Development of a multi-locus sequence typing scheme for the molecular typing of *Mycoplasma pneumoniae*.

Rebecca J. Brown[a,b], Matthew T.G. Holden[c], O. Brad Spiller[a] and Victoria J. Chalker[b]#

Institutions:

[a] Cardiff University, School of Medicine, Department of Child Health, University Hospital of Wales, Cardiff, UK; [b] Vaccine Preventable Bacteria Reference Unit, Public Health England, London, UK; [c] University of St Andrews, School of Medicine, Medical & Biological Sciences, North Haugh, St Andrews, UK.

Running title: *Mycoplasma pneumoniae* MLST

#Corresponding Author: Dr. Victoria Chalker, Vaccine Preventable Bacteria Reference Unit, Public Health England, London, UK

Phone: +44 (0)20 8327 6636 e-mail: vicki.chalker@phe.gov.uk

Abstract

Mycoplasma pneumoniae is a major human respiratory pathogen causing both upper and lower respiratory disease in humans of all ages, and can also result in other serious extra-pulmonary sequelae. A multi-locus sequence typing (MLST) scheme for *M. pneumoniae* was developed, based on the sequence of eight housekeeping genes (*ppa*, *pgm*, *gyrB*, *gmk*, *glyA*, *atpA*, *arcC*, and *adk*) and applied to 55 *M. pneumoniae* clinical isolates and the two type strains M129 and FH. A total of 12 sequence types (STs) resulted for 57 *M. pneumoniae* isolates tested; with a discriminatory index of 0.21 STs per isolate. The MLST loci used in this scheme were shown to be stable in ten strains following ten sequential sub-culture passages. Phylogenetic analysis of concatenated sequences of the eight loci indicated two distinct genetic clusters which could be directly linked to multi-locus variable-number tandem repeat analysis (MLVA) type. Genetic MLST clustering was confirmed by genomic sequence analysis, indicating that the MLST scheme developed in this study is representative of the genome. Furthermore, this MLST scheme was shown to be more discriminatory than both MLVA and P1 typing for the *M. pneumoniae* isolates examined, providing a method for further and more detailed analysis of observed epidemic peaks of *M. pneumoniae* infection. This scheme is supported by a public web-based database (**http://pubmlst.org/mpneumoniae**).

Introduction

Mycoplasma pneumoniae is a common cause of community-acquired pneumonia (CAP) transmitted by aerosol or close contact (1). *M. pneumoniae* may cause other serious extra-pulmonary sequelae such as encephalitis (2). The pathogen is found in all age groups, with higher prevalence in children aged 5-14 years (3, 4). Admissions to a UK hospital in patients with CAP that were attributed to *M. pneumoniae* were estimated at 18% in 1982 and 4% in 1999 (5). Major increases and decreases in *M. pneumoniae* infection have occurred periodically in the United Kingdom; historically, epidemics have occurred at approximately four yearly intervals and have lasted 12-15 months, concurrent with sporadic infection at a lower level and seasonal peaks December to February (4, 6). However, globally, peaks of infection have been observed in either summer or autumn, with no obvious explanation for this seasonal variation (7-10).

Typing of clinical isolates by molecular methods is of importance for the understanding of the epidemiology of *M. pneumoniae* infection and for analysis of endemic outbreaks. It is generally considered that molecular typing of *M. pneumoniae* is hampered by the fact that the pathogen is a genetically homologous species (11). Initial molecular typing targeted the gene encoding the major surface protein (P1) of *M. pneumoniae*. PCR-restriction fragment length polymorphism (RFLP) analysis of the P1 gene, encoding a major adhesion, is the most common genotyping method. This enables the separation of isolates into two types, type 1 and 2 (11-13). More recent studies utilise the repetitive regions, RepMp2/3 and RepMp4 which can be found in the P1 gene, for molecular typing and have resulted in the identification of an additional subtype and three variants of these subtypes (14, 15). Multi-locus variable-number tandem-repeat (VNTR) analysis (MLVA) has also been used, based on the variation in the copy number of tandem repeated sequences, called VNTRs, found at different loci across the genome. The variation of the copy

number of these tandem repeats (TRs) depends on the isolate tested. Initially, 265 strains were grouped into twenty-six MLVA types, based on five VNTR loci (Mpn1, Mpn13-16) and an additional 18 novel types have since been reported (16-18). However, locus Mpn1 is unstable in both clinical strains and in laboratory passages, and most of the novel types came from variations in Mpn1, therefore there is international consensus that this locus should be removed from the typing scheme (19).

Multi-locus sequence typing (MLST) was previously attempted for the molecular typing of *M. pneumoniae* however, due to the homogeneity of the *M. pneumoniae* species, very little polymorphism was found in the housekeeping genes examined and it was previously concluded that the use of an MLST with housekeeping and structural genes was not useful for molecular typing (20). Only three housekeeping genes were thoroughly examined for polymorphisms across 30 isolates of either P1 type 1, 2, or a variant strain. The other genes selected for analysis were examined against a single representative strain from each subtype. In this study an MLST scheme was developed with a high discriminatory ability to differentiate *M. pneumoniae* isolates based on sequence polymorphisms within eight housekeeping genes, improving on all published typing methods for *M. pneumoniae*.

Materials and Methods

***Mycoplasma pneumoniae* strains, culture conditions and sample preparation** The strains analysed in this study are listed in Table 1. Fifty five *M. pneumoniae* strains were submitted to Public Health England, UK for clinical diagnostic purposes and the two *M. pneumoniae* type strains, FH (NCTC 10119; ATCC 15531) and M129 (ATCC 29342) were obtained from National Collection of Type Cultures (NCTC; held by Public Health England). All strains were

78 triple cloned on Mycoplasma Agar (Mycoplasma Experience; Surrey, UK) and confirmed as *M.*
79 *pneumonia*e by amplification of *p1* gene (21).

80 All strains were subsequently cultured in Mycoplasma Liquid Medium (MLM; Mycoplasma
81 Experience; Surrey, UK). For genomic sequencing, strains were grown in 100 ml broth culture
82 and the genomic DNA was extracted using the GenElute^{TM} Bacterial Genomic DNA Kit (Sigma;
83 Dorset, UK). PCR amplification was performed on bacterial DNA from a 500 µl, four day
84 culture that was released by boiling lysis (95°C for 10 minutes) following centrifugation at
85 17000 *xg* for 10 minutes, removal of all MLM, and re-suspension in 50 µl sterile water.

86 **Multi-locus Sequence Typing** Housekeeping genes considered conserved in other bacterial
87 species under a low rate of selective pressure were chosen for analysis (Table 2). Locus
88 sequences were selected using the available genome sequences of *M. pneumoniae* FH and M129
89 (FH: NC_017504.1; M129: NC_000912.1) and available whole genome sequence of 35 clinical
90 isolates. Ten genes were included for initial analysis: recA protein (*recA*), inorganic phosphatase
91 (*ppa*), phosphoglycerate mutase (*pgm*), DNA gyrase subunit B (*gyrB*), guanylate kinase (*gmk*),
92 serine hydroxymethyltransferase (*glyA*), elongation factor P (*efp*), ATP synthase subunit α
93 (*atpA*), carbamate kinase (*arcC*), and adenylate kinase (*adk*); however, *recA* and *efp* were
94 excluded from the resulting MLST scheme. Locus regions for PCR amplification were selected
95 based on areas of the CDS containing nucleotide polymorphisms.

