R. Marks, M. M. Pettigrew and A. P. Hakansson (2014). *Streptococcus pneumoniae biofilm formation and dispersion during colonization and disease*. Front Cell Infect Microbiol, 4, 194.

Chewapreecha, C., S. R. Harris, N. J. Croucher, C. Turner, P. Marttinen, L. Cheng, A. Pessia, D. M. Aanensen, A. E. Mather, A. J. Page*, et al.* (2014). *Dense genomic sampling identifies highways of pneumococcal recombination*. Nat Genet, 46(3), 305-309.

Choi, Y. H., M. Jit, N. Gay, N. Andrews, P. A. Waight, A. Melegaro, R. George and E. Miller (2011). *7-Valent pneumococcal conjugate vaccination in England and Wales: is it still beneficial despite high levels of serotype replacement?* PLoS One, 6(10), e26190.

Chopra, I. and M. Roberts (2001). *Tetracycline antibiotics: mode of action, applications, molecular biology, and epidemiology of bacterial resistance*. Microbiol Mol Biol Rev, 65(2), 232-60 ; second page, table of contents.

Claverys, J. P. and L. S. Havarstein (2002). *Extracellular-peptide control of competence for genetic transformation in Streptococcus pneumoniae*. Front Biosci, 7, d1798-814.

Claverys, J. P., B. Martin and P. Polard (2009). *The genetic transformation machinery: composition, localization, and mechanism*. FEMS Microbiol Rev, 33(3), 643-56.

Claverys, J. P., M. Prudhomme and B. Martin (2006). *Induction of competence regulons as a general response to stress in gram-positive bacteria*. Annu Rev Microbiol, 60, 451-75.

Cornejo, O. E., L. McGee and D. E. Rozen (2010). *Polymorphic competence peptides do not restrict recombination in Streptococcus pneumoniae*. Mol Biol Evol, 27(3), 694-702.

Cornick, J. E., S. R. Harris, C. M. Parry, M. J. Moore, C. Jassi, A. Kamng'ona, B. Kulohoma, R. S. Heyderman, S. D. Bentley and D. B. Everett (2014). *Genomic identification of a novel co-trimoxazole resistance genotype and its prevalence amongst Streptococcus pneumoniae in Malawi*. J Antimicrob Chemother, 69(2), 368-74.

Courvalin, P. and C. Carlier (1986). *Transposable multiple antibiotic resistance in Streptococcus pneumoniae*. Mol Gen Genet, 205(2), 291-7.

Croucher, N. J., S. R. Harris, C. Fraser, M. A. Quail, J. Burton, M. van der Linden, L. McGee, A. von Gottberg, J. H. Song, K. S. Ko*, et al.* (2011). *Rapid pneumococcal evolution in response to clinical interventions*. Science, 331(6016), 430-4.

Curiel, J. A., B. de Las Rivas, J. M. Mancheno and R. Munoz (2011). *The pURI family of expression vectors: a versatile set of ligation independent cloning plasmids for producing recombinant His-fusion proteins*. Protein Expr Purif, 76(1), 44-53.

Dagan, R. (2009a). *Impact of pneumococcal conjugate vaccine on infections caused by antibiotic-resistant Streptococcus pneumoniae*. Clin Microbiol Infect, 15 Suppl 3, 16-20.

Dagan, R. (2009b). *Serotype replacement in perspective*. Vaccine, 27 Suppl 3, C22-4.

Dagkessamanskaia, A., M. Moscoso, V. Henard, S. Guiral, K. Overweg, M. Reuter, B. Martin, J. Wells and J. P. Claverys (2004). *Interconnection of competence, stress and CiaR regulons in Streptococcus pneumoniae: competence triggers stationary phase autolysis of ciaR mutant cells*. Mol Microbiol, 51(4), 1071-86.

Dang-Van, A., G. Tiraby, J. F. Acar, W. V. Shaw and D. H. Bouanchaud (1978). *Chloramphenicol resistance in Streptococcus pneumoniae: enzymatic acetylation and possible plasmid linkage*. Antimicrob Agents Chemother, 13(4), 577-83.

Dawid, S., A. M. Roche and J. N. Weiser (2007). *The blp bacteriocins of Streptococcus pneumoniae mediate intraspecies competition both in vitro and in vivo*. Infect Immun, 75(1), 443-51.

Del Grosso, M., J. G. Northwood, D. J. Farrell and A. Pantosti (2007). *The macrolide resistance genes erm(B) and mef(E) are carried by Tn2010 in dual-gene Streptococcus pneumoniae isolates belonging to clonal complex CC271*. Antimicrob Agents Chemother, 51(11), 4184-6.

Del Grosso, M., A. Scotto d'Abusco, F. Iannelli, G. Pozzi and A. Pantosti (2004). *Tn2009, a Tn916-like element containing mef(E) in Streptococcus pneumoniae*. Antimicrob Agents Chemother, 48(6), 2037-42.

Diekema, D. J. and R. N. Jones (2001). *Oxazolidinone antibiotics*. Lancet, 358(9297), 1975-82.

Dobson, L., I. Remenyi and G. E. Tusnady (2015). *CCTOP: a Consensus Constrained TOPology prediction web server*. Nucleic Acids Res, 43(W1), W408-12.

Donati, C., N. L. Hiller, H. Tettelin, A. Muzzi, N. J. Croucher, S. V. Angiuoli, M. Oggioni, J. C. Dunning Hotopp, F. Z. Hu, D. R. Riley*, et al.* (2010). *Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species*. Genome Biol, 11(10), R107.

du Plessis, M., A. von Gottberg, S. A. Madhi, O. Hattingh, L. de Gouveia and K. P. Klugman (2008). *Serotype 6C is associated with penicillin-susceptible meningeal infections in human immunodeficiency virus (HIV)-infected adults among invasive pneumococcal isolates previously identified as serotype 6A in South Africa*. Int J Antimicrob Agents, 32 Suppl 1, S66-70.

El Moujaber, G., M. Osman, R. Rafei, F. Dabboussi and M. Hamze (2017). *Molecular mechanisms and epidemiology of resistance in Streptococcus pneumoniae in the Middle East region*. J Med Microbiol, 66(7), 847-858.

Elberse, K., S. Witteveen, H. van der Heide, I. van de Pol, C. Schot, A. van der Ende, G. Berbers and L. Schouls (2011). *Sequence diversity within the capsular genes of Streptococcus pneumoniae serogroup 6 and 19*. PLoS One, 6(9), e25018.

Eldholm, V., O. Johnsborg, K. Haugen, H. S. Ohnstad and L. S. Havarstein (2009). *Fratricide in Streptococcus pneumoniae: contributions and role of the cell wall hydrolases CbpD, LytA and LytC*. Microbiology, 155(Pt 7), 2223-34.

Enright, M. C. and B. G. Spratt (1999). *Extensive variation in the ddl gene of penicillin-resistant Streptococcus pneumoniae results from a hitchhiking effect driven by the penicillin-binding protein 2b gene*. Mol Biol Evol, 16(12), 1687-95.

Evans, B. A. and D. E. Rozen (2013). *Significant variation in transformation frequency in Streptococcus pneumoniae*. ISME J, 7(4), 791-9.

Farrell, D. J., I. Morrissey, S. Bakker, S. Buckridge and D. Felmingham (2004). *In vitro activities of telithromycin, linezolid, and quinupristin-dalfopristin against Streptococcus pneumoniae with macrolide resistance due to ribosomal mutations*. Antimicrob Agents Chemother, 48(8), 3169-71.

Feng, J., A. Lupien, H. Gingras, J. Wasserscheid, K. Dewar, D. Legare and M. Ouellette (2009). *Genome sequencing of linezolid-resistant Streptococcus pneumoniae mutants reveals novel mechanisms of resistance*. Genome Res, 19(7), 1214-23.

Fijarczyk, A. and W. Babik (2015). *Detecting balancing selection in genomes: limits and prospects*. Mol Ecol, 24(14), 3529-45.

Filipe, S. R. and A. Tomasz (2000). *Inhibition of the expression of penicillin resistance in Streptococcus pneumoniae by inactivation of cell wall muropeptide branching genes*. Proc Natl Acad Sci U S A, 97(9), 4891-6.

Francisco, A. P., M. Bugalho, M. Ramirez and J. A. Carrico (2009). *Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach*. BMC Bioinformatics, 10, 152.

Froger, A. and J. E. Hall (2007). *Transformation of plasmid DNA into E. coli using the heat shock method*. J Vis Exp(6), 253.

Garcia-Bustos, J. and A. Tomasz (1990). *A biological price of antibiotic resistance: major changes in the peptidoglycan structure of penicillin-resistant pneumococci*. Proc Natl Acad Sci U S A, 87(14), 5415-9.

Garvey, M. I., A. J. Baylay, R. L. Wong and L. J. Piddock (2011). *Overexpression of patA and patB, which encode ABC transporters, is associated with fluoroquinolone resistance in clinical isolates of Streptococcus pneumoniae*. Antimicrob Agents Chemother, 55(1), 190-6.

Gasteiger, E., C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel and A. Bairoch (2005). Protein Identification and Analysis Tools on the ExPASy

Server. The Proteomics Protocols Handbook J. M. Walker, Humana Press**:** 571-607.

Gillespie, S. H., L. L. Voelker, J. E. Ambler, C. Traini and A. Dickens (2003). *Fluoroquinolone resistance in Streptococcus pneumoniae: evidence that gyrA mutations arise at a lower rate and that mutation in gyrA or parC predisposes to further mutation*. Microb Drug Resist, 9(1), 17-24.

