Genome Biology and Evolution Advance Access published March 24, 2016 doi:10.1093/gbe/evw066

Genome Biology and Evolution

SMBE

Title:

Comparative genomics of carriage and disease isolates of *Streptococcus pneumoniae* serotype 22F reveals lineage specific divergence and niche adaptation

Authors:

Cleary DW[1,2], Devine VT[1], Jefferies JMC[1,2], Webb JS[2,3,4], Bentley SD[5], Gladstone RA[5], Faust SN[1,3,6] and Clarke SC[1,2,3,6*]

1. Academic Unit of Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, Southampton, UK

2. Institute for Life Sciences, University of Southampton, Southampton, UK

3. Southampton NIHR Respiratory Biomedical Research Unit, University Hospital Southampton Foundation NHS Trust, Southampton, UK

4. Centre for Biological Sciences, Faculty of Natural and Environmental Sciences, University of Southampton, Southampton, UK

5. Pathogen Genomics, Wellcome Trust Sanger Institute, Cambridge, UK

6. NIHR Southampton Wellcome Trust Clinical Research Facility, University Hospital Southampton Foundation NHS Trust, Southampton, UK

*Author for Correspondence: Dr Stuart C. Clarke, Academic Unit of Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, UK. Tel. +44 (0)2381 206652 E-mail S.C.Clarke@soton.ac.uk

Abstract:

Streptococcus pneumoniae is a major cause of meningitis, sepsis and pneumonia worldwide. Pneumococcal conjugate vaccines (PCV) have been part of the UK's childhood immunisation programme since 2006 and have significantly reduced the incidence of disease due to vaccine efficacy in reducing carriage in the population. Here we isolated two clones of 22F (an emerging serotype of clinical concern, multilocus sequence types (MLST) 433 and 698) and conducted comparative genomic analysis on four isolates, paired by ST with one of each pair being derived from carriage and the other disease (sepsis). The most compelling observation was of non-synonymous mutations in *pgdA*, encoding peptidoglycan *N*-acetylglucosamine deacetylase A, which were found in the carriage isolates of both ST433 and 698. Deacetylation of pneumococcal peptidoglycan is known to enable resistance to lysozyme upon invasion. Whilst no other clear genotypic signatures related to disease or carriage could be determined, additional intriguing comparisons between the two STs were possible. These include the presence of an intact prophage, in addition to numerous additional phage insertions, within the carriage isolate of ST433. Contrasting gene repertoires related to virulence and colonisation, including: bacteriocins, lantibiotics, and toxin-antitoxin systems, were also observed.

Keywords: *Streptococcus pneumoniae*; Genome sequencing; Invasive pneumococcal disease (IPD)

Introduction:

Streptococcus pneumoniae is a Gram positive commensal of the human nasopharynx. A highly recombinogenic bacterium, there are currently >90 known serotypes as defined by antisera cross-reactivity with the capsular polysaccharide. It is a leading global cause of pneumonia, bacteraemia and meningitis. Invasive pneumococcal disease (IPD) exhibits high levels of mortality and morbidity (over 900, 000 deaths in 2010 from pneumococcal pneumonia and pneumococcal meningitis (Lozano *et al.*, 2010)) and is of particular issue to children under five years of age and the elderly (O'Brien *et al.*, 2009). Carriage, a prerequisite for invasive disease, is approximately 30% for children under five in the UK and has remained such before and during vaccine implementation (Tocheva *et al.*, 2011). Not all carried serotypes are equally associated with IPD with some at least ten-fold more likely to cause disease (Brueggemann *et al.*, 2003).

Conjugate vaccine efficacy is now recognised to primarily act via population (herd) protection (Haber *et al.*, 2007). The changing epidemiology of pneumococcal serotypes and their association with IPD has been driven in the UK by the introduction of the polysaccharide-protein conjugate vaccines, PCV7, in 2006 and PCV13 in 2010, although fluctuations in the diversity of serotype and genotype have been demonstrated in the absence of PCV (Jefferies *et al.*, 2010). Although overall carriage of *S. pneumoniae* has remained consistent (Gladstone *et al.*, 2015), IPD caused by vaccine serotypes (VT) has dropped markedly (Feiken *et al.*, 2013).

As carriage rates have remained constant (Gladstone *et al.*, 2015), vigilance is still required to monitor IPD related to non-VTs - a process known as serotype-replacement. This occurs as a consequence of vaccines targeting a relatively small subset of serotypes thereby creating a vacant niche that has been filled by those not included in PCVs. Serotype 22F is one such example of serotype replacement that has been observed in the Southampton paediatric pneumococcal carriage study (Gladstone *et al.*, 2015). The most marked increase in carriage occurred between 2006 and 2009 (0.3%, n=1 to 2.3%, n= 9). Since then levels have stabilised at approximately 1% (unpublished data, personal communication). Of the 31 22F isolates obtained from the paediatric carriage study to date, 48% were multilocus sequence type (MLST) 433 and 29% ST698. Data from the European Centre for Disease Control's (ECDC) annual surveillance of invasive bacterial diseases has also shown year-on-year increases for IPD caused by this serotype (ECDC, 2013). For example 22F exhibited the largest proportional increase from 2010 (4.37 %) to 2011 (6.23 %) and in 2012 was the fifth most common serotype observed in IPD (7.4 % of cases, n=963) (ECDC, 2015). To address the gap in 22F genome data, we present here the first pairwise comparison of serotype 22F pneumococci that were isolated either from carriage or disease.

Methods

***S. pneumoniae* Serotype 22F Isolates:** Non-IPD isolates (referred to herein as 3326 and 3298) were isolated during the 2008/9 sampling period of the Southampton paediatric pneumococcal carriage study (REC No. 06/Q1704/105, RHM MED 0704). IPD isolates (09M852950S and 08M333175U, herein referred to as 09 and 08) were isolated from a 58 yr old male and 52 yr old female in 2009 and 2008 respectively.