96 PCR utilising the primers listed in Table 3 were used to amplify the target genes from a further
97 20 *M. pneumoniae* clinical isolates. Amplification of each of the locus sequences were
98 performed in a DNA thermocycler (Techne Prime; Stone, UK) in 50 µl reactions containing:
99 1 x GoTaq Fexi Buffer (Promega; Southampton, UK), 1.5 mM $MgCl_2$, 0.2 mM deoxynucleoside

triphosphates, 0.5 pmol/µl of each primer, 1.56 units of GoTaq DNA Polymerase (Promega; Southampton, UK), and 2.5 µl template DNA. PCR reactions consisted of an initial denaturation step of three minutes at 94°C, followed by 35 cycles of 60 seconds at 94°C, 60 seconds at 60°C and 60 seconds at 72°C. A final extension step was maintained for 10 minutes at 72°C. Primer sequences and PCR product sizes are shown in Table 3. The PCR products were analysed on 1.5% agarose gels with ethidium bromide visualisation. All PCR reactions were performed in duplicate.

PCR amplicons were purified using a Qiagen MiniPrep kit (Qiagen Inc.; Hilden, Germany) as per manufacturer's instructions and sequenced using the amplification primers, performed by MWG Eurofins (Ebersberg, Germany). The sequences obtained from each corresponding forward and reverse primer were assembled and trimmed for double-stranded, high quality sequence. All the sequences obtained for each locus were aligned using ClustalW (Vector NTI; Paisley, UK) and different allelic types (AT; sequences with at least a one-nucleotide difference) were assigned sequential numbers. The combination of the eight alleles determined a strain's allelic profile, and each unique allelic profile was designated a unique sequence type (ST). Open-reading frame amino acid sequences were identified using Expasy translation tool (*mycoplasma* setting; web.expasy.org/translate/) for each AT. Deduced amino acid sequences were aligned using ClustalW (Vector NTI; Paisley, UK) for each locus and synonymous changes were identified.

MLVA and P1-typing MLVA type was determined as described by Dégrange *et al*. (16), excluding the VNTR locus Mpn-1 and using international nomenclature consensus (19). P1 type was determined as described by Dumke *et al.* (15).

Genomic sequencing Genomic sequence data for 35 isolates was obtained using the Illumina Nextera XT sample prep kit (Illumina; Cambridge, UK) and sequenced on an Illumina HiSeq 2500 platform with TruSeq Rapid SBS kits (200 cycles; Illumina) and cBOT for cluster generation (Illumina). Fastq reads were trimmed using trimmomatic 0.32 with the parameters: LEADING: 30; TRAILING: 30; SLIDINGWINDOW: 10:30; MINLEN: 50 (20). Illumina reads were assembled to the M129 type strain (NC_000912.1) using SPAdes version 2.5.0 (21) and mapped to M129 using Genious® version 8.0.4. Sequencing yielded at least one contig of between 99,047 bp and 324,397 bp with homology to M129 type strain (NC_000912.1) passing quality and coverage checks. Identity as *M. pneumoniae* from genomic data was confirmed with 16S rRNA sequence analysis. Illumina reads for all the isolates were mapped against the reference chromosome M129 (EMBL accession code U00089) using SMALT (http://www.sanger.ac.uk/resources/software/smalt/) in order to identify SNPs as previously described (22). Regions of recombination in the whole chromosomes of the isolates were analysed for using Genealogies Unbiased By recomBinations In Nucleotide Sequences (GUBBINS) (23).

Phylogenetic analysis The locus sequences corresponding to each strain were concatenated head-to-tail for diversity analysis. Sequence analyses and tree construction were performed using MEGA 6.0. Neighbour-joining trees were constructed for each individual locus and concatenated sequences using Kimura's two-parameter model (26, 27). Maximum-likelihood trees were constructed for each individual locus using the Jukes-Cantor model of sequence evolution (28). Maximum-likelihood trees were constructed from concatenated sequences of the eight MLST loci using the generalised time-reversible (GTR) model of sequence evolution with uniform rates of variation (29). Bootstrap analyses with 1000 replicates were performed for every phylogenetic

tree (30). Relatedness between STs was analysed based on allelic profiles using eBURST version3. Maximum-likelihood trees were constructed from genomic sequences after the removal of areas of recombination. In total 1854 SNP sites were identified in comparison to the M129 reference chromosome. Three regions were predicted to contain SNP sites that had arisen by recombination, and these contained 28 SNP sites.

Results

MLST of *M. pneumoniae* Initial examination of ten gene targets in the two type strains M129 and FH and genomic sequence from 35 *M. pneumoniae* clinical isolates identified variation, SNP differences, in eight out of the ten genes. Genes *recA* and *efp* were 100% conserved in all sequences analysed and were therefore excluded from the MLST scheme. Genomic sequence analysis and additional PCR and sequencing of a further 20 clinical isolates of all eight targets resolved a total of 12 STs. The discriminatory typing ability for *M. pneumoniae* was 0.21 ST per isolate. The number of SNPs observed within each individual locus and the percentage of polymorphic sites are indicated in Table 3, with *pgm* having the highest number of SNPs (10 SNPs) and the highest percentage of polymorphic sites corrected for sequence length (0.93%). The number of alleles per locus ranged from two (*ppa, gyrB, gmk* and *arc*) to four (atpA) (Table 3). Examination of the Hunter-Gaston diversity index (DI; which ranges from 0.0 = no diversity to 1.0 = complete diversity) indicated moderate diversity between the STs (DI: 0.784; 95% CI: 0.716-0.852) with the greatest individual diversity shown in *pgm* (DI: 0.620; 95% CI: 0.566-0.674) and the lowest diversity in *arcC* (DI: 0.069; 95% CI: 0.000-0.158).

Neighbour-joining and maximum-likelihood trees constructed from concatenated sequences of the eight loci for the 57 *M. pneumoniae* isolates (Figure 1) illustrated two genetically distinct

clusters which were confirmed by eBURST examination of relatedness (Figure 2). The two clusters, clonal complexes (CC) designated CC1 and CC2, contained ST1, ST3, ST5, ST9 and ST11, and ST2, ST4, ST6, ST7, ST8 and ST10, respectively. ST12 located distal to the two main clusters, however, phylogenetic analysis revealed closer positioning to CC1. Neighbour-joining and maximum-likelihood trees were constructed for the eight loci individually (data not shown) and topology of both neighbour-joining and maximum-likelihood trees was consistent for all loci and concatenated sequences.

Five homogenous strains (MPN13-MPN17) originating from nose and throat swabs of the same patient with Stevens-Johnson syndrome had identical STs (ST3). Additionally, two clinical isolates (MPN104 and MPN106) originating from separate sputum samples from a patient with bronchopneumonia taken four days apart also had identical STs (ST4). This indicates a single, clonal population responsible for infection in these cases.

