Streptococcus pneumoniae Clonal Complex 199: Genetic Diversity and Tissue-Specific Virulence

Jonathan C. Thomas[1,2], Marisol Figueira[3], Kristopher P. Fennie[4], Alison S. Laufer[1], Yong Kong[5], Michael E. Pichichero[6], Stephen I. Pelton[3], Melinda M. Pettigrew[1]*

1 Division of Epidemiology of Microbial Diseases, Yale School of Public Health, Yale University School of Medicine, New Haven, Connecticut, United States of America, **2** Department of Microbiology, University of Mississippi Medical Center, Jackson, Mississippi, United States of America, **3** Boston University School of Medicine and Public Health, Boston Medical Center, Boston, Massachusetts, United States of America, **4** Yale University School of Nursing, New Haven, Connecticut, United States of America, **5** Department of Molecular Biophysics and Biochemistry, W. M. Keck Foundation Biotechnology Resource Laboratory, New Haven, Connecticut, United States of America, **6** Rochester General Hospital Research Institute, Rochester, New York, United States of America

Abstract

Streptococcus pneumoniae is an important cause of otitis media and invasive disease. Since introduction of the heptavalent pneumococcal conjugate vaccine, there has been an increase in replacement disease due to serotype 19A clonal complex (CC)199 isolates. The goals of this study were to 1) describe genetic diversity among nineteen CC199 isolates from carriage, middle ear, blood, and cerebrospinal fluid, 2) compare CC199 19A (n = 3) and 15B/C (n = 2) isolates in the chinchilla model for pneumococcal disease, and 3) identify accessory genes associated with tissue-specific disease among a larger collection of *S. pneumoniae* isolates. CC199 isolates were analyzed by comparative genome hybridization. One hundred and twenty-seven genes were variably present. The CC199 phylogeny split into two main clades, one comprised predominantly of carriage isolates and another of disease isolates. Ability to colonize and cause disease did not differ by serotype in the chinchilla model. However, isolates from the disease clade were associated with faster time to bacteremia compared to carriage clade isolates. One 19A isolate exhibited hypervirulence. Twelve tissue-specific genes/regions were identified by correspondence analysis. After screening a diverse collection of 326 isolates, spr0282 was associated with carriage. Four genes/regions, SP0163, SP0463, SPN05002 and RD8a were associated with middle ear isolates. SPN05002 also associated with blood and CSF, while RD8a associated with blood isolates. The hypervirulent isolate's genome was sequenced using the Solexa paired-end sequencing platform and compared to that of a reference serotype 19A isolate, revealing the presence of a novel 20 kb region with sequence similarity to bacteriophage genes. Genetic factors other than serotype may modulate virulence potential in CC199. These studies have implications for the long-term effectiveness of conjugate vaccines. Ideally, future vaccines would target common proteins to effectively reduce carriage and disease in the vaccinated population.

Citation: Thomas JC, Figueira M, Fennie KP, Laufer AS, Kong Y, et al. (2011) *Streptococcus pneumoniae* Clonal Complex 199: Genetic Diversity and Tissue-Specific Virulence. PLoS ONE 6(4): e18649. doi:10.1371/journal.pone.0018649

Editor: Frank R. DeLeo, National Institute of Allergy and Infectious Diseases, National Institutes of Health, United States of America

Received October 12, 2010; **Accepted** March 14, 2011; **Published** April 14, 2011

Funding: This work was supported by the National Institute of Allergy and Infectious Diseases (R01 A1068043) to MMP. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Melinda.Pettigrew@yale.edu

Introduction

Streptococcus pneumoniae asymptomatically colonizes the upper respiratory tract of approximately half of all healthy children and is a leading cause of acute otitis media, pneumonia and meningitis globally [1]–[3]. Although our comprehension of the epidemiology, pathogenesis, and virulence factors of *S. pneumoniae* has improved in recent years, the basis for whether colonization with a specific strain establishes asymptomatic colonization or produces local or invasive diseases requires further elucidation.

Encapsulated strains of *S. pneumoniae* express one of at least 93 distinct capsular polysaccharides [4], [5]. Since the introduction of the heptavalent pneumococcal conjugate vaccine (PCV7; serotypes 4, 6B, 9V, 14, 18C, 19F and 23F) in 2000, a large overall decrease in invasive disease continues to be observed [6], [7]. However, non-vaccine serotypes have increased in prevalence [8]–[11]. Studies show a significant increase in the number of otitis media and invasive disease cases due to serotype 19A [8], [12]–[14]. In the United States and Europe, clonal complex (CC)199 was a major clonal lineage throughout the period of expansion of serotype 19A [9], [14]–[17]. A small percentage of these CC199 19A isolates are associated with antimicrobial resistance [14]. CC199 strains may also express the 15B/C capsule. Serogroup 15 isolates have recently increased in prevalence, albeit to a lesser extent than serotype 19A [18], [19]. Serotype 19A is now included in the expanded pneumococcal conjugate vaccine (PCV13), while serotype 15B/C is not.

Differences in virulence have been reported between pneumococcal serotypes [11], [20]–[22]. Furthermore, differences in virulence within a serotype have been reported in animal models [23] and population based studies [24]. Even closely related isolates belonging to the same clone or sequence type (ST) can differ in their capacity to cause disease [23].

The pneumococcal genome exhibits high plasticity and may be categorized into the core genome, consisting of genes conserved between all *S. pneumoniae* isolates, and the accessory genome, consisting of genes that are variably present throughout the *S. pneumoniae* population [25]–[27]. Between 21–32% of genes in a given strain belong to the accessory genome [26]. It has been hypothesized that some genes may engender tissue-specific advantages [23], [28], such as providing the isolate with an increase in fitness or capacity to invade a given niche. In the identification of tissue-specific genes, studies have often focussed on established pneumococcal virulence factors or genes that are differentially expressed *in vivo* [29]–[31]. Few studies have examined the non-core component of the pneumococcal genome to identify genes that provide a biological basis for tissue-specificity. Comparative genome studies have focussed on invasive disease rather than otitis media as a disease outcome [23], [27], [32].

This study sought to evaluate genetic variation among CC199 isolates, and to identify genes associated with strains isolated from a particular tissue source. CC199 isolates of the same genetic background but different serotype (serotype 19A and 15B/C) were compared in the chinchilla model of pneumococcal disease to assess the relationship between genetic diversity and capsular serotype on virulence. Importantly, this model allowed us to ascertain virulence for otitis media. These strains also produced bloodstream infection, permitting assessment of relative virulence for invasive disease. The initial analysis of related isolates from the same clonal complex reduced the genetic variation identified, and consequently the level of noise encountered while identifying tissue-associated genes among CC199. To gain additional insight into *S. pneumoniae* tissue-associated genes, we identified genetic regions associated with specific tissue sources among CC199 isolates and used these to screen a larger, genetically diverse collection of pneumococcal isolates. We reasoned that if the identified genes were truly important for tissue-specific virulence, then they would also occur more frequently among diverse pneumococcal sequence types and serotypes from these tissue sites.

