Comparative Genomics of Mycoplasma pneumoniae isolated from Children with Pneumonia: South Korea, 2010-2016

Joon Kee Lee
Chungbuk National University Hospital https://orcid.org/0000-0001-8191-0812
Eun Hwa Choi (eunchoi@snu.ac.kr)
Moon-Woo Seong
Seoul National University Hospital
Youbin Yeon
Seoul National University Hospital
Sung Im Cho
Seoul National University Hospital
Sung Sup Park
Seoul National University Hospital

Research article

Keywords: Mycoplasma pneumoniae, whole genome analysis, comparative genomics

Posted Date: June 3rd, 2019

DOI: https://doi.org/10.21203/rs.2.9997/v1

Version of Record: A version of this preprint was published on November 29th, 2019. See the published version at https://doi.org/10.1186/s12864-019-6306-9.

Abstract

Background Mycoplasma pneumoniae is a common cause of respiratory tract infections in children and adults. This study applied high-throughput whole genome sequencing (WGS) technologies to analyze the genomes of 30 M. pneumoniae strains isolated from children with pneumonia in South Korea during the two epidemics from 2010 to 2016 in comparison with a global collection of 48 M. pneumoniae strains which includes seven countries ranging from 1944 to 2017. Results The 30 Korean strains had approximately 40% GC content and ranged from 815,686 to 818,669 base pairs, coding for a total of 809 to 828 genes. Overall, BRIG revealed 99% to > 99% similarity among strains. The genomic similarity dropped to approximately 95% in the P1 type 2 strains when aligned to the reference M129 genome, which corresponded to the region of the p1 gene. MAUVE detected four subtype-specific insertions (three in P1 type 1 and one in P1 type 2), of which were all hypothetical proteins except for one tRNA insertion in all P1 type 1 strains. The phylogenetic associations of 30 strains were generally consistent with the multilocus sequence typing results. eBURST analysis demonstrated two clonal complexes which are accordant with the known P1 typing, with higher diversity among P1 type 2 strains. The phylogenetic tree constructed with 78 genomes including 48 genomes outside Korea formed three clusters, in which the sequence type 3 strains from Korea were divided into two P1 type 1 clusters. Conclusions The comparative genome analysis of the 78 M. pneumoniae strains including 30 strains from Korea by WGS reveals structural diversity and phylogenetic associations, even though the similarity across the strains was very high.

Background

M. pneumoniae is an important cause of respiratory tract infections in children and adults, ranging from mild upper respiratory infections to life-threatening conditions (1). *M. pneumoniae* infections are more common among children 5 years of age or older than among younger children (2). Mild upper respiratory infections are common with a considerable portion of asymptomatic patients, but 3 to 10% develop pneumonia with a wide spectrum of radiologic findings (3-5). Extrapulmonary abnormalities are an important part of *M. pneumoniae* diseases both in diagnosis and treatment. The spectrum of manifestations includes extrapulmonary symptoms such as skin rash, hemolytic anemia, arthritis, and neurologic abnormalities (1).

P1 adhesin (P1), a 170-kD surface protein located at the tip-like structure of virulent *M. pneumoniae*, mediates its cytoadherence to the surface of respiratory epithelial cells (6). As P1 adhesin protein plays a critical step in the infection process, studies regarding the genetics of *M. pneumoniae* focused mainly on P1 types and subtypes (7, 8). P1 typing was the only available tool that could be applied in the past to determine genotype. Although P1 typing can separate *M. pneumoniae* into two types and an additional six variants, it did not always convey information regarding epidemiologic characteristics or clinical severity. New genetic analysis techniques, such as multilocus variable-number tandem-repeat analysis (MLVA) and multilocus sequence typing (MLST), have been applied to *M. pneumoniae* (9, 10).

Despite the evolution of molecular microbiology and advanced classifications beyond P1 typing, research to understand the entire genome structures of *M. pneumoniae* in regard to molecular epidemiology has

remained much behind that of other bacteria such as *Streptococcus pneumoniae* and *Escherichia coli*. Recent advances in molecular microbiology and bioinformatics have made it possible to analyze *M. pneumoniae* through high-throughput sequencing technologies such as Illumina dye sequencing, pyrosequencing, and single-molecule real-time (SMRT) sequencing (11). The whole genome of *M. pneumoniae* is ≈820 kb and has up to 700 coding operons (12). The comparably short size of the whole genome and limited number of operons are challenges in the genomic investigation of *M. pneumoniae*.

This study aims to analyze genomes of 30 *M. pneumoniae* strains isolated from children with pneumonia in South Korea during two epidemics from 2010 to 2016 and compare with a global collection of 48 *M. pneumoniae* strains which includes seven countries ranging from 1944 to 2017.

Results

Strain characteristics

The strains were isolated from nasopharyngeal samples obtained from children with pneumonia. Thirty-seven and 45 *M. pneumoniae* strains were collected in 2010-12 and 2014-16, respectively. Thirty *M. pneumoniae* strains were chosen for the current study (Additional file 1). Eighteen strains and twelve strains were selected from 2010-12 and 2014-16 epidemic years, respectively. Twenty-four (80.0%) P1 type 1 strains, five (16.7%) P1 type 2c strains and a P1 type 2a strain (3.3%) were included. Five sequence types (STs) were included: ST1 (n=2, 6.7%), ST3 (n=20, 66.7%), ST14 (n=5, 16.7%), ST17 (n=2, 6.7%), and ST33 (n=1, 3.3%).

Genome assembly

The characteristics of the assemblies and the background information are found in Table 1. The resulting contigs were mapped to the M129 reference genome and joined via PCR. The thirty genomes had all contigs joined to form a single, continuous (circular) contig. Following assembly and editing, the genomes underwent automated gene annotation. With approximately 40% GC content and ranging from 815,686 to 818,669 bp, the genomes coded for a total of 809 to 828 genes.

Overall comparison

The 30 sequenced genomes were aligned to the reference M129 genome using BLAST Ring Image Generator (BRIG). Overall, the genomes were 99% to > 99% identical. The similarity dropped to approximately 95% in the type 2 strains, which corresponded to the area of the *p1* gene (Fig. 1).

Genomic structural comparison

For the detection of large chromosomal rearrangements, deletions, and duplications, MAUVE was applied to the 30 sequenced genomes with 6 reference genomes. All genomes fell into three locally collinear blocks (LCBs), which are conserved segments. The three LCBs were in the same order without any rearrangement. MAUVE detected four subtype-specific insertions (Fig. 2): three type 1-specific insertions (M129 numbering;

169-170 kb, 178-179 kb, and 558-560 kb) and a type 2-specific insertion (M129 numbering; 708 kb). Type 1 insertions were all annotated as hypothetical proteins (MPN130, MPN137, MPN138, and MPN457-459) except for the tRNA gene (MPNt26) positioned at 558635 to 558723 (M129 numbering). The proteins of the type 2 insertion (6 kbp) were annotated as hypothetical proteins without exception (BIX66_03340, 03345, 03350, 03355, and 03360).