The possibility of synonymous sequence changes (indicating a pressure to conserve amino acid sequence and protein structure) was investigated by comparing predicted translated sequences for each locus. Analysis of deduced amino acid sequences of the eight loci for the 57 strains indicated that both synonymous and non-synonymous SNPs occurred of which approximately 44% resulted in an amino acid change. Non-synonymous SNPs are highlighted in Figure A2. Amino acid sequences for ArcC, Gmk and GyrB yielded homologous sequences for all ATs, numbering at two ATs for each locus. In comparison, Pgm analysis resulted in the largest number of non-synonymous changes in amino acid sequence, with four changes in the sequence between three ATs.

The MLST scheme was applied to the published complete genome sequences of *M. pneumoniae* available from NCBI: 309 (NC_016807.1), M129-B7 (CP003913.2), M29 (NZ_CP008895.1), PO1 (GCA_000319655.1), PI 1428 (GCA_000319675.1) and 19294 (GCA_000387745.1). These strains were determined as ST2, ST1, ST3, ST2, ST1 and ST7, respectively.

The stability of each MLST locus was assessed in ten *M. pneumoniae* isolates. Isolates were re-typed following short-term passage (ten sequential sub-culture passages) in liquid medium. All loci were found to be completely stable, with no SNPs in comparison to the original isolate.

Genomic sequence analysis Three regions of SNPs were predicted to have arisen by homologous recombination in the chromosomes of the 35 clinical *M. pneumoniae* isolates (Figure 3); one of which distinguished the genomic clade (GC) GC1 from GC2; and the other two occur within GC1. Area one was predicted to occur in all strains in GC1, area two in ten strains, and a single strain MPN113 had a single additional predicted area of recombination, area 3. Following removal of predicted areas of recombination two distinct genetic clades were identified, GC1 and GC2 (Figure 3). Excellent parity was found using this method and concatenated MLST sequences with all strains co-locating to the corresponding CC and GC.

Comparison to other typing methods There was no obvious link between the MLST ST and the year when the strains were collected, the patient's age and the sample origin; however, limited numbers of strains were available per year and for some years there were no strains. Indeed, multiple STs can be observed in a single year. Furthermore, MLST ST was unrelated to P1 type, with multiple P1 types observed within a ST (Table 1). However, in the two most common STs, the majority strains were P1 type 2 and P1 type 1 for ST2 and ST3 respectively. In comparison, this MLST scheme was more comparable to MLVA typing. The two major clusters

observed, CC1 and CC2, could be directly linked to MLVA type; CC1 contained MLVA type 4572 whereas CC2 contained MLVA types 3662 and 3562. Each ST contained only one MLVA type with the exception of ST2 which contained both 3662 and 3562 and ST11 which contained 4572, 3662 and 3562 (Table 1). Distribution of MLVA type, P1 type and MLST ST can be observed in Figure 4, indicating that P1 type 1, MLST ST2 and MLVA types 3662 and 4572 were the most frequently occurring in the isolates tested.

In the isolates tested in this scheme, MLST was deemed to be more discriminatory than both MLVA typing and P1 typing; resulting in 0.21, 0.05 and 0.07 types per isolate, respectively. This was confirmed by examination of Hunter-Gaston DI indicating larger discriminatory ability for the MLST scheme (DI: 0.784; 95% CI: 0.716-0.852) than the current MLVA scheme (DI: 0.633; 95% CI: 0.583-0.683) and P1 typing (DI: 0.551; 95% CI: 0.485-0.617).

Online database A *Mycoplasma pneumoniae* MLST online database was created for both MLST allele and profile definitions and isolate data (31); website **http://pubmlst.org/mpneumoniae**

Discussion

MLST has been used to genotype several species of bacteria, including several *mycoplasma* species; *Mycoplasma agalactiae, Mycoplasma bovis* and *Mycoplasma hyorhinis* (32-34). This study has described the successful development of a novel *M. pneumoniae* MLST scheme to allow the characterisation of clinical isolates. This scheme was successfully used to discriminate 55 clinical isolates of *M. pneumoniae* from British patients (with the exception of two USA isolates) within the reference laboratory collection, from respiratory and extra-pulmonary sites and the two type strains M129 and FH. Eight housekeeping genes were identified as suitable

targets for the scheme and these were used to genotype *M. pneumoniae* isolates by either PCR followed by sequencing or whole genome sequence analysis. *gyrB* contains a quinolone resistance-determining region (QRDR) with documented *in vitro* mutations at amino acid positions 443, 464 and 483. Clinical use of quinolones may increase selective pressure *in vivo* resulting in a high mutation rate (35). However, the *gyrB* locus sequence amplified in this MLST scheme is in a different region of the gene from the QRDR and is therefore considered a suitable MLST target. The stability of the eight loci was evaluated *in vitro* and was confirmed before and after ten repeated passages of ten strains in liquid medium. However, stability over a larger number of passages in liquid medium and evaluation of stability using *in vitro* tissue culture was not assessed.

The discriminatory power of this MLST scheme with the eight loci was 0.784 for the collection of 57 isolates. In comparison, the Hunter-Gaston DI of the P1-typing method for the 57 isolates was 0.551 and the DI of the MLVA scheme was 0.633; therefore this MLST scheme was more discriminatory for the isolates tested. However, it has previously been shown that the established MLVA method is more discriminatory than P1 typing (16), confirmed in this study. The allelic diversity of each of the MLST loci varied significantly at each locus, with the *pgm*, *glyA*, *atpA*, *gyrB*, *gmk* and *ppa* loci being more discriminatory than the *adk* and *arcC* loci. The association of this set of markers with variable Hunter-Gaston DIs makes this MLST, in theory, more optimal for epidemiological studies than other existing methodologies.

Analysis of *M. pneumoniae* infection at an individual patient level was possible using this scheme. Multiple clinical isolates were available from two of 50 patients: five from a patient with Stevens-Johnson syndrome (MPN013-MPN017) and two from a patient with bronchopneumonia taken four days apart. In both cases the MLST ST, MLVA type and P1 type

255 remained the same, indicating a single clonal isolate was responsible for infection. Recurring or
256 re-infection of *M. pneumoniae* could be determined using this scheme. Recurring infection
257 would have the same ST as the original infection whereas re-infection with *M. pneumoniae*
258 would likely be a different ST. Genetic MLST instability in isolates could occur however, in this
259 study this was not seen over ten passages.

260 The eBURST analysis illustrates the relationship of STs on the basis on the number of MLST
261 loci that differ between two STs. Analysis of this population modelling indicates that the two
262 clusters, CC1 and CC2, differed by more than one locus, but within each cluster the STs did not
263 differ by more than one locus. Within a cluster, this highlights the homogenous nature of the
264 *M. pneumoniae* species, however a definitive split can be observed between the two clusters in
265 both MLST ST and MLVA type. A possible divergent clade with ST12 from CC1 is also
266 apparent, however more isolates need to be typed by this method to confirm this observation.
267 Few typing methods have previously been able to detect significant differences between strains,
268 including one previous attempt to subtype *M. pneumoniae* by MLST with housekeeping and
269 structural genes (12, 15, 22). The previous MLST was determined to be not sufficiently
270 discriminatory to be used for epidemiological purposes. However, the MLST scheme developed
271 in this study was able to discriminate between *M. pneumoniae* isolates and resulted in two
272 genetically distinct clusters, indicating significant differences between strains.