Results

Genomic Diversity within Clonal Complex 199

Comparative genome hybridization (CGH) was used to evaluate genetic diversity among nineteen CC199 isolates. One hundred and twenty seven genes were found to be variable within CC199 (Table S1). The CGH results identified four regions of diversity (RD), using the criteria specified by previous studies investigating pneumococcal genomic diversity [23]. These included RD2, RD6, and RD24 [23], and RD8 [27] (Table 1). The regions of diversity differed from those previously described by Silva *et al.* RD2 identified within CC199 isolates did not contain SP0114 and SP0115 and was approximately 0.7 kb smaller; RD24, identified within CC199 isolates, was slightly larger due to the addition of SP1947.

A phylogeny of the isolates was constructed by hierarchical clustering based on the $\log_2$ ratio matrix (Figure 1). The phylogeny obtained by hierarchical clustering closely agreed with the dendrogram obtained by Dollo parsimony of the presence and absence matrix (data not shown). CC199 splits into two main clades, one of which consists of all but one of the disease isolates, while the other comprises all but one of the carriage isolates (Figure 1). The carriage clade is further split, based on serotype, while the disease clade exhibits no additional structure based on serotype.

Chinchilla Model of Pneumococcal Disease

We next sought to compare the virulence of carriage and disease clade isolates in the chinchilla model of disease. Five representative CC199 strains were selected for testing. In addition to a genome sequenced reference 19A strain (CDC3059-06), one serotype 15B/C and one serotype 19A isolate were selected from both the carriage and invasive clades (Figure 1). None of the CC199 isolates exhibited a significantly increased ability to colonize (Figure 2a) or cause otitis media (Figure 2b). There was little difference between serotype 19A and serotype 15B/C isolates' ability to cause otitis media (Figure 2b) or in the time taken for the isolate to cause bacteremia ($p = 0.63$).

Table 1. Regions of genetic diversity identified in clonal complex 199 isolates.

Region of Diversity	Genes	Size*	Encodes§
2	SP0109	~4.2 kb	Putative Bacteriocin
	-		Putative Amino Acid ABC Transporters
	SP0113		Hypothetical Protein
6	SP0394	~5.3 kb	Putative transcriptional regualtor
	-		Mannitol Phosphate Dehydrogenase
	SP0397		Phophotransferase System
8	SP1315	~31.5 kb	Sodium ATP Synthase
	-		Oxidoreductase
	SP1351		Putative neuraminidase
			N-Acetylneuraminate Lyase
			Putative N-Acetylmannosamine-6-P Epimerase
			Putative Phosphosugar-Binding Transcriptional Regulator
			Methyltransferase
			Transposase
			ABC Transporter/ATP Binding Protein
			Drug Efflux ABC Transporter
			Prolyl Oligopeptidase Family Protein
			Putative Membrane Protein
			Hypothetical Proteins
24	SP1947	~9.0 kb	Putative Bacteriocin Formation Protein
	-		Toxin Secretion ABC Transporter
	SP1955		Serine Protease
			Hypothetical Proteins

*Based on gene sizes taken from TIGR4 genome.
§Based on annotation taken from TIGR4 genome.
doi:10.1371/journal.pone.0018649.t001

While the isolates did not differ in their virulence potential by serotype, isolates belonging to the disease clade of the CC199 phylogeny did exhibit a decreased duration of colonization before causing bacteremia compared to carriage clade isolates ($p = 0.04$). Of the isolates from the disease clade, one serotype 19A isolate (06-01-003 MEF) demonstrated an increased invasive virulence potential beyond that of the other isolates, resulting in chinchilla mortality within 48–72 hours, significantly more rapidly than other disease isolates ($p<0.0001$).

Tissue-Specific Genes

The separation of the phylogeny of CC199 based predominantly on isolate tissue source (i.e. carriage vs. disease), combined with the similar fitness of isolates by serotype, indicates that genetic factors other than serotype impact on fitness for tissue-specific

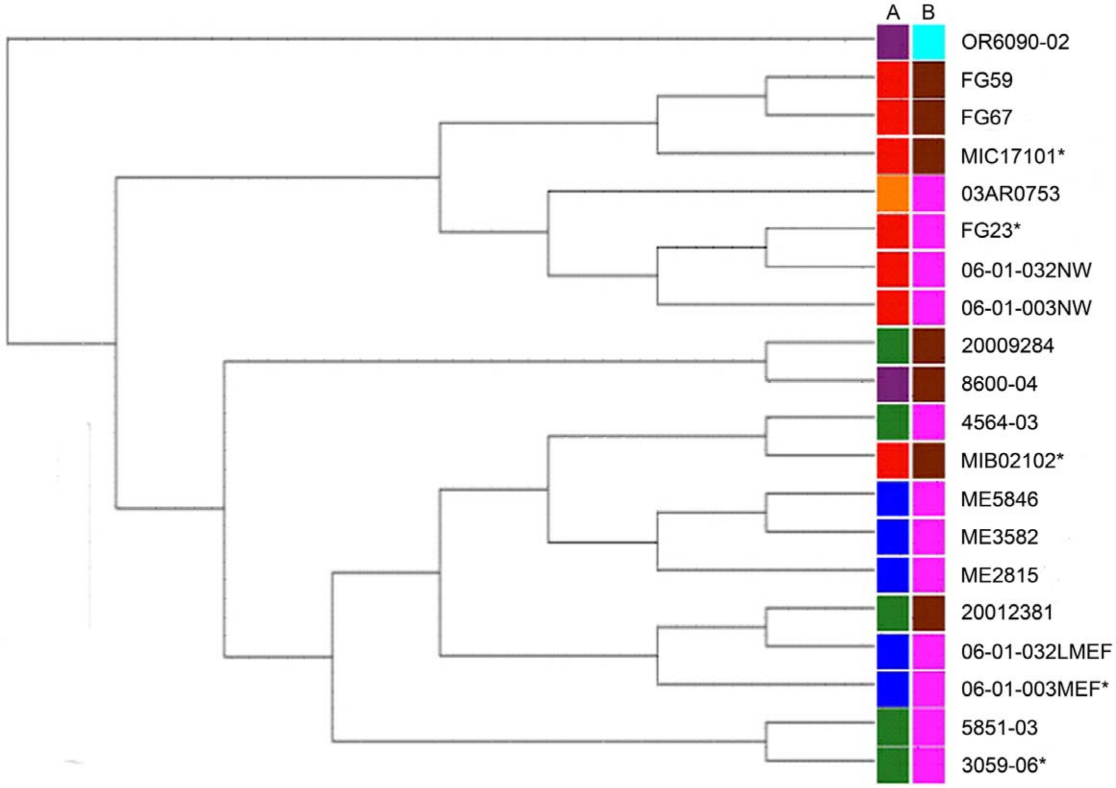

Figure 1. Hierarchical clustering dendrogram of clonal complex 199 isolates. *Strain tested in the chinchilla model of disease Coloured column one relates to tissue source. Red = carriage, Blue = Middle ear fluid, Green = Blood, Purple = Cerebrospinal fluid, Orange = Pleural Fluid. Coloured column two relates to serotype. Bright blue = 3, Brown = 15B/C, Pink = 19A.
doi:10.1371/journal.pone.0018649.g001

virulence. The microarray analysis was not able to identify any single genetic factor as being responsible for the difference between the carriage and disease clades. Therefore, we sought to identify non-core genes that might influence tissue-specific virulence.

