Global Genome Diversity and Recombination in *Mycoplasma pneumoniae*

Yu-Chia Hsieh, Shiao-Wen Li, Yi-Yin Chen, Ching-Chia Kuo, Yin-Cheng Chen, Ian Yi-Feng Chang, Yi-Jiun Pan, Ting-Hsuan Li, Ruei-Lin Chiang, Ya-Yu Huang, Wei-Chao Liao

Genomic changes in *Mycoplasma pneumoniae* caused by adaptation to environmental or ecologic pressures are poorly understood. We collected *M. pneumoniae* from children who had confirmed pneumonia in Taiwan during 2017–2020. We used whole-genome sequencing to compare these isolates with a worldwide collection of current and historical clinical strains for characterizing population structures. A phylogenetic tree for 284 strains showed that all sequenced strains consisted of 5 clades: T1–1 (sequence type [ST]1), T1–2 (mainly ST3), T1–3 (ST17), T2–1 (mainly ST2), and T2–2 (mainly ST14). We identified a putative recombination block containing 6 genes (MPN366–371). Macrolide resistance involving 23S rRNA mutations was detected for each clade. Clonal expansion of macrolide resistance occurred mostly within subtype 1 strains, of which clade T1–2 showed the highest recombination rate and genome diversity. Functional characterization of recombined regions provided clarification of the biologic role of these recombination events in the evolution of *M. pneumoniae*.

Mycoplasma pneumoniae, a major pathogen causing community-acquired pneumonia, has a reduced genome of ≈800 kb, which has been relatively stable over time and geographic distance (*1*). Widespread use of macrolides has driven an increase and spread of macrolide-resistant *M. pneumoniae* in several countries, including Taiwan (*2*–*4*). High numbers of repetitive DNA elements (RepMPs), designated RepMP1, RepMP2/3, RepMP4, and RepMP5, comprise ≈8% of the *M. pneumoniae* genome and play essential roles in survival niches of *M. pneumoniae* by engaging in recombination events to generate surface antigen diversity (*5*). Epidemiologic data show that epidemic peaks of *M. pneumoniae* occur every 3–7 years, as indicated by 3 outbreaks in 2011–2012, 2014–2015, and 2015–2016 in Japan and Europe (*6*,*7*). Genotype shifts from 1 P1 adhesin subtype to another occurred repeatedly at an interval of 10 years (*7*).

The relationship between epidemic periodicity and genotype shifts is unclear. P1 adhesin causes antigenic variation between clinical strains as a result of homologous recombination between RepMP2/3 and RepMP4 domains located within their open reading frames and at repetitive DNA elements at other sites in the bacterial genome (*8*). However, P1 adhesin type is not always the determinant factor for cyclic outbreaks because cocirculation of both P1 subtypes and multiple variants in endemic and epidemic settings has been documented (*9*). Given the existence of multiple copies of specific RepMPs dispersed across chromosomes, whether other genetic regions showing recombination diversity involved in *M. pneumoniae* circulation remains unclear.

To study how the *M. pneumoniae* lineage has evolved as it has spread, we characterized the genomes of 99 *M. pneumoniae* strains isolated in Taiwan, together with a global collection of 185 *M. pneumoniae* genomes deposited in public databases. At the same time, we sought to identify major putative recombination hot spots on a genome wide scale to comprehensively understand the significance of recombination on the evolutionary dynamics of *M. pneumoniae*.

Materials and Methods

Isolation of *M. pneumoniae*

We prospectively enrolled children <18 years of age who were hospitalized because of clinical and radiographic pneumonia during April 2017–January 2020 at Chang Gung Memorial Hospital Lin Kou Branch (Taoyuan, Taiwan), Chang Gung Memorial

Author affiliations: Chang Gung University, Taoyuan, Taiwan (Y.-C. Hsieh, S.-W. Li, Y.-C. Chen, I.Y.-.F. Chang, R.-L. Chiang, W.-C. Liao); Chang Gung Children's Hospital, Taoyuan (Y.-C. Hsieh, Y.-Y. Chen, C.-C. Kuo, T.-H. Li, Y.-Y. Huang); Linkou Chang Gung Memorial Hospital, Taoyuan (I.Y.-.F. Chang, W.-C. Liao); China Medical University, Taichung, Taiwan (Y.-J. Pan)

DOI: https://doi.org/10.3201/eid2801.210497

Hospital Kaohsiung Branch (Kaohsiung, Taiwan), and Saint Paul's Hospital (Taoyuan, Taiwan). We obtained written informed consent from patients or their parents. Throat swab specimens were then collected by pediatricians who used sterile swabs (FLOQSwabs; Copan, https://www.copanusa.com). All throat swab specimens were sent to a laboratory at the Chang Gung Memorial Hospital Lin Kou Branch. We inoculated throat swab specimens into SP4 medium and observed spherical *M. pneumoniae* colonies by using microscopy. This study protocol was approved by the research ethics committee of Chang Gung Memorial Hospital, Taiwan (approval nos. 201900420A3 and 202000687B0).

Whole-Genome Sequencing

We obtained whole-genome sequencing (WGS) data for 99 clinical isolates. We cultured *M. pneumoniae* in 75-cm^2 tissue culture flasks (Techno Plastic Products AG, https://www.tpp.ch) containing 10 mL of SP4 medium for 96 hours at 37°C. We extracted *M. pneumoniae* genomic DNA directly from cultures by using the QIAamp DNA Mini Kit (QIAGEN, https://www.qiagen.com), and prepared libraries for WGS by using NEBNext Ultra II DNA Library Prep Kits (Illumina, https://www.illumina.com).

Prediction of Recombination Sites by Phylogenetic Analysis

We combined 185 isolates that had available WGS data, downloaded from the RefSeq database of the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov), and our assembled 99 clinical isolates (Appendix 1 Table, https://wwwnc.cdc.gov/EID/article/28/1/21-0497-App1.xlsx). We used all 284 *M. pneumoniae* isolates to analyze homologous recombination and genetic diversity as described (*10*).

Clade-Specific Recombination Gene Sequence Analysis

We used Gubbins (https://sanger-pathogens.github.io/gubbins) to identify areas that were likely introduced by homologous recombination involving *M. pneumoniae* clades. We scored each gene with the number of predicted recombination events in each clade, and used R version 3.6.3 (R Foundation for Statistical Computing, https://cran.r-project.org) to draw a heatmap by using the pheatmap package.