273 Comparison between genomic sequence analysis after the removal of predicated areas of
274 recombination and phylogenetic analysis of concatenated MLST sequences showed similar
275 topology and the same distinct genetic clustering. This indicates that this MLST scheme is
276 representative of the genome and confirms *M. pneumoniae* can be subdivided into two distinct
277 genetic lineages.

Typing of clinical *M. pneumoniae* isolates is becoming increasingly important due to the global increase in *M. pneumoniae* infections and the increase in macrolide-resistant strains (36, 37). This scheme provides a more discriminatory method than both the MLVA and P1 typing methods currently in use, allowing further and more detailed analysis of observed epidemic peaks of *M. pneumoniae* infection. Community outbreaks of pneumonia caused by *M. pneumoniae* have been described worldwide (38-40), and it would be interesting to evaluate this MLST scheme in such epidemic situations. The level of discrimination of this typing method and usefulness in epidemic analysis should be confirmed by comparing outbreak-related strains to a set of control strains that were isolated from a similar time period and geographical area but that are not epidemiologically related. More severe or adverse infections with *M. pneumoniae* are seen in some patients. The reason for this is not clear however, it can be postulated that this is due to specific microbe pathogenicity (identified through genetic markers) or variance in host susceptibility. This method could assist in determining if this is a strain specific phenomenon. One advantage of MLST is that it is PCR based and does not require the growth of bacteria, which can be a lengthy process for *M. pneumoniae* and it does not limit investigation through requirement of specialist methodology. However, there is a large amount of sequencing required for this method which can be laborious and expensive; therefore, adaptation for wide-spread use directly on clinical specimens would be beneficial.

In conclusion, this study presents a robust MLST scheme that has proven discriminatory for *M. pneumoniae*, providing isolate characterisation and a higher level of discrimination than MLVA and P1-typing methods. In addition, phylogenetic analysis of both MLST STs and whole genome sequence data revealed two genetically distinct clusters. Crucially, this scheme for

300 *M. pneumoniae* is also supported by a public web-based database
301 (**http://pubmlst.org/mpneumoniae**).

302 **References**

303 1. **Waites K, Talkington D.** 2004. Mycoplasma pneumoniae and its role as a human
304 pathogen. Clinical Microbiology Reviews **17**:697-728.
305 2. **Meyer Sauteur P, Jacobs B, Spuesens E, Jacobs E, Nadal D, Vink C, van Rossum A.**
306 2014. Antibody responses to Mycoplasma pneumoniae: role in pathogenesis and
307 diagnosis of Encephalitis? PLOS Pathogens **10**:e1003983.
308 3. **Polkowska A, Harjunpaa A, Toikkanen S, Lappalainen M, Vuento R, Vuorinen T,**
309 **Kauppinen J, Flinck H, Lyytikainen O.** 2012. Increased incidence of Mycoplasma
310 pneumoniae infection in Finland, 2010-2011. Euro surveillance : bulletin Europeen sur
311 les maladies transmissibles = European communicable disease bulletin **17**.
312 4. **Chalker VJ, Stocki T, Mentasti M, Fleming D, Sadler C, Ellis J, Bermingham A,**
313 **Harrison TG.** 2011. Mycoplasma pneumoniae infection in primary care investigated by
314 real-time PCR in England and Wales. European journal of clinical microbiology &
315 infectious diseases : official publication of the European Society of Clinical
316 Microbiology **30:**915-921.
317 5. **Howard LS, Sillis M, Pasteur MC, Kamath AV, Harrison BD.** 2005. Microbiological
318 profile of community-acquired pneumonia in adults over the last 20 years. The Journal of
319 infection **50:**107-113.
320 6. **Chalker V, Stocki T, Litt D, Bermingham A, Watson J, Fleming D, Harrison T.**
321 2012. Increased detection of Mycoplasma pneumoniae infection in children in England

and Wales, October 2011 to January 2012. Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin **17**.

7. **Jacobs E.** 2012. Mycoplasma pneumoniae: now in the focus of clinicians and epidemiologists. Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin **17**.

8. **Liu J, Ai H, Xiong Y, Li F, Wen Z, Liu W, Li T, Qin K, Wu J, Liu Y.** 2015. Prevalence and correlation of infectious agents in hospitalized children with acute respiratory tract infections in central china. PloS one **10:**e0119170.

9. **Rastawicki W, Kaluzewski S, Jagielski M, Gierczyski R.** 1998. Epidemiology of Mycoplasma pneumoniae infections in Poland : 28 years of surveillance in Warsaw 1970-1997. Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin **3:**99-100.

10. **Tjhie JH, Dorigo-Zetsma JW, Roosendaal R, Van Den Brule AJ, Bestebroer TM, Bartelds AI, Vandenbroucke-Grauls CM.** 2000. Chlamydia pneumoniae and Mycoplasma pneumoniae in children with acute respiratory infection in general practices in The Netherlands. Scandinavian journal of infectious diseases **32:**13-17.

11. **Cousin-Allery A, Charron A, de Barbeyrac B, Fremy G, Skov Jensen J, Renaudin H, Bebear C.** 2000. Molecular typing of Mycoplasma pneumoniae strains by PCR-based methods and pulsed-field gel electrophoresis. Application to French and Danish isolates. Epidemiology and infection **124:**103-111.

12. **Dorigo-Zetsma JW, Dankert J, Zaat SA.** 2000. Genotyping of Mycoplasma pneumoniae clinical isolates reveals eight P1 subtypes within two genomic groups. Journal of clinical microbiology **38:**965-970.

13. **Sasaki T, Kenri T, Okazaki N, Iseki M, Yamashita R, Shintani M, Sasaki Y, Yayoshi M.** 1996. Epidemiological study of Mycoplasma pneumoniae infections in japan based on PCR-restriction fragment length polymorphism of the P1 cytadhesin gene. Journal of clinical microbiology **34:**447-449.

14. **Dumke R, Von Baum H, Luck PC, Jacobs E.** 2010. Subtypes and variants of Mycoplasma pneumoniae: local and temporal changes in Germany 2003-2006 and absence of a correlation between the genotype in the respiratory tract and the occurrence of genotype-specific antibodies in the sera of infected patients. Epidemiology and infection **138:**1829-1837.

15. **Dumke R, Luck PC, Noppen C, Schaefer C, von Baum H, Marre R, Jacobs E.** 2006. Culture-independent molecular subtyping of Mycoplasma pneumoniae in clinical samples. Journal of clinical microbiology **44:**2567-2570.