Correspondence analysis identified 11 genes as well as the region of diversity RD8a as being associated with CC199 isolates obtained from a particular tissue source. Next, we screened a larger, genetically diverse collection of *S. pneumoniae* isolates (n = 326). Eight of the 11 genes were significantly associated with a specific tissue source after screening of the larger collection. Two of these genes encoded choline-binding proteins (*cbpC* and *cbpJ*), which were negatively associated with invasive strains, compared to carriage, as determined by PCR (OR [95% CI]): 0.46 [0.23, 0.90] and 0.33 [0.17, 0.61] for *cbpC* and *cbpJ*, respectively. These findings could not be confirmed by hybridization due to the cross-hybridization of the probe with other choline binding protein-encoding genes in the genome.

Based on PCR, dot blot, and Southern hybridization screening, seven genetic regions were found to be significant in their association with either a particular tissue source, or with either invasive or non-invasive isolates. The prevalence data for the seven genetic regions within CC199 are in Table 2. It should be noted that correspondence analysis does not provide a statistical test of the strength of the association and a limited number of CC199 strains were examined. Among the larger collection of *S. pneumoniae* isolates, the prevalence ranged from 29% (RD8 in carriage isolates) to 81% (SP1793 in CSF isolates) (Table 2). The odds ratios and 95% CI for each genetic region by tissue source are presented in Table 3. SPN05002, which encodes a hypothetical protein, was present significantly more frequently in strains isolated from middle ear fluid, blood, and CSF than from those strains isolated from carriage. The genetic region RD8a was also present significantly more frequently in strains isolated from middle ear fluid and blood than carriage. Conversely, spr0282, which encodes a phosphotransferase system sugar-specific EII component, was present significantly more frequently in strains isolated from carriage than from middle ear fluid or blood.

Hypervirulent Isolate

As described above, one isolate 06-01-003MEF (CC199, 19A) was found to have an increased invasive virulence potential in the chinchilla model. No single genetic variation could be found by microarray to explain the increased virulence observed for the serotype 19A isolate. While current microarray slides comprise probes for genes from several completely sequenced genomes, the high degree of plasticity in *S. pneumoniae* genomes is such that microarrays will not possess the full complement of probes for a given isolate being tested. Therefore, a comparison was made of the genome sequences of the hypervirulent isolate and the isolate CDC3059-06 (CC199, 19A), which was representative of the virulence potential of the CC199 isolates in this study.

The genome of the hypervirulent isolate 06-01-003MEF was found to contain approximately 35 kb of genetic material not possessed by the reference isolate CDC3059-06, while CDC3059-06 contained approximately 42 kb of sequence not found in the genome of 06-01-003MEF. The three remaining CC199 isolates tested in the chinchilla model of pneumococcal disease were screened for the presence or absence of these genetic regions, both by PCR and by examining the microarray data. A single genetic

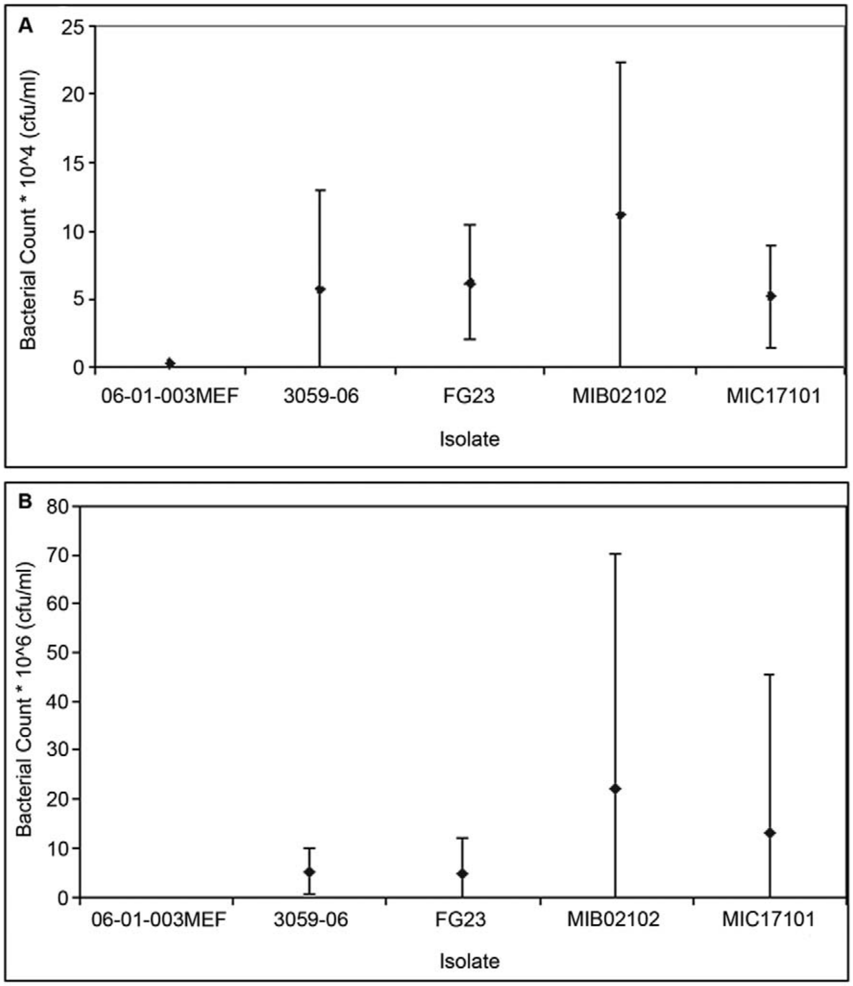

Figure 2. Bacterial counts for each of the five CC199 pneumococcal isolates tested in the chinchilla model of pneumococcal disease. A. Bacterial counts from nasopharyngeal washes obtained on day 1. B. Bacterial counts from middle ear fluid samples obtained on day 8.
doi:10.1371/journal.pone.0018649.g002

Table 2. Distribution of putative tissue-specific genes among *S. pneumoniae* isolate collections.