Estimation of Synonymous and Nonsynonymous Substitution Rates

We annotated 284 *M. pneumoniae* by using Prokka version 1.14.6 (*11*) and identified core genes by using Roary version 3.13.0 (*12*) with default settings. We aligned sequences of genes by using ClustalW2 (*13*) and converted protein and DNA multiple sequence alignments into codon alignments by using PAL2NAL (*14*). We used the output data file to estimate synonymous and nonsynonymous substitution rates (K_a/K_s) in different branches of the phylogenetic tree by using a branch-specific algorithm in the codeml program from PAML version 4.8 (*15*), for which we provide additional details (Appendix 2, https://wwwnc.cdc.gov/EID/article/28/1/21-0497-App2.pdf). We used M129 as a reference genome to describe the location of RepMPs based on the NCBI annotation file.

Tip-Dating Analysis

We used whole-genome alignments to construct a phylogenetic tree by using RAxML (https://cme.h-its.org/exelixis/software.html) with the GTRCAT model. We performed root-to-tip regression analysis by using TempEst version 1.5.3 (*16*). To estimate the phylodynamics of *M. pneumoniae*, we performed temporal analysis using the BEAST version 1.10.4 software package (*17*). We used a strict clock exponential model and a Hasegawa–Kishono–Yano nucleotide substitution model with a discretized gamma distribution to measure rate heterogeneity across sites. We summarized tree data to generate a maximum clade credibility tree by using TreeAnnotator (https://beast.community/treeannotator#user-interface) and visualized it by using FigTree version 1.4.4 (*1*).

Population Structure Analysis

We used PLINK version 1.9 (*18*) to generate a binary output from the vcf file. We performed population structure inference by using ADMIXTURE version 1.3.0 (*19*).

Results

Genotype Distribution

From the NCBI database, we collected *M. pneumoniae* strains from China (n = 23), Japan (n = 70), South Korea (n = 30), the United States (n = 29), the United Kingdom (n = 4), Denmark (n = 3), Spain (n = 1), France (n = 15), Germany (n = 2), Guatemala (n = 1), Egypt (n = 2), Kenya (n = 3), and Tunisia (n = 2) obtained during 1944–2016 (Figure). These strains included P1 subtype 1 sequence type (ST) 3, prevalent worldwide; P1 subtype 2 ST2 from Japan, the United States, and countries in Europe; P1 subtype 1 ST1 from the United States, South Korea, China, and Tunisia; P1 subtype 1 ST17 from South Korea; P1 subtype 2 ST7 from Japan; and P1 subtype 2 ST14 from South Korea, Japan, France, and the United States.

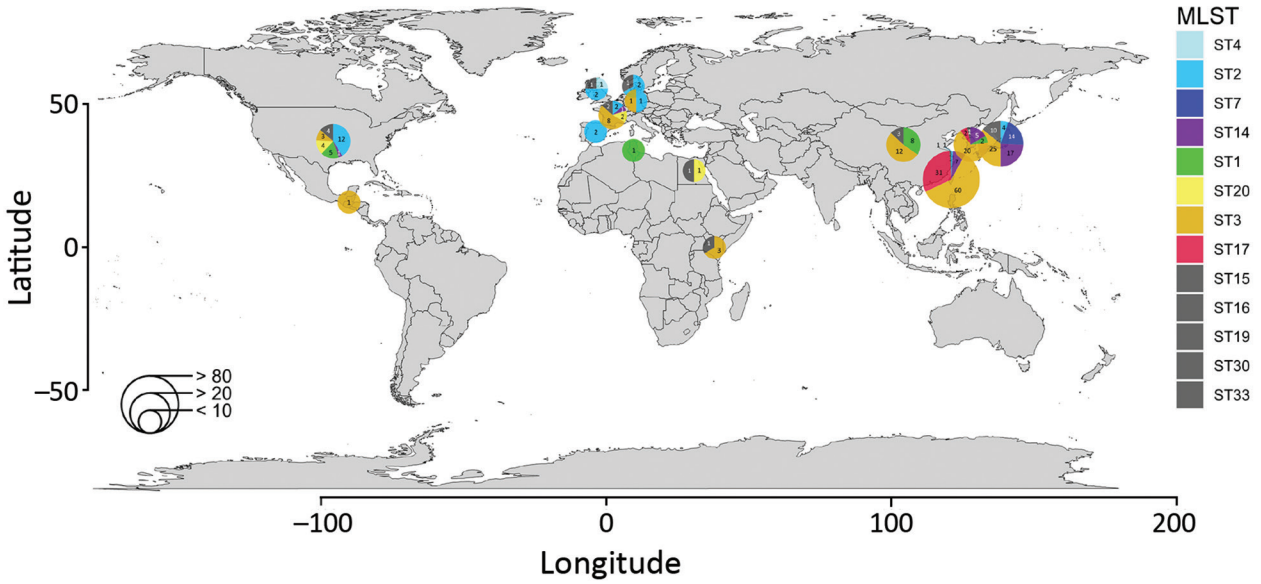


Figure. Genotype and origin of *Mycoplasma pneumoniae* genome data compared in study of global genome diversity and recombination. Pie charts indicate number of isolates in indicated countries. Detailed information regarding each sample is provided in Appendix 1 Table 1 (https://wwwnc.cdc.gov/EID/article/28/1/21-0497-App1.xlsx). MLST, multilocus sequence typing; ST, sequence type.

We observed macrolide resistance with the A2063G mutation among ST3 strains from China, South Korea, Japan, and France; ST1 strains from China; ST7 strains from Japan; and ST14 strains from South Korea and Japan. A total of 99 strains from Taiwan collected during 2017–2020 consisted of P1 subtype 1 ST3 (n = 60, 60.6%), P1 subtype 1 S17 (n = 31, 31.3%), P1 subtype 2 ST14 (n = 7, 7.1%), and P1 subtype 2 ST2 (n = 1, 1%). Macrolide resistance with A2063G was also observed in ST3 (n = 59), ST17 (n = 14), and ST14 (n = 3). A2063T was observed in ST17 (n = 10), A2064G in ST3 (n = 1).

Recombination Landscape

We inferred genomic regions identified by Gubbins as having densities of single-nucleotide variants (SNVs) that were different from background SNV densities in the core genome to be homologous recombination events. By examining all 284 *M. pneumoniae* genomes in this study, we detected 108 putative recombination blocks (Appendix 2 Figure 1, panel A) spanning an average of 1.3 kb/recombination event (95% CI 0.9–1.7 kb) and covering an average of 10 kb/isolate (1.3% of the genome).