Antibiotic Resistance: Susceptibility to penicillin, erythromycin and tetracycline was determined using E-test strips (AB Biodisk, Solna, Sweden) graduated from 0.016 to 256 μg mL^{-1}. Bacterial suspensions were prepared in saline from overnight cultures on blood agar to a density of 0.5 McFarland Standard. A swab from this suspension was used to inoculate Mueller-Hinton plates supplemented with 5 % sheep's blood (Becton Dickinson, Cockeysville, Md.). E-strips were then placed onto the plates whereupon they were incubated at 35 ^{o}C in 5 % CO_2 for 24 h. Minimal inhibitory concentrations (MICs) were read as the intersection between the ellipse edge and strip. Intermediate E-test MICs were adjusted to the next highest doubling-dilution value.

DNA Extraction and Genome Sequencing: Isolates of *S. pneumoniae* were cultured from STGG stocks stored at -70^{o}C on CBA plates incubated at 37 ^{o}C + 5 % CO_2 for 18 h. Genomic DNA was then extracted from ten combined colony picks using a QIAmp DNA Mini kit (Qiagen, UK) as per manufacturer's instructions. Concentration of genomic DNA was determined using Qubit™ 2.0 fluorometric quantitation (Thermo-Fisher, UK). Sequencing was done using

454™ 8kb and MiSeq 2x250 (V2) paired-end (PE) chemistry. 454™ was done at the Centre for Genomic Research, University of Liverpool. Illumina sequencing was performed at the University of Southampton.

Genomic Analysis: Hybrid assembly of Illumina and 454™ PE reads was done using MIRA v4.0.2 (Chevreux *et al.*, 1999). Read mapping for MLST, virulence and antibiotic resistance gene identification was done using SRST2 v0.1.3 using standard parameters (Inouye *et al.*, 2014). Virulence and antibiotic resistance genes were identified using the databases obtained from VFDB (http://www.mgc.ac.cn/VFs/main.html) and ResFinder (https://cge.cbs.dtu.dk/services/ResFinder/) respectively. Annotation of assemblies was performed using PROKKA v1.10 (Seeman, 2014), and RAST (Overbeek *et al.*, 2014). Genome comparisons were undertaken using the sequence comparison tool in RAST (Overbeek *et al.*, 2014), and BLAST Ring Image Generator (BRIG) (Alikhan *et al.*, 2011). Core genome SNP phylogeny was constructed using Wombac v2.0 (https://github.com/tseemann/wombac) with resultant trees visualised using FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/). Pangenome analysis was done using Roary (Page *et al.*, 2015). Breseq was used to identify non-synonomous mutations between the carriage and disease isolates (Deatherage & Barrick, 2014). Putative genomic islands containing prophage sequences were identified using PHAST (http://phast.wishartlab.com/) (Zhou *et al.*, 2011). Integrated conjugative elements were identified through a blastn search using BLOSUM62 in ICEberg (Bi *et al.*, 2012). Unless otherwise stated figures were produced using R Studio v0.98.994.

GenBank Submission: Genome sequences have been deposited in GenBank under accession numbers LSFU00000000, LSFV00000000, LSFW00000000, LSFX00000000. The versions described in this paper are versions LSFU01000000, LSFV01000000, LSFW01000000 and LSFX01000000.

Results and Discussion

Genome Sequencing: Assembly and annotation results are shown in Table 1. For isolate 3326 the assembly using only Illumina PE data generated fewer contigs (97 contigs with an N50 of 68 184) compared to the hybrid assembly using additional 454™ data and was thus used in further analyses. Annotation revealed between 1972 and 2154 CDS features.

The core and accessory gene content for these four isolates was determined using Roary (Page *et al.*, 2015). Analysis revealed a core and pan genome size of 1349 and 2416 CDS respectively (Figure 1). This core genome size is smaller than previous estimates of 1427 (Kulohoma *et al.*, 2015), 1553 (Obert *et al.*, 2006), and 1454 (Hiller *et al.*, 2007), but larger than the 1194 described from a similar surveillance study in Massachusetts, USA (Croucher *et al.*, 2013). It is not possible to determine whether this is a feature common to serotype 22F given the small number of genomes examined in the present study. Including additional STs of 22F (only two are examined here) would reduce the core genome size as the number of shared orthologous gene clusters would be less in a more genetically disparate collection.

The phylogeny of these four isolates, in the context of the broader species diversity for the pneumococcus, was determined using a maximum-likelihood neighbour-joining tree of core genome SNPs made with an additional 29 *S. pneumoniae* genomes (Figure 2). The overall tree topology is concordant with that shown previously (Donati *et al.*, 2010) with clustering of serotypes 1, 2 and 3 in particular but divergence of isolates of serotype 14 and 19. The isolates from the present study do not cluster as a serotype but do so by sequence type. No phylogenetic placement based on provenance in terms of disease or carriage was noted.