Gilley, R. P. and C. J. Orihuela (2014). *Pneumococci in biofilms are non-invasive: implications on nasopharyngeal colonization*. Front Cell Infect Microbiol, 4, 163.

Grabenstein, J. D. and K. P. Klugman (2012). *A century of pneumococcal vaccination research in humans*. Clin Microbiol Infect, 18 Suppl 5, 15-24.

Grebe, T. W. and J. B. Stock (1999). *The histidine protein kinase superfamily*. Adv Microb Physiol, 41, 139-227.

Griffith, F. (1928). *The Significance of Pneumococcal Types*. J Hyg (Lond), 27(2), 113-59.

Grohs, P., P. Trieu-Cuot, I. Podglajen, S. Grondin, A. Firon, C. Poyart, E. Varon and L. Gutmann (2012). *Molecular basis for different levels of tet(M) expression in Streptococcus pneumoniae clinical isolates*. Antimicrob Agents Chemother, 56(10), 5040-5.

Guiral, S., T. J. Mitchell, B. Martin and J. P. Claverys (2005). *Competence-programmed predation of noncompetent cells in the human pathogen Streptococcus pneumoniae: genetic requirements*. Proc Natl Acad Sci U S A, 102(24), 8710-5.

Hansman, D. and M. M. Bullen (1967). *A resistant pneumococcus*. The Lancet, 290(7509), 264-265.

Hausdorff, W. P. (2002). *Invasive pneumococcal disease in children: geographic and temporal variations in incidence and serotype distribution*. Eur J Pediatr, 161 Suppl 2, S135-9.

Havarstein, L. S., G. Coomaraswamy and D. A. Morrison (1995). *An unmodified heptadecapeptide pheromone induces competence for genetic transformation in Streptococcus pneumoniae*. Proc Natl Acad Sci U S A, 92(24), 11140-4.

Havarstein, L. S., P. Gaustad, I. F. Nes and D. A. Morrison (1996). *Identification of the streptococcal competence-pheromone receptor*. Mol Microbiol, 21(4), 863-9.

Havarstein, L. S., R. Hakenbeck and P. Gaustad (1997). *Natural competence in the genus Streptococcus: evidence that streptococci can change pherotype by interspecies recombinational exchanges*. J Bacteriol, 179(21), 6589-94.

Havarstein, L. S., B. Martin, O. Johnsborg, C. Granadel and J. P. Claverys (2006). *New insights into the pneumococcal fratricide: relationship to clumping and identification of a novel immunity factor*. Mol Microbiol, 59(4), 1297-307.

Henriques-Normark, B. and E. I. Tuomanen (2013). *The pneumococcus: epidemiology, microbiology, and pathogenesis*. Cold Spring Harb Perspect Med, 3(7).

Henriques Normark, B., R. Novak, A. Ortqvist, G. Kallenius, E. Tuomanen and S. Normark (2001). *Clinical isolates of Streptococcus pneumoniae that exhibit tolerance of vancomycin*. Clin Infect Dis, 32(4), 552-8.

Heron, M. (2012). *Deaths: leading causes for 2009*. Natl Vital Stat Rep, 61(7), 1-94.

Hooper, D. C. (2000). *Mechanisms of action and resistance of older and newer fluoroquinolones*. Clin Infect Dis, 31 Suppl 2, S24-8.

Horácio, A. N., J. Diamantino-Miranda, S. I. Aguiar, M. Ramirez and J. Melo-Cristino (2012). *Serotype changes in adult invasive pneumococcal infections*

in Portugal did not reduce the high fraction of potentially vaccine preventable infections. Vaccine, 30(2), 218-24.

Horácio, A. N., J. Diamantino-Miranda, S. I. Aguiar, M. Ramirez and J. Melo-Cristino (2013). *The majority of adult pneumococcal invasive infections in Portugal are still potentially vaccine preventable in spite of significant declines of serotypes 1 and 5*. PLoS One, 8(9), e73704.

Horácio, A. N., C. Silva-Costa, J. Diamantino-Miranda, J. P. Lopes, M. Ramirez and J. Melo-Cristino (2016). *Population Structure of Streptococcus pneumoniae Causing Invasive Disease in Adults in Portugal before PCV13 Availability for Adults: 2008-2011*. PLoS One, 11(5), e0153602.

Hudson, R. R. (2000). *A new statistic for detecting genetic differentiation*. Genetics, 155(4), 2011-4.

Hudson, R. R., D. D. Boos and N. L. Kaplan (1992a). *A statistical test for detecting geographic subdivision*. Mol Biol Evol, 9(1), 138-51.

Hudson, R. R., M. Slatkin and W. P. Maddison (1992b). *Estimation of levels of gene flow from DNA sequence data*. Genetics, 132(2), 583-9.

Hui, F. M. and D. A. Morrison (1991). *Genetic transformation in Streptococcus pneumoniae: nucleotide sequence analysis shows comA, a gene required for competence induction, to be a member of the bacterial ATP-dependent transport protein family*. J Bacteriol, 173(1), 372-81.

Hui, F. M., L. Zhou and D. A. Morrison (1995). *Competence for genetic transformation in Streptococcus pneumoniae: organization of a regulatory locus with homology to two lactococcin A secretion genes*. Gene, 153(1), 25-31.

Iannelli, F., M. R. Oggioni and G. Pozzi (2005). *Sensor domain of histidine kinase ComD confers competence pherotype specificity in Streptoccoccus pneumoniae*. FEMS Microbiol Lett, 252(2), 321-6.

Ichihara, H., K. Kuma and H. Toh (2006). *Positive selection in the ComC-ComD system of Streptococcal Species*. J Bacteriol, 188(17), 6429-34.

Jensen, A., O. Valdorsson, N. Frimodt-Moller, S. Hollingshead and M. Kilian (2015). *Commensal streptococci serve as a reservoir for beta-lactam resistance genes in Streptococcus pneumoniae*. Antimicrob Agents Chemother, 59(6), 3529-40.

Jin, P., F. Kong, M. Xiao, S. Oftadeh, F. Zhou, C. Liu, F. Russell and G. L. Gilbert (2009). *First report of putative Streptococcus pneumoniae serotype 6D among nasopharyngeal isolates from Fijian children*. J Infect Dis, 200(9), 1375-80.

Johnsborg, O., V. Eldholm and L. S. Havarstein (2007). *Natural genetic transformation: prevalence, mechanisms and function*. Res Microbiol, 158(10), 767-78.

Johnsborg, O. and L. S. Havarstein (2009). *Regulation of natural genetic transformation and acquisition of transforming DNA in Streptococcus pneumoniae*. FEMS Microbiol Rev, 33(3), 627-42.

Johnsborg, O., P. E. Kristiansen, T. Blomqvist and L. S. Havarstein (2006). *A hydrophobic patch in the competence-stimulating Peptide, a pneumococcal competence pheromone, is essential for specificity and biological activity*. J Bacteriol, 188(5), 1744-9.

Kausmally, L., O. Johnsborg, M. Lunde, E. Knutsen and L. S. Havarstein (2005). *Choline-binding protein D (CbpD) in Streptococcus pneumoniae is essential for competence-induced cell lysis*. J Bacteriol, 187(13), 4338-45.

Kawaguchiya, M., N. Urushibara and N. Kobayashi (2015). *High prevalence of genotype 6E (putative serotype 6E) among noninvasive/colonization isolates*

of Streptococcus pneumoniae in northern Japan. Microb Drug Resist, 21(2), 209-14.

Kilian, M., K. Poulsen, T. Blomqvist, L. S. Havarstein, M. Bek-Thomsen, H. Tettelin and U. B. Sorensen (2008). *Evolution of Streptococcus pneumoniae and its close commensal relatives*. PLoS One, 3(7), e2683.

Klugman, K. P. (2001). *Efficacy of pneumococcal conjugate vaccines and their effect on carriage and antimicrobial resistance*. Lancet Infect Dis, 1(2), 85-91.

Lau, G. W., S. Haataja, M. Lonetto, S. E. Kensit, A. Marra, A. P. Bryant, D. McDevitt, D. A. Morrison and D. W. Holden (2001). *A functional genomic analysis of type 3 Streptococcus pneumoniae virulence*. Mol Microbiol, 40(3), 555-71.

Leclercq, R. and P. Courvalin (1991). *Bacterial resistance to macrolide, lincosamide, and streptogramin antibiotics by target modification*. Antimicrob Agents Chemother, 35(7), 1267-72.

Lee, M. S. and D. A. Morrison (1999). *Identification of a new regulator in Streptococcus pneumoniae linking quorum sensing to competence for genetic transformation*. J Bacteriol, 181(16), 5004-16.

Lin, A. H., R. W. Murray, T. J. Vidmar and K. R. Marotti (1997). *The oxazolidinone eperezolid binds to the 50S ribosomal subunit and competes with binding of chloramphenicol and lincomycin*. Antimicrob Agents Chemother, 41(10), 2127-31.

Loman, N. J., R. A. Gladstone, C. Constantinidou, A. S. Tocheva, J. M. Jefferies, S. N. Faust, L. O'Connor, J. Chan, M. J. Pallen and S. C. Clarke (2013). *Clonal expansion within pneumococcal serotype 6C after use of seven-valent vaccine*. PLoS One, 8(5), e64731.

Long, K. S. and B. Vester (2012). *Resistance to linezolid caused by modifications at its binding site on the ribosome*. Antimicrob Agents Chemother, 56(2), 603-12.