Phylogenetic analysis using core genome single-nucleotide polymorphisms (SNPs) identified 5 clades linked to multilocus sequence typing structures and associated with P1 subtypes, which resembled previous findings (*1,7,20*). Clade T1–1 consisted of ST1; clade T1–2 consisted of ST3, ST19, ST20, and ST30; clade T1–3 consisted of ST17; clade T2–1 consisted of ST2, ST4, ST7, and ST16; clade T2–2 consisted of ST2, ST14, ST15, and ST33. Clades T1–1, T1–2, and T1–3 belonged to P1 subtype 1, and clades T2–1 and T2–2 belonged to P1 subtype 2.

With regard to switching from subtype 1 to subtype 2, two regions of SNPs were predicted to have arisen by homologous recombination: first, 183717–186553 bp sequence coordinates containing P1 adhesin (MPN141) for T2–1 and 183381–187741 bp for T2–2; and second, 437080–443621 bp sequence coordinates containing hypothetical proteins (MPN366–367), RepMP1, DUF16 domain-containing protein (MPN368), lipoprotein (MPN369), putative adhesin (MPN370), and MPN371 for both clades T2–1 and T2–2. Among strains of subtype 1, clade T1–2 had 3 regions of SNPs predicted to have arisen by homologous recombination: first 34,0697–340757 bp sequence coordinates containing *prrB* (MPN285); second, 493781–493791 bp sequence coordinates containing adhesin (MPN409); and third, 570767–570790 bp sequence coordinates containing adhesin (MPN468) (Appendix 1 Table 2; Appendix 2 Figure 1, panel A).

The MPN409 recombined region was more frequently detected in most macrolide-resistant clade T1–2 strains than in macrolide-susceptible clade T1–2 strains. We ran Gubbins independently for subtype 2 (88 isolates) based on the core genome sequence alignment involving *M. pneumoniae* subtype 2 reference strain FH (GenBank accession no. CP010546.1). Among

strains of subtype 2, only 1 recombination hotspot was predicted: 179477–183466 bp sequence coordinates containing P1 adhesin (equal to *MPN141* of M129) (Appendix 1 Table 2; Appendix 2 Figure 1, panel B).

We estimated 2 measures of the recombination rate in the core genome alignment by using Gubbins ρ/θ and r/m ratios (Table). We provide boxplots of putative recombination statistics for *M. pneumoniae* clades (Appendix 2 Figure 2). ρ/θ estimates the relative frequency of occurrence of recombination and mutation in the history of the clade. The overall ρ/θ ratio for each clade ranged from 0 (clade T2–1) to 0.03 (clade T1–2), indicating that putative recombination events have occurred less frequently than vertically inherited mutations in all 5 clades, and lower than in most microbes. The r/m index assesses the relative effect of recombination and mutation on genetic diversification of the clade. Predicted recombination occurred more frequently in clade T1–2, for which the r/m rate was 4–13 times higher than for the other 4 clades.

Clade-Specific Gene Recombination

We defined clade-specific genes involved in these putative recombination events by clustering of genes by recombination frequency in each clade (Appendix 2 Figure 3, panel A). These data can guide identification of multiple genes under different selective pressures in each clade. A set of 8 genes (*MPN141, MPN142, MPN366, MNP368, MPN369, MPN370, MPN_RS02085*, and *MPN_RS02055*) appeared to increase the recombination frequency in *M. pneumoniae* subtype 2, and extensive DNA repetitive elements are located in each of these genes. *MPN366, MPN368, MPN369*, and *MPN370* were under even greater selective pressure than *MPN141* and *MPN142* (Appendix 2 Figure 3, panel A).

We investigated adhesin P1 sequence variation by calculating homology with the reference M129 strain DNA sequence (Appendix 2 Figure 3, panel B). Two variable regions provided most variation between *M. pneumoniae* subtypes 1 and 2, and were located at ≈3000 bp and ≈1000 bp within *MPN141*. In contrast, these 2 regions were highly conserved within clades T1–1, T1–2, and T1–3. For subtype 2, clade T2–2 had less conserved sequence than clade T2–1.

Two additional regions of nucleotide sequence differed among strains belonging to clade T2–2 compared with M129. Sequence analysis of the P1 gene showed divergent sequences from 5 clades at position 2000 bp, which might be an AGT trinucleotide variable-number tandem repeat (*21*).

Genome Diversity within Subclades

To clarify this potentially confounding within–clade mutation, we analyzed the genetic diversity of the collection by estimating the nonsynonymous to synonymous (K_a/K_s) mutation ratio. We assessed divergence along the genome between clades by determining synonymous substitution rates, a near-neutral indicator of genetic divergence. Data showed that different genomic regions had higher divergence among those 5 clades. Clade T2–1 had more divergent regions than clade T2–2, and clade T2–2 had higher K_s values in less divergent regions. The genomic divergence pattern of clade T1–2 was observed across the genome. This result indicated that clade T1–2 was the most divergent in *M. pneumoniae* subtype 1, and the genome of clade T1–3 was stable without evidence of divergence (Appendix 2 Figure 4, panel A). We also identified numerous diverse regions located near RepMP elements across the genome. We further investigated the function of these genomic regions that showed higher divergence. From a global perspective, numerous positions are located in genes that are annotated as adhesin and MgPa (*mgp*) operon of *M. genitalium* coding for proteins (Appendix 2 Figure 4, panel B).

Positive Selection Detection Involving Orthologous Genes

To track genomic evolutionary footprints, we used codeml analysis to test genes for evidence of positive selection. We detected 42 genes as being under positive selection by using likelihood ratio tests (Appendix 1 Table 3). Of these positively selected genes, 21 were core orthologous loci present among all 284 genomes investigated. To determine the relationship of genes with certain evolutionary features, we classified these genes into different functional categories by using clusters of orthologous groups of proteins

Table. Putative recombination statistics for *Mycoplasma pneumoniae* clades for study of global genome diversity and recombination*