An all-by-all BLAST comparison of the four 22F isolates with *S. pneumoniae* TIGR4 (NC_003028) and the serotype 23F isolate, ATCC 700669 of the ST81 lineage (GCA_000026665.1) was done using BRIG (Alikhan *et al.*, 2011) and is shown in Figure 3. Rather than a single reference, two divergent genomes were chosen to enable broader comparisons within the two 22F sequence types. Regions where genome content differed notably from the two reference isolates are annotated in red. Although many of these regions contain large numbers of hypothetical genes, there are differences of note. Within the capsular polysaccharide synthesis locus (*cps*) for example the repertoire of Glycosyl transferases, accessory secretory proteins and translocases is shown to be heterogeneous for all four 22F isolates. This is in agreement with previous work that showed the presence of novel glycosyl and acetyltransferase genes at this locus in serotype 22F isolates (Salter *et al.*, 2012). Comparisons using both BRIG (Alikhan *et al.*, 2011), and RAST (Overbeek *et al.*, 2014), revealed no large regions of contrasting genomic content between the disease and carriage

isolates. However, between the STs there were some interesting contrasts. In particular were genes or loci that are linked to colonisation and niche competition which included toxin-antitoxins (TAs) systems as well as bacteriocin and lantibiotic synthesis gene clusters. Loci encoding TAs consist of two genes organised as an operon where the antitoxin encodes a repressor of the gene for a toxin protein. When under stress conditions, the repressor**,** which is labile, is degraded more rapidly allowing the toxin to bind cellular targets and halt essential cellular processes. Previous studies have shown there to be between four and ten of these TAs in *S. pneumoniae* although only *relBE2*, *yefM-yoeB*, *pezAT* and *phd-doc* are considered to be true TAs (Chan *et al.*, 2014). Both isolates of ST698 (09 and 3298) had the *phd-doc* and *yefM-yoeB* TAs whilst these were absent in 08 and 3326 of ST433. Both STs were found to harbour both the *relBE2* and *pezAT* systems. The former locus has been postulated to allow *S. pneumoniae* to colonise when conditions for unrestricted growth are unfavourable (Nieto *et al.*, 2010). Whether the lack of these additional systems hinders the stress responses of ST433 in particular thereby leading to a lower capacity for colonisation in comparison to ST698 remains to be seen.

Lanthionine, or lantibiotic, biosynthesis genes *lanL* and *lanM* were identified in ST433 isolates 08 and 3326 but not in ST698 isolates. Lantibiotics are antimicrobial compounds that exhibit a broad activity against Gram positive bacteria (Willey & van der Donk, 2007). Often associated with transposable elements (Croucher *et al.*, 2011) and previously observed in other expanding serotypes (Loman *et al.*, 2013), it is tempting to consider that the presence of these might affect a competitive advantage in niche colonisation for these isolates. Similarly, both isolates of ST433 were found to harbour additional

genes belonging to the Bacteriocin-like peptide family, notably *blpI* and *blpJ*. These small heat-labile proteins are common in Gram positive bacteria and have an established colonisation impact through elimination of competitor strains (Lux *et al.*, 2007).

We recently demonstrated accumulation of mutations within the DNA-directed RNA polymerase delta subunit (RpoE) of biofilm-derived small colony variants (SCV) of isolate 3326 (ST433) (our unpublished data). We therefore undertook a comparison of this locus in these four isolates. Whilst this did not reveal any difference in predicated amino acid sequence**,** there was a synonymous mutation at nucleotide position 69 that distinguished isolates of ST433 and 698.

Genomic Islands: Phylogenetic clusters of *S. pneumoniae* contain diverse genomic islands (GIs) that contribute to the genomic and phenotypic diversity of the species (Croucher *et al.*, 2014). At least three types of mobile genetic elements (MGEs) have been characterised in *S. pneumoniae*; phages, most commonly of the *Siphoviridae* family (Romero *et al.*, 2007); plasmids, of which just two cryptic examples are known (Smith & Guild, 1979; Romero *et al.*, 2009); and integrative and conjugative elements (ICEs).

Bacteriophage that have integrated within a host bacterial genome, termed prophages, are known to facilitate virulence gene transfer between numerous bacterial species via transduction; a well-established phenomenon in bacterial pathogens (Wagner & Waldor, 2002). No prophage sequences were found in disease isolate 08. Remnant operons were found in both ST698 isolates 3298 and 09. An intact prophage was found in isolate 3326 belonging to ST433. This

prophage is 53.1 kb in length constituting 68 CDS with a GC content of 39.61 %. A further three incomplete prophage regions of 9.3, 18.7 and 19.5 kb in length and containing a further 73 CDS were also identified. The gene content of these prophages is supplied in supplementary data (S1). Genes homologous to platelet-binding protein B (PblB) were found in these prophages. These have been shown to be important for virulence in *Streptococcus mitis* but their role in *S. pneumoniae* virulence has yet to be established (Harvey *et al.*, 2011).

ICEs are mobile genetic elements capable of being transmitted between bacteria through conjugative transfer, doing so autonomously using encoded conjugative elements. For each isolate the total number of ICEs identified were 94 and 89 for isolate 09 and 3298 (ST698), and 85 and 95 for 08 and 3326 (ST433). Seven ICEs that have previously been identified in *S. pneumoniae* were unevenly distributed among the four isolates and included ICESpn11930-2 and ICESpn23771 (3326 and 09 only), ICESpn9409, ICESpn11928 and MalM6, (3326, 3298 and 09) ICESpn8140 (08, 3326 and 3298) and HO34800032 (3326 and 3298). A further six elements were identified in all four isolates. These were ICE6094, ICESpn23FST81, ICESpn11876, ICESpn11930, *Tn*1545 (partial) and the pneumococcal pathogenicity island-1 (PPI-1), a ~30kb ICE common in *S. pneumoniae* and which contains genes that are essential for virulence (Harvey *et al.*, 2011). The PPI-1 region was found to differ between the two sequence types examined with ST433 in particular harboring a number of additional genes of unknown function in the 3' region as well as an ABC-2 transporter permease (Figure 3). These differences however did not relate to the key components of this GI previously shown to be essential for full virulence in murine systemic and

pulmonary models of infection (Brown *et al.*, 2004). These include the *piaABCD,* lipoprotein components of an iron ABC transport system, and Orf 9 and 10 of Tn*5252* constituting the conjugative machinery genes of a relaxase and a MobC-domain protein.