Luo, P., H. Li and D. A. Morrison (2003). *ComX is a unique link between multiple quorum sensing outputs and competence in Streptococcus pneumoniae*. Mol Microbiol, 50(2), 623-33.

Luo, P., H. Li and D. A. Morrison (2004). *Identification of ComW as a new component in the regulation of genetic transformation in Streptococcus pneumoniae*. Mol Microbiol, 54(1), 172-83.

Luo, P. and D. A. Morrison (2003). *Transient association of an alternative sigma factor, ComX, with RNA polymerase during the period of competence for genetic transformation in Streptococcus pneumoniae*. J Bacteriol, 185(1), 349-58.

Lupien, A., H. Gingras, M. G. Bergeron, P. Leprohon and M. Ouellette (2015). *Multiple mutations and increased RNA expression in tetracycline-resistant Streptococcus pneumoniae as determined by genome-wide DNA and mRNA sequencing*. J Antimicrob Chemother, 70(7), 1946-59.

Lysenko, E. S., A. J. Ratner, A. L. Nelson and J. N. Weiser (2005). *The role of innate immune responses in the outcome of interspecies competition for colonization of mucosal surfaces*. PLoS Pathog, 1(1), e1.

Marimon, J. M., M. Ercibengoa, E. Tamayo, M. Alonso and E. Pérez-Trallero (2016). *Long-Term Epidemiology of Streptococcus pneumoniae Serogroup 6 in a Region of Southern Europe with Special Reference to Serotype 6E*. PLoS One, 11(2), e0149047.

Marimón, J. M., M. Ercibengoa, E. Tamayo, M. Alonso and E. Pérez-Trallero (2016). *Long-Term Epidemiology of Streptococcus pneumoniae Serogroup 6*

in a Region of Southern Europe with Special Reference to Serotype 6E. PLoS One, 11(2), e0149047.

Martin, B., C. Granadel, N. Campo, V. Henard, M. Prudhomme and J. P. Claverys (2010). *Expression and maintenance of ComD-ComE, the two-component signal-transduction system that controls competence of Streptococcus pneumoniae*. Mol Microbiol, 75(6), 1513-28.

Mavroidi, A., D. Godoy, D. M. Aanensen, D. A. Robinson, S. K. Hollingshead and B. G. Spratt (2004). *Evolutionary genetics of the capsular locus of serogroup 6 pneumococci*. J Bacteriol, 186(24), 8181-92.

McDougal, L. K., F. C. Tenover, L. N. Lee, J. K. Rasheed, J. E. Patterson, J. H. Jorgensen and D. J. LeBlanc (1998). *Detection of Tn917-like sequences within a Tn916-like conjugative transposon (Tn3872) in erythromycin-resistant isolates of Streptococcus pneumoniae*. Antimicrob Agents Chemother, 42(9), 2312-8.

McGee, L., L. McDougal, J. Zhou, B. G. Spratt, F. C. Tenover, R. George, R. Hakenbeck, W. Hryniewicz, J. C. Lefévre, A. Tomasz*, et al.* (2001). *Nomenclature of major antimicrobial-resistant clones of Streptococcus pneumoniae defined by the pneumococcal molecular epidemiology network*. J Clin Microbiol, 39(7), 2565-71.

Melo-Cristino, J., M. Ramirez, N. Serrano and T. Hanscheid (2003). *Macrolide resistance in Streptococcus pneumoniae isolated from patients with community-acquired lower respiratory tract infections in Portugal: results of a 3-year (1999-2001) multicenter surveillance study*. Microb Drug Resist, 9(1), 73-80.

Millar, E. V., F. C. Pimenta, A. Roundtree, D. Jackson, G. Carvalho Mda, M. J. Perilla, R. Reid, M. Santosham, C. G. Whitney, B. W. Beall*, et al.* (2010). *Pre- and post-conjugate vaccine epidemiology of pneumococcal serotype 6C invasive disease and carriage within Navajo and White Mountain Apache communities*. Clin Infect Dis, 51(11), 1258-65.

Miller, E., N. J. Andrews, P. A. Waight, M. P. Slack and R. C. George (2011). *Herd immunity and serotype replacement 4 years after seven-valent pneumococcal conjugate vaccination in England and Wales: an observational cohort study*. Lancet Infect Dis, 11(10), 760-8.

Miller, E. L., B. A. Evans, O. E. Cornejo, I. S. Roberts and D. E. Rozen (2017). *Pherotype Polymorphism in Streptococcus pneumoniae Has No Obvious Effects on Population Structure and Recombination*. Genome Biol Evol, 9(10), 2546-2559.

Moraes, I., G. Evans, J. Sanchez-Weatherby, S. Newstead and P. D. Stewart (2014). *Membrane protein structure determination - the next generation*. Biochim Biophys Acta, 1838(1 Pt A), 78-87.

Moreno-Gámez, S., R. A. Sorg, A. Domenech, M. Kjos, F. J. Weissing, G. S. van Doorn and J. W. Veening (2017). *Quorum sensing integrates environmental cues, cell density and cell history to control bacterial competence*. Nat Commun, 8(1), 854.

Mortier-Barriere, I., A. de Saizieu, J. P. Claverys and B. Martin (1998). *Competence-specific induction of recA is required for full recombination proficiency during transformation in Streptococcus pneumoniae*. Mol Microbiol, 27(1), 159-70.

Moscoso, M. and J. P. Claverys (2004). *Release of DNA into the medium by competent Streptococcus pneumoniae: kinetics, mechanism and stability of the liberated DNA*. Mol Microbiol, 54(3), 783-94.

Mosser, J. L. and A. Tomasz (1970). *Choline-containing teichoic acid as a structural component of pneumococcal cell wall and its role in sensitivity to lysis by an autolytic enzyme*. J Biol Chem, 245(2), 287-98.

Nahm, M. H., J. Lin, J. A. Finkelstein and S. I. Pelton (2009). *Increase in the prevalence of the newly discovered pneumococcal serotype 6C in the nasopharynx after introduction of pneumococcal conjugate vaccine*. J Infect Dis, 199(3), 320-5.

Naucler, P., J. Darenberg, E. Morfeldt, A. Ortqvist and B. Henriques Normark (2013). *Contribution of host, bacterial factors and antibiotic treatment to mortality in adult patients with bacteraemic pneumococcal pneumonia*. Thorax, 68(6), 571-9.

Neu, H. C. and T. D. Gootz (1996). *Antimicrobial Chemotherapy*.

Novak, R., B. Henriques, E. Charpentier, S. Normark and E. Tuomanen (1999). *Emergence of vancomycin tolerance in Streptococcus pneumoniae*. Nature, 399(6736), 590-3.

Nunes, S., C. Valente, R. Sá-Leão and H. de Lencastre (2009). *Temporal trends and molecular epidemiology of recently described serotype 6C of Streptococcus pneumoniae*. J Clin Microbiol, 47(2), 472-4.

Nuorti, J. P. and C. G. Whitney (2010). *Prevention of pneumococcal disease among infants and children - use of 13-valent pneumococcal conjugate vaccine and 23-valent pneumococcal polysaccharide vaccine - recommendations of the Advisory Committee on Immunization Practices (ACIP)*. MMWR Recomm Rep, 59(RR-11), 1-18.

Nurkka, A., H. Ahman, M. Yaich, J. Eskola and H. Kayhty (2001). *Serum and salivary anti-capsular antibodies in infants and children vaccinated with octavalent pneumococcal conjugate vaccines, PncD and PncT*. Vaccine, 20(1-2), 194-201.

O'Brien, K. L., L. J. Wolfson, J. P. Watt, E. Henkle, M. Deloria-Knoll, N. McCall, E. Lee, K. Mulholland, O. S. Levine and T. Cherian (2009). *Burden of disease caused by Streptococcus pneumoniae in children younger than 5 years: global estimates*. Lancet, 374(9693), 893-902.

Obregón, V., P. García, E. García, A. Fenoll, R. López and J. L. García (2002). *Molecular peculiarities of the lytA gene isolated from clinical pneumococcal strains that are bile insoluble*. J Clin Microbiol, 40(7), 2545-54.

Padayachee, T. and K. P. Klugman (1999). *Novel expansions of the gene encoding dihydropteroate synthase in trimethoprim-sulfamethoxazole-resistant Streptococcus pneumoniae*. Antimicrob Agents Chemother, 43(9), 2225-30.

Park, I. H., S. Park, S. K. Hollingshead and M. H. Nahm (2007a). *Genetic basis for the new pneumococcal serotype, 6C*. Infect Immun, 75(9), 4482-9.

Park, I. H., D. G. Pritchard, R. Cartee, A. Brandao, M. C. Brandileone and M. H. Nahm (2007b). *Discovery of a new capsular serotype (6C) within serogroup 6 of Streptococcus pneumoniae*. J Clin Microbiol, 45(4), 1225-33.

Perez-Dorado, I., A. Gonzalez, M. Morales, R. Sanles, W. Striker, W. Vollmer, S. Mobashery, J. L. Garcia, M. Martinez-Ripoll, P. Garcia*, et al.* (2010). *Insights into pneumococcal fratricide from the crystal structures of the modular killing factor LytC*. Nat Struct Mol Biol, 17(5), 576-81.