Clade	No. strains	Total SNPs	No. SNPs inside recombination	No. SNPs outside recombination	No. recombination blocks	r/m	ρ/θ
T1–1	17	10.29 ± 23.44	5.65 ± 20.62	4.65 ± 6.78	0.18 ± 0.53	0.03 ± 0.14	0.01 ± 0.02
T1–2	146	9.78 ± 12.41	1.76 ± 4.06	8.02 ± 10.88	0.24 ± 0.55	0.25 ± 0.91	0.03 ± 0.12
T1–3	33	3.00 ± 5.70	0.18 ± 1.04	2.82 ± 5.53	0.03 ± 0.17	0.04 ± 0.21	0.01 ± 0.03
T2–2	46	14.46 ± 13.65	0.33 ± 1.71	14.13 ± 13.28	0.04 ± 0.21	0.02 ± 0.07	0 ± 0.01
T2–1	42	12.93 ± 16.71	1.98 ± 5.13	10.95 ± 12.59	0.26 ± 0.63	0.07 ± 0.17	0.01 ± 0.02

*Values are mean ± SD. r/m assesses the relative effect of recombination and mutation on genetic diversification of the clade. ρ/θ estimates the relative frequency of occurrence of recombination and mutation in the history of the clade. SNP, single-nucleotide polymorphism.

annotation. There are a large number of genes in the functional categories translation, ribosomal structure, and biogenesis and cell wall/membrane/envelope biogenesis. The cell wall/membrane/envelope biogenesis category includes membrane-associated genes under positive selection, consistent with observations for other bacterial species (*22*). The translation, ribosomal structure, and biogenesis category, including ribosomal biogenesis genes, might be correlated with drug resistance.

Evolution Timeline

We calculated the temporal history of *M. pneumoniae* by using Bayesian evolutionary analysis sampling trees to place the most recent common ancestor in ≈1300 (95% highest posterior density 1230–1392) by exponential growth analysis (Appendix 2 Figure 5). The maximum clade credibility tree also showed 2 distinct clades that are consistent with the 2 subtypes for P1 gene variation. This analysis suggested that the 2 subtypes might have diverged at approximately the same time as the most recent common ancestor arose. The evolutionary divergence time estimates for subtypes 1 and 2 for the 5 clades were from 1800 to 1900 (subtype 1, 1823–1865; subtype 2, 1811–1855). *M. pneumoniae* subtypes 1 and 2 spread rapidly within 20 years (through the year 2000).

Population Structures and Sequences

Genetic history can be established by examining the patterns of shared genetic variation between ≥2 populations. Admixture proportion inference is an algorithm used to infer the proportions of ancestry from each source population based on shared genetic drift. These proportions can produce visual summaries that identify a population structure. We used ADMIXTURE to fit a model which the genome is composed of sites from $K = 2$ to $K = 10$ ($K = n$, where n is the number of ancestral generations) for ancestral populations. We provide ADMIXTURE analysis with estimated proportions of 7 ancestral populations (Appendix 2 Figure 6).

We provide unsupervised models assuming $K = 4$ and $K = 7$ ancestral clusters (Appendix 2 Figure 6, panel A). At $K = 7$, it is difficult to discern the major contributors to each clade, which is likely to result in a disrupted relationship between clades T2–1 and T1–1 and other clades. Because each column represents 1 isolate, we show data for admixed strains observed particularly in clades T1–2 and T2–2, reflecting the divergence of the 2 clades (Appendix 2 Figure 6, panel A).

To further clarify the genetic track, we mapped populations to geographic region (Appendix 2 Figure 6, panel B). Each pie chart shows the mean composition of 7 populations in each region. The distribution shows genetic diversity related to geography. Thus, *M. pneumoniae* has the potential to be transmitted efficiently and spread regionally from Europe to America, and subsequently emerge rapidly in Asia. Genome sequences generated during this study were deposited into the NCBI BioProject database (accession no. PRJNA699672).

Discussion

According to our predictions, *M. pneumoniae*, a major cause of atypical bacterial pneumonia described in the 1930s, might have originated in Europe, and has existed for >200 years. Our analysis showed that *M. pneumoniae* genome evolution by recombination varies by clade, as well as recombined regions dependent on clades. In addition to the well-known P1gene locus, there are recombined regions focused around surface protein genes, which are believed to be targets of natural selection and are involved in affecting *M. pneumoniae* population structures. Adherence to cells of the respiratory tract is considered a prerequisite for colonization and pathogenesis by *M. pneumoniae*, which adhere to host target cells by a polar structure known as the attachment organelle (*23*). On the surface of this organelle, P1 and 2 accessory proteins P40/P90 form a transmembrane adhesion complex, which is directly involved in receptor binding (*23*).

Our clinical and epidemiologic information for *M. pneumoniae* is based primarily on typing of the P1 gene. Two major P1 genotype strains (1 and 2) represent evolutionarily divergent clades. In Japan, genotype shifts between P1 subtypes 1 and 2 driven by the human immune system repeatedly occurred at 10-year intervals with transitionary 2–3 years intervals during the past 40 years, including a recent shift from P1 subtype 1 in the 2000s to subtype 2 in the later 2010s (*7*). In this study, by comparing subtype strains 1 and 2, we identified a new predicted recombination block located on *MPN366–371*. Apart from *MPN366* and *MPN367*, which were predicted as pseudogenes, MPN369 lipoprotein and MPN370 are predicted to be expressed as cell-surface proteins.

Theoretically, the increase in frequency of recombination involving these regions might be caused by selective advantages offered by divergent sequences introduced by putative recombination events. The biologic role of putative recombination events in *MPN366–371* is worthy of further investigation. These loci might be associated with antigenic variation and

adherence to promote the persistent circulation of *M. pneumoniae* lineages in human populations.

RecA, single-stranded DNA-binding protein, RecU, RuvA, and RuvB are enzymes responsible for DNA recombination in *M. pneumoniae* based on sequence homology analysis (*24*). Sluijter et al. found that subtype 1 strains harbor a nonsense protein–truncating codon in the *recU* gene. This gene is fully expressed in subtype 2 strains, but there is no detectable enzyme activity under various conditions (*24*). Given that there are more P1 subtype variants among strains of subtype 2 compared with strains of subtype 1 determined from epidemiologic data (*25*), RecU in subtype 2 strains might show enzyme activity under ≥1 cofactors (*24*).

We also found predicted recombination hot spots only in regions flanking the P1 gene but not in other genomic regions among strains of subtype 2. Among strains of subtype 1, we found hot spots in 3 non-P1 gene regions in clade T1–2 (*MPN285*, *MPN409*, and *MPN468*), despite no recombination events involving the P1 gene region. These 3 genes were predicted to express the type I restriction enzyme EcoKI-specificity protein, adhesin, and adhesin P1 family proteins, respectively. Clade T1–2 showed the highest apparent recombination rate compared with the other clades. The protein machinery required for DNA repeat sequence rearrangement might differ by clade, which would account for the observed dissimilar recombination kinetics involving repetitive genetic elements.