SNP and indel analysis

SNPs and indels were compared for the identification of sequence level differences against the reference genome. The results are shown in Table 2. As expected, P1 type 1 strains showed fewer variant numbers (140-455) than P1 type 2 strains (1778-1796), showing a clear distinction.

Proteins and functional analysis

The Protein Family Sorter tool at Pathosystems Resource Integration Center (PATRIC) allows the selection of a set of genomes of interest and the examination of the distribution of protein families across genomes. An interactive heatmap viewer provides a comprehensive view of the distribution of the protein families across multiple genomes, with clustering and anchoring functions to show relative conservation of synteny and to identify lateral transfers. Based on gene annotation from PATRIC, a heatmap of all proteins was produced along with the reference genome *M. pneumoniae* M129 (Fig. 3). Unsurprisingly, when genomes were classified into P1 types 1 and 2, distinction between genomes was apparent. Nevertheless, most of the genomes that showed different expressions between them were hypothetical proteins with uncertain significance.

Phylogenetic associations

Thirty genomes were aligned with MAFFT, and a phylogenetic tree was generated (Additional file 2). The phylogenetic tree was divided into two clades in accordance with the P1 typing. In general, the STs of the 30 strains were consistent with the phylogenetic relationship. This was prominent among the P1 type 1 strains, while a ST33 strain cut into the P1 type 2 strains.

All 78 strains, including strains from this study and NCBI, were aligned with CLC Phylogeny Module, and a phylogenetic tree was generated (Fig. 4). In general, the strains in this study were scattered throughout the entire phylogenic tree, along with the expansion of certain clades. Surprisingly, strains were grouped primarily into three clusters, forming two separated clusters in the P1 type 1 clade. A few sub-clusters were noticed. ST1 strains from China from 2015 to 2016 and ST20 strains from USA from 2006 to 2012 along with a strain from Egypt in 2010 formed distinct groups. The findings from the MAFFT phylogenetic tree based on the 30 strains in this study were consistent with the phylogenetic tree based on extended strains.

Comparative genomics with global strains-MLST

For the comparative genome analysis of global strains, 48 *M. pneumoniae* genomes were accessed from NCBI. For the analysis, typing of P1 types and MLST types was performed (Table 3, Additional file 1). An

eBURST diagram was constructed in two sequential procedures (Fig. 5). First, all of the known STs (from ST1 to ST33) were applied for the inspection of the association of the STs as well as the evolution of STs. Second, 30 strains from this study and 48 strains from NCBI were added to the diagram to observe the prevalence and proportional weight of each ST in global.

The initial diagram showed two clonal complexes with two singletons of ST12 and ST22. The founder ST of CC1 was identified as ST3 with no double locus variants (DLVs). The founder ST of CC2 was recognized as ST2 with multiple subgroup founders (ST7, ST14 and ST24), multiple single locus variants (SLVs) and DLVs. When examining the eBURST diagram of global strains, ST3 and ST1 from CC1, ST2 and ST14 from CC2 were the main STs, respectively.

Discussion

The comparative genome analysis of 30 *M. pneumoniae* strains prevalent in South Korea during two epidemics with 48 global strains reveals structural diversity and phylogenetic associations between and within the strains in geographic regions.

M. pneumoniae is known as a 'difficult-to-culture' organism (1). Thus, unlike ordinary bacterial pathogens, the aid of molecular biology in the diagnosis of *M. pneumoniae* is critical (13). With the burden of disease caused by this organism and diverse extrapulmonary clinical manifestations, it seems natural that *M. pneumoniae* has drawn the attention of researchers. Nevertheless, in addition to the molecular diagnosis of *M. pneumoniae* by the P1 adhesin, P1 typing has been the sole method for classification for decades (14). However, because the size of the *M. pneumoniae* genome is short compared to that of other bacteria and because the P1 adhesin is the only apparently diverse part of the whole gene, it may be reasonable for researchers to continue to focus on the P1 adhesin. Despite these efforts, P1 was not sufficient for the explanation of epidemics or for the explanation of clinical severity (15, 16).

Recent advances in molecular microbiology have widened the scope of the implementation of sophisticated techniques, such as MLVA and MLST (9, 10). New classifications developed by such new technologies have expanded P1 classification with enhanced distinction. Nevertheless, epidemics still cannot be clearly explained by the new technologies, and there are reports that chest X-rays are the most predictive clue in the course of infection regardless of the molecular genetics (4). Nevertheless, attempts to utilize molecular biology by availing MLVA or MLST have shown some successful insights and the possibility of further investigations (17-19).

Although not extensively applied, high-throughput technologies have been applied to the investigation of *M. pneumoniae*. A study conducted by Xiao *et al.* analyzed 15 *M. pneumoniae* genomes obtained by Illumina sequencing, including 11 clinical isolates and 4 reference strains (20). Although approximately 1500 SNP and indel variants exist between type 1 and type 2 strains, an overall high degree of sequence similarity was found among the strains (> 99% identical to each other). The study concluded that the *M. pneumoniae* genome is extraordinarily stable over time and geographic distances across the globe, with a striking lack of evidence of horizontal gene transfer.

One of the most recent NGS studies performed by Diaz *et al.* demonstrated WGS analysis of 107 *M. pneumoniae* isolates, including 67 newly sequenced isolates, using the Pacific BioSciences RS II and/or Illumina MiSeq sequencing platforms (21). Population structure analysis done by this study supported the existence of six distinct subgroups. Although this study included the largest collection of *M. pneumoniae* isolates ever, only a few strains were included from Asian regions where the unique epidemiologic features (for example, high rate of macrolide-resistance among *M. pneumoniae*) are noticed.

Comparative genome analysis was performed using BRIG, MAUVE, MAFFT and CLC Phylogeny Module. Unsurprisingly, the genomes were classified mainly by the legendary P1. BRIG clearly distinguished P1 types 1 and 2, but no further information could be found, as separate genes could not be visualized (22). MAUVE utilizes LCBs, which are conserved segments that appear to be internally free from genome rearrangements (23). The result from MAUVE showed that large rearrangements (e.g., plasmids, phage or resistance genes) were not observed among *M. pneumoniae*. Specific insertions were noted in both P1 types. Nevertheless, the translated proteins of the inserted genes were generally hypothetical proteins with the exception of a tRNA. This is consistent with a previous report by Xiao *et al.,* but the two insertions at 169-170 kb and 178-179 kb have not been described previously (20). The heatmap generated by PATRIC confirmed the P1 classification by differences in protein production. This is consistent with additional studies that applied NGS technology.

The SNP approach is widely used in the study of antimicrobial resistance and genetic diversity and is not limited to *M. pneumoniae* (21-23). This study is consistent with previous studies investigating SNPs within *M. pneumoniae.* Variant calling against M129 of P1 subtypes showing substantially fewer variants compared to P1 type 2 in both nonsynonymous SNPs and total variants is a natural result.