Peterson, S. N., C. K. Sung, R. Cline, B. V. Desai, E. C. Snesrud, P. Luo, J. Walling, H. Li, M. Mintz, G. Tsegaye*, et al.* (2004). *Identification of competence pheromone responsive genes in Streptococcus pneumoniae by use of DNA microarrays*. Mol Microbiol, 51(4), 1051-70.

Pikis, A., J. M. Campos, W. J. Rodriguez and J. M. Keith (2001). *Optochin resistance in Streptococcus pneumoniae: mechanism, significance, and clinical implications*. J Infect Dis, 184(5), 582-90.

Pilishvili, T., C. Lexau, M. M. Farley, J. Hadler, L. H. Harrison, N. M. Bennett, A. Reingold, A. Thomas, W. Schaffner, A. S. Craig*, et al.* (2010). *Sustained reductions in invasive pneumococcal disease in the era of conjugate vaccine*. J Infect Dis, 201(1), 32-41.

Piotrowski, A., P. Luo and D. A. Morrison (2009). *Competence for genetic transformation in Streptococcus pneumoniae: termination of activity of the alternative sigma factor ComX is independent of proteolysis of ComX and ComW*. J Bacteriol, 191(10), 3359-66.

Pozzi, G., L. Masala, F. Iannelli, R. Manganelli, L. S. Havarstein, L. Piccoli, D. Simon and D. A. Morrison (1996). *Competence for genetic transformation in encapsulated strains of Streptococcus pneumoniae: two allelic variants of the peptide pheromone*. J Bacteriol, 178(20), 6087-90.

Ramirez, M., D. A. Morrison and A. Tomasz (1997). *Ubiquitous distribution of the competence related genes comA and comC among isolates of Streptococcus pneumoniae*. Microb Drug Resist, 3(1), 39-52.

Ribeiro-Gonçalves, B., A. P. Francisco, C. Vaz, M. Ramirez and J. A. Carriço (2016). *PHYLOViZ Online: web-based tool for visualization, phylogenetic inference, analysis and sharing of minimum spanning trees*. Nucleic Acids Res, 44(W1), W246-51.

Rimini, R., B. Jansson, G. Feger, T. C. Roberts, M. de Francesco, A. Gozzi, F. Faggioni, E. Domenici, D. M. Wallace, N. Frandsen*, et al.* (2000). *Global analysis of transcription kinetics during competence development in Streptococcus pneumoniae using high density DNA arrays*. Mol Microbiol, 36(6), 1279-92.

Roberts, A. P. and P. Mullany (2009). *A modular master on the move: the Tn916 family of mobile genetic elements*. Trends Microbiol, 17(6), 251-8.

Rolo, D., C. Ardanuy, L. Calatayud, R. Pallares, I. Grau, E. García, A. Fenoll, R. Martín and J. Liñares (2011a). *Characterization of invasive Pneumococci of serogroup 6 from adults in Barcelona, Spain, in 1994 to 2008*. J Clin Microbiol, 49(6), 2328-30.

Rolo, D., A. Fenoll, C. Ardanuy, L. Calatayud, M. Cubero, A. G. de la Campa and J. Liñares (2011b). *Trends of invasive serotype 6C pneumococci in Spain: emergence of a new lineage*. J Antimicrob Chemother, 66(8), 1712-8.

Sá-Leão, R., S. Nunes, A. Brito-Avô, N. Frazão, A. S. Simões, M. I. Crisóstomo, A. C. Paulo, J. Saldanha, I. Santos-Sanches and H. de Lencastre (2009). *Changes in pneumococcal serotypes and antibiotypes carried by vaccinated and unvaccinated day-care centre attendees in Portugal, a country with widespread use of the seven-valent pneumococcal conjugate vaccine*. Clin Microbiol Infect, 15(11), 1002-7.

Sá-Leao, R., F. Pinto, S. Aguiar, S. Nunes, J. A. Carriço, N. Frazão, N. Gonçalves-Sousa, J. Melo-Cristino, H. de Lencastre and M. Ramirez (2011). *Analysis of invasiveness of pneumococcal serotypes and clones circulating in Portugal before widespread use of conjugate vaccines reveals heterogeneous behavior of clones expressing the same serotype*. J Clin Microbiol, 49(4), 1369-75.

Sanchez, D., M. Boudes, H. van Tilbeurgh, D. Durand and S. Quevillon-Cheruel (2015). *Modeling the ComD/ComE/comcde interaction network using small angle X-ray scattering*. FEBS J, 282(8), 1538-53.

Schmitz, F. J., M. Perdikouli, A. Beeck, J. Verhoef and A. C. Fluit (2001). *Resistance to trimethoprim-sulfamethoxazole and modifications in genes*

coding for dihydrofolate reductase and dihydropteroate synthase in European Streptococcus pneumoniae isolates. J Antimicrob Chemother, 48(6), 935-6.

Scott, D. A., S. F. Komjathy, B. T. Hu, S. Baker, L. A. Supan, C. A. Monahan, W. Gruber, G. R. Siber and S. P. Lockhart (2007). *Phase 1 trial of a 13-valent pneumococcal conjugate vaccine in healthy adults*. Vaccine, 25(33), 6164-6.

Serrano, I., J. Melo-Cristino, J. A. Carriço and M. Ramirez (2005). *Characterization of the genetic lineages responsible for pneumococcal invasive disease in Portugal*. J Clin Microbiol, 43(4), 1706-15.

Serrano, I., M. Ramirez and J. Melo-Cristino (2004). *Invasive Streptococcus pneumoniae from Portugal: implications for vaccination and antimicrobial therapy*. Clin Microbiol Infect, 10(7), 652-6.

Shi, W., K. Yao, M. He, S. Yu and Y. Yang (2014). *Population biology of 225 serogroup 6 Streptococcus pneumoniae isolates collected in China*. BMC Infect Dis, 14, 467.

Shoemaker, N. B., M. D. Smith and W. R. Guild (1979). *Organization and transfer of heterologous chloramphenicol and tetracycline resistance genes in pneumococcus*. J Bacteriol, 139(2), 432-41.

Siegel, S. J., A. M. Roche and J. N. Weiser (2014). *Influenza promotes pneumococcal growth during coinfection by providing host sialylated substrates as a nutrient source*. Cell Host Microbe, 16(1), 55-67.

Siegel, S. J. and J. N. Weiser (2015). *Mechanisms of Bacterial Colonization of the Respiratory Tract*. Annu Rev Microbiol, 69, 425-44.

Siguier, P., J. Perochon, L. Lestrade, J. Mahillon and M. Chandler (2006). *ISfinder: the reference centre for bacterial insertion sequences*. Nucleic Acids Res, 34(Database issue), D32-6.

Simões, A. S., L. Pereira, S. Nunes, A. Brito-Avô, H. de Lencastre and R. Sá-Leão (2011). *Clonal evolution leading to maintenance of antibiotic resistance rates among colonizing Pneumococci in the PCV7 era in Portugal*. J Clin Microbiol, 49(8), 2810-7.

Smith, A. M. and K. P. Klugman (2001). *Alterations in MurM, a cell wall muropeptide branching enzyme, increase high-level penicillin and cephalosporin resistance in Streptococcus pneumoniae*. Antimicrob Agents Chemother, 45(8), 2393-6.

Sorensen, U. B. (1993). *Typing of pneumococci by using 12 pooled antisera*. J Clin Microbiol, 31(8), 2097-100.

Steinmoen, H., E. Knutsen and L. S. Havarstein (2002). *Induction of natural competence in Streptococcus pneumoniae triggers lysis and DNA release from a subfraction of the cell population*. Proc Natl Acad Sci U S A, 99(11), 7681-6.

Steinmoen, H., A. Teigen and L. S. Havarstein (2003). *Competence-induced cells of Streptococcus pneumoniae lyse competence-deficient cells of the same strain during cocultivation*. J Bacteriol, 185(24), 7176-83.

Straume, D., G. A. Stamsas and L. S. Havarstein (2015). *Natural transformation and genome evolution in Streptococcus pneumoniae*. Infect Genet Evol, 33, 371-80.

Sung, C. K. and D. A. Morrison (2005). *Two distinct functions of ComW in stabilization and activation of the alternative sigma factor ComX in Streptococcus pneumoniae*. J Bacteriol, 187(9), 3052-61.

Sung, H., H. B. Shin, M. N. Kim, K. Lee, E. C. Kim, W. Song, S. H. Jeong, W. G. Lee, Y. J. Park and G. M. Eliopoulos (2006). *Vancomycin-tolerant Streptococcus pneumoniae in Korea*. J Clin Microbiol, 44(10), 3524-8.

Sutcliffe, J., A. Tait-Kamradt and L. Wondrack (1996). *Streptococcus pneumoniae and Streptococcus pyogenes resistant to macrolides but sensitive to clindamycin: a common resistance pattern mediated by an efflux system*. Antimicrob Agents Chemother, 40(8), 1817-24.

Swaney, S. M., H. Aoki, M. C. Ganoza and D. L. Shinabarger (1998). *The oxazolidinone linezolid inhibits initiation of protein synthesis in bacteria*. Antimicrob Agents Chemother, 42(12), 3251-5.

Syrogiannopoulos, G. A., I. N. Grivea, A. Tait-Kamradt, G. D. Katopodis, N. G. Beratis, J. Sutcliffe, P. C. Appelbaum and T. A. Davies (2001). *Identification of an erm(A) erythromycin resistance methylase gene in Streptococcus pneumoniae isolated in Greece*. Antimicrob Agents Chemother, 45(1), 342-4.