16. **Degrange S, Cazanave C, Charron A, Renaudin H, Bebear C, Bebear CM.** 2009. Development of multiple-locus variable-number tandem-repeat analysis for molecular typing of Mycoplasma pneumoniae. Journal of clinical microbiology **47:**914-923.

17. **Dumke R, Jacobs E.** 2011. Culture-independent multi-locus variable-number tandem-repeat analysis (MLVA) of Mycoplasma pneumoniae. Journal of microbiological methods **86:**393-396.

18. **Zhao F, Liu G, Cao B, Wu J, Gu Y, He L, Meng F, Zhu L, Yin Y, Lv M, Zhang J.** 2013. Multiple-locus variable-number tandem-repeat analysis of 201 Mycoplasma pneumoniae isolates from Beijing, China, from 2008 to 2011. Journal of clinical microbiology **51:**636-639.

19. **Chalker VJ, Pereyre S, Dumke R, Winchell J, Khosla P, Sun H, Yan C, Vink C, Bébéar C, ESGMI.** 2015. International Mycoplasma pneumoniae typing study: the interpretation of Mycoplasma pneumoniae multilocus variable-number tandem-repeat analysis. New Microbes and New Infections doi:10.1016/j.nmni.2015.05.005

20. **Bolger AM, Lohse M, Usadel B.** 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30:**2114-2120.

21. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner.** 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Jounal of Computational Biology **19:**455-477.

22. **Dumke R, Catrein I, Pirkil E, Herrmann R, Jacobs E.** 2003. Subtyping of Mycoplasma pneumoniae isolates based on extended genome sequencing and on expression profiles. International journal of medical microbiology : IJMM **292:**513-525.

23. **Pitcher D, Chalker VJ, Sheppard C, George RC, Harrison TG.** 2006. Real-time detection of Mycoplasma pneumoniae in respiratory samples with an internal processing control. Journal of medical microbiology **55:**149-155.

24. **Hsu LY, Harris S, Chlebowicz M, Lindsay J, Koh TH, Kristnan P, Tan TY, Hon PY, Grubb W, Bentley S, Parkhill J, Peacock S, Holden M.** 2015. Evolutionary dynamics of methicillin-resistant Staphylococcus aureus within a healthcare system. Genome biology **16:**81.

25. **Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR.** 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic acids research **43:**e15.

26. **Saitou N, Nei M.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution **4:**406-425.

27. **Kimura M.** 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of molecular evolution **16:**111-120.

28. **Jukes T, Cantor C.** 1969. Evolution of protein molecules, p. 21-132. *In* Munro H (ed.), Mammalian Protein Metabolism. Academic Press, New York.

29. **Nei M, Kumar S.** 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New Yprk.

30. **Felsenstein J.** 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution **39:**783-791.

31. **Jolley KA, Maiden MC.** 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC bioinformatics **11:**595.

32. **Manso-Silvan L, Dupuy V, Lysnyansky I, Ozdemir U, Thiaucourt F.** 2012. Phylogeny and molecular typing of Mycoplasma agalactiae and Mycoplasma bovis by multilocus sequencing. Veterinary microbiology **161:**104-112.

33. **Tocqueville V, Ferre S, Nguyen NH, Kempf I, Marois-Crehan C.** 2014. Multilocus sequence typing of Mycoplasma hyorhinis strains identified by a real-time TaqMan PCR assay. Journal of clinical microbiology **52:**1664-1671.

34. **Rosales RS, Churchward CP, Schnee C, Sachse K, Lysnyansky I, Catania S, Iob L, Ayling RD, Nicholas RA.** 2015. Global multilocus sequence typing analysis of Mycoplasma bovis isolates reveals two main population clusters. Journal of clinical microbiology **53:**789-794.

35. **Gruson D, Pereyre S, Renaudin H, Charron A, Bébéar C, Bébéar CM.** 2005. In vitro development of resistance to six and four fluoroquinolones in Mycoplasma pneumoniae and Mycoplasma hominis, respectively. Antimicrobial Agents and Chemotherapy **49**:1190-1193.

36. **Diaz MH, Benitez AJ, Winchell JM.** 2015. Investigations of Mycoplasma pneumoniae infections in the United States: trends in molecular typing and macrolide resistance from 2006 to 2013. Journal of clinical microbiology **53:**124-130.

37. **Zhou Z, Li X, Chen X, Luo F, Pan C, Zheng X, Tan F.** 2015. Macrolide-resistant Mycoplasma pneumoniae in adults in Zhejiang, China. Antimicrobial agents and chemotherapy **59:**1048-1051.

38. **Chen FQ, Yang YZ, Yu LL, Bi CB.** 2015. Prevalence of Mycoplasma pneumoniae: A cause for community-acquired infection among pediatric populaztion. Nigerian journal of clinical practice **18:**354-358.

39. **Klement E, Talkington DF, Wasserzug O, Kayouf R, Davidovitch N, Dumke R, Bar-Zeev Y, Ron M, Boxman J, Lanier Thacker W, Wolf D, Lazarovich T, Shemer-Avni Y, Glikman D, Jacobs E, Grotto I, Block C, Nir-Paz R.** 2006. Identification of risk factors for infection in an outbreak of Mycoplasma pneumoniae respiratory tract disease. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America **43:**1239-1245.

40. **Meyer Sauteur PM, Bleisch B, Voit A, Maurer FP, Relly C, Berger C, Nadal D, Bloemberg GV.** 2014. Survey of macrolide-resistant Mycoplasma pneumoniae in children with community-acquired pneumonia in Switzerland. Swiss medical weekly **144:**w14041.

Acknowledgements

This work was funded by Public Health England. Senior authorship for this manuscript is shared between OBS and VJC. These studies were supported by funding initiatives by the National Institute for Social Care and Health Research (NISCHR; research support from the Welsh Government) via the registered research group Microbial and Infection Translational Research Group (MITReG) and Children and Young Persons Research Network (CYPRN).

The authors have no conflicts of interest to declare.

Table 1. Description of *Mycoplasma pneumoniae* strains used in this study, their sequence type (ST) and allelic profile, and their MLVA and P1 types. Strains isolated from the same patient are indicated by grey shading.