Since 2010, macrolide-resistant *M. pneumoniae* have caused several epidemics in Asia. Macrolide resistance involving point mutations in 23S rRNA occurred in each clade. However, in countries that had high rates of macrolide resistance, clonal expansion of ST3 was seen most commonly (*2*). Clade T1–2 (ST3 strains), which appeared during 1960, has predominantly circulated in Europe, the United States, and Asia. Our analysis showed that clade T1–2 had the highest K_s value and greatest genome diversity, which should be associated with the reason why it became an epidemiologically successful genome worldwide even under the selective pressure exerted by macrolide overuse.

Overall, homologous recombination played only a minor role in shaping diversity within *M. pneumoniae*, based on the low ratio of r/m, estimated to be 0.05. However, the highest r/m ratio of clade T1–2 indicated that probable recombination between genetic repetitive elements across chromosomes makes greater contributions to clonal diversification in clade T1–2. Our phylogenetic analysis indicated an additional clade T1–3 (ST17 strains) compared with those obtained in a study in South Korea conducted by Lee et al. (*1*). ST17 is another smaller macrolide-resistant clone that spread in Taiwan (*4*). It is hypothesized that predominance of macrolide-resistant *M. pneumoniae* in subtype 1 is associated with macrolide overuse when subtype 1 is circulating (*7*). During 2017, subtype 2 replaced subtype 1 as the dominant subtype and had decreased macrolide resistance in Japan (*7*), but increased macrolide resistance was detected in China (*26*). It is imperative to continuously monitor spread and evolution of macrolide-resistant subtype 2 strains to obtain additional perspectives on macrolide resistance.

Although this study demonstrates the need for WGS for studying genetic diversity, the first limitation is that the sample size of the genome sequence for studying evolution of *M. pneumoniae* was small. Second, populations of *M. pneumoniae* can be found worldwide, but only 284 isolates were sequenced in our study, which represented only a few countries and lacked global representation, geographic distribution, and evolutionary spread between regions. Third, we predicted recombination events by using Gubbins, which identified loci containing elevated densities of base substitutions while concurrently constructing a phylogeny based on putative point mutations outside these regions. Homologous recombination identified in our study might be a hotspot of SNV occurrence by natural mechanisms. Because detection of recombination is challenging, these results only predict the possibility of recombination by using tools. However, these regions have the potential to undergo recombination events.

In summary, this study comprehensively demonstrated detailed insights into the recombination dynamics within *M. pneumoniae*. Further delineation of the role of homologous recombination in the virulence and adaptation of *M. pneumoniae* to modern environments will provide useful information for public health issues.

Acknowledgments

We thank the Next Generation Sequencing Core, the Bioinformatics Core, and the Molecular Medicine Research Center, Chang Gung University, Taoyuan, Taiwan, for analyzing whole-genome sequences and providing assistance.

This study was supported by the Chang Gung Memorial Hospital, Taiwan (grants CMRPG3K1141 and CMRPG3J0751-0752). Whole-genome sequencing was supported by grants CLRPD1J0012-0013 and EMRPD1L0351.

About the Author

Dr. Hsieh is a professor in the Department of Pediatrics, Chang Gung Children's Hospital; Chang Gung Memorial Hospital; and Chang Gung University College of

Medicine, Taoyuan, Taiwan. Her primary research interest is infectious diseases, particularly the pathogenesis of pneumococcal disease.

References

1. Lee JK, Seong MW, Shin D, Kim JI, Han MS, Yeon Y, et al. Comparative genomics of *Mycoplasma pneumoniae* isolated from children with pneumonia: South Korea, 2010–2016. BMC Genomics. 2019;20:910. https://doi.org/10.1186/s12864-019-6306-9
2. 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. Emerg Infect Dis. 2018;24:1465–71. https://doi.org/10.3201/eid2408.180081
3. Okada T, Morozumi M, Tajima T, Hasegawa M, Sakata H, Ohnari S, et al. Rapid effectiveness of minocycline or doxycycline against macrolide-resistant *Mycoplasma pneumoniae* infection in a 2011 outbreak among Japanese children. Clin Infect Dis. 2012;55:1642–9. https://doi.org/10.1093/cid/cis784
4. Hung HM, Chuang CH, Chen YY, Liao WC, Li SW, Yi-Feng Chang I, et al. Clonal spread of macrolide-resistant *Mycoplasma pneumoniae* sequence type-3 and type-17 with recombination on non-P1 adhesin among children in Taiwan. Clin Microbiol Infect. 2020.
5. Rocha EP, Blanchard A. Genomic repeats, genome plasticity and the dynamics of *Mycoplasma* evolution. Nucleic Acids Res. 2002;30:2031–42. https://doi.org/10.1093/nar/30.9.2031
6. Beeton ML, Zhang XS, Uldum SA, Bébéar C, Dumke R, Gullsby K, et al.; ESCMID Study Group for Mycoplasma and Chlamydia Infections (ESGMAC) Mycoplasma pneumoniae subgroup members not listed as an individual author. *Mycoplasma pneumoniae* infections, 11 countries in Europe and Israel, 2011 to 2016. Euro Surveill. 2020;25:1900112. https://doi.org/10.2807/1560-7917.ES.2020.25.2.1900112
7. Kenri T, Suzuki M, Sekizuka T, Ohya H, Oda Y, Yamazaki T, et al. Periodic genotype shifts in clinically prevalent *Mycoplasma pneumoniae* strains in Japan. Front Cell Infect Microbiol. 2020;10:385. https://doi.org/10.3389/fcimb.2020.00385
8. Dandekar T, Huynen M, Regula JT, Ueberle B, Zimmermann CU, Andrade MA, et al. Re-annotating the *Mycoplasma pneumoniae* genome sequence: adding value, function and reading frames. Nucleic Acids Res. 2000;28:3278–88. https://doi.org/10.1093/nar/28.17.3278
9. Diaz MH, Benitez AJ, Winchell JM. Investigations of *Mycoplasma pneumoniae* infections in the United States: trends in molecular typing and macrolide resistance from 2006 to 2013. J Clin Microbiol. 2015;53:124–30. https://doi.org/10.1128/JCM.02597-14
10. Hung HM, Chuang CH, Chen YY, Liao WC, Li SW, Chang IY, et al. Clonal spread of macrolide-resistant *Mycoplasma pneumoniae* sequence type-3 and type-17 with recombination on non-P1 adhesin among children in Taiwan. Clin Microbiol Infect. 20210;27:1169e1–6. https://doi.org/10.1016/j.cmi.2020.09.035
11. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9. https://doi.org/10.1093/bioinformatics/btu153
12. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31:3691–3. https://doi.org/10.1093/bioinformatics/btv421
13. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22:4673–80. https://doi.org/10.1093/nar/22.22.4673
14. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 20064;34(Suppl_2):W609–12.
15. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. 2007;24:1586–91. https://doi.org/10.1093/molbev/msm088
16. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2016;2:vew007. https://doi.org/10.1093/ve/vew007
17. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 2018;4:vey016. https://doi.org/10.1093/ve/vey016
18. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7. https://doi.org/10.1186/s13742-015-0047-8
19. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64. https://doi.org/10.1101/gr.094052.109
20. Diaz MH, Desai HP, Morrison SS, Benitez AJ, Wolff BJ, Caravas J, et al. Comprehensive bioinformatics analysis of *Mycoplasma pneumoniae* genomes to investigate underlying population structure and type-specific determinants. PLoS One. 2017;12:e0174701. https://doi.org/10.1371/journal.pone.0174701
21. 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. J Clin Microbiol. 2011;49:3000–3. https://doi.org/10.1128/JCM.00105-11
22. Cao P, Guo D, Liu J, Jiang Q, Xu Z, Qu L. Genome-wide analyses reveal genes subject to positive selection in *Pasteurella multocida*. Front Microbiol. 2017;8:961. https://doi.org/10.3389/fmicb.2017.00961
23. Vizarraga D, Kawamoto A, Matsumoto U, Illanes R, Pérez-Luque R, Martín J, et al. Immunodominant proteins P1 and P40/P90 from human pathogen *Mycoplasma pneumoniae*. Nat Commun. 2020;11:5188.
24. Sluijter M, Kaptein E, Spuesens EB, Hoogenboezem T, Hartwig NG, Van Rossum AM, et al. The *Mycoplasma genitalium* MG352-encoded protein is a Holliday junction resolvase that has a non-functional orthologue in *Mycoplasma pneumoniae*. Mol Microbiol. 2010;77:1261–77. https://doi.org/10.1111/j.1365-2958.2010.07288.x
25. Waites KB, Xiao L, Liu Y, Balish MF, Atkinson TP. *Mycoplasma pneumoniae* from the respiratory tract and beyond. Clin Microbiol Rev. 2017;30:747–809. https://doi.org/10.1128/CMR.00114-16
26. Zhao F, Liu J, Shi W, Huang F, Liu L, Zhao S, et al. Antimicrobial susceptibility and genotyping of *Mycoplasma pneumoniae* isolates in Beijing, China, from 2014 to 2016. Antimicrob Resist Infect Control. 2019;8:18. https://doi.org/10.1186/s13756-019-0469-7