Virulence Genes: Although the polysaccharide capsule is the principal virulence determinant, *S. pneumoniae* also harbour additional factors important for disease. These include autolysins, LPXTG-anchored cell surface proteins such as hyaluronidase and serine protease, the pneumococcal pilus and choline binding proteins (Mitchell & Mitchell, 2010). In order to determine the presence of these known virulence factors in each of the four *S. pneumoniae* serotype 22F isolates, raw sequence reads were mapped against a gene database. Visualisation of this mapping against known alleles of these genes is shown in Figure 4. Similar to that observed with the core SNP analysis, isolates are shown to group by ST and not by phenotype i.e. disease or carriage. However notable differences in the presence (blue, where light-blue boxes represent novel alleles) and absence (black) of common virulence determinants were observed within each ST. Absences were confirmed by examining the annotated genomes. Only two virulence genes were found to be absent in comparisons between isolates of ST698. These were *rmlD* that was absent in isolate 3298 and *rfbD* that was absent in 09. Isolate 08 of ST433 revealed a divergent gene repertoire compared to 3326 lacking various alleles involved in capsule biosynthesis (*rmlD*, *rfbD*, *rfbB* and *wzh*). In addition, compared to 3326, 08 was shown to lack the *iga*, *plr*/*gapA* and *srtA* alleles, where the former are involved in cleavage of opsonising IgA1 and latter in adhesion respectively. Although gene content conservation across

the isolates was high it is important to highlight that there was a substantial degree of allelic (up to 10% sequence dissimilarity) for 20 of the 37 genes examined. Given the reported high levels of recombination and transformation for *S. pneumoniae* this allelic variation was to be expected.

Non-synonymous (NS) mutations: In addition to determining the presence of genes previously identified as important for virulence, the impact of single nucleotide polymorphisms was also considered. Here Breseq (Deatherage & Barrick, 2014) was used to compare the disease and carriage isolates for ST433 (Table 2) and ST698 (Table 3). Few NS mutations were observed between the disease and carriage isolates although it is interesting to note their presence within known virulence determinants such as autolysin, IgA protease and neuraminidase. The most intriguing observation was the presence of ns mutations of both carriage isolates in peptidoglycan-N-acetylglucosamine deacetylase A *(pgdA)*. These mutations were found in different regions of the *pgdA* CDS, occurring in 100 % of mapped reads with depths of 37x and 66x for the carriage isolates of ST433 and 698 respectively. Acetylation of pneumococcal peptidoglycan has previously been shown to increase sensitivity to lysozyme, a first line defence against bacterial invasion and thus, one might speculate, maintaining a deacetylated state may be of reduced necessity in those bacteria that colonise rather than go on to cause bacteraemia (Vollmer & Tomasz, 2002).

Antibiotic Resistance: Antibiotic resistance was assessed through *in silico* prediction and *in vitro* sensitivity testing using E-tests strips (AB Biodisk, Solna,

Sweden). No antibiotic resistance genes were identified from the sequence data. Additionally, all four isolates were shown to be susceptible to the three antibiotics tested with MIC's of < 0.25 μg mL^{-1} for tetracycline and erythromycin, and < 0.06 μg mL^{-1} for penicillin.

Conclusion: *S. pneumoniae* causing invasive disease remains a significant global challenge. Whilst the introduction of PCV has measurably altered pneumococcal molecular epidemiology, given the dynamics of serotype replacement (Jefferies *et al.*, 2010; Gladstone *et al.*, 2015) it remains vital to explore the genome repertoire of emerging serotypes of clinical significance. In this first-ever examination of serotype 22F genomes we have shown a number of lineage specific characteristics of significance to disease and carriage potential. These included the presence of an intact prophage within the carriage isolate of ST433, with contrasting gene repertoires related to both virulence and niche colonisation. The presence of non-synonymous mutations in *pgdA,* which were found in the carriage isolates of both ST433 and 698, were also of note given the previously observed role in pneumococcal resistance to lysozyme (Vollmer & Tomasz, 2002).

References

Alikhan NF, Petty NK, Zakour NL, Beatson SA. 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics, 12:402. doi: 10.1186/1471-2164-12-402

Bi D. 2012. ICEberg: a web-based resource for integrative and conjugative elements found in Bacteria. Nucleic Acids Res. 40, D621-D626. doi: 10.1093/nar/gkr846.

Brown JS, Gilliland SM, Spratt BG, Holden DW. 2004. A Locus Contained within a Variable Region of Pneumococcal Pathogenicity Island 1 Contributes to Virulence in Mice. Infect Immun. 72(3): 1587–1593 doi: 10.1128/IAI.72.3.1587-1593.2004

Brueggemann AB, et al. 2003. Clonal relationships between invasive and carriage *Streptococcus pneumoniae* and serotype- and clone-specific differences in invasive disease potential. J Infect Dis 187: 1424–1432.

Chan WT, Yeo CC, Sadowy E, Espinosa M. 2014. Functional validation of putative toxin-antitoxin genes from the Gram-positive pathogen *Streptococcus pneumoniae*: *phd-doc* is the fourth *bona-fide* operon. Front Microbiol. 5: 677 doi: 10.3389/fmicb.2014.00677

Chevreux B, Wetter T, Suhai S. 1999. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) 99:45-56.