Tait-Kamradt, A., J. Clancy, M. Cronan, F. Dib-Hajj, L. Wondrack, W. Yuan and J. Sutcliffe (1997). *mefE is necessary for the erythromycin-resistant M phenotype in Streptococcus pneumoniae*. Antimicrob Agents Chemother, 41(10), 2251-5.

Valente, C., H. De Lencastre and R. Sá-Leão (2012). *Pherotypes of co-colonizing pneumococci among Portuguese children*. Microb Drug Resist, 18(6), 550-4.

van der Linden, M., N. Winkel, S. Kuntzel, A. Farkas, S. R. Perniciaro, R. R. Reinert and M. Imohl (2013a). *Epidemiology of Streptococcus pneumoniae serogroup 6 isolates from IPD in children and adults in Germany*. PLoS One, 8(4), e60848.

van der Linden, M., N. Winkel, S. Küntzel, A. Farkas, S. R. Perniciaro, R. R. Reinert and M. Imöhl (2013b). *Epidemiology of Streptococcus pneumoniae serogroup 6 isolates from IPD in children and adults in Germany*. PLoS One, 8(4), e60848.

van Tonder, A. J., J. E. Bray, L. Roalfe, R. White, M. Zancolli, S. J. Quirk, G. Haraldsson, K. A. Jolley, M. C. Maiden, S. D. Bentley, *et al.* (2015). *Genomics Reveals the Worldwide Distribution of Multidrug-Resistant Serotype 6E Pneumococci*. J Clin Microbiol, 53(7), 2271-85.

Varaldo, P. E., M. P. Montanari and E. Giovanetti (2009). *Genetic elements responsible for erythromycin resistance in streptococci*. Antimicrob Agents Chemother, 53(2), 343-53.

Vestrheim, D. F., P. Gaustad, I. S. Aaberge and D. A. Caugant (2011). *Pherotypes of pneumococcal strains co-existing in healthy children*. Infect Genet Evol, 11(7), 1703-8.

Vijayakumar, M. N., S. D. Priebe and W. R. Guild (1986). *Structure of a conjugative element in Streptococcus pneumoniae*. J Bacteriol, 166(3), 978-84.

Watson, D. A., D. M. Musher, J. W. Jacobson and J. Verhoef (1993). *A brief history of the pneumococcus in biomedical research: a panoply of scientific discovery*. Clin Infect Dis, 17(5), 913-24.

Ween, O., P. Gaustad and L. S. Havarstein (1999). *Identification of DNA binding sites for ComE, a key regulator of natural competence in Streptococcus pneumoniae*. Mol Microbiol, 33(4), 817-27.

Weiser, J. N. (2009). *The pneumococcus: why a commensal misbehaves*. J Mol Med (Berl), 88(2), 97-102.

Werno, A. M. and D. R. Murdoch (2008). *Medical microbiology: laboratory diagnosis of invasive pneumococcal disease*. Clin Infect Dis, 46(6), 926-32.

Weyder, M., M. Prudhomme, M. Bergé, P. Polard and G. Fichant (2018). *Dynamic Modeling of Streptococcus pneumoniae Competence Provides Regulatory Mechanistic Insights Into Its Tight Temporal Regulation*. Front Microbiol, 9, 1637.

Whatmore, A. M., V. A. Barcus and C. G. Dowson (1999). *Genetic diversity of the streptococcal competence (com) gene locus*. J Bacteriol, 181(10), 3144-54.

Wilson, K. (2001). *Preparation of genomic DNA from bacteria*. Curr Protoc Mol Biol, Chapter 2, Unit 2 4.

Yang, Y., B. Koirala, L. A. Sanchez, N. R. Phillips, S. R. Hamry and Y. Tal-Gan (2017). *Structure-Activity Relationships of the Competence Stimulating Peptides (CSPs) in Streptococcus pneumoniae Reveal Motifs Critical for Intra-group and Cross-group ComD Receptor Activation*. ACS Chem Biol, 12(4), 1141-1151.

Yun, K. W., E. Y. Cho, E. H. Choi and H. J. Lee (2014). *Capsular polysaccharide gene diversity of pneumococcal serotypes 6A, 6B, 6C, and 6D*. Int J Med Microbiol, 304(8), 1109-17.

Yunis, A. A. (1988). *Chloramphenicol: relation of structure to activity and toxicity*. Annu Rev Pharmacol Toxicol, 28, 83-100.

Zhou, C. C., S. M. Swaney, D. L. Shinabarger and B. J. Stockman (2002). *1H nuclear magnetic resonance study of oxazolidinone binding to bacterial ribosomes*. Antimicrob Agents Chemother, 46(3), 625-9.

Zhu, L. and G. W. Lau (2011). *Inhibition of competence development, horizontal gene transfer and virulence in Streptococcus pneumoniae by a modified competence stimulating peptide*. PLoS Pathog, 7(9), e1002241.

Appendix I. Published article

PLOS ONE

RESEARCH ARTICLE

Clonal and serotype dynamics of serogroup 6 isolates causing invasive pneumococcal disease in Portugal: 1999-2012

Jorge Diamantino-Miranda, Sandra Isabel Aguiar¤, João André Carriço, José Melo-Cristino, Mário Ramirez*

Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Av. Prof. Egas Moniz, Lisboa, Portugal

¤ Current address: Centro de Investigação Interdisciplinar em Sanidade Animal (CIISA), Faculdade de Medicina Veterinária da Universidade de Lisboa, Avenida da Universidade Técnica, Lisboa, Portugal
* ramirez@medicina.ulisboa.pt.

OPEN ACCESS

Citation: Diamantino-Miranda J, Aguiar SI, Carriço JA, Melo-Cristino J, Ramirez M (2017) Clonal and serotype dynamics of serogroup 6 isolates causing invasive pneumococcal disease in Portugal: 1999-2012. PLoS ONE 12(2): e0170354. doi:10.1371/journal.pone.0170354

Editor: Herminia de Lencastre, Rockefeller University, UNITED STATES

Received: September 20, 2016

Accepted: January 3, 2017

Published: February 2, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: J. Diamantino-Miranda and S.I. Aguiar were supported by grants SFRH/BD/81766/2011 and SFRH/BPD/78376/2011, respectively, from Fundação para a Ciência e Tecnologia, Portugal. The work was partly supported by Fundação para a Ciência e Tecnologia, Portugal (PTDC/DTP-EPI/1759/2012 and PTDC/DTP-EPI/1555/2014) and an

Abstract

Although serogroup 6 was among the first to be recognized among *Streptococcus pneumoniae*, several new serotypes were identified since the introduction of pneumococcal conjugate vaccines (PCVs). A decrease of the 6B-2 variant among invasive pneumococcal disease (IPD), but not 6B-1, was noted post conjugate vaccine introduction, underpinned by a decrease of CC273 isolates. Serotype 6C was associated with adult IPD and increased in this age group representing two lineages (CC315 and CC395), while the same lineages expressed other serogroup 6 serotypes in children. Taken together, these findings suggest a potential cross-protection of PCVs against serotype 6C IPD among vaccinated children but not among adults. Serotype 6A became the most important serogroup 6 serotype in children but it decreased in adult IPD. No other serogroup 6 serotypes were detected, so available phenotypic or simple genotypic assays remain adequate for distinguishing serotypes within serogroup 6 isolates.

Introduction

Although historically only two serotypes were recognized within serogroup 6 (6A, 6B), five more were recently discovered (6C, 6D, 6F, 6G and 6H) and shown to be the result of alterations in the *wciN* and *wciP* genes [1,2]. An additional serotype "6E" was proposed based on divergent capsular loci (including SNPs and indels), most frequently of serotype 6B isolates. However, the polysaccharide of this putative serotype, henceforth designated 6B-2, was recently shown to be identical of that of isolates with canonical serotype 6B loci (designated 6B-1) [2].

The 7-valent PCV (PCV7), which includes serotype 6B, was introduced in Portugal in 2001. However, PCV7 was not included in the National Immunization Plan so its uptake increased slowly. Previous work from our laboratory identified a period when no PCV7 attributable changes in serotypes causing invasive pneumococcal disease (IPD) occurred (Pre-PCV7,

unrestricted Investigator initiated project from Pfizer. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: JMC has received research grants administered through his university and received honoraria for serving on the speakers bureaus of Pfizer, Bial, GlaxoSmithKline and Novartis. MR has received honoraria for serving on speakers bureau of Pfizer. The other authors declare no conflict of interest. No company or financing body had any interference in the decision to publish. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

1999–2002), early-PCV7 (2003–2006) and late-PCV7 (2007–2009) periods [3–8]. By mid-2009 and start of 2010, 10-valent (PCV10) and 13-valent (PCV13) PCVs were introduced replacing PCV7. While PCV10 was used only briefly and does not include additional serogroup 6 serotypes, PCV13 includes serotype 6A. This study aimed to evaluate which serogroup 6 serotypes were causing IPD in Portugal, the genetic lineages associated and their antimicrobial resistance as well as any changes during a period when PCVs were being introduced.

Materials and methods

Isolates and serotyping

In 1999–2012, n = 4812 isolates causing IPD (n = 985 in children <18 yrs and n = 3847 in adults ≥18 yrs) were identified and characterized regarding serotype and antimicrobial susceptibility [3–8]. Isolates were serotyped by the standard capsular reaction test (Statens Serum Institut, Copenhagen, Denmark). Only isolates expressing serogroup 6 (n = 242, 5%) were retained for further study. To confirm the phenotypically determined serotypes and to identify the most recent serotypes and locus class, three PCR reactions were used (Table 1), complemented by sequencing of *wciN* and *wciP*.