Generation of the phylogenetic tree by MAFFT and CLC Phylogeny Module revealed a few interesting findings. First, based on the phylogenetic tree of the 30 strains in this study, clear distinction was identified according to P1 type. Each ST type was grouped by the same branch, which reconfirms the advancement of MLST distinction. The P1 classification was still valid when a phylogenetic tree was generated with the 30 sequenced genomes plus the 48 NCBI genomes, including 6 reference genomes. Unexpectedly, the phylogenetic tree was divided into three clades, with an additional branch harboring the S355 reference genome, which originated from China in 2012. Because the strains from the current study were dispersed throughout the phylogenetic tree, it is not convincing that clonal expansion of certain strains has occurred. Nevertheless, as ST3 strains from this study are divided into and enriching two clades, it may be possible that clonal expansion occurred in both clades. It is a quite interesting result that despite the higher diversity shown by eBURST in the P1 type 2 strains, oppositely, associations revealed by phylogenetic tree shows higher diversity rather in the P1 type 1 strains.

The comparative genomics of the 78 *M. pneumoniae* strains including 30 strains from Korea by WGS reveals structural diversity and phylogenetic associations, even though the similarity across the strains was very high.

Methods

M. pneumoniae strains

This study comprised *M. pneumoniae* strains detected from children with pneumonia at two hospitals during two consecutive outbreaks of *M. pneumoniae* pneumonia in South Korea in 2010–2012 and 2014–2016. Epidemic periods were previously defined by an interval spanning an increase of >5 cases/2 months over a 4-month period to a decrease of <5 cases/2 months over a 4-month period in the primary site of this study (24, 25). *M. pneumoniae* pneumonia was diagnosed using the following criteria: 1) the presence of rales during auscultation or infiltration of the lung demonstrated with a chest radiograph and 2) isolation of *M. pneumoniae* in culture. Specimens were obtained from Seoul National University Children's Hospital (Seoul) and Seoul National University Bundang Hospital (Seongnam).

Cultivation

Cultivation of *M. pneumoniae* was performed at the Seoul National University Children's Hospital. Reference strain M129 (ATCC 29342) was cultured in parallel with the clinical samples using pleuropneumonia-like organism (PPLO) broth and agar. Two hundred microliters of the nasopharyngeal specimen were serially diluted 64-fold. The broth medium was composed of 70 mL of PPLO broth, 20 mL of horse serum, 10 mL of 25% yeast extract, 2.5 mL of 20% glucose, 200 μL of 1% phenol red, 1 mL of 2.5% thallium acetate, 0.5 mL of 200,000 units/mL penicillin G potassium, and 0.5 mL of 20,000 μg/mL cefotaxime. The agar was prepared with the same components as the broth medium except that cefotaxime was omitted and 1.2% agar powder was added instead of broth powder. The broth and the agar media were incubated aerobically at 37 °C for 6 weeks.

DNA preparation

The plates were observed daily to identify color changes in the broth medium from red to transparent orange. Upon color change, 10 μL were subcultured onto agar plates. Spherical *M. pneumoniae* colonies were observed under a microscope at 100X magnification. DNA was extracted directly from cultivated *M. pneumoniae* using an extraction kit (DNeasy Kit; QIAGEN, Hilden, Germany) according to the manufacturer's instructions. The *p1* gene was amplified by PCR for the confirmation of *M. pneumoniae*.

MLST analysis and P1 typing

MLST was performed on the *M. pneumoniae* DNA samples as previously described. Each allele was assigned to the 8 housekeeping genes (*ppa*, *pgm*, *gyrB*, *gmk*, *glyA*, *atpA*, *arc*, and *adk*), and a corresponding ST was given for each sample (28). P1 typing was performed by sequencing 2 of the repetitive elements located in the *p1* gene of the *M. pneumoniae* genome: *RepMP2/3* and *RepMP4*. P1 subtypes and each subtype variant were assigned by comparison with previously published data (26).

Selection of strains for Whole-genome analysis

A total of 30 strains were selected for the whole-genome sequencing (WGS) investigation. Thirty-seven strains from the 2010-12 epidemic year and 45 strains from 2014-16 were candidates for WGS.

Next-generation sequencing (NGS)

NGS of all *M. pneumoniae* strains was performed using the Illumina MiSeq desktop sequencer (Illumina, San Diego, CA, USA). Illumina NGS workflows include four basic steps: library preparation, cluster amplification, sequencing and alignment. The NGS library is prepared by fragmenting a genomic DNA sample and ligating specialized adapters to both fragment ends. The library is loaded into a flow cell, and the fragments are hybridized to the flow cell surface. Each bound fragment is clonally amplified through bridge amplification. Sequencing repeats, including fluorescently labeled nucleotides, are added, and the first base is incorporated. The flow cell is imaged, and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated 'n' times to create a read length of 'n' bases. In this study, paired-end 250-bp reads were used with an average depth (coverage) of 442.93 (ranging from 172.95 to 795.39). Instead of directly aligning the reads to a reference sequence, *de novo* assembly was performed.

Genome assembly and annotation

NGS reads were assembled *de novo* using SPAdes (27). The number of contigs generated ranged from 3 to 8 per strain. These contigs were mapped to the M129 reference genome using the BLAST-like alignment tool (BLAT) and visualized using Integrative Genomics Viewer (IGV) (28-30). This mapping was used to develop PCR primers to join the contigs. High fidelity PCRs and Sanger sequencing were performed using standard methods. Overlapping and joining of the contigs were performed manually with Sequencher version 5.4.6 (Gene Codes Corporation, Ann Arbor, MI, USA). The initial NGS reads were aligned to the *de novo* assembled genome for the correction of errors. The corrected and completed circular genomes were annotated using Rapid Annotation using Subsystem Technology (RAST) (31).

Comparative genomics

Completed genomes were aligned using BRIG for the overall sequence similarity between the strains (32). MAUVE was used to detect large chromosomal rearrangements, deletions, and duplications (33). For phylogenetic tree generation and visualization, MAFFT and CLC Phylogeny Module was used (Qiagen, Venlo, Netherlands). For the extended phylogenetic analysis along with global strains, 48 strains downloaded from the National Center for Biotechnology Information (NCBI) were included. eBURST version 3 software (http://eburst.mlst.net/) was used to estimate the relationships among the strains and to assign strains to a clonal complex (CC) (34).

Single nucleotide polymorphism (SNP) and insertion/deletion (indel) analysis

To call SNPs and indels, completed genomes were first broken into 10-kb "reads" at 1-kb intervals and then aligned to the M129 reference strain (NCBI Accession Number NC_000912) using BWA v0.7.7 (35). Variant calling was performed using Samtools (36). The effects of the SNPs and indels in the resulting VCF files were evaluated and annotated using SnpEff v3.3 (37).