	No. of isolates positive for given gene (%)						
Collection	**SP0163**	**SP0395**	**SP0463**	**SP1793**	**SPN05002**	**spr0282**	**RD8a**
CC199							
Carriage (n = 7)	6 (86)	3 (43)	2 (29)	3 (43)	4 (57)	0 (0)	7 (100)
Middle Ear (n = 5)	3 (60)	2 (40)	3 (60)	3 (60)	1 (20)	0 (0)	1 (20)
Blood (n = 5)	4 (80)	1 (20)	3 (60)	4 (80)	2 (40)	1 (20)	2 (40)
Larger Collection							
Carriage (n = 92)	45 (49)	59 (64)	35 (38)	56 (61)	31 (34)	66 (72)	27 (29)
Middle Ear (n = 94)	68 (72)	43 (46)	58 (62)	53 (56)	60 (64)	50 (53)	48 (51)
Blood (n = 83)	44 (53)	56 (67)	39 (47)	52 (63)	60 (72)	46 (55)	41 (49)
CSF (n = 57)	27 (47)	44 (77)	19 (33)	46 (81)	44 (77)	34 (60)	25 (44)
Total (n = 326)	184 (56)	202 (62)	151 (46)	207 (63)	195 (60)	196 (60)	141 (43)

*RD8 was present in the CSF isolate. SP0163, SP0395, spr0282, and RD8a were present in the pleural fluid isolate.
doi:10.1371/journal.pone.0018649.t002

Table 3. Adjusted odds ratios for each of the tissue-specific genes identified by correspondence analysis.

Locus	Product	Odds Ratios				
		Middle ear vs Carriage	Blood vs Carriage	CSF vs Carriage	Disease vs Carriage	Invasive vs Non
SP0163	Transcriptional Regulator	**2.73 (1.48, 5.02)**	1.18 (0.65, 2.13)	0.94 (0.48, 1.82)	1.53 (0.94, 2.48)	0.66 (0.43, 1.04)
SP0395	Transcriptional Regulator	**0.45 (0.25, 0.82)**	1.23 (0.65, 2.30)	1.89 (0.89, 4.01)	0.88 (0.53, 1.44)	**2.18 (1.37, 3.49)**
SP0463	Cell Wall Surface Anchor Protein	**2.49 (1.38, 4.47)**	1.44 (0.79, 2.64)	0.81 (0.41, 1.63)	1.57 (0.96, 2.58)	0.72 (0.46, 1.12)
SP1793	Hypothetical Protein	0.83 (0.46, 1.49)	1.08 (0.59, 1.99)	**2.69 (1.23, 5.86)**	1.24 (0.76, 2.05)	**1.65 (1.04, 2.62)**
SPN05002	Hypothetical Protein	**3.13 (1.72, 5.67)**	**4.89 (2.57, 9.32)**	**6.35 (2.99, 13.47)**	**4.27 (2.56, 7.12)**	**3.01 (1.88, 4.85)**
spr0282	Phosphotransferase System	**0.43 (0.23, 0.78)**	**0.49 (0.26, 0.92)**	0.58 (0.29, 1.17)	**0.48 (0.29, 0.81)**	**0.51 (0.33, 0.79)**
RD8a	Region of Diversity	**2.41 (1.32, 4.39)**	**2.35 (1.26, 4.38)**	1.88 (0.94, 3.75)	**2.25 (1.34, 3.77)**	1.34 (0.86, 2.09)

Odds ratios in bold indicate those that were significant.

doi:10.1371/journal.pone.0018649.t003

region, approximately 20 kb in size, was present in the genome of the hypervirulent isolate 06-01-003MEF, but absent in CDC3059-06 and all three of the other isolates tested in the chinchilla model of pneumococcal disease. The region comprised at least 25 predicted genes, with ~65% of the sequence comprising coding sequence and a GC content of ~39%. When queried against the NCBI BLAST database, 14 of the 25 predicted genes were of unknown function, while the remainder corresponded to bacteriophage structural genes.

Discussion

The data presented in this study supports previous research that has identified *S. pneumoniae* as a genetically highly heterogeneous species, even within closely related strains [23], [26]. The ability to colonize and cause disease did not differ depending on the isolate's serotype when tested in the chinchilla model. However, isolates in the disease clade were associated with a decreased time to cause bacteremia compared to isolates in the carriage clade. One hypervirulent serotype 19A isolate was associated with a short time to bacteremia and high mortality. The genome of this strain contains a 20 kb genetic region that is absent in the other CC199 isolates tested in the chinchilla. This hypervirulent strain may have acquired a bacteriophage with novel virulence factors, although further analysis is required to identify the precise role of this sequence, if any, in the observed hypervirulence of isolate 06-01-003MEF. Taken together, these data indicate that genetic factors, other than capsule polysaccharide, modulate virulence potential in CC199. Serotype 15B/C is not included in PCV13 [32]. Given the similar genetic background and similar pathogenicity in the chinchilla, CC199 strains of serotype 15B/C may increasingly cause acute otitis media in children.

We hypothesize that CC199 strains are successful at the population level because they are a generalist lineage adept at carriage, otitis media, and invasive disease. Within CC199, isolates of both serotype 15B/C and 19A were equally able to colonize and cause otitis media in the chinchilla model. Furthermore, there was little difference in the time taken to cause invasive disease by serotype. These data expand upon a previous study that indicated similar fitness for otitis media when serotype 19A and 15B/C isolates were inoculated together in competition [34]. While isolates belonging to the carriage clade have the potential to cause disease, isolates from the disease clade were associated with faster development of invasive disease. It remains to be seen if the identified pattern of separate carriage and disease clades in CC199 is repeated for other established clones, such as CC156 or CC176.

This study has identified genes of potential importance in tissue-specific virulence, which will require additional study to assess their potential and function as virulence factors. A larger collection more representative of the pneumococcal population was screened to determine whether those associations identified in CC199 isolates would be maintained in a more genetically diverse collection. Given that *S. pneumoniae* is a highly recombinogenic species, tissue-specific associations that are preserved throughout the population as a whole would be expected to be of biological importance to the species. However, it is worth noting that for two of the genetic loci, spr0282 and RD8a, the tissue-specific associations differed in the overall pneumococcal population when compared to the CC199 data, although no statistical significance was given to the CC199 data due to small sample size. This would seem to indicate that the effect of a gene may be dependent on the combination of genes found in the remainder of the pneumococcal genome.