Multilocus Sequence Typing (MLST)

MLST was performed as previously described [11]. Briefly, seven housekeeping genes were sequenced and compared to the pneumococcal MLST database (http://pubmlst.org/spneumoniae) to identify the alleles and respective sequence types (STs). PHYLOViZ [12] was used to define clonal complexes (CCs) at double locus variant (DLV) level using only the STs found in this study.

Antimicrobial susceptibility testing

Etest strips (AB Biodisk, Solna, Sweden) were used to determine the minimal inhibitory concentration (MIC) for penicillin, cefotaxime and levofloxacin. Susceptibility to erythromycin, clindamycin, vancomycin, linezolid, trimethoprim-sulfamethoxazole, tetracycline and chloramphenicol was tested by the Kirby-Bauer disk diffusion technique according to CLSI procedures. Unless otherwise stated, the interpretative criteria prior to 2008 [13] were used. These criteria were selected to enable comparison with previous studies. Multidrug resistance (MDR) was defined by non-susceptibility to at least three classes of antibiotics.

Statistical analysis

Genetic diversity was evaluated using Simpson's index of diversity (SID) and respective 95% confidence intervals ($CI_{95\%}$) [14]. The Cochran-Armitage test (CA) was used to evaluate the temporal trends of serotypes, STs and CCs. Only STs and CCs with ≥5 isolates were considered for the CA analysis. Fisher's exact test (FET) was used to test association of serotypes, STs and CCs with age group. The false discovery rate (FDR) correction for multiple testing [15] was used in both tests. A $p<0.05$ was considered significant for all tests.

Results and discussion

The serotypes identified were 6A (n = 80), 6B (n = 79) and 6C (n = 83). All class 2 loci were identified among serotype 6B isolates (6B-2, n = 52; 6B-1, n = 27). No representatives of serotypes 6D, 6F, 6G and 6H were found, confirming their rarity in Europe [16,17]. The proportion of serogroup 6 isolates was higher in children than in adults [6.4% and 4.7%, respectively, FET p = 0.032] and differences in serotype distribution were also observed (Fig 1 and S1 and

Table 1. PCRs used for serotype and class identification.

Purpose of the PCR	Target gene	Primer	Sequence (5'-3')	Target serotype	Product size	Reference
Multiplex PCR for identification of 6A, 6B, 6C and 6D serotypes	*wzy*	wzy-f	CGACGTAACAAAGAACTAGGTGCTGAAAC	Serogroup 6	220 bp	[9]
		wzy-r	AAGTATATAACCACGCTGTAAAACTCTGAC			
	wciP	wciP-f1	ATATGTAGAAGAACTGGCTCAGGGTAG	6A, 6C and 6F	128 bp	This study
		wciP-r1	GATGACTAGATGGTACATTATGTCCAT			
	wciN	wciN-f1	CATTTTAGTGAAGTTGGCGGTGGAGTT	6C and 6D	727 bp	This study
		wciN-r1	AGCTTCGAAGCCCATACTCTTCAATTA			
Class identification	*wzh*	wzh-f	TGATATTCATTCGCACATTGTC	Class 2 sequences	578 bp	[10]
		wzh-r	TATGAACCAAATCACGCTCCAAG			
	wze	wze-f	CTCACAGGCAAAATTGGATTC	Class 1 sequences	217 bp	[10]
		wze-r	AACAGAATTGCGAATATCTC			
wciN sequencing	*wciN*	wciN-f2	TGGAAAGATATTGAAATTTT	Serogroup 6	1.4 kb (6A/6B-1/6F/6G)	This study
					1.2 kb (6C/6D)	
		wciN-r2	GTT TTTCTTTCAATATCTTTA		1.7 kb (6B-2)	
wciP sequencing	*wciP*	wciP-f2	CGATTAATTTTTTATTAATG	Serogroup 6	1.0 kb	This study
		wciP-r2	ATATGAATAAGAAATTTAAAAG			

doi:10.1371/journal.pone.0170354.t001

S2 Tables). Most 6C pneumococci were recovered from adults (n = 79/83) resulting in an association with this age group [FET $p<0.001$, significant after FDR correction]. There was no association between serotype and isolate source. Despite no overall change in frequency of serogroup 6 pneumococci following the introduction of PCVs, there were changes in proportion of serotypes (Fig 1). In the pre-PCV7 period, serotype 6B (6B-2) was the most frequent in contrast to neighboring Spain where 6B-1 predominated [16], but it subsequently decreased in both adults and children, similarly to Spain [supported only in children: CA $p<0.001$, significant after FDR; CA $p = 0.160$, in adults]. Class 6B-1 accounted for a small proportion of IPD in both age groups before vaccine introduction and, although there was an increase in the proportion of 6B-1 isolates in both age groups, this was supported only in adults and before FDR correction (CA $p = 0.041$). So although both 6B-1 and 6B-2 express the same polysaccharide, they showed different dynamics following vaccination. It was hypothesized that the two classes could present differences in capsule expression [2] potentially explaining their contrasting variations seen here. Serotype 6A increased to become the most frequent serotype in children in late-PCV7 and PCV13 periods. In adults, although 6A was an important serotype up to 2009, a decrease was seen afterwards (CA $p = 0.006$, significant after FDR), coinciding with PCV13 introduction in children (Fig 1B). While the proportion of isolates expressing serotype 6C remained low and stable in children, in adults an increase in serotype 6C was noted after 2008, although not statistically supported (CA $p = 0.392$), establishing it as the most frequent serotype of serogroup 6 in adults.

All isolates were characterized by multilocus sequence typing (MLST) to complement the information already available [18] and PHYLOViZ was used to define CCs (Fig 1C). Serotype 6A was the most diverse (27 STs and 13 CCs) followed by 6B-2 (21 STs and 12 CCs), 6C (17 STs and 9 CCs) and 6B-1 (5 STs and 2 CCs). Still, the majority of serogroup 6 isolates (n = 210, 87%) were distributed into only 7 genetic lineages, three of which included PMEN clones [19]:

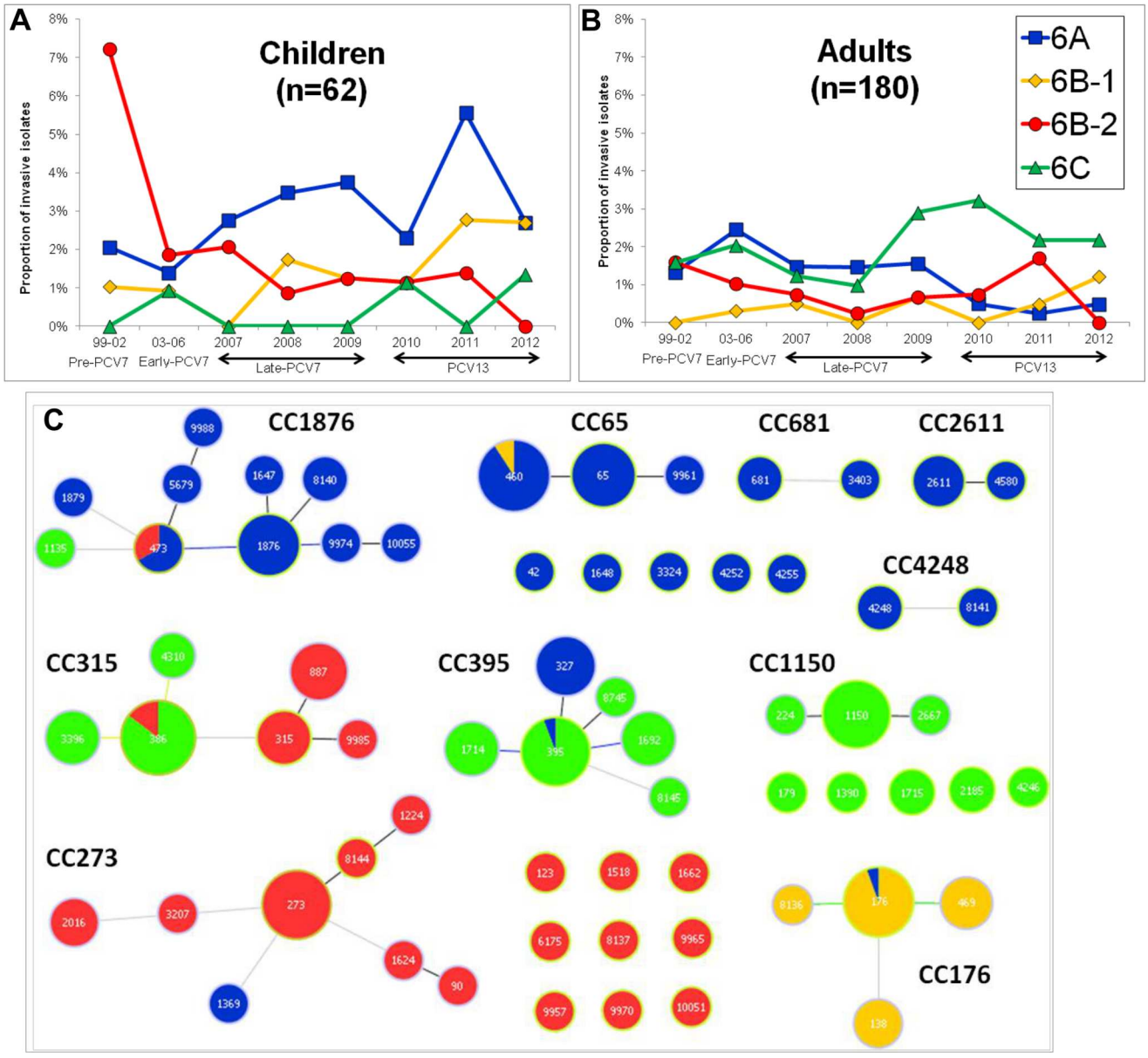


Fig 1. Clonal composition and changes in serogroup 6 serotypes among invasive pneumococci recovered in Portugal (1999–2012). (A) Shows the variations of the serotypes and serotype classes in children and (B) in adults. The years before PCV7 introduction (1999–2002) and subsequent periods were defined as described in the text. (C) shows STs and CCs identified colored by serotype. Each circle represents an ST and the diameter represents its frequency in a logarithmic scale. Grey lines connect STs that are double-locus variants, while lines of other colors connect STs that are single-locus variants according to the PHYLOViZ tie-break rule reached. STs that are linked belong to the same CC. This data set can be explored online at http://bit.do/PHYLOViZ_sero6.