Croucher NJ, et al. 2011. Rapid pneumococcal evolution in response to clinical interventions. Science 331 (6016):430 – 434 doi: 10.1126/science.1198545

Croucher NJ, et al. 2013. Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat Gen. 45 656-663 doi:10.1038/ng.2625

Croucher NJ, et al. 2014. Diversification of bacterial genome content through distinct mechanisms over different timescales. Nat Comms. 5:5471 doi:10.1038/ncomms6471

Deatherage DE, Barrick JE. 2014. Identification of mutations in laboratory evolved microbes from next-generation sequencing data using *breseq*. Methods Mol. Biol. 1151:165-188 doi: 10.1007/978-1-4939-0554-6_12

Donati C, et al. 2010. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. Genome Biol. 11:R107 doi: 10.1186/gb-2010-11-10-r107

European Centre for Disease Prevention and Control. Surveillance of invasive bacterial diseases in Europe, 2011. 2013. Stockholm: ECDC

European Centre for Disease Prevention and Control. Annual epidemiological report 2014 – Vaccine-preventable diseases – invasive bacterial diseases. 2015. Stockholm: ECDC

Feikin DR, et al. 2013. Serotype-Specific Changes in Invasive Pneumococcal Disease after Pneumococcal Conjugate Vaccine Introduction: A Pooled Analysis of Multiple Surveillance Sites. PLoS Med 10(9): e1001517 doi: 10.1371/journal.pmed.1001517

Gladstone RA, et al. 2015. Five winters of pneumococcal serotype replacement in UK carriage following PCV introduction. Vaccine 33:17 2015-21 doi: 10.1016/j.vaccine.2015.03.012

Haber M, et al. 2007. Herd immunity and pneumococcal conjugate vaccine: A quantitative model. Vaccine 25:29 5390-5398 doi:10.1016/j.vaccine.2007.04.088

Harvey RM, *et al*. 2011. A Variable Region within the Genome of *Streptococcus pneumoniae* Contributes to Strain-Strain Variation in Virulence. PLoS ONE 6(5): e19650. doi:10.1371/journal.pone.0019650

Hiller NL, et al. 2007. Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. J Bacteriol 189: 8186-8195 doi: 10.1128/JB.00690-07

Inouye M, *et al*. 2014. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome Med. 6(11): 90.

Jefferies JM, et al. 2010. Temporal analysis of invasive pneumococcal clones from Scotland illustrates fluctuations in diversity of serotype and genotype in the absence of pneumococcal conjugate vaccine. J Clin Microbiol. 8:1 87-96. Epub 2009 Nov 18. doi: 10.1128/JCM.01485-09.

Kulohoma BW, et al. 2015. Comparative genomic analysis of meningitis- and bacteremia-causing pneumococci identifies a common core genome. Infect. Immun. 83 (10): 4165. doi:10.1128/IAI.00814-15

Loman NJ, et al. 2013. Clonal Expansion within Pneumococcal Serotype 6C after Use of Seven-Valent Vaccine. PLoS ONE. 8(5): e64731. doi:10.1371/journal.pone.0064731

Lozano R, et al. 2012. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. The Lancet. 380: 9859. 2095-2128 doi:10.1016/S0140-6736(12)61728-0

Lux T, Nuhn M, Hakenbeck R, Reichmann P. 2007. Diversity of Bacteriocins and Activity Spectrum in *Streptococcus pneumoniae*. J Bacteriol. 189(21): 7741–7751 doi: 10.1128/JB.00474-07

Mitchell AM, Mitchell TJ. 2010. *Streptococcus pneumoniae*: virulence factors and variation. Clin Microbiol Infect. 16: 411-418 doi: 10.1111/j.1469-0691.2010.03183.x

Nieto C, Sadowy E, de la Campa AG, Hryniewicz W, Espinosa M. 2010. The *relBE2Spn* Toxin-Antitoxin System of *Streptococcus pneumoniae*: Role in Antibiotic Tolerance and Functional Conservation in Clinical Isolates PLoS One. 5(6): e11289. doi: 10.1371/journal.pone.0011289

Obert C, et al. 2006. Identification of a Candidate *Streptococcus pneumoniae* core genome and regions of diversity correlated with invasive pneumococcal disease. Infect. Immun. 74: 4766-4777 doi: 10.1128/IAI.00316-06

O'Brien KL, et al. 2009. Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates. Lancet 374:9693 893-902 doi:10.1016/S0140-6736(09)61204-6

Overbeek R, *et al*. 2014. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. 42(Database issue): D206–D214. doi: 10.1093/nar/gkt1226

Page AJ, *et al*. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 31 (22): 3691-3693. doi: 10.1093/bioinformatics/btv421

Romero P, et al. 2007. Isolation and characterisation of a new plasmid pSpnP1 from a multidrug-resistant clone of *Streptococcus pneumoniae*. Plasmid 58: 51-60 doi:10.1016/j.plasmid.2006.12.006

Romero P, et al. 2009. Comparative genomic analysis of ten *Streptococcus pneumoniae* temperate bacteriophages. J. Bacteriol. 191: 4854-4862 doi: 10.1128/JB.01272-08.

Salter SJ et al. 2012. Variation at the capsule locus, cps, of mistyped and non-typable *Streptococcus pneumoniae isolates*. Microbiology. 158: 1560-1569 doi:10.1099/mic.0.056580-0

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30(14): 2068-9. doi: 10.1093/bioinformatics/btu153

Smith MD, Guild WR. 1979. A plasmid in *Streptococcus pneumoniae*. J. Bacteriol. 137:735-739

Tocheva AS, et al. 2011. Declining serotype coverage of new pneumococcal conjugate vaccines relating to the carriage of *Streptococcus pneumoniae* in young children. Vaccine 29:26 4400‑4404 doi: 10.1016/j.vaccine.2011.04.004.