Address for correspondence: Wei-Chao Liao, Molecular Medicine Research Center, Chang Gung University, No. 259, Wenhua 1st Rd, Guishan District, Taoyuan City 333, Taiwan; email: pettliao@mail.cgu.edu.tw

Article DOI: https://doi.org.10.3201/eid2801.210497

Global Genome Diversity and Recombination in *Mycoplasma pneumoniae*

Appendix 2

Supplementary Methods

Whole-Genome Sequencing

Library samples were multiplexed and sequenced on an Illumina (https://www.illumina.com) NextSeq 500 Platform with 2 × 150-bp paired-end reads. Adaptor and poor quality sequences were trimmed by using BBduk (*1*), and genomes were assembled de novo by using SPAdes version 3.13.0 (*2*) with k-mer values of 127 and in careful mode. This genome assembly, which was evaluated by using QUAST (*3*), produced an average of 18 contigs with an N50 length of 143,785 bp (Appendix 1, Table 4, https://wwwnc.cdc.gov/EID/article/28/1/21-0497-App1.xlsx). In silico multilocus sequence typing, a typing technique more discriminatory than multiple locus variable-number tandem repeat analysis and P1 typing (*4*), was performed by using a BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi)–based tool (https://github.com/tseemann/mlst) on de novo genome assemblies (*5*).

Prediction of Recombination Sites by Phylogenetic Analysis

Variants were called by using the *Mycoplasma pneumoniae* M129 genome (GenBank accession no. NC_000912.1), a macrolide-sensitive sequence type 1 strain isolated in the United States during 1968, as reference by using Snippy version 4.4.5 (*6*), which uses BWA Mem (https://github.com/lh3/bwa) to map the reads or sequences to the reference, and then subsequently calls single-nucleotide polymorphisms and insertions/deletions (indels) with FreeBayes (https://github.com). Reference-based

alignments of the core genome output from Snippy (Snippy-Core) were used for downstream phylogenetic analysis. Recombination sites were assessed by using Gubbins version 2.3.4 (*7*), and a maximum-likelihood tree was generated by using RAxML version 8.2.10 (*8*). RAxML was implemented by using the GTRCAT model, with no rate heterogeneity or ascertainment bias, rapid hill climbing, and default parameters. After 4 iterations, Gubbins reached a stable tree topology, including inferred regions of hypothetical genetic recombination events. The resulting phylogenetic tree, isolate metadata, core genome single-nucleotide polymorphisms, and recombination sites were visualized by using Phandango version 1.3.0 (*9*).

Lineage-Specific Recombination Gene Sequence Analysis

In this study, to explore *MPN141* gene divergence, 148 isolates could be extracted from the *MPN141* sequence and aligned by using MUSCLE version 3.8.31 (*10*). Pairwise differences were calculated by comparison to the reference M129 stain and used a 25-bp sliding window with incremental 1-bp step sizes.

Estimation of Synonymous and Nonsynonymous Substitution Rates

Codon frequencies were fixed by using the F3X4 model and a branch-site alternative model (model = 2, NS sites = 2) with settings according to Huang et al. (*11*). To detect genes under positive selection, we applied the codeml algorithm to estimate a single ω for all sites in the alignment. Three selection models were generated with initial ω (K_a/K_s) values of 0.5, 1, and 1.5, and we fixed ω at 1 as the neutral model. To identify genes with significance, we performed likelihood ratio tests between the neutral model (M1) and selection models (M2) with adjusted p values <0.05. To visualize the distribution of K_s divergence as genome-wide views involving 5 lineages, we calculated the mean value of by K_s by using a 25-gene window.