Proteins and functional analysis

For the analysis of proteins and functional annotation, PATRIC was used, and a heatmap was generated based on annotations (38). Gene translation, multiple sequence alignment and visualization of proteins were performed using Clustal Omega (39). Annotation of any hypothetical genes was performed using a BLAST search against the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (40, 41).

References genomes

Six reference genomes were included in each analysis as appropriate. *M. pneumoniae* M129, FH, 309, KCH-402 and K405 are representatives of each P1 type and subtype. *M. pneumoniae* S355 is included, as this strain is one of the earliest strains that was fully sequenced and expressed macrolide resistance.

List Of Abbreviations

ST: sequence type; MLVA: multilocus variable-number tandem-repeat analysis; MLST: multilocus sequence typing; BRIG: BLAST Ring Image Generator; PATRIC: Pathosystems Resource Integration Center; SLV: single locus variant; DLV: double locus variant; CC: clonal complex

Declarations

Ethics approval and consent to participate

The institutional review board of Seoul National University Hospital approved the study protocol (IRB no. H-1012–007–341). Informed consent was exempted because nasopharyngeal aspirates were obtained as a standard of patient care to identify the etiologic agents of acute pneumonia.

Consent for publication: Not applicable

Availability of data and material

All data generated or analyzed during this study are included in this published article. The gene sequences are deposited in NCBI database under the accession numbers SAMN11472195, SAMN11472196, SAMN11472197, SAMN11472198, SAMN11472199, SAMN11472200, SAMN11472201, SAMN11472202, SAMN11472203, SAMN11472204, SAMN11472205, SAMN11472206, SAMN11472207, SAMN11472208, SAMN11472209, SAMN11472210, SAMN11472211, SAMN11472212, SAMN11472213, SAMN11472214, SAMN11472215, SAMN11472216, SAMN11472217, SAMN11472218, SAMN11472219, SAMN11472220, SAMN11472221, SAMN11472222, SAMN11472223, SAMN11472224.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by the 2017 Seoul National University Hospital Research Fund (0320170230), Seoul, Korea.

Authors' contributions

JKL conducted the experiments and contributed writing the manuscript. MWS and SSP helped in bioinformatics analysis. YY and SIC had contributed for raw data analysis. EHC had designed the work plan and contributed writing the manuscript. All authors read and approved the final manuscript.

Acknowledgements: Not applicable

Additional files

Additional file 1: P1 type and MLST type of the 30 strains from this study and 48 strains from NCBI (XLSX, 13.5 kb)

Additional file 2: Phylogenetic tree based on whole genome alignment of the 30 sequenced strains. (PDF, 84.7 kb)

References

1. Waites KB, Xiao L, Liu Y, Balish MF, Atkinson TP. *Mycoplasma pneumoniae* from the Respiratory Tract and Beyond. Clinical microbiology reviews. 2017;30(3):747-809.

2. Jain S, Williams DJ, Arnold SR, Ampofo K, Bramley AM, Reed C, et al. Community-acquired pneumonia requiring hospitalization among U.S. children. N Engl J Med. 2015;372(9):835-45.

3. Mansel JK, Rosenow EC, 3rd, Smith TF, Martin JW, Jr. *Mycoplasma pneumoniae* pneumonia. Chest. 1989;95(3):639-46.

4. Yoon IA, Hong KB, Lee HJ, Yun KW, Park JY, Choi YH, et al. Radiologic findings as a determinant and no effect of macrolide resistance on clinical course of *Mycoplasma pneumoniae* pneumonia. BMC infectious diseases. 2017;17(1):402.

5. Spuesens EB, Fraaij PL, Visser EG, Hoogenboezem T, Hop WC, van Adrichem LN, et al. Carriage of *Mycoplasma pneumoniae* in the upper respiratory tract of symptomatic and asymptomatic children: an observational study. PLoS medicine. 2013;10(5):e1001444.

6. Su CJ, Chavoya A, Dallo SF, Baseman JB. Sequence divergency of the cytadhesin gene of *Mycoplasma pneumoniae*. Infection and immunity. 1990;58(8):2669-74.

7. Su CJ, Chavoya A, Baseman JB. Regions of *Mycoplasma pneumoniae* cytadhesin P1 structural gene exist as multiple copies. Infection and immunity. 1988;56(12):3157-61.

8. Kenri T, Okazaki N, Yamazaki T, Narita M, Izumikawa K, Matsuoka M, et al. Genotyping analysis of *Mycoplasma pneumoniae* clinical strains in Japan between 1995 and 2005: type shift phenomenon of M. pneumoniae clinical strains. Journal of medical microbiology. 2008;57(Pt 4):469-75.

9. Degrange S, Cazanave C, Charron A, Renaudin H, Bebear C, Bebear CM. Development of multiple-locus variable-number tandem-repeat analysis for molecular typing of *Mycoplasma pneumoniae*. Journal of clinical microbiology. 2009;47(4):914-23.

10. Brown RJ, Holden MT, Spiller OB, Chalker VJ. Development of a Multilocus Sequence Typing Scheme for Molecular Typing of *Mycoplasma pneumoniae*. Journal of clinical microbiology. 2015;53(10):3195-203.

11. Mukhopadhyay R. DNA sequencers: the next generation. Anal Chem. 2009;81(5):1736-40.

12. Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res. 1996;24(22):4420-49.

13. Loens K, Ursi D, Goossens H, Ieven M. Molecular diagnosis of *Mycoplasma pneumoniae* respiratory tract infections. Journal of clinical microbiology. 2003;41(11):4915-23.

14. Diaz MH, Winchell JM. The Evolution of Advanced Molecular Diagnostics for the Detection and Characterization of *Mycoplasma pneumoniae*. Frontiers in microbiology. 2016;7:232.

15. Jacobs E, Ehrhardt I, Dumke R. New insights in the outbreak pattern of *Mycoplasma pneumoniae*. International journal of medical microbiology : IJMM. 2015;305(7):705-8.

16. Waller JL, Diaz MH, Petrone BL, Benitez AJ, Wolff BJ, Edison L, et al. Detection and characterization of *Mycoplasma pneumoniae* during an outbreak of respiratory illness at a university. Journal of clinical microbiology. 2014;52(3):849-53.

17. Lee JK, Lee JH, Lee H, Ahn YM, Eun BW, Cho EY, et al. Clonal Expansion of Macrolide-Resistant Sequence Type 3 *Mycoplasma pneumoniae*, South Korea. Emerging infectious diseases. 2018;24(8):1465-71.

18. Ando M, Morozumi M, Adachi Y, Ubukata K, Iwata S. Multilocus Sequence Typing of *Mycoplasma pneumoniae*, Japan, 2002-2016. Emerging infectious diseases. 2018;24(10):1895-901.

19. Sun H, Xue G, Yan C, Li S, Zhao H, Feng Y, et al. Changes in Molecular Characteristics of *Mycoplasma pneumoniae* in Clinical Specimens from Children in Beijing between 2003 and 2015. PloS one. 2017;12(1):e0170253.