Vollmer W, Tomasz A. 2002. Peptidoglycan *N*-Acetylglucosamine Deacetylase, a Putative Virulence Factor in *Streptococcus pneumoniae***.** Infect Immun. 70(12): 7176–7178. doi: 10.1128/IAI.70.12.7176-7178.2002

Wagner PL, Waldor MK. 2002. Bacteriophage control of bacterial virulence. Infect. Immun. 70: 3985-3993. doi: 10.1128/IAI.70.8.3985-3993.2002

Willey JM, van der Donk WA. 2007. Lantibiotics: peptides of diverse structure and function. Annu Rev Microbiol. 61:477–501 doi: 10.1146/annurev.micro.61.080706.093501

Zhou Y, Liang Y, Lynch K, Dennis JJ, Wishart DS. 2011. PHAST: A Fast Phage Search Tool. Nucl. Acids Res. 39(suppl 2): W347-W352 doi: 10.1093/nar/gkr485

Acknowledgments

This work was supported by a small project grant from the MRC/CGR Small Research Projects Initiative. The authors wish to acknowledge the Centre for Genomic Research, University of Liverpool for providing the 454™ sequence data. We also wish to thank Rebecca Anderson and Andrew C. Tuck for assistance with culture of the isolates and generation of Illumina sequence data.

Table Legends

Table 1: Assembly and annotation statistics for *Streptococcus pneumoniae* serotype 22F disease isolates 08, 09 and carriage isolates 3298 and 3326.

Table 2: Non-synonymous mutations identified between disease and carriage isolates of *Streptococcus pneumoniae* serotype 22F, ST433

Table 3: Non-synonymous mutations identified between disease and carriage isolates of *Streptococcus pneumoniae* serotype 22F, ST698

Figure Legends

Figure 1: Pangenome size as defined by shared genes (left) and number of unique genes (right) within the four *S. pneumoniae* serotype 22F isolates examined. The dashed blue line represents the core genome size of 1349.

Figure 2: Maximum likelihood neighbor-joining tree showing phylogenetic placement of the *S. pneumoniae* serotype 22F isolates of ST698 (purple) and ST433 (green).

Figure 3: Genome comparisons of four *S. pneumoniae* serotype 22F isolates against TIGR4 and ATCC70669 23F. Rings representing ST433 isolates (08 and 3326) are coloured green with ST698 isolates (09 and 3298) coloured purple. The red labels show regions of gene absences, gene classifications for which were derived from TIGR4 annotations in RAST.

Figure 4: Distribution of common virulence genes found in *S. pneumoniae* serotype 22F isolates as determined by SRST2. Key: Dark blue – known allele, light blue – novel allele (10 % sequence dissimilarity to known allele), black – not identified. Columns are colour-coded by ST: green – ST433, purple - ST698.

Table 1

Isolate	Contigs	N50	Depth	CDS	tRNA	rRNA
08	110	53 893	63.66	2016	56	5
09	64	86 298	41.06	1975	48	5
3298	76	78 042	58.61	1972	50	6
3326	64	68 184	34.28	2154	48	6