Previous studies have identified individual genes associated with otitis media, pneumonia, and meningitis [31], [35]–[37]. We identified RD8a as more common among our middle ear and blood isolates in comparison to carriage. Obert *et al* previously described the correlation between the region of diversity RD8a and invasive isolates [27]. SP0463 (*rrgB*) is on the pilus locus [38] and has previously been identified in a minority of invasive pneumococcal isolates [39]. Our finding that SP0463 is found significantly more often in middle ear isolates than carriage isolates seems at variance with other data that pneumococcal isolates do not contain genes encoding pili at a greater frequency in otitis media isolates than in invasive disease isolates [40]. It is likely that the disparity arises through the specific isolate collections examined or the difference in target genes used by each study. Moschioni *et al* [40] targeted entire pilus encoding islets, this study focussed on a single gene within the islet.

The specific combination of putative tissue-specific genes differed for each of the isolates tested in the chinchilla model. Furthermore, we did not identify a single genetic factor from the accessory genome, present in all strains from a specific tissue source. A supragenome pool exists for *S. pneumoniae*, which consists of the total number of genes available to the species [26], [41], [42]. A given gene may associate with *S. pneumoniae* isolates from a specific tissue-source. Furthermore, a given gene may enhance fitness for tissue-specific disease, yet an isolate lacking this gene can still cause disease. These data may be explained by high levels of recombination and functional redundancy amongst the accessory regions within the pneumococcal genome. Blomberg *et al* did not identify a unique pattern of accessory regions among invasive disease isolates and concluded that redundancy existed amongst the accessory regions required to cause invasive pneumococcal disease [32]. Our data indicate that this redundancy holds for otitis media as well as invasive disease.

One hypervirulent CC199 19A (06-01-003 MEF) strain caused rapid mortality. The bacteriophage sequence within this region may encode virulence determinants. A majority of *S. pneumoniae* strains contain lysogenic phage [43], [44]. Researchers have speculated that phage-encoded proteins modulate *S. pneumoniae* fitness [45]. Further analysis of this sequence is required to identify if this is indeed the basis for the differences in virulence observed between the isolates.

This study had limitations. Genomic diversity of CC199 was evaluated using CGH. As mentioned previously, microarrays are limited in that genetic regions that are not included on the array will be missed. Genes that are highly variable in sequence may also fail to hybridize and will be counted as absent. Additional unmeasured factors, such as level of capsule expression, may also contribute to differences in virulence. As with any animal model, our data may not be fully reflective of *S. pneumoniae* carriage and disease in humans. A small number of chinchillas were used in each experiment and the chinchilla model does not use a genetically pure line. However, the advantage to the chinchilla model is it closely mirrors pneumococcal disease aetiology where colonization is established before disease.

In summary, the approval of PCV13 by the FDA in 2010 [33] means that isolates belonging to serotype 19A should be protected against in the future. However, this expanded vaccine does not include protection against serotype 15B/C. Recent studies have indicated an increase in the prevalence of serogroup 15 isolates amongst carriage and otitis media [10], [18]. Small but significant increases in the proportion of serotype 15B/C isolates have been observed among invasive disease cases [7]. While immunization with PCV13 is likely to result in disease reduction, the similarities in terms of virulence between 19A and 15B/C isolates suggest that a vaccine based on genetic factors other than serotype is necessary, especially for otitis media and nonbacteremic pneumonia. Such genetic factors could belong to the core genome or comprise several from the accessory genome, which together cover the entire pneumococcal population. Alternatively, future vaccines could be targeted towards tissue-specific genes, allowing physicians to protect patients from invasive pneumococcal disease, while not affecting those pneumococcal isolates that exist among carriage isolates in the nasopharynx [46].

Materials and Methods

Ethics Statement

The Institutional Animal Care and Use Committee at Boston University Medical Center approved our animal care protocols as being consistent with humane treatment of laboratory animals and with standards set forth in the Guide for the Care and Use of Laboratory Animals and the Animal Welfare Act (Animal Welfare Assurance approval number A-3316-01). Thus, our study was carried out in strict accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health. All efforts were made to minimize suffering. Procedures to obtain nasopharyngeal lavages, middle ear cultures and phlebotomy were performed under Ketamine and Xylazine anaesthesia. All animals were examined daily for activity levels, fluid intake, and temperature. Early euthanasia was used when labyrinthitis was present or when animals exhibited a decrease in weight greater than 10%. Animals were euthanized using CO_2.

Bacterial Isolates, Growth Conditions, and DNA Extraction

A collection of 20 *S. pneumoniae* isolates was selected for analysis by CGH. Nineteen CC199 isolates were selected from collections previously analyzed by multilocus sequence typing (MLST) [28], [47]. Isolates were obtained from a variety of tissue sources, including the nasopharynx or throat (n = 7), middle ear fluid (n = 5), blood (n = 5), cerebrospinal fluid (CSF) (n = 1), and pleural fluid (n = 1). CC199 isolates were serotype 19A (n = 7) or 15B/C (n = 12). A serotype 3 isolate (ST180) was included as an outgroup for phylogeny construction. All isolates were obtained from the United States. With only two exceptions, strains were isolated since 2000 (isolates ME2815 and ME3582 were isolated in 1996). Isolate 3059-06 came from a patient 47 years of age and isolate 03AR0753 came from a patient of unknown age. The remaining isolates were from patients less than 48 months of age. Isolates to be analyzed by CGH were grown overnight in Todd Hewitt, 5% yeast extract at 37°C with 5% CO_2. DNA was extracted using QIAGEN 100/G genomic tips.

A larger genetically diverse collection of *S. pneumoniae* isolates [28], was screened to ascertain the prevalence of genes associated with CC199 isolates from a given tissue source. This genetically diverse collection was comprised of nasopharyngeal or throat isolates from healthy children (collectively analyzed as the "carriage group") (n = 92), middle ear (n = 94), blood (n = 83), and cerebrospinal fluid (n = 57). *S. pneumoniae* strains represented a range of serotypes and sequence types. DNA extractions for isolates to be screened by PCR involved overnight growth on trypticase soy agar plates with 5% sheep blood and inoculation of a colony into a 96-well plate containing 50 µl of Tris-EDTA buffer followed by boiling for 10 min.

DNA Labelling and Microarray Hybridization

DNA was labelled using a BioPrime® Plus Array CGH Indirect Genomic Labelling System (Invitrogen). DNA labelling and microarray hybridization was conducted by staff at the W.M. Keck Facility at Yale.

Microarray Analysis

TM4: Microarray Software Suite (TIGR) [48] and SAS v9.1 (SAS Institute, Cary, NC) were used in data analyses. LOWESS (locally weighted scatterplot smoothing) was used to normalize raw data in a single experiment, with 50% of the data being used for smoothing. The LOWESS normalization was used as part of a pipeline in MIDAS, which includes total intensity normalization, LOWESS normalization, standard deviation regularization and low intensity filtering [48]. Hybridization spots with a signal to noise ratio of less than 2.0 were removed from further analysis. Partek® Genomics Suite^{TM} 6.4 (Partek Inc., St. Louis, MO) was used to remove variation in the data due to batch effects.