20. Xiao L, Ptacek T, Osborne JD, Crabb DM, Simmons WL, Lefkowitz EJ, et al. Comparative genome analysis of *Mycoplasma pneumoniae*. BMC genomics. 2015;16:610.

21. Ramanathan B, Jindal HM, Le CF, Gudimella R, Anwar A, Razali R, et al. Next generation sequencing reveals the antibiotic resistant variants in the genome of Pseudomonas aeruginosa. PloS one. 2017;12(8):e0182524.

22. Lee JY, Na IY, Park YK, Ko KS. Genomic variations between colistin-susceptible and -resistant Pseudomonas aeruginosa clinical isolates and their effects on colistin resistance. J Antimicrob Chemother. 2014;69(5):1248-56.

23. Li SL, Sun HM, Zhu BL, Liu F, Zhao HQ. Whole Genome Analysis Reveals New Insights into Macrolide Resistance in *Mycoplasma pneumoniae*. Biomedical and environmental sciences : BES. 2017;30(5):343-50.

24. Hong KB, Choi EH, Lee HJ, Lee SY, Cho EY, Choi JH, et al. Macrolide resistance of *Mycoplasma pneumoniae*, South Korea, 2000-2011. Emerging infectious diseases. 2013;19(8):1281-4.

25. Eun BW, Kim NH, Choi EH, Lee HJ. *Mycoplasma pneumoniae* in Korean children: the epidemiology of pneumonia over an 18-year period. The Journal of infection. 2008;56(5):326-31.

26. Zhao F, Cao B, Li J, Song S, Tao X, Yin Y, et al. Sequence analysis of the p1 adhesin gene of *Mycoplasma pneumoniae* in clinical isolates collected in Beijing in 2008 to 2009. Journal of clinical microbiology. 2011;49(8):3000-3.

27. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455-77.

28. Kent WJ. BLAT–the BLAST-like alignment tool. Genome Res. 2002;12(4):656-64.

29. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics. 2013;14(2):178-92.

30. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nature biotechnology. 2011;29(1):24-6.

31. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. 2014;42(Database issue):D206-14.

32. Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC genomics. 2011;12:402.

33. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PloS one. 2010;5(6):e11147.

34. Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. Journal of

bacteriology. 2004;186(5):1518-30.

35. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589-95.

36. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.

37. . !!! INVALID CITATION !!! {}.

38. Wattam AR, Brettin T, Davis JJ, Gerdes S, Kenyon R, Machi D, et al. Assembly, Annotation, and Comparative Genomics in PATRIC, the All Bacterial Bioinformatics Resource Center. Methods in molecular biology. 2018;1704:79-101.

39. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology. 2011;7:539.

40. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389-402.

41. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353-D61.

Tables

Table 1. Genome lengths and contigs determined from the initial assembly with complete genome structures annotated by RAST

Strain	Contigs	L50	N50	Min Length	Max Length	Total Length	%GC	Genes		
								CDS	RNA	Total
10-980	6	2	152,732	14,538	390,907	816,424	40.0	776	40	816
10-1048	6	2	152,735	14,538	392,185	816,465	40.0	777	40	817
10-1059	7	2	98,837	14,538	392,164	816,681	40.0	776	40	816
10-1110	8	2	152,733	20,993	388,970	816,522	40.0	775	40	815
10-1213	5	1	451,397	14,538	451,397	816,521	40.0	772	40	812
10-1257	3	1	702,439	14,562	702,439	816,333	40.0	776	40	816
10-1385	9	3	95,255	14,577	297,117	817,191	40.0	780	39	819
11-107	5	2	249,794	14,538	389,683	816,346	40.0	773	40	813
11-129	6	2	152,693	14,538	392,172	816,432	40.0	775	40	815
11-174	6	2	258,682	13,367	282,196	815,686	40.0	776	39	815
11-212	7	2	152,734	14,538	389,655	816,503	40.0	778	40	818
11-473	6	2	152,734	14,538	389,647	816,518	40.0	778	40	818
11-634	7	2	152,735	14,775	391,525	816,551	40.0	777	40	817
11-949	6	2	258,658	13,367	283,608	817,102	40.0	784	39	823
11-994	5	2	249,776	14,538	389,685	816,304	40.0	776	40	816
11-1384	6	2	258,694	13,367	283,575	818,669	40.0	787	39	826
12-060	6	2	152,734	14,538	392,205	816,506	40.0	775	40	815
12-091	6	2	152,734	14,538	391,968	816,510	40.0	777	40	817

14-637	6	2	156,124	60,136	298,090	818,560	40.0	789	39	828
15-215	6	2	152,734	14,561	392,183	816,388	40.0	775	40	815
15-885	6	2	152,734	14,561	389,671	816,420	40.0	776	40	816
15-969	6	2	152,735	14,538	392,144	816,389	40.0	780	40	820
15-982	5	2	156,554	14,538	390,947	816,495	40.0	769	40	809
16-002	6	2	152,736	14,538	389,658	816,530	40.0	773	40	813
16-004	6	2	152,736	14,538	392,133	816,561	40.0	777	40	817
16-032	6	2	152,734	14,538	392,119	816,471	40.0	772	40	812
16-118	5	1	443,549	14,538	443,549	816,467	40.0	775	40	815
16-462	5	2	152,735	57,889	392,162	816,525	40.0	776	40	816
16-710	7	2	152,734	14,538	392,162	816,537	40.0	773	40	813
16-734	6	2	258,694	13,367	283,522	818,445	40.0	784	39	823

Table 2. Variant patterns relative to the nucleotide and amino acid structure of M129 reference strain

	upstream	synonymous	missense	splice	start/stop	in-frame	frameshift	Total
10-980	37	32	48		4	3	16	140
10-1048	89	105	153		13	6	25	391
10-1059	93	100	149		11	7	29	389
10-1110	56	31	49		5	2	16	159
10-1213	93	102	154		16	7	25	397
10-1257	92	95	151		15	5	25	383
10-1385	518	480	659	1	56	9	55	1778
11-107	114	107	172		15	9	23	440
11-129	96	113	160		13	6	28	416
11-174	518	479	658	1	57	11	54	1778
11-212	118	108	154		13	7	25	425
11-473	116	97	141		15	5	25	399
11-634	110	103	154		16	6	25	414
11-949	521	489	665	1	53	9	55	1793
11-994	92	99	151		12	7	24	385
11-1384	519	490	668	1	53	9	56	1796
12-060	119	104	160		15	7	25	430
12-091	130	104	162		16	7	27	446
14-637	518	483	657	1	51	11	59	1782

15-215	95	106	155		13	7	27	403
15-885	130	108	170		15	7	25	455
15-969	114	104	157		14	8	25	422
15-982	142	108	157		14	8	25	454
16-002	92	104	156		12	8	25	397
16-004	116	114	163		14	8	27	442
16-032	121	106	166		17	6	25	441
16-118	126	100	156		14	7	25	428
16-462	128	101	159		14	7	25	434
16-710	115	100	158		14	7	25	419
16-734	519	486	660	1	54	10	55	1785