Figure 1

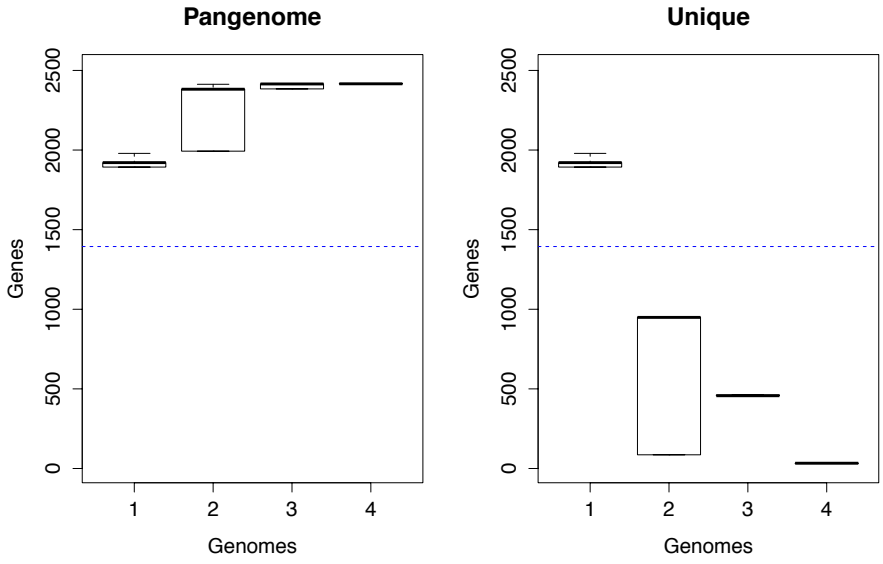


Figure 2

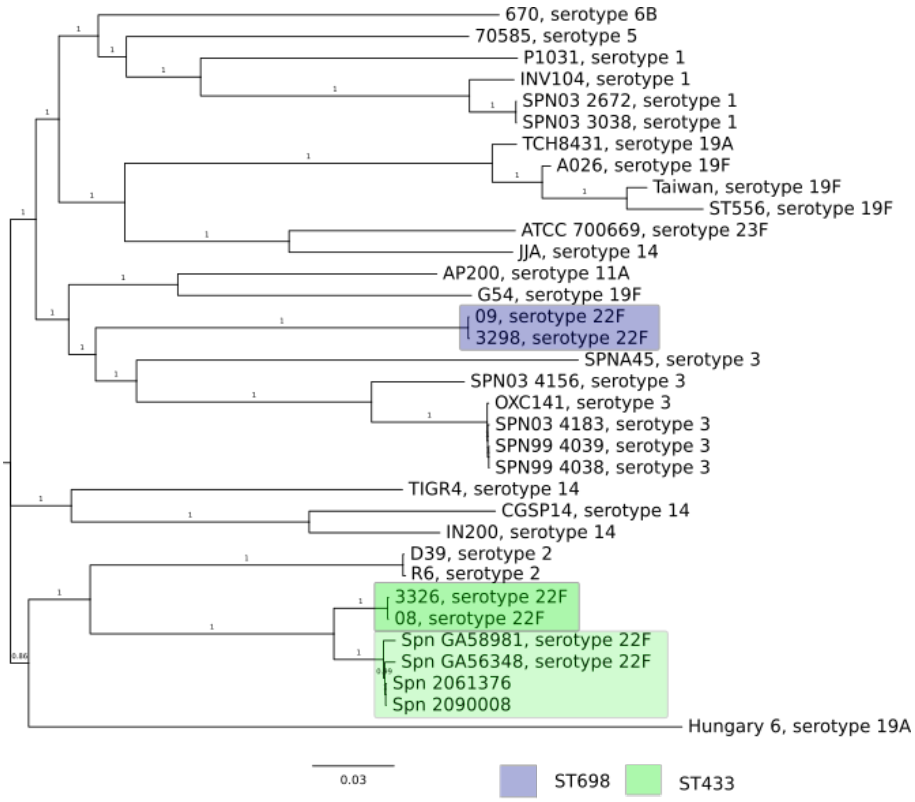


Figure 3

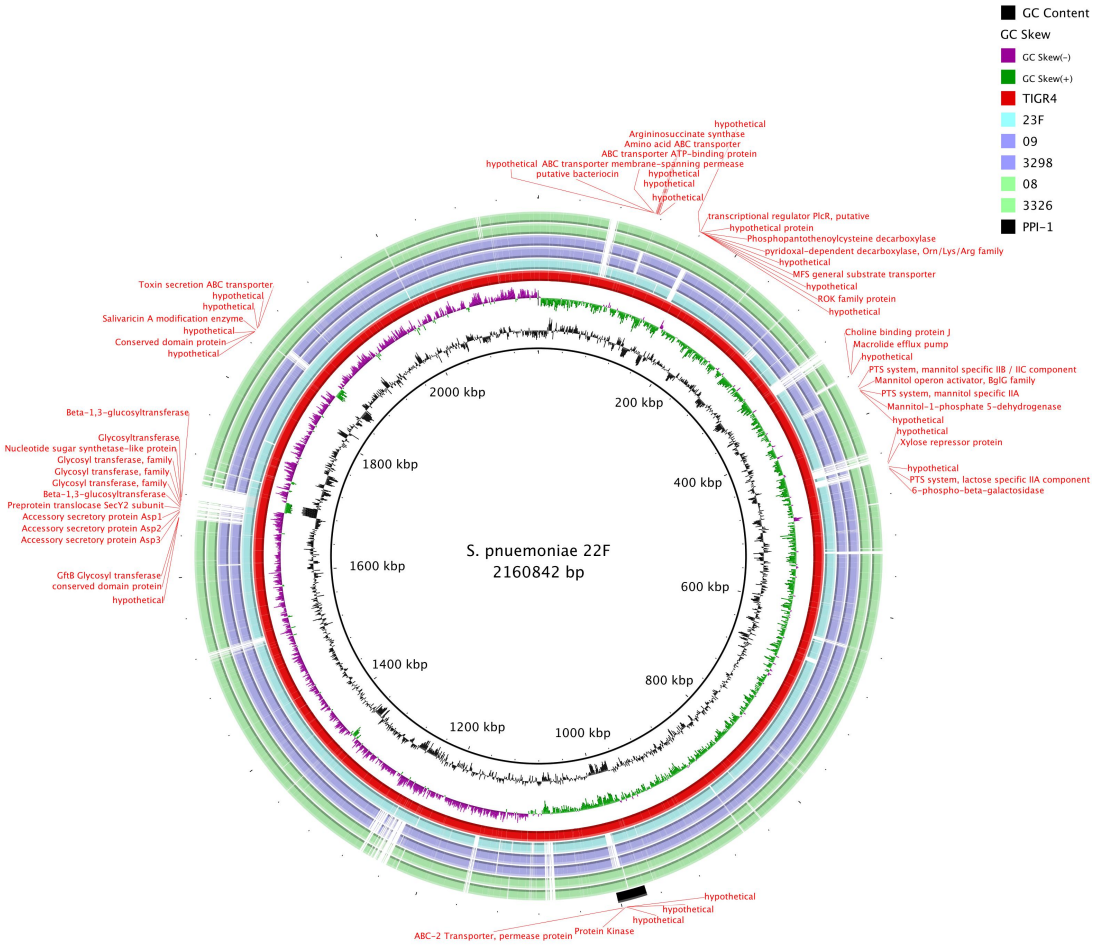
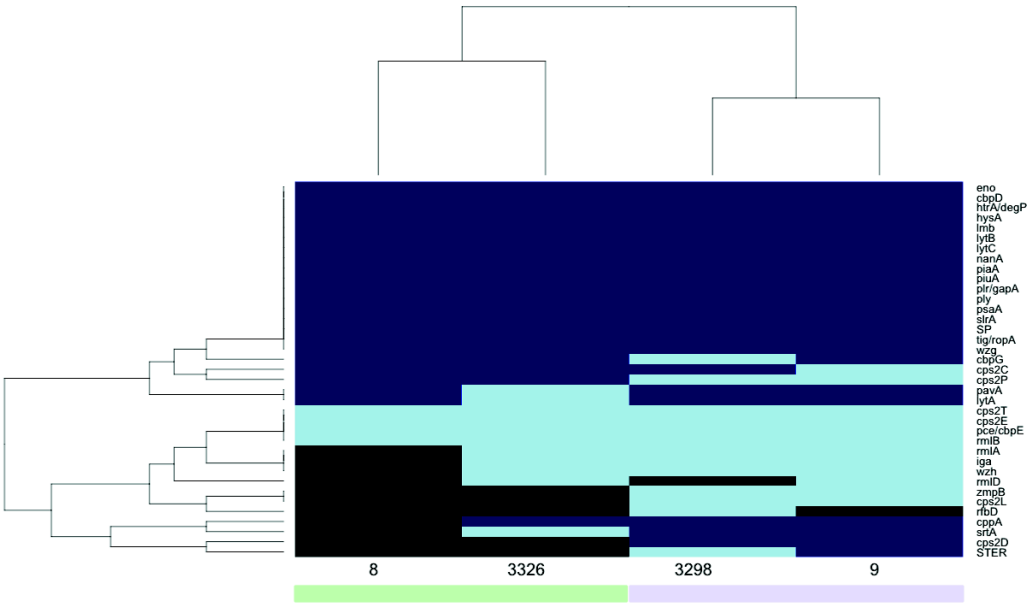