doi:10.1371/journal.pone.0170354.g001

CC315 (Poland[6B]315, n = 47), CC395 (Portugal[6A]327, n = 37), CC65 (n = 35), CC176 (n = 26), CC273 (Greece[6B]273 and Spain[6B]90, n = 26), CC1876 (n = 22) and CC1150 (n = 17) (S3 and S4 Tables). Mostly, each ST presented only one serotype, with the exception of STs ST176, ST460, ST473, ST386 and ST395, possibly reflecting capsular switching events (Fig 1C). The 6B-2 ST90 lineage dominant in Asia [10,20] was represented by a single isolate.

The genetic diversity of serogroup 6 isolates recovered from both children and adults was similar when considering both STs and CCs, with all seven major lineages present in both age

groups (S3 and S4 Tables). Although CC176 was found more frequently in children (FET p = 0.017, not supported after FDR), no association of particular STs or CCs with age group was observed. While in adults the majority of CC315 and CC395 isolates expressed serotype 6C (n = 28/39 and n = 27/31, respectively), in children these lineages expressed mostly other serotypes: CC315/6B-2 (n = 6/8) and CC395/6A (n = 5/6). Serotype 6C was found in 3.0% of carriage isolates in children [21] but in only 0.4% of IPD (S1 Table). Taken together these observations suggest that children may be particularly protected against serotype 6C IPD, potentially through cross-protection due to the presence of polysaccharides 6A and 6B in PCV13. CC315 increased in adults (CA p = 0.012, significant after FDR). This CC included both 6C and 6B-2 isolates but only the 6C drove the increase (CA p = 0.001). In contrast, and in agreement with the decrease of 6B-2, there was a decrease in the major lineage expressing this class (CC273) in both children and adults (CA $p<0.001$ and p = 0.002, respectively, both significant after FDR) (S1 Fig and S3 and S4 Tables). Increase of the prevalence of serotype 6C pneumococci was reported in several regions [17,22,23]. However, the genetic lineage that increased was not always CC315. For example, in Southampton, England, the increase of serotype 6C pneumococci in carriage was due to the clonal expansion of CC395 [22]. In Spain, the authors associated the increase of serotype 6C isolates with spread of CC1150 [23], although isolates of CC315 emerged in 2007, coinciding with the data from IPD and carriage in Portugal.

Sequencing of the capsular genes *wciN* and *wciP* identified 5 and 14 alleles, respectively (S5 Table). There was a strong correspondence between the alleles at these loci and serotype and locus type but two isolates of ST460 presented unusual capsular loci. In one isolate a point mutation seems to have occurred in an allele characteristic of serotype 6A switching it to 6B (6B-1). The other possibly resulted from horizontal DNA transfer, presenting a hybrid locus including class 1 sequences associated with 6A and 6B-1 (wciN-1) and class 2 sequences (wciP-8). These observations confirm that serotype switching within serogroup 6 may occur in the wild by both point mutation and recombination.

The antimicrobial susceptibility of the isolates is indicated in Table 2. Serotype 6B-2 presented the highest proportion of multidrug resistant isolates (67%) followed by 6C (36%) (MDR, defined as non-susceptibility to at least 3 antimicrobial classes). The majority of isolates non-susceptible to penicillin (PNSP) [13] expressed low level resistance (MIC = 0.12–1 μg/mL), with the exception of a single 6B-2 isolate that expressed high level resistance (MIC = 2 μg/mL). None of the PNSP isolates was 6B-1. Considering the current CLSI guidelines [24], 6 cerebrospinal fluid isolates (6B-2, n = 3; 6C, n = 2; 6A, n = 1) would be considered resistant and all remaining isolates would be considered fully susceptible to penicillin using the non-meningitis breakpoints. Erythromycin resistance (ERP) was identified in 77 isolates and 53 isolates were simultaneously PNSP and ERP (serotype 6C, n = 29; serotype 6B-2, n = 20; and serotype 6A, n = 4). CC315 and CC273 presented the highest proportions of non-susceptible isolates (Table 2). The proportion of isolates representing these CCs changed over time and this was the basis of changes in resistance within serotypes 6C and 6B-2. Among 6C, the proportions of PNSP ($p<0.001$), ERP (p = 0.001), clindamycin (p = 0.001), tetracycline ($p<0.001$) and MDR (p = 0.001) increased (CA, all significant after FDR), changes driven mostly by increases in CC315. In fact, the remaining two CCs representing almost all serotype 6C isolates were CC1150, including mostly PNSP, and CC395 including mostly susceptible isolates. In 6B-2, decreases of ERP (p = 0.014), clindamycin (p = 0.005), tetracycline (p = 0.031) and MDR (p = 0.024) resistance were observed (CA, all significant after FDR correction), reflecting the decrease in CC273.

The rarity of serotypes 6D, 6F, 6G and 6H indicates that currently available phenotypic or simple genotypic assays remain adequate for distinguishing serotypes within serogroup 6 isolates. Although the 6B-2 class seems to have been preferentially affected by vaccination in

Table 2. Antimicrobial resistance of serogroup 6 isolates responsible for invasive infections in Portugal (1999–2012).

	No. of non-susceptible isolates (%)											
	Serotype				Clonal complex							
Antimicrobial [a]	6A	6B-1	6B-2	6C	CC315	CC395	CC65	CC273	CC176	CC1876	CC1150	Other[b]
MDR	4 (5.0)	3 (11.1)	35 (67.3)	30 (36.1)	47 (100.0)	0 (0)	1 (2.9)	17 (65.4)	3 (12.0)	1 (4.5)	0 (0)	3 (9.4)
PEN	10 (12.5)	0 (0)	26 (50.0)	42 (50.6)	45 (95.7)	1 (2.7)	2 (5.7)	9 (34.6)	0 (0)	4 (18.2)	13 (76.5)	4 (12.5)
MIC_{50}	0.023	0.016	0.047	0.064	0.125	0.016	0.023	0.023	0.016	0.016	0.094	0.023
MIC_{90}	0.064	0.032	0.19	0.19	0.19	0.032	0.047	0.38	0.023	0.19	0.125	0.064
ERY	4 (5.0)	9 (33.3)	34 (65.4)	30 (36.1)	47 (100.0)	0 (0)	0 (0)	16 (61.5)	9 (34.6)	3 (13.6)	0 (0)	2 (6.3)
CLI	1 (1.3)	4 (14.8)	33 (63.5)	30 (36.1)	47 (100.0)	0 (0)	0 (0)	16 (61.5)	4 (15.4)	0 (0)	0 (0)	1 (3.1)
LEV	0 (0)	0 (0)	0 (0)	1 (1.2)	1 (2.1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
SXT	10 (12.5)	6 (22.2)	27 (51.9)	3 (3.6)	6 (12.8)	3 (8.1)	4 (11.4)	16 (61.5)	5 (19.2)	1 (4.5)	2 (11.8)	9 (26.1)
TET	5 (6.3)	3 (11.1)	34 (65.4)	32 (38.6)	44 (93.6)	0 (0)	1 (2.9)	21 (80.8)	3 (11.5)	0 (0)	1 (5.9)	4 (12.5)
CHL	1 (1.3)	0 (0)	7 (13.5)	2 (2.4)	2 (4.3)	0 (0)	1 (2.9)	6 (23.1)	0 (0)	0 (0)	0 (0)	1 (3.1)

[a]All isolates were susceptible to cefotaxime, vancomycin and linezolid. MDR: multidrug resistance, PEN: penicillin, MIC: minimum inhibitory concentration, ERY: erythromycin, CLI: clindamycin, LEV: levofloxacin, SXT: trimethoprim-sulfamethoxazole, TET: tetracycline, CHL: chloramphenicol. Isolates presenting PEN MIC≥0.12μg/ml were considered resistant and isolates presenting PEN MIC<0.12μg/ml were considered susceptible.