Table 3. P1 type and MLST type distribution of 30 strains from this study and 48 NCBI strains

		No. of strains (%)		
		This Study	NCBI	Total
P1 type	1	24 (80)	25 (52.1)	49 (62.8)
	2		13 (27.1)	13 (16.7)
	2a	1 (3.3)	4 (8.3)	5 (6.4)
	2b		5 (10.4)	5 (6.4)
	2c	5 (16.7)	1 (2.1)	6 (7.7)
MLST type	ST1	2 (6.7)	12 (25)	14 (17.9)
	ST2		14 (29.2)	14 (17.9)
	ST3	20 (66.7)	7 (14.6)	27 (34.6)
	ST4		1 (2.1)	1 (1.3)
	ST7		3 (6.3)	3 (3.8)
	ST14	5 (16.7)	3 (6.3)	8 (10.3)
	ST15		1 (2.1)	1 (1.3)
	ST16		1 (2.1)	1 (1.3)
	ST17	2 (6.7)		2 (2.6)
	ST19		1 (2.1)	1 (1.3)
	ST20		5 (10.4)	5 (6.4)
	ST33	1 (3.3)		1 (1.3)

Figures

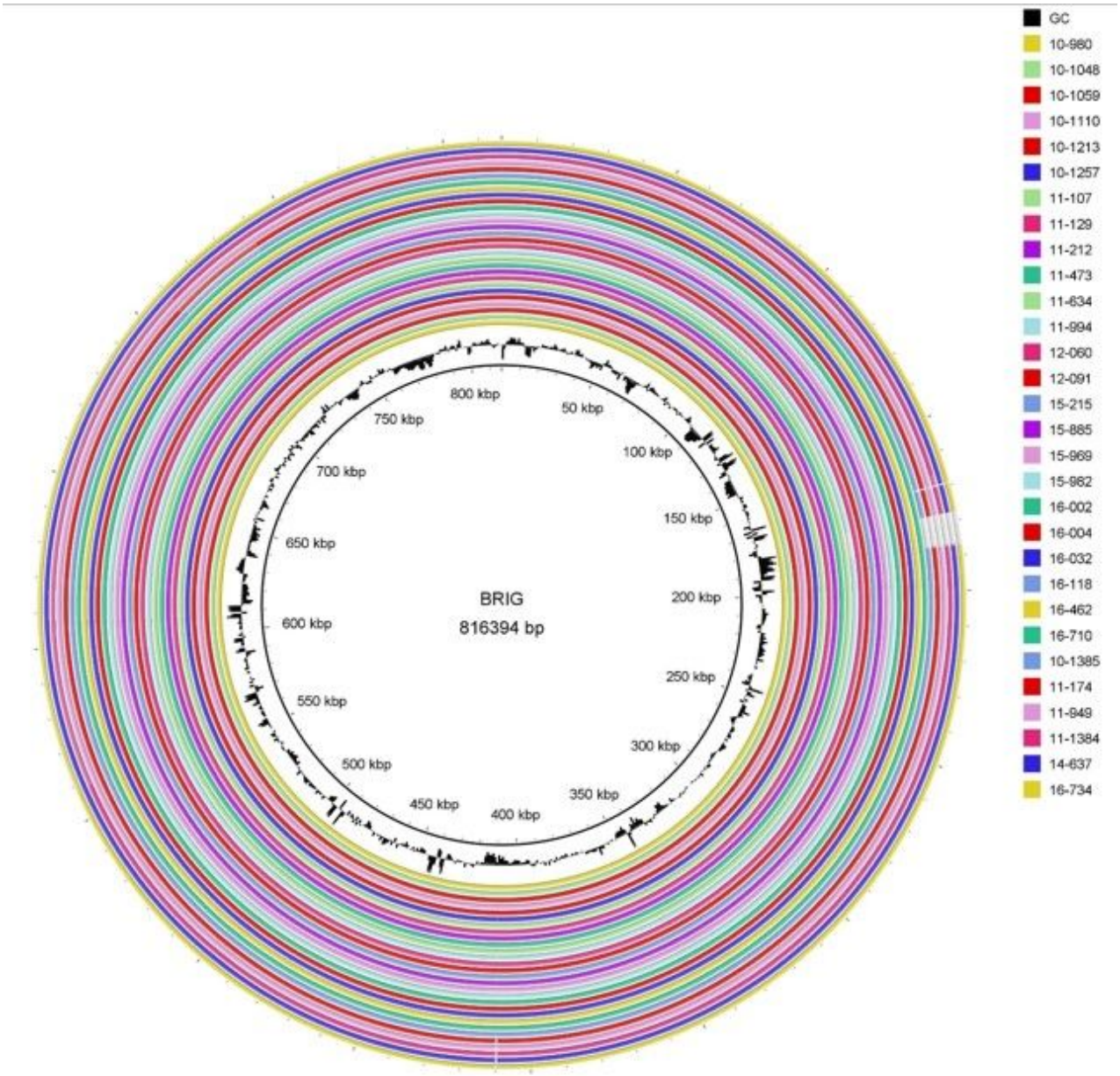

Figure 1

Overall sequence identity of the 30 sequenced strains with the reference M129 genome.