Strain	Year of isolation	Country of isolation	Isolation site	ST	Allelic profile								MLVA type	P1 type
					ppa	*pgm*	*gyrB*	*gmk*	*glyA*	*atpA*	*arcC*	*adk*		
M129 (ATCC 29342)	1969	USA	Unknown	1	1	2	1	1	1	3	2	1	4572	1
MPN135	1986	USA	Unknown	1	1	2	1	1	1	3	2	1	4572	V1
FH (ATCC 15531)	1944	USA	Sputum	2	2	3	2	2	2	4	1	1	3662	2
MPN007	1978	UK	Throat swab	2	2	3	2	2	2	4	1	1	NT[a]	1
MPN021	1983	UK	Unknown	2	2	3	2	2	2	4	1	1	3662	NT[a]
MPN022	2010	UK	Sputum	2	2	3	2	2	2	4	1	1	3562	2c
MPN023	1983	UK	Sputum	2	2	3	2	2	2	4	1	1	3662	2
MPN101	1978	UK	Unknown	2	2	3	2	2	2	4	1	1	3562	1
MPN102	1981	UK	Brain frontal lobe	2	2	3	2	2	2	4	1	1	3662	2
MPN107	1983	UK	Sputum	2	2	3	2	2	2	4	1	1	3562	1
MPN114	1983	UK	Sputum	2	2	3	2	2	2	4	1	1	3662	1
MPN117	1982	UK	Sputum	2	2	3	2	2	2	4	1	1	3562	2
MPN119	1982	UK	Sputum	2	2	3	2	2	2	4	1	1	3562	2
MPN121	1983	UK	Sputum	2	2	3	2	2	2	4	1	1	3662	2c
MPN123	1983	UK	Sputum	2	2	3	2	2	2	4	1	1	3662	2
MPN125	1983	UK	Sputum	2	2	3	2	2	2	4	1	1	3562	2
MPN126	1979	UK	Unknown	2	2	3	2	2	2	4	1	1	3662	2
MPN128	1976	USA	Unknown	2	2	3	2	2	2	4	1	1	3662	1
MPN132	1982	UK	Sputum	2	2	3	2	2	2	4	1	1	3562	2
MPN133	1982	UK	Sputum	2	2	3	2	2	2	4	1	1	3662	2
MPN134	1982	UK	Sputum	2	2	3	2	2	2	4	1	1	3662	2
MPN005	1983	UK	Sputum	3	1	2	1	1	1	3	1	1	4572	1
MPN006	1982	UK	Sputum	3	1	2	1	1	1	3	1	1	4572	NT[a]
MPN013	2009	UK	Nose & throat swabs	3	1	2	1	1	1	3	1	1	4572	1
MPN014	2009	UK	Nose & throat swabs	3	1	2	1	1	1	3	1	1	4572	1
MPN015	2009	UK	Nose & throat swabs	3	1	2	1	1	1	3	1	1	4572	1
MPN016	2009	UK	Nose & throat swabs	3	1	2	1	1	1	3	1	1	4572	1
MPN017	2009	UK	Nose & throat swabs	3	1	2	1	1	1	3	1	1	4572	1

MPN020	1982	UK	Sputum	3	1	2	1	1	1	3	1	1	4572	2
MPN103	1982	UK	Sputum	3	1	2	1	1	1	3	1	1	4572	1
MPN105	1983	UK	Sputum	3	1	2	1	1	1	3	1	1	4572	1
MPN108	1983	UK	Sputum	3	1	2	1	1	1	3	1	1	4572	1
MPN109	1982	UK	Sputum	3	1	2	1	1	1	3	1	1	4572	2
MPN113	1967	UK	Unknown	3	1	2	1	1	1	3	1	1	4572	1
MPN116	1982	UK	Sputum	3	1	2	1	1	1	3	1	1	4572	1
MPN118	1996	UK	Sputum	3	1	2	1	1	1	3	1	1	4572	1
MPN120	1982	UK	Sputum	3	1	2	1	1	1	3	1	1	4572	1
MPN122	1982	UK	Sputum	3	1	2	1	1	1	3	1	1	4572	1
MPN136	1982	UK	Sputum	3	1	2	1	1	1	3	1	1	4572	1
MPN004	1981	UK	Sputum	4	2	1	2	2	2	4	1	1	3662	1
MPN104	1981	UK	Sputum	4	2	1	2	2	2	4	1	1	3662	2
MPN106	1981	UK	Sputum	4	2	1	2	2	2	4	1	1	3662	2
MPN110	1981	UK	Sputum	4	2	1	2	2	2	4	1	1	3662	2
MPN124	1981	UK	Sputum	4	2	1	2	2	2	4	1	1	3662	2
MPN131	1981	UK	Sputum	4	2	1	2	2	2	4	1	1	3662	1
MPN111	1968	UK	Unknown	5	1	2	1	1	1	2	1	1	4572	1
MPN011	1983	UK	Sputum	6	2	3	2	2	2	1	1	1	3662	1
MPN112	1983	UK	Sputum	6	2	3	2	2	2	1	1	1	3662	1
MPN127	1982	UK	Sputum	7	2	3	2	2	2	4	1	2	3662	2
MPN129	1983	UK	Sputum	8	2	3	2	2	2	4	1	3	3662	2
MPN130	1983	UK	Sputum	9	1	2	1	1	1	3	1	4	4572	1
MPN008	1981	UK	Sputum	10	2	1	2	2	2	4	1	2	3662	2
MPN018	1981	UK	Sputum	10	2	1	2	2	2	4	1	2	3662	2
MPN010	1983	UK	Sputum	11	1	2	1	1	3	3	1	1	3662	1
MPN003	1983	UK	Sputum	11	1	2	1	1	3	3	1	1	4572	1
MPN012	1981	UK	Brain cyst	11	1	2	1	1	3	3	1	1	3562	NT[a]
MPN019	1983	UK	Sputum	12	2	2	1	1	3	3	1	4	4572	1

[a]NT *M. pneumoniae* not classified by MLVA/P1 typing

445

446

Table 2. MLST loci used in established bacterial MLST schemes also present in *M. pneumoniae*.

Bacterial Species	MLST Loci [a]									
	reca	***ppa***	***pgm***	***gyrb***	***gmk***	***glya***	***efp***	***atpa***	***arcc***	***adk***
Bacillus cereus					✓					
Chlamydia trachomatis						✓				
Campylobacter jejuni			✓			✓				
Escherichia coli	✓			✓						✓
Enterococcus faecium								✓		✓
Haemophilus influenzae	✓									✓
Helicobacter pylori		✓					✓	✓		
Moraxella catarrhalis		✓					✓			✓
Neisseria meningitidis			✓							✓
Staphylococcus aureus					✓				✓	
Staphylococcus epidermidis									✓	
Streptococcus suis	✓									
Vibrio vulnificus				✓						
Yersinia pseudotuberculosis										✓

[a] MLST loci were chosen based on the frequency of use in other bacterial MLST schemes (http://www.mlst.net/) and the presence of the gene in the published M129 and FH whole genomes.

Table 3. Primer pairs developed in this study and variability of the different loci.