Population Structure Analysis

To determine the ideal number of populations, we assumed K from 2 to 10, and calculated cross-validation and likelihood statistics. The K that yielded the smallest cross-validation value and the highest likelihood index was chosen to execute the admixture.

References

1. Bushnell B. BBMap [cited 2021 Sep 2]. https://sourceforge.net/projects/bbmap

2. 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;19:455–77. PubMed https://doi.org/10.1089/cmb.2012.0021

3. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072–5. PubMed https://doi.org/10.1093/bioinformatics/btt086

4. Brown RJ, Holden MT, Spiller OB, Chalker VJ. Development of a multilocus sequence typing scheme for molecular typing of *Mycoplasma pneumoniae.* J Clin Microbiol. 2015;53:3195–203. PubMed https://doi.org/10.1128/JCM.01301-15

5. Seemann T. Scan contig files against PubMLST typing schemes: tseemann/mlst, 2018 [cited 2021 Sep 2]. https://githubcom/tseemann/mlst

6. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60. PubMed https://doi.org/10.1093/bioinformatics/btp324

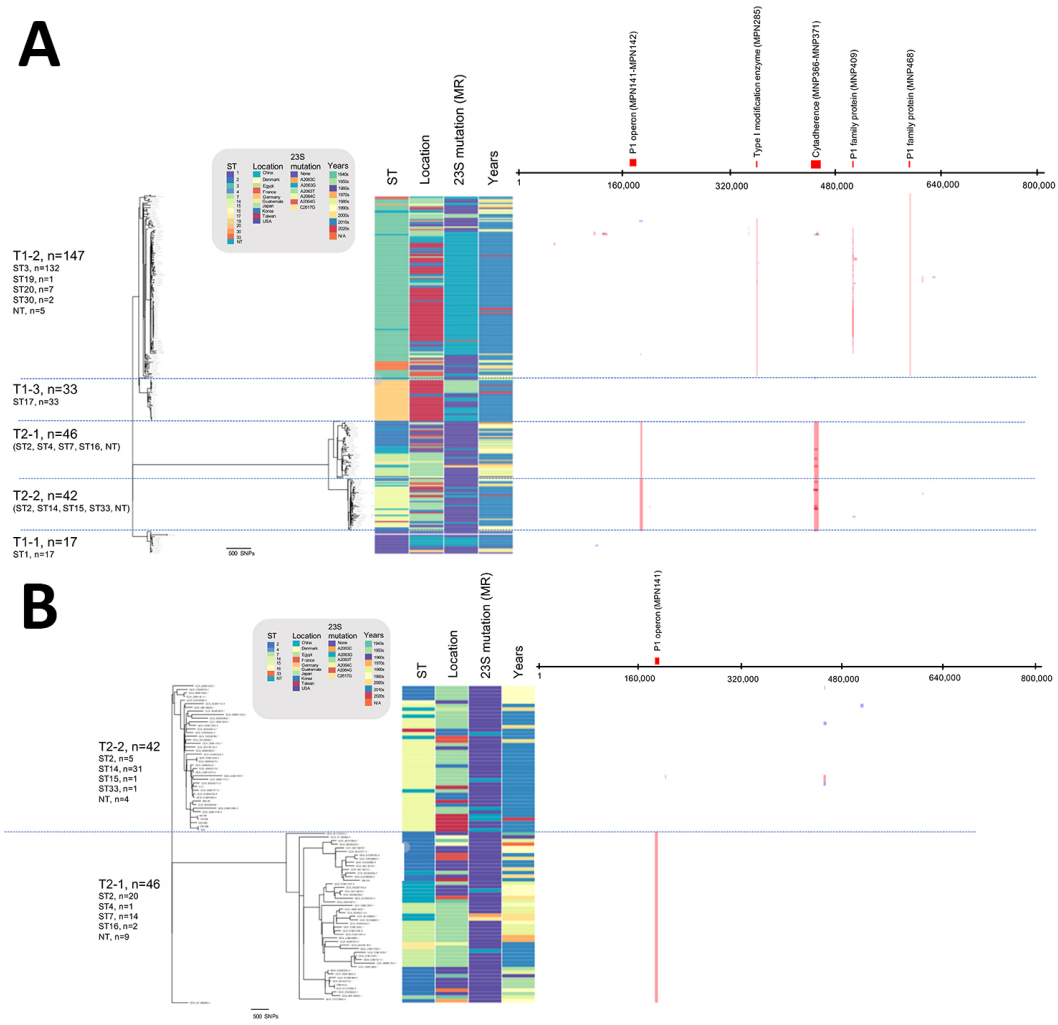
7. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015;43:e15. PubMed https://doi.org/10.1093/nar/gku1196

8. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006;22:2688–90. PubMed https://doi.org/10.1093/bioinformatics/btl446