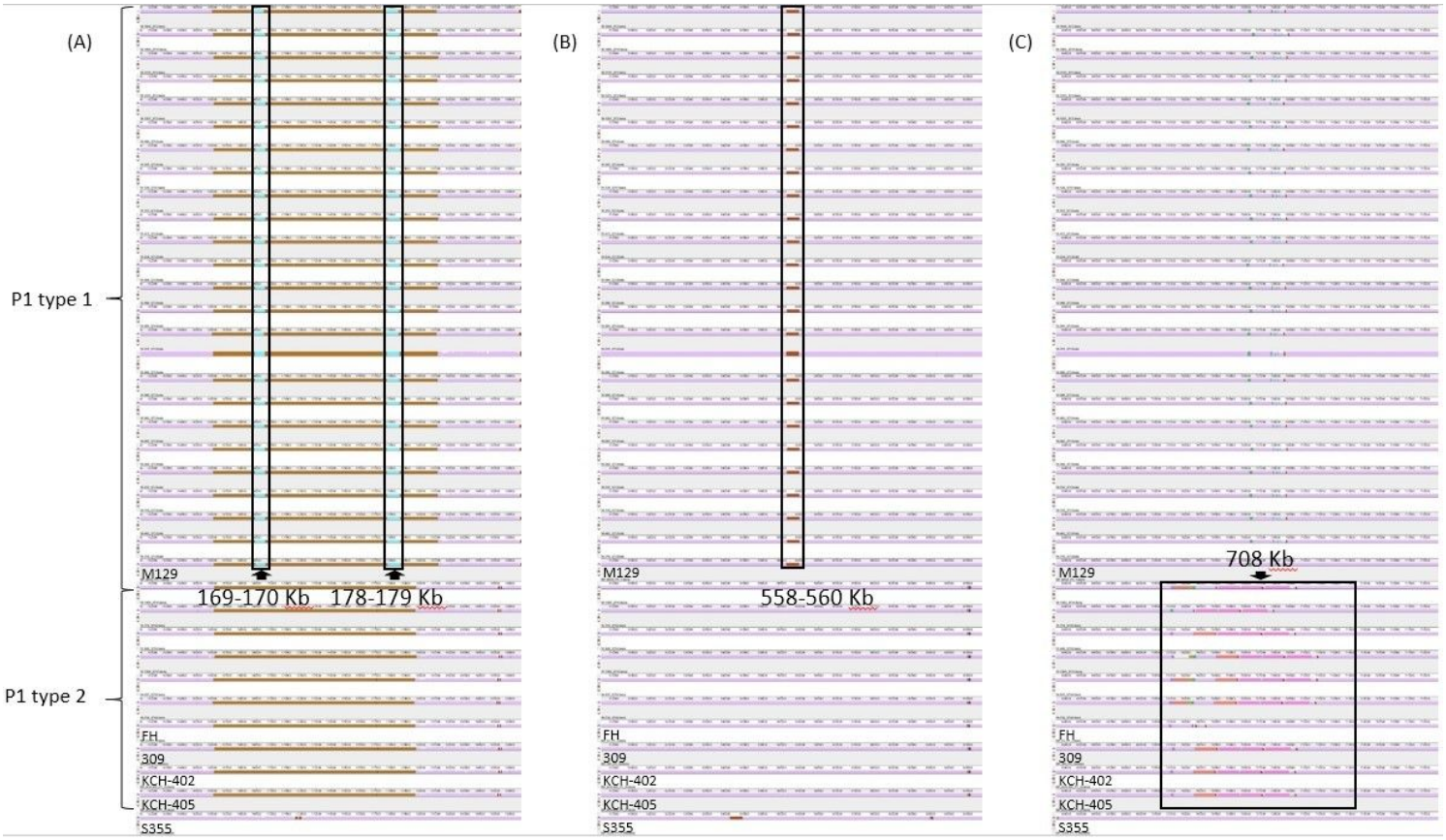

Figure 2

Whole genome alignment of the 30 sequenced strains with 6 reference sequences using MAUVE.

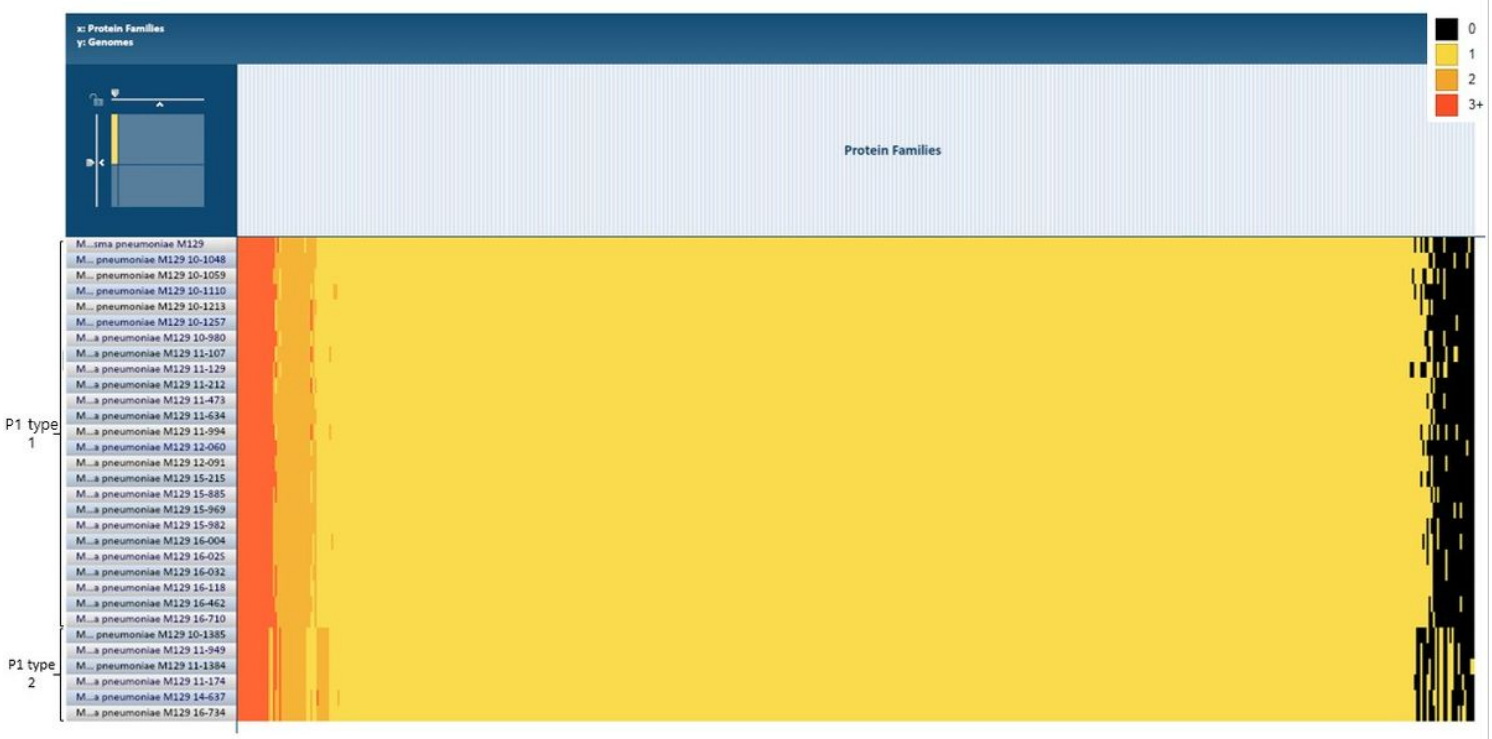

Figure 3

Heatmap of protein families of 30 sequenced genomes with reference genome M. pneumoniae M129.