[b]Other CCs or STs not included in those discriminated. MDR: CC4248 (n = 1), ST1518 (n = 1), ST1662 (n = 1); PEN: CC4248 (n = 1), ST1518 (n = 1), ST1662 (n = 1), ST4255 (n = 1); ERY: ST1518 (n = 1), ST1662 (n = 1); CLI: ST1662 (n = 1); SXT: CC681 (n = 1), CC4248 (n = 1), ST42 (n = 1), ST1518 (n = 1), ST1662 (n = 1), ST3324 (n = 1), ST9957 (n = 1), ST9965 (n = 1), ST9970 (n = 1); TET: CC4248 (n = 2), ST1715 (n = 1), ST3324 (n = 1); CHL: ST1662 (n = 1).

doi:10.1371/journal.pone.0170354.t002

Portugal, this was in contrast to what was documented elsewhere [20]. Further studies are necessary to clarify potential differences between the two classes and their response to vaccination. Expansion of resistant 6C lineages in adults, but not in children where the same lineages express other serogroup 6 serotypes, may be a hallmark of the post-PCV period and suggest that vaccination and antimicrobial use are the two major forces currently shaping pneumococcal populations.

Supporting information

S1 Fig. Temporal changes of the proportion of CC273 and CC315 among all invasive pneumococci in Portugal (1999–2012).
(PDF)

S1 Table. No. of isolates of each serotype of serogroup 6 responsible for invasive infections in children (<18 years) in Portugal (1999–2012).
(PDF)

S2 Table. No. of isolates of each serotype of serogroup 6 responsible for invasive infections in adults (≥18 years) in Portugal (1999–2012).
(PDF)

S3 Table. No. of isolates of STs and CCs of serogroup 6 responsible for invasive infections in children (<18 years) in Portugal (1999–2012).
(PDF)

S4 Table. No. of isolates of STs and CCs of serogroup 6 responsible for invasive infections in adults (≥18 years) in Portugal (1999–2012).
(PDF)

S5 Table. Allelic profiles of genes *wciN* and *wciP* and respective serotype and CC/ST.
(PDF)

Author contributions

Conceptualization: JDM SIA JMC MR.

Formal analysis: JDM JAC MR.

Funding acquisition: JMC.

Investigation: JDM SIA.

Methodology: JDM MR.

Project administration: MR.

Resources: JMC.

Supervision: MR.

Writing – original draft: JDM.

Writing – review & editing: JDM JMC MR.

References

1. Bratcher PE, Park IH, Oliver MB, Hortal M, Camilli R, Hollingshead SK, et al. Evolution of the capsular gene locus of *Streptococcus pneumoniae* serogroup 6. Microbiology. 2011; 157: 189–198. doi: 10.1099/mic.0.043901-0 PMID: 20929956
2. Burton RL, Geno KA, Saad JS, Nahm MH. Pneumococcus with the "6E" *cps* locus produces serotype 6B capsular polysaccharide. J Clin Microbiol. 2016; 54: 967–971. doi: 10.1128/JCM.03194-15 PMID: 26818670
3. Serrano I, Ramirez M, The Portuguese Surveillance Group for the Study of Respiratory Pathogens, Melo-Cristino J. Invasive *Streptococcus pneumoniae* from Portugal: implications for vaccination and antimicrobial therapy. Clin Microbiol Infect. 2004; 10: 652–6. doi: 10.1111/j.1469-0691.2004.00869.x PMID: 15214879
4. Aguiar SI, Serrano I, Pinto FR, Melo-Cristino J, Ramirez M. Changes in *Streptococcus pneumoniae* serotypes causing invasive disease with non-universal vaccination coverage of the seven-valent conjugate vaccine. Clin Microbiol Infect. 2008; 14: 835–843. doi: 10.1111/j.1469-0691.2008.02031.x PMID: 18844684
5. Aguiar SI, Brito MJ, Gonçalo-Marques J, Melo-Cristino J, Ramirez M. Serotypes 1, 7F and 19A became the leading causes of pediatric invasive pneumococcal infections in Portugal after 7 years of heptavalent conjugate vaccine use. Vaccine. 2010; 28: 5167–5173. doi: 10.1016/j.vaccine.2010.06.008 PMID: 20558247
6. Horácio AN, Diamantino-Miranda J, Aguiar SI, Ramirez M, Melo-Cristino J, the Portuguese Group for the Study of Streptococcal Infections. Serotype changes in adult invasive pneumococcal infections in Portugal did not reduce the high fraction of potentially vaccine preventable infections. Vaccine. 2012; 30: 218–224. doi: 10.1016/j.vaccine.2011.11.022 PMID: 22100892
7. Horácio AN, Diamantino-Miranda J, Aguiar SI, Ramirez M, Melo-Cristino J, the Portuguese Group for the Study of Streptococcal Infections. The majority of adult pneumococcal invasive infections in Portugal are still potentially vaccine preventable in spite of significant declines of serotypes 1 and 5. PLoS ONE. 2013; 8: e73704. doi: 10.1371/journal.pone.0073704 PMID: 24066064
8. Aguiar SI, Brito M, Horácio AN, Lopes J, Ramirez M, Melo-Cristino J, et al. Decreasing incidence and changes in serotype distribution of invasive pneumococcal disease in persons aged under 18 years since introduction of 10-valent and 13-valent conjugate vaccines in Portugal, July 2008 to June 2012. Euro Surveill. 2014; 19: pii: 20750.
9. Brito DA, Ramirez M, de Lencastre H. Serotyping *Streptococcus pneumoniae* by multiplex PCR. J Clin Microbiol. 2003; 41: 2378–84. doi: 10.1128/JCM.41.6.2378-2384.2003 PMID: 12791852

10. Kawaguchiya M, Urushibara N, Kobayashi N. High prevalence of genotype 6E (putative serotype 6E) among noninvasive/colonization isolates of *Streptococcus pneumoniae* in Northern Japan. Microb Drug Resist. 2015; 21: 209–214. doi: 10.1089/mdr.2014.0181 PMID: 25361198

11. Enright MC, Spratt BG. A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. Microbiology. 1998; 144: 3049–60. doi: 10.1099/00221287-144-11-3049 PMID: 9846740

12. Nascimento M, Sousa A, Ramirez M, Francisco AP, Carriço JA, Vaz C. PHYLOViZ 2.0: Providing scalable data integration and visualization for multiple phylogenetic inference methods. Bioinformatics. 2016;In press.

13. Clinical and Laboratory Standards Institute. Performance standards for antimicrobial susceptibility testing—seventeenth informational supplement. Wayne, PA: Clinical and Laboratory Standards Institute; 2007.

14. Carriço JA, Silva-Costa C, Melo-Cristino J, Pinto FR, de Lencastre H, Almeida JS, et al. Illustration of a common framework for relating multiple typing methods by application to macrolide-resistant *Streptococcus pyogenes*. J Clin Microbiol. 2006; 44: 2524–32. doi: 10.1128/JCM.02536-05 PMID: 16825375

15. Benjamini Y, Hochberg Y. Controlling the false discovery rate—a practical and powerful approch to multiple testing. J R Stat Soc Ser B Stat Methodol. 1995; 57: 289–300.

16. Marimón JM, Ercibengoa M, Tamayo E, Alonso M, Pérez-Trallero E. Long-term epidemiology of *Streptococcus pneumoniae* serogroup 6 in a region of Southern Europe with special reference to serotype 6E. PloS One. 2016; 11: e0149047. doi: 10.1371/journal.pone.0149047 PMID: 26863305

17. van der Linden M, Winkel N, Küntzel S, Farkas A, Perniciaro SR, Reinert RR, et al. Epidemiology of *Streptococcus pneumoniae* serogroup 6 isolates from IPD in children and adults in Germany. PloS One. 2013; 8: e60848. doi: 10.1371/journal.pone.0060848 PMID: 23593324

18. Horácio AN, Silva-Costa C, Diamantino-Miranda J, Lopes JP, Ramirez M, Melo-Cristino J, et al. Population structure of *Streptococcus pneumoniae* causing invasive disease in adults in Portugal before PCV13 availability for adults: 2008–2011. PloS One. 2016; 11: e0153602. doi: 10.1371/journal.pone.0153602 PMID: 27168156

19. McGee L, McDougal L, Zhou J, Spratt BG, Tenover FC, George R, et al. Nomenclature of major antimicrobial-resistant clones of *Streptococcus pneumoniae* defined by the pneumococcal molecular epidemiology network. J Clin Microbiol. 2001; 39: 2565–71. doi: 10.1128/JCM.39.7.2565-2571.2001 PMID: 11427569

20. Baek JY, Park IH, Song J-H, Ko KS. Prevalence of isolates of *Streptococcus pneumoniae* putative serotype 6E in South Korea. J Clin Microbiol. 2014; 52: 2096–2099. doi: 10.1128/JCM.00228-14 PMID: 24719436

21. Nunes S, Valente C, Sá-Leão R, de Lencastre H. Temporal trends and molecular epidemiology of recently described serotype 6C of *Streptococcus pneumoniae*. J Clin Microbiol. 2009; 47: 472–474. doi: 10.1128/JCM.01984-08 PMID: 19073873

22. Loman NJ, Gladstone RA, Constantinidou C, Tocheva AS, Jefferies JMC, Faust SN, et al. Clonal expansion within pneumococcal serotype 6C after use of seven-valent vaccine. PloS One. 2013; 8: e64731. doi: 10.1371/journal.pone.0064731 PMID: 23724086

23. Rolo D, Fenoll A, Ardanuy C, Calatayud L, Cubero M, de la Campa AG, et al. Trends of invasive serotype 6C pneumococci in Spain: emergence of a new lineage. J Antimicrob Chemother. 2011; 66: 1712–1718. doi: 10.1093/jac/dkr193 PMID: 21628304

24. Clinical and Laboratory Standards Institute. Performance standards for antimicrobial susceptibility testing—twenty-fourth informational supplement. Wayne, PA: Clinical and Laboratory Standards Institute; 2014.