Two different methods were applied to the microarray data to determine if a given gene was present or absent. The program GACK was used to convert the data from each microarray experiment to present or absent [49]. Alternatively, arbitrary cutoff values were selected. Hybridization spots with a $\log_2$ ratio above 1, 1.5, 2 or 2.5 were denoted as present, while those with a ratio below −1, −1.5, −2 or −2.5 were identified as absent. A range of cut-off value combinations were assessed.

Hierarchical clustering of the $\log_2$ ratio matrix, and visualization of the resulting dendrogram, was performed in Partek® Genomics Suite^{TM} 6.4, using Pearson's Dissimilarity to calculate row dissimilarity and Ward's method for row clustering. A phylogeny of the isolates was also constructed using Dollo parsimony based on the presence and absence matrix of genes [50] implemented in PAUP4.0b10. Hybridization spots that produced missing data for more than 20% of the isolates were removed before phylogeny construction. Both dendrograms were rooted with a serotype 3, ST180 isolate.

All CGH data generated by this study were fully annotated and deposited at the Gene Expression Omnibus (GEO; Accession numbers GSM591013–GSM591034). Data uploaded to the GEO database is MIAME compliant, representing the final data from the normalisation pipeline. Subsequent data generated via batch effect removal could not be uploaded due to file formatting issues, but is available upon request.

PCR Validation of Microarray Results

Eighteen genes, including both core and variable genes, were chosen to validate CGH results. Primer sequences were designed for each gene (Table S2). PCR involved an initial denaturation step of 95°C for 3 min; 35 cycles of 95°C for 30 sec, the relevant annealing temperature for 30 sec (Table S2), and 72°C for 30 sec; and a final extension of 72°C for 10 min.

Chinchilla Model of Pneumococcal Disease

Five pneumococcal isolates were tested in the chinchilla model of pneumococcal disease. Female chinchillas (*Chinchilla laniger*) with no prior evidence of middle ear infection were used. One serotype 15B/C and one serotype 19A isolate were selected from each clade of the CC199 phylogeny. Isolates MIB02102 and MIC17101 were serotype 15B/C, and 06-01-003MEF and FG23 were serotype 19A, from the invasive and carriage clades, respectively. A genome sequenced reference strain (isolate CDC3059-06, serotype 19A) was also included.

Isolates were grown as monocultures and used to inoculate the left nare of two to eight healthy chinchillas [34], [51]. The *S. pneumoniae* strains were allowed to establish nasopharyngeal colonization. Nasopharyngeal washes were collected from all animals on day 1 and day 5 by lavage with Hanks buffer. Barotrauma, which creates negative pressure in the middle ear cavity by aspiration of up to 250 µl of air with a 25-guage needle, was performed unilaterally on day 5 to prompt development of otitis media. Chinchillas were monitored daily by otomicroscopy and tympanometry. Once the animal developed signs of otitis media, the middle ear cavity was accessed through a small hole in the bullar bone. Nasopharyngeal and middle ear samples were collected for quantitative microbiology. Total viable counts (CFU/ml) were obtained by plating on blood agar and incubation at 37°C overnight. Total viable counts were calculated from day 1, representing the initial colonization phase, and day 8, the first middle ear fluid sample obtained.

Differences in the total viable counts for each isolate were tested for by one-way analysis of covariance (ANCOVA), with Tukey, Bonferroni and Duncan post-tests being used for multiple comparisons. Weight of the individual chinchilla and the inoculation bacterial load were controlled for during these tests. The assumption of normality was not met. However non-parametric and parametric bivariate analyses provided similar results, suggesting the ANCOVA is robust enough to handle the violation of assumptions, and also allow us to control for weight and inoculation load. An isolate's propensity to cause invasive disease was determined through the length of time between colonization and the development of bacteremia. Differences in the time to bacteremia were tested using survival analysis (Kaplan-Meier survival curves, and log rank tests). The assumption of proportionality was met.

Identification of Tissue-Specific Genes

Correspondence analysis was implemented in SAS v9.1 (SAS Institute, Cary, NC) to narrow the list of genes associated with isolates obtained from a particular tissue source (carriage, middle ear, or blood isolates). CSF and pleural fluid isolates were not included due to low numbers. For any given gene, its frequency amongst the isolates obtained from each tissue source was calculated, and used for each comparison. Correspondence analysis does not provide a measure of significance. Therefore the list of genes was verified and narrowed further by removal of genes that were equally present between tissue sources.

The larger isolate collection [29] was screened by PCR to ascertain the prevalence of genes identified as associated with a particular tissue source by correspondence analysis. Primers are detailed in Table S3. With the exception of primers to amplify *cbp*G [52], all primers were designed within this study. PCR involved an initial denaturation step of 95°C for 3 min; 35 cycles of 95°C for 30 sec, the relevant annealing temperature for 30 sec (Table S3), and 72°C for 30 sec; and a final extension of 72°C for 10 min. The presence or absence of each gene was confirmed for PCR-negative isolates by dot blot and Southern hybridization, using previously described methods [28]. Briefly, PCR amplification of TIGR4 genomic DNA was used to generate a gene-specific probe, using the gene-specific primer pairs listed in Table S3. The Gene Images AlkPhos Direct Labeling and ECF chemifluorescence detection system (Amersham Biosciences, Piscataway, N.J.) was used for labelling, hybridization, washes, and signal detection.

Tissue-Specific Gene Analysis

Statistical analyses were carried out using SAS version 9.1 (SAS Institute, Cary, NC). The distribution of *S. pneumoniae* STs, clonal groups, and tissue-specific genes in the isolate collection were described using simple descriptive statistics. Logistic regression was used to calculate odds ratios and 95% confidence intervals for each gene for the isolate populations of each tissue source versus the carriage population. In addition, statistics were calculated for disease (middle ear, blood, and CSF) versus carriage populations as well as invasive (blood and CSF) versus non-invasive (carriage and middle ear) isolate populations.

Complete Genome Sequencing

The chromosomal DNA of isolate 06-01-003MEF (CC199, serotype 19A) was sequenced using the Solexa paired-end sequencing platform (Illumina, San Diego, CA). Seventy-five bp reads were generated, resulting in an on average coverage of 16X. The paired-end sequences were trimmed based on quality score by bTrim (Kong, unpublished, http://graphics.med.yale.edu/trim/readme). Sequences that passed trimming were assembled by Velvet [53]. Different parameters of bTrim and Velvet were evaluated to optimize the final contigs with respect to n50 and the greatest contig length. This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession AEGF00000000. The version described in this paper is the first version, AEGF01000000.