Figure 4



Non-synonymous mutation (disease→carriage)	Gene	Function
V-A (GTT→GCT)	*bglA_1*	Aryl-phospho-beta-D-glucosidase BglA
H-R (CAC→CGC)	*fucI*	L-fucose isomerase
V-F (GTT→TTT)	-	Glycosyl hydrolase family 20, catalytic domain
L-M (TTG→ATG)	*ssb_1*	Single-stranded DNA-binding protein ssb
Y-H (TAT→CAT)	*lytA_4*	Autolysin
I-V (ATT→GTT)	*lytA_4*	Autolysin
K-R (AAG→AGG)	-	hypothetical protein
M-I (ATG→ATA)	*pgdA*	Peptidoglycan-N-acetylglucosamine deacetylase
C-G (TGC→GGC)	*azr_1*	FMN-dependent NADPH-azoreductase
G-R (GGA→AGA)	*azr_1*	FMN-dependent NADPH-azoreductase
F-S (TTT→TCT)	*licT_1*	Transcription antiterminator LicT
R-H (CGT→CAT)	*sarA_2*	Oligopeptide-binding protein SarA precursor
V197I (GTT→ATT)	*yecS_2*	Inner membrane amino-acid ABC transporter permease protein YecS
G-C (GGT→TGT)	*yheS_2*	putative ABC transporter ATP-binding protein YheS
R-S (CGT→AGT)	*prmC*	Release factor glutamine methyltransferase
H-Q (CAT→CAG)	-	lineage-specific thermal regulator protein
S-I (AGT→ATT)	*ugd_2*	UDP-glucose 6-dehydrogenase
C-Y (TGC→TAC)	-	Ribonuclease J 1
I-L (ATC→CTC)	*mnmG*	tRNA uridine 5-carboxymethylaminomethyl modification enzyme MnmG
Y-D (TAT→GAT)	*strH_2*	Beta-N-acetylhexosaminidase precursor
L-F (TTA→TTC)	*treA*	Trehalose-6-phosphate hydrolase
T-I (ACC→ATC)	*stkP*	Serine/threonine-protein kinase StkP
A-S (GCC→TCC)	-	hypothetical protein
P-S (CCA→TCA)	*cshA_2*	DEAD-box ATP-dependent RNA helicase CshA
E-D (GAG→GAT)	*hlyB_1*	Alpha-hemolysin translocation ATP-binding protein HlyB
C-Y (TGT→TAT)	-	Bacteriocin class II with double-glycine leader peptide
V-A (GTA→GCA)	*rplU*	50S ribosomal protein L21
S-L (TCA→TTA)	*manZ_4*	Mannose permease IID component
L-S (TTG→TCG)	*ugl*	Unsaturated chondroitin disaccharide hydrolase

Non-synonomous mutation (disease→carriage)	Gene	Function
*-Q (TAG→CAG)	*iga_2*	Immunoglobulin A1 protease precursor
*-S (TAA→TCA)	-	hypothetical protein
A-S (GCT→TCT)	*rnjB*	Ribonuclease J 2
A-V (GCT→GTT)	*recG*	ATP-dependent DNA helicase RecG
A-V (GCA→GTA)	-	hypothetical protein
A-T (GCT→ACT)	-	hypothetical protein
C-Y (TGC→TAC)	-	hypothetical protein
D-A (GAT→GCT)	*trpF*	N-(5'-phosphoribosyl)anthranilate isomerase
G-D (GGT→GAT)	*pgdA*	Peptidoglycan-N-acetylglucosamine deacetylase
G-E (GGG→GAG)	*glnH_1*	Glutamine-binding periplasmic protein precursor
G-V (GGG→GTG)	*nanA_2*	Sialidase A precursor
G-C (GGT→TGT)	-	hypothetical protein
I-T (ATT→ACT)	*ybjI*	Flavin mononucleotide phosphatase YbjI
K-E (AAA→GAA)	*pflA_1*	Pyruvate formate-lyase-activating enzyme
L-F (TTA→TTC)	*kanE*	Alpha-D-kanosaminyltransferase
L-F (CTT→TTT)	*leuS*	Leucine--tRNA ligase
P=S (CCC→TCC)	*pstA_2*	Phosphate transport system permease protein PstA
Q-K (CAA→AAA)	*rpoC*	DNA-directed RNA polymerase subunit beta'
Q-P (CAA→CCA)	*thiD*	Hydroxymethylpyrimidine/phosphomethylpyrimidine kinase
R-* (CGA→TGA)	-	hypothetical protein
T-I (ACA→ATA)	-	hypothetical protein
T-A (ACT→GCT)	*panT*	Pantothenic acid transporter PanT
V-A (GTC→GCC)	-	hypothetical protein
V-A (GTA→GCA)	-	hypothetical protein
V-M (GTG→ATG)	-	hypothetical protein
Y-D (TAC→GAC)	*thiD*	Hydroxymethylpyrimidine/phosphomethylpyrimidine kinase
A-V (GCT→GTT)	*tmk*	Thymidylate kinase