Name		Primer sequence (5'-3')	Amplicon (bp)	MLST locus location	No. of alleles	No. polymorphic sites	% polymorphic sites	Average G + C content (%)	Hunter-Gaston Diversity Index [a]	95% Confidence Interval
ppa	F	CGCTGACCAAGCCTTTCTAC	256	192-440	2	1	0.39	38.4	0.501	0.470-0.533
	R	CACTCCAAACTTTGCACTCCC								
pgm	F	AGCACCTTGCACGATGAAGA	1072	456-1652	3	10	0.93	43.7	0.620	0.566-0.674
	R	CCTGCGCCTTCGTTAATTGG								
gyrB	F	TTGTCCCGGACTTTACCGTG	429	524-952	2	2	0.47	39.9	0.505	0.482-0.528
	R	TGTTTTCGACAGCAAAGCGG								
gmk	F	GAGCGGTGTTGGCAAAAGTA	394	189-582	2	1	0.25	40.1	0.505	0.482-0.528
	R	TGCATCCTCGTCATTACGCTT								
glyA	F	CAGAGAACTATGTGAGTAGGGACA	676	74-749	3	2	0.30	45.6	0.560	0.493-0.627
	R	TGACAACCCGGAAAGACACC								
atpA	F	GTCGCTGATGGCATTGCTAAG	796	100-895	4	3	0.38	44.8	0.557	0.502-0.612
	R	CCAGTAAACGCGAGTGCAAG								
arcC	F	CCCCATCAAGCCGTGTACTT	570	304-873	2	1	0.18	45.5	0.069	0.000-0.158
	R	TTGGGCAATAATGGCCGTCT								
adk	F	GTAGCCAACACCACCGGATT	473	70-542	4	3	0.63	47.5	0.199	0.063-0.335
	R	ACGGTGTCTTCGTAAAGCGT								

[a] Hunter-Gaston diversity index (DI, ranges from 0.0 indicates no diversity to 1.0 indicates complete diversity)

Figure Legends

Figure 1. Phylogenetic trees based on concatenated sequences of 8 MLST loci. Phylogenetic trees were constructed based on concatenated sequences of eight housekeeping loci for 12 unique STs using Maximum likelihood (A) and Neighbour-joining (B) methods. Bootstrap support values of over 70% are shown. STs are indicated by differential shading.

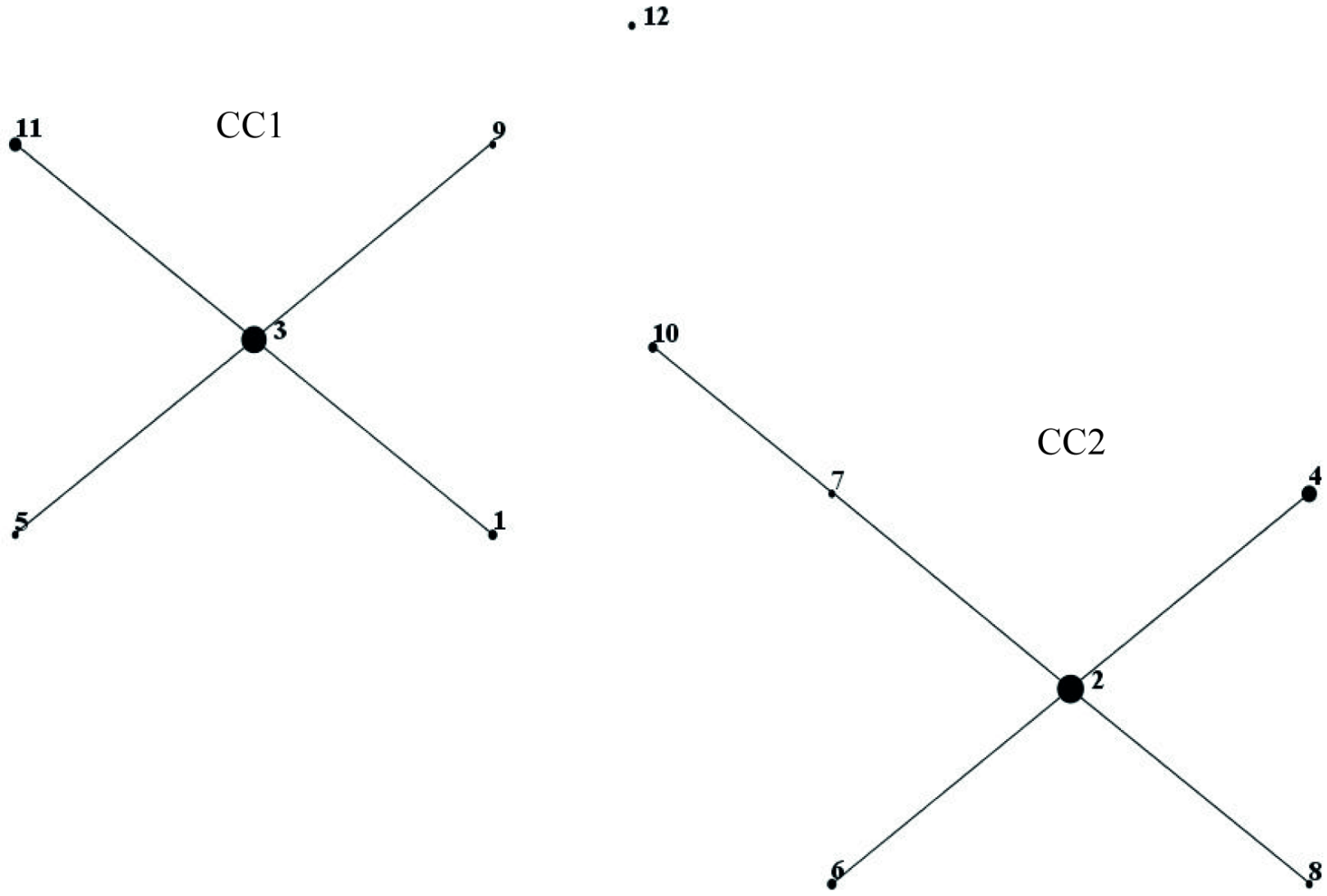
Figure 2. Analysis of *M. pneumoniae* using eBURST. eBURST version 3 was used to analyse the 12 unique STs resolved for all 57 *M. pneumoniae* isolates. Two main clonal complexes (CC) were defined. The size of each dot is proportional to the number of isolates included in the analysis for each ST.