The genome sequence of 06-01-003MEF was compared to the publicly available genome of isolate CDC3059-06 (CC199, serotype 19A) (accession number NZ_ABGG00000000). Sequences unique to each genome were identified via subtractive BLAST analysis on the assembled sequence data sets. Sequences found to have a matching sequence in the corresponding genome were removed from both genomes. Genes were subsequently identified in the remaining unique sequences using Glimmer3 [54]. Sequence similarity with publicly available sequences was assessed using the NCBI BLAST database (http://www.ncbi.nlm.nih.gov/BLAST).

Supporting Information

Table S1 Genes found to be variably present within CC199.
(XLS)

Table S2 Primer sequences used in the validation of CGH data.
(XLS)

Table S3 Primer sequences used to test for the presence/absence of tissue-specific genes identified by correspondence analysis.
(XLS)

Acknowledgments

We thank Irina Tikhonova and the W.M. Keck Facility for the labelling of DNA and hybridization to microarray slides. We also acknowledge the NIAID sponsored Pathogen Functional Genomics Resource Center for use of *S. pneumoniae* arrays. We thank Loc Truong, Abbie Stevenson, and Jonathan Gaskins (Boston Medical Center) for their assistance with the chinchilla model.

Author Contributions

Conceived and designed the experiments: JCT SIP MMP. Performed the experiments: JCT MF. Analyzed the data: JCT KPF ASL YK. Contributed reagents/materials/analysis tools: YK MEP SIP. Wrote the paper: JCT MMP.

References

1. Fine MJ, Smith MA, Carson CA, Mutha SS, Sankey SS, et al. (1996) Prognosis and outcomes of patients with community-acquired pneumonia. A meta-analysis. JAMA 275: 134–141.
2. Rodgers GL, Arguedas A, Cohen R, Dagan R (2009) Global serotype distribution among *Streptococcus pneumoniae* isolates causing otitis media in children: Potential implications for pneumococcal conjugate vaccines. Vaccine 27: 3802–3810.
3. Greene CM, Kyaw MH, Ray SM, Schaffner W, Lynfield R, et al. (2006) Preventability of invasive pneumococcal disease and assessment of current polysaccharide vaccine recommendations for adults: United States, 2001–2003. Clin Infect Dis 43: 141–150.
4. Park I, Pritchard DG, Cartee R, Brandao A, Brandileone MCC, et al. (2007) Discovery of a new capsular serotype (6C) within serogroup 6 of *Streptococcus pneumoniae*. J Clin Microbiol 45: 1225–33.
5. Calix JJ, Nahm MH (2010) A new pneumococcal serotype, 11E, has a variably inactivated *wcj*E gene. J Infect Dis 202: 29–38.
6. Poehling KA, Talbot TR, Griffin MR, Craig AS, Whitney CG, et al. (2006) Invasive pneumococcal disease among infants before and after introduction of pneumococcal conjugate vaccine. JAMA 295: 1668–1674.
7. Pilishvili T, Lexau C, Farley MM, Hadler J, Harrison LH, et al. (2010) Sustained reductions in invasive pneumococcal disease in the era of conjugate vaccine. J Infect Dis 2010: 201: 32–41.
8. Hsu HE, Shutt KA, Moore MR, Beall BW, Bennett NM, et al. (2009) Effect of pneumococcal conjugate vaccine on pneumococcal meningitis. N Engl J Med 360: 244–256.
9. Pelton SI, Huot H, Finkelstein JA, Bishop CJ, Hsu KK, et al. (2007) Emergence of 19A as virulent and multidrug resistant pneumococcus in Massachusetts following universal immunization of infants with pneumococcal conjugate vaccine. Pediatr Infect Dis J 26: 468–472.
10. Huang SS, Platt R, Rifas-Shiman SL, Pelton SI, Goldmann D, et al. (2005) Post-PCV7 changes in colonizing pneumococcal serotypes in 16 Massachusetts communities, 2001 and 2004. Pediatrics 116: e408–13.
11. Byington C, Samore M, Stoddard G, Barlow S, Daly J, et al. (2005) Temporal trends of invasive disease due to *Streptococcus pneumoniae* among children in the Intermountain West: Emergence of nonvaccine serogroups. Clin Infect Dis 41: 21–29.
12. Hwa Choi E, Hee Kim S, Wook Eun B, Jung Kim S, Hee Kim N, et al. (2008) *Streptococcus pneumoniae* serotype 19A in children, South Korea. Emerg Infect Dis 14: 275–281.
13. Munoz-Almagro C, Jordan I, Gene A, Latorre C, Garcia-Garcia JJ, et al. (2008) Emergence of invasive pneumococcal disease caused by nonvaccine serotypes in the era of 7-valent conjugate vaccine. Clin Infect Dis 46: 174–182.
14. van Gils EJM, Veenhoven RH, Hak E, Rodenburg GD, Keijzers WCM, et al. (2010) Pneumococcal conjugate vaccination and nasopharyngeal acquisition of pneumococcal serotype 19A strains. JAMA 304: 1099–1106.
15. Pai R, Moore MR, Pilishvili T, Gertz RE, Whitney CG, et al. (2005) Postvaccine genetic structure of *Streptococcus pneumoniae* serotype 19A from children in the United States. J Infect Dis 192: 1988–1995.
16. Gertz RE, McEllistrem MC, Boxrud DJ, Li Z, Sakota V, et al. (2003) Clonal distribution of invasive pneumococcal isolates from children and selected adults in the United States prior to the 7-valent conjugate vaccine introduction. J Clin Microbiol 41: 4194–4216.
17. Moore MR, Gertz RE Jr., Woodbury RL, Barkocy-Gallagher GA, Schaffner W, et al. (2008) Population snapshot of emergent *Streptococcus pneumoniae* serotype 19A in the United States, 2005. J Infect Dis 197: 1016–1027.
18. Casey JR, Adlowitz DG, Pichichero ME (2010) New patterns in the otopathogens causing acute otitis media six to eight years after introduction of pneumococcal conjugate vaccine. Pediatr Infect Dis J 29: 304.
19. Gonzalez BE, Hulten KG, Lamberth L, Kaplan SL, Mason EO Jr. (2006) *Streptococcus pneumoniae* serogroups 15 and 33: An increasing cause of pneumococcal infections in children in the United States after the introduction of the pneumococcal 7-valent conjugate vaccine. Pediatr Infect Dis J 25: 301.
20. Hanage WP, Kaijalainen T, Herva E, Saukkoriipi A, Syrjanen R, et al. (2005) Using multilocus sequence data to define the pneumococcus. J Bacteriol 187: 6223–6230.
21. Harboe ZB, Thomsen RW, Riis A, Valentiner-Branth P, Christensen JJ, et al. (2009) Pneumococcal serotypes and mortality following invasive pneumococcal disease: A population-based cohort study. PLoS Med 6: e1000081.
22. Brueggemann AB, Griffiths D, Meats E, Peto T, Crook DW, et al. (2003) Clonal relationships between invasive and carriage *Streptococcus pneumoniae*. J Infect Dis 187: 1424–1432.
23. Silva NA, McCluskey J, Jefferies JMC, Hinds J, Smith A, et al. (2006) Genomic diversity between strains of the same serotype and multilocus sequence type among pneumococcal clinical isolates. Infect Immun 74: 3513–3518.
24. Sandgren A, Sjostrom K, Olsson Liljequist B, Christensson B, Samuelsson A, et al. (2004) Effect of clonal and serotype specific-properties on the invasive capacity of *Streptococcus pneumoniae*. J Infect Dis 189: 785–796.
25. Brückner R, Nuhn M, Reichmann P, Weber B, Hakenbeck R (2004) Mosaic genes and mosaic chromosomes-genomic variation in *Streptococcus pneumoniae*. Int J Med Microbiol 294: 157–168.
26. Hiller NL, Janto B, Hogg JS, Boissy R, Yu S, et al. (2007) Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: Insights into the pneumococcal supragenome. J Bacteriol 189(22): 8186–8195.
27. Obert C, Sublett J, Kaushal D, Hinojosa E, Barton T, 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.
28. Pettigrew MM, Fennie KP, York MP, Daniels J, Ghaffar F (2006) Variation in the presence of neuraminidase genes among *Streptococcus pneumoniae* isolates with identical sequence types. Infect Immun 74: 3360–3365.
29. LeMessurier KS, Ogunniyi AD, Paton JC (2006) Differential expression of key pneumococcal virulence genes *in vivo*. Microbiology 152: 305–311.
30. Mahdi LK, Ogunniyi AD, LeMessurier KS, Paton JC (2008) Pneumococcal virulence gene expression and host cytokine profiles during pathogenesis of invasive disease. Infect Immun 76: 646.
31. Orihuela CJ, Gao G, Francis KP, Tuomanen EI (2004) Tissue-specific contributions of pneumococcal virulence factors to pathogenesis. J Infect Dis 190: 1661–1669.
32. Blomberg C, Dagerhamn J, Dahlberg S, Browall S, Fernebro J, et al. (2009) Pattern of accessory regions and invasive disease potential in *Streptococcus pneumoniae*. J Infect Dis 199: 1032–1042.
33. Complete PCV (2010) Licensure of a 13-valent pneumococcal conjugate vaccine (PCV13) and recommendations for use among children—advisory committee on immunization practices (ACIP), 2010. MMWR Morb Mortal Wkly Rep 59: 258–261.
34. Laufer AS, Thomas JC, Figueira M, Gent JF, Pelton SI, et al. (2010) Capacity of serotype 19A and 15B/C *Streptococcus pneumoniae* isolates for experimental otitis media: Implications for the conjugate vaccine. Vaccine 28: 2450–7.
35. Chen H, Ma Y, Yang J, O'Brien CJ, Lee SL, et al. (2007) Genetic requirement for pneumococcal ear infection. PLoS ONE 3: e2950.
36. Pettigrew MM, Fennie KP (2005) Genomic subtraction followed by dot blot screening of *Streptococcus pneumoniae* clinical and carriage isolates identifies genetic differences associated with strains that cause otitis media. Infect Immun 73: 2805–2811.
37. Oggioni MR, Memmi G, Maggi T, Chiavolini D, Iannelli F, et al. (2003) Pneumococcal zinc metalloprotease *Zmp*C cleaves human metalloproteinase 9 and is a virulence factor in experimental pneumonia. Mol Microbiol 49: 795–805.
38. LeMieux J, Hava DL, Basset A, Camilli A (2006) RrgA and RrgB are components of a multisubunit pilus encoded by the *Streptococcus pneumoniae* rlrA pathogenicity islet. Infect Immun 74: 2453–2456.
39. Aguiar SI, Serrano I, Pinto FR, Melo-Cristino J, Ramirez M (2008) The presence of the pilus locus is a clonal property among pneumococcal invasive isolates. BMC Microbiol 8: 41.
40. Moschioni M, De Angelis G, Melchiorre S, Masignani V, Leibovitz E, et al. (2010) Prevalence of pilus-encoding islets among acute otitis media *Streptococcus*