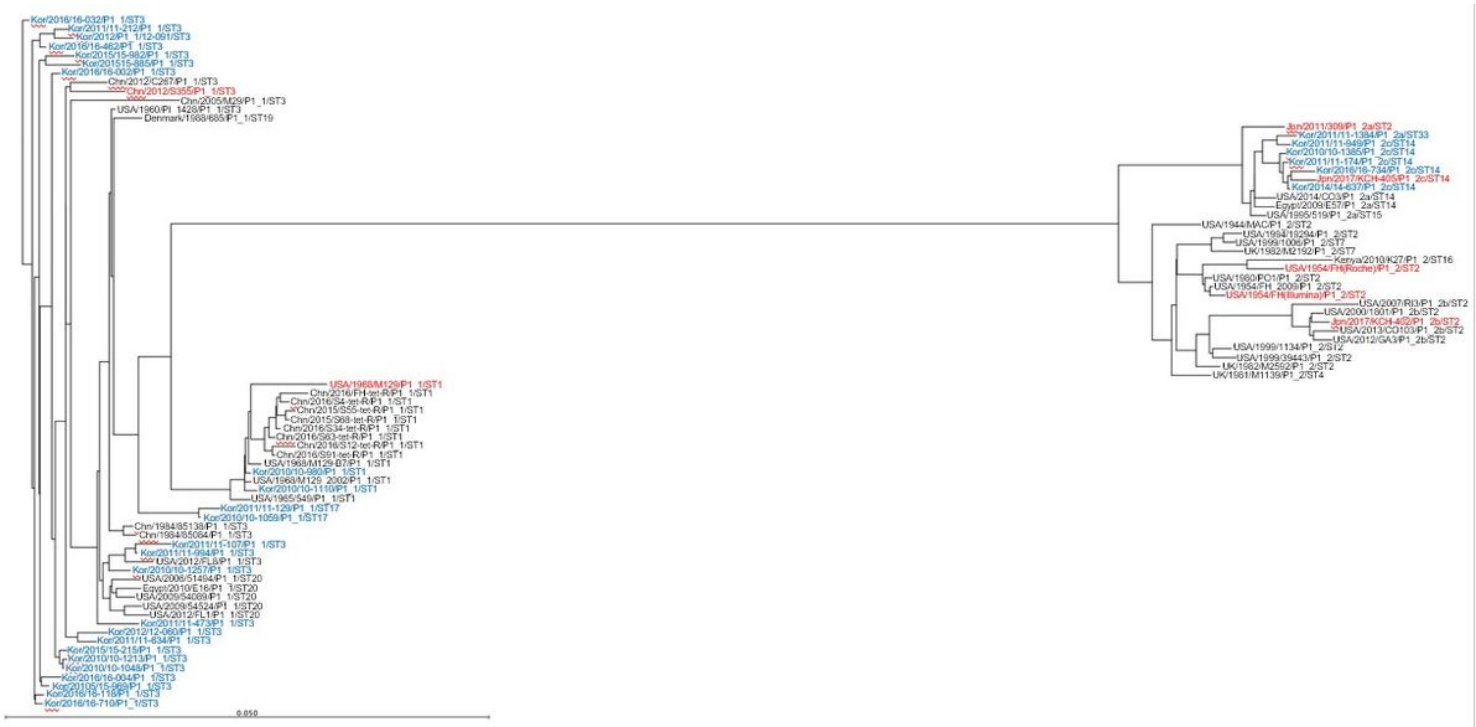

Figure 4

Phylogenetic tree based on whole genome alignment of the 30 sequenced strains with 48 M. pneumoniae genomes accessed from NCBI.

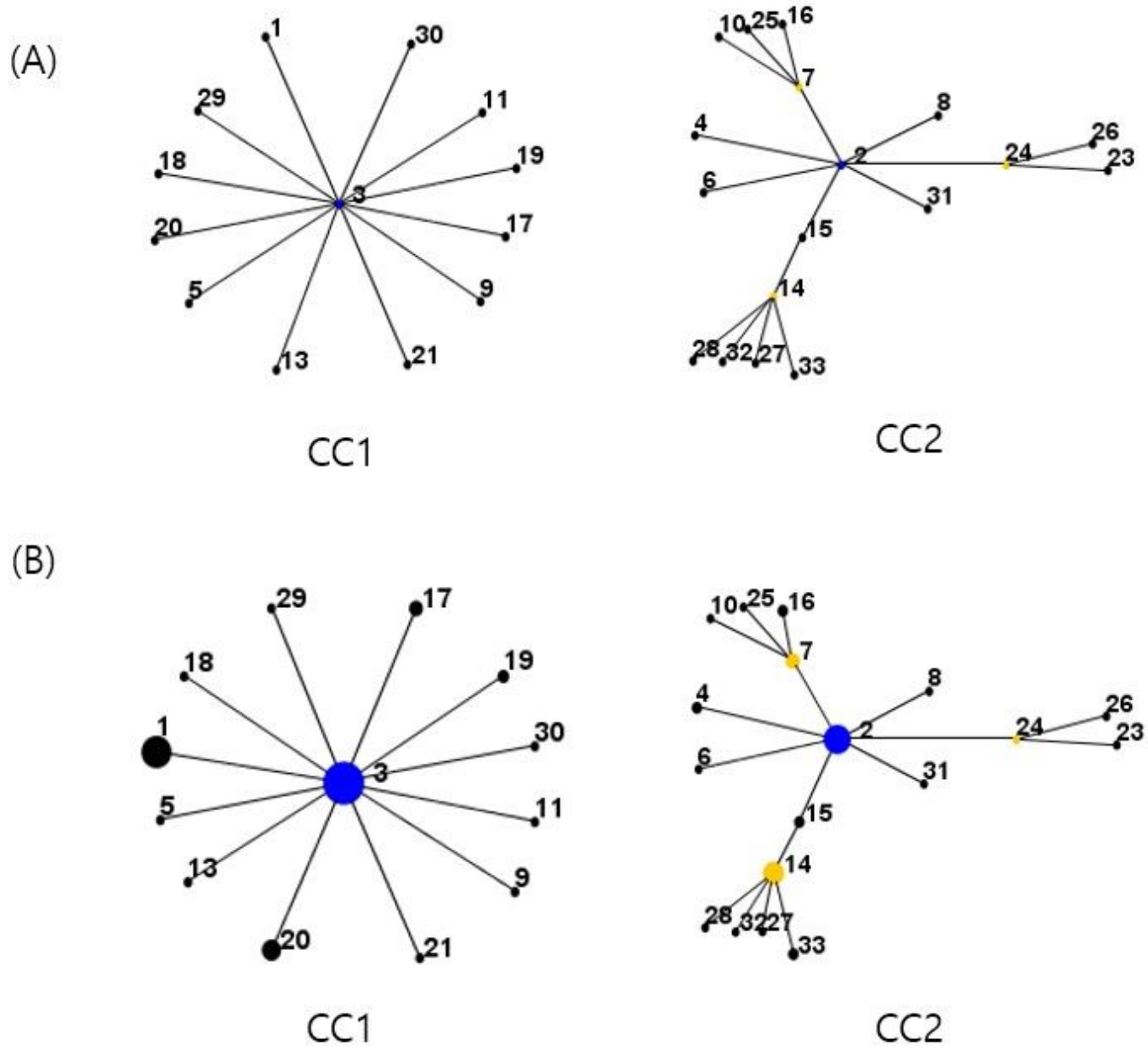

Figure 5

Mycoplasma pneumoniae sequence type (ST) relationship by eBURST analysis.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplement1.xlsx
- supplement2.pdf