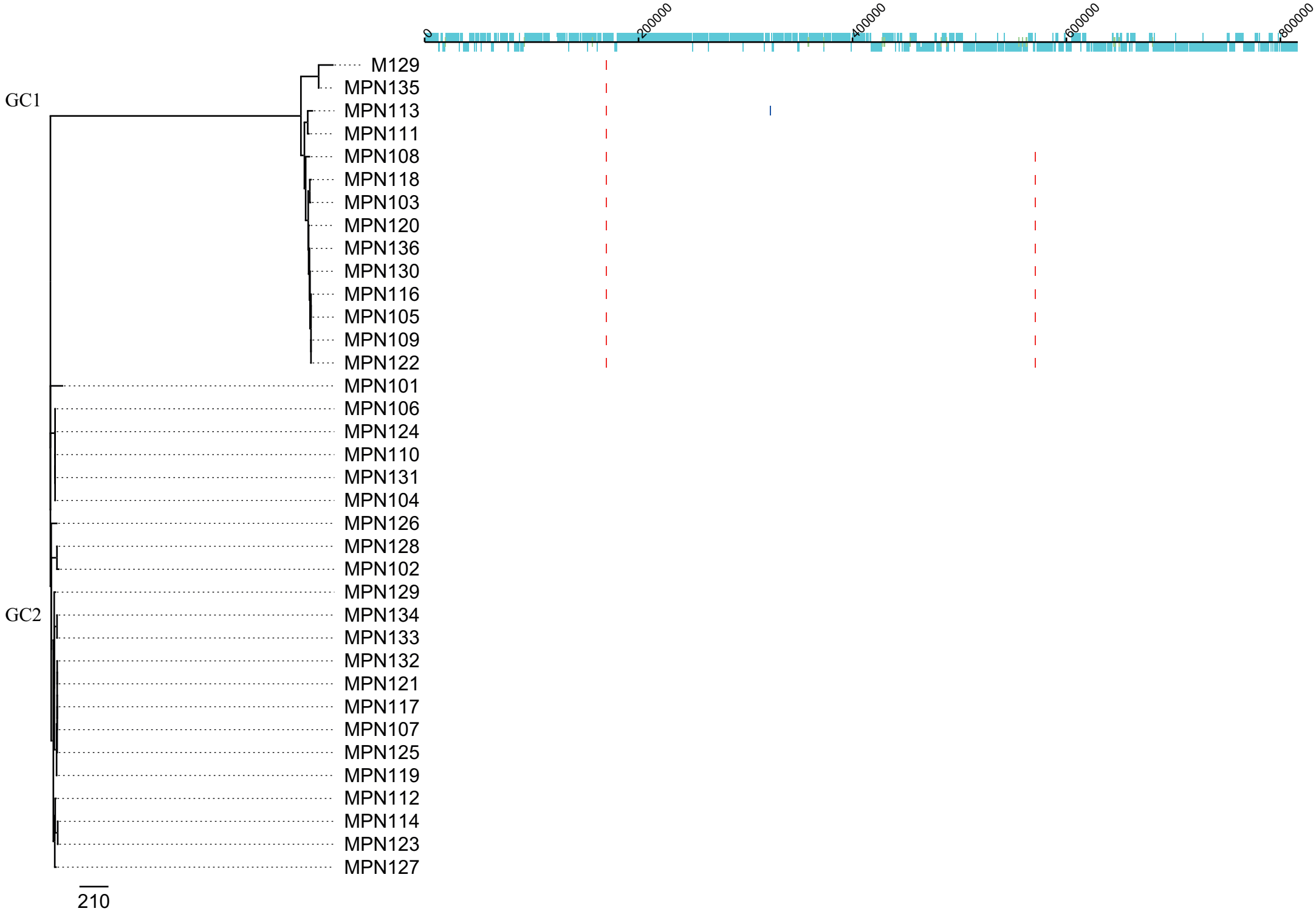
Figure 3. Prediction of recombination in the *M. pneumoniae* isolates chromosomes. Regions of variation in the genomes of the 35 clinical *M. pneumoniae* isolates and the type strain M129 which are predicted to have arisen by homologous recombination are shown in the panel on the right. Red blocks indicated recombination predicted to have occurred on internal nodes, blue indicates taxa-specific recombination). Isolates are ordered according to the phylogenetic tree displayed on the left. The track along the top of the figure displays the M129 chromosome and annotation, where protein coding sequences (CDS) are indicated in light blue.

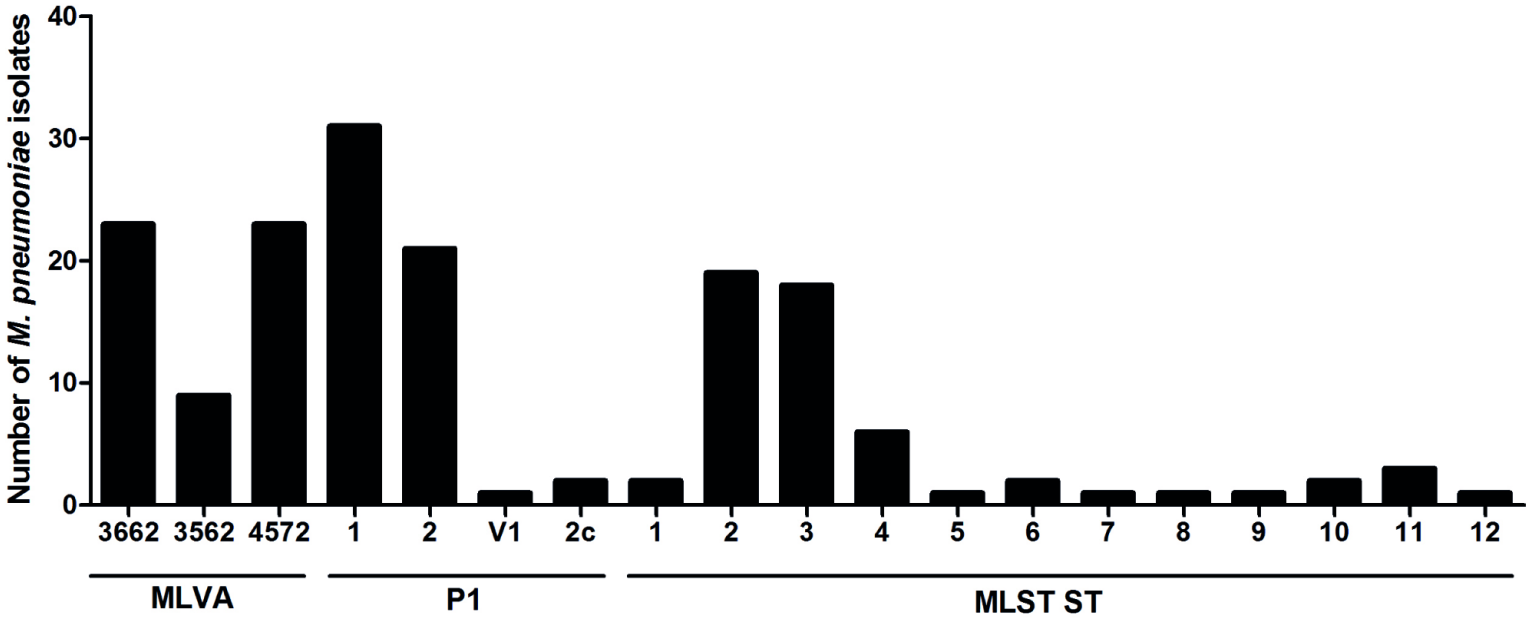
Figure 4. Distribution of MLVA, P1 type and MLST ST for 57 *M. pneumoniae* isolates. The 57 isolates were separated independently for MLVA type, P1 type and MLST type (each group defined by line).

A.

B.


12
CC1
11
9
3
5
1
10
CC2
7
4
2
6
8


0
200000
400000
600000
800000
GC1
GC2
M129
MPN135
MPN113
MPN111
MPN108
MPN118
MPN103
MPN120
MPN136
MPN130
MPN116
MPN105
MPN109
MPN122
MPN101
MPN106
MPN124
MPN110
MPN131
MPN104
MPN126
MPN128
MPN102
MPN129
MPN134
MPN133
MPN132
MPN121
MPN117
MPN107
MPN125
MPN119
MPN112
MPN114
MPN123
MPN127
210


Number of M. pneumoniae isolates
40
30
20
10
0
3662
3562
4572
1
2
V1
2c
1
2
3
4
5
6
7
8
9
10
11
12
MLVA
P1
MLST ST