pneumoniae isolates from Israel. Clin Microbiol Infect;doi: 10.1111/j.1469-0691.2010.03105.x.

41. Ehrlich GD, Hu FZ, Shen K, Stoodley P, Post JC (2005) Bacterial plurality as a general mechanism driving persistence in chronic infections. Clin Orthop 437: 20–4.
42. Feil EJ, Smith JM, Enright MC, Spratt BG (2000) Estimating recombinational parameters in *Streptococcus pneumoniae* from multilocus sequence typing data. Genetics 154: 1439–1450.
43. Romero P, Croucher NJ, Hiller NL, Hu FZ, Ehrlich GD, et al. (2009) Comparative genomic analysis of ten *Streptococcus pneumoniae* temperate bacteriophages. J Bacteriol 191: 4854–4862.
44. Romero P, Garcia E, Mitchell TJ (2009) Development of a prophage typing system and analysis of prophage carriage in *Streptococcus pneumoniae*. Appl Environ Microbiol 75: 1642–1649.
45. Loeffler JM, Fischetti VA (2006) Lysogeny of *Streptococcus pneumoniae* with MM1 phage: Improved adherence and other phenotypic changes. Infect Immun 74: 4486–4495.
46. Glover DT, Hollingshead SK, Briles DE (2008) *Streptococcus pneumoniae* surface protein PcpA elicits protection against lung infection and fatal sepsis. Infect Immun 76: 2767–2776.
47. Xu Q, Pichichero ME, Casey JR, Zeng M (2009) Novel type of *Streptococcus pneumoniae* causing multidrug-resistant acute otitis media in children. Emerg Infect Dis 15: 547.
48. Sioson AA, Mane SP, Li P, Sha W, Heath LS, et al. (2006) The statistics of identifying differentially expressed genes in expresso and TM4: A comparison. BMC Bioinformatics 7: 215.
49. Kim CC, Joyce EA, Chan K, Falkow S (2002) Improved analytical methods for microarray-based genome-composition analysis. Genome Biol 3: 0065.1–0065.17.
50. Farris JS (1977) Phylogenetic analysis under dollo's law. Syst Zool 26: 77–88.
51. Sabharwal V, Ram S, Figueira M, Park IH, Pelton SI (2009) Role of complement in host defense against pneumococcal otitis media. Infect Immun 77: 1121.
52. Mann B, Orihuela C, Antikainen J, Gao G, Sublett J, et al. (2006) Multifunctional role of choline binding protein G in pneumococcal pathogenesis. Infect Immun 74: 821–829.
53. Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de bruijn graphs. Genome Res 18: 821.
54. Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) Identifying bacterial genes and endosymbiont DNA with glimmer. Bioinformatics 23: 673.