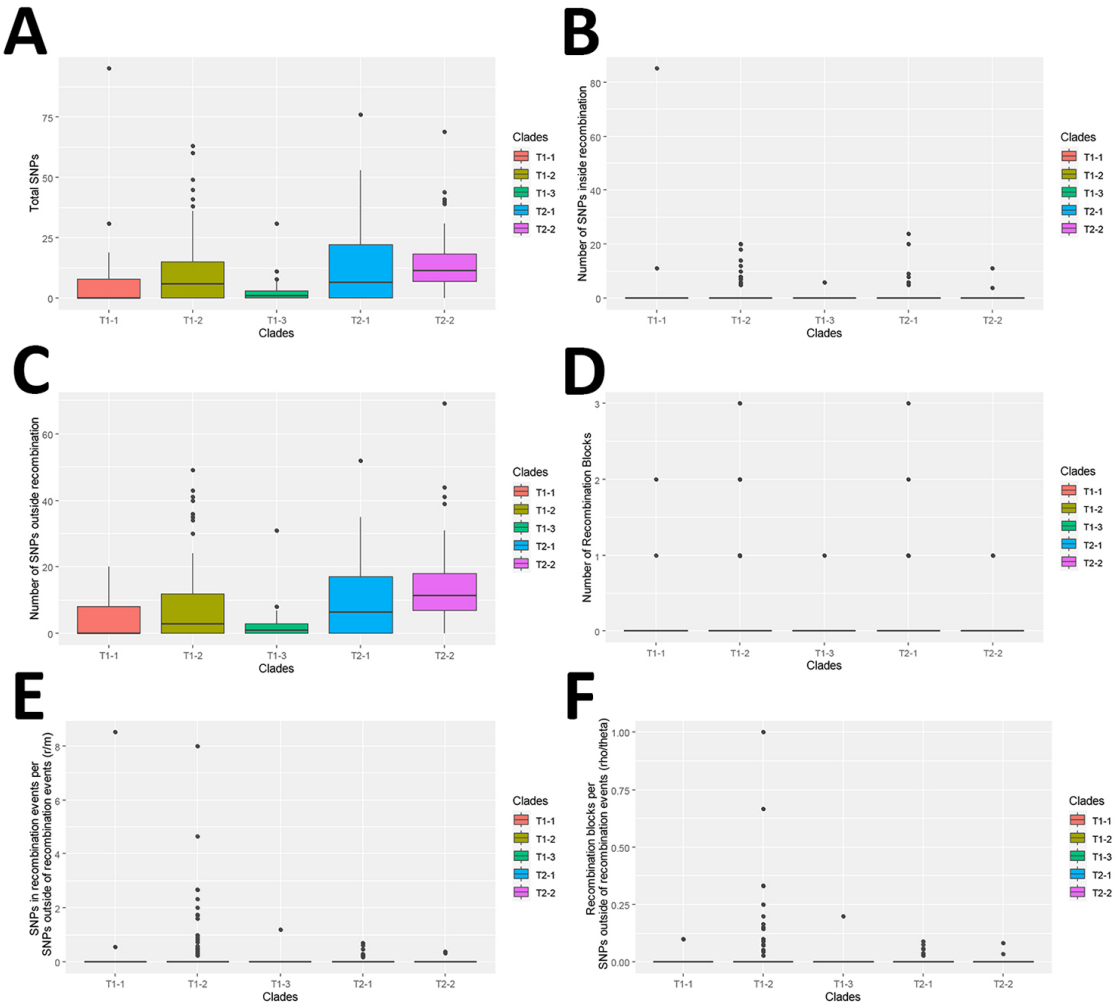
9. Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, Harris SR. Phandango: an interactive viewer for bacterial population genomics. Bioinformatics. 2018;34:292–3. PubMed https://doi.org/10.1093/bioinformatics/btx610

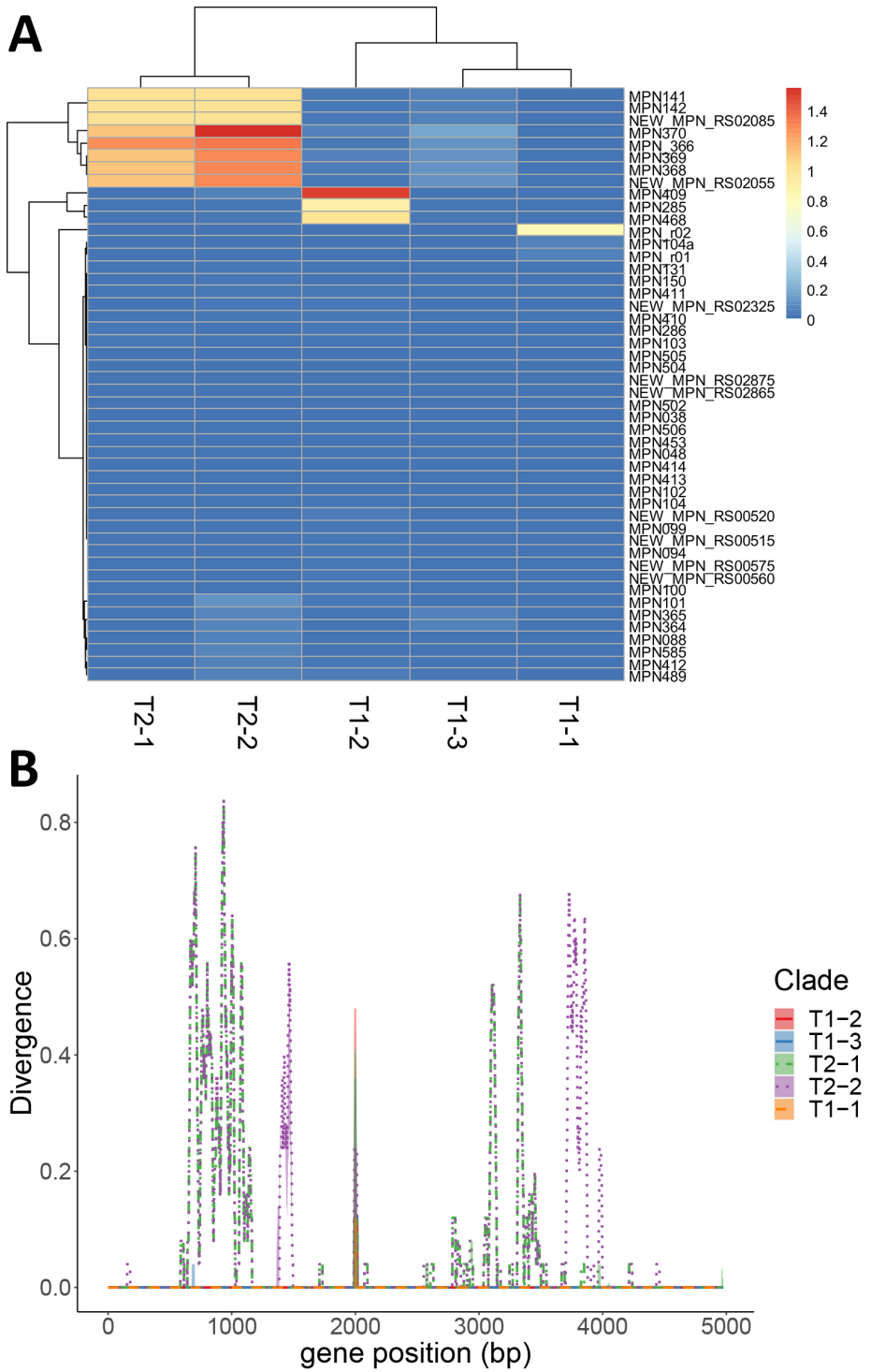
10. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7. PubMed https://doi.org/10.1093/nar/gkh340

11. Huang CL, Pu PH, Huang HJ, Sung HM, Liaw HJ, Chen YM, et al. Ecological genomics in *Xanthomonas*: the nature of genetic adaptation with homologous recombination and host shifts. BMC Genomics. 2015;16:188. PubMed https://doi.org/10.1186/s12864-015-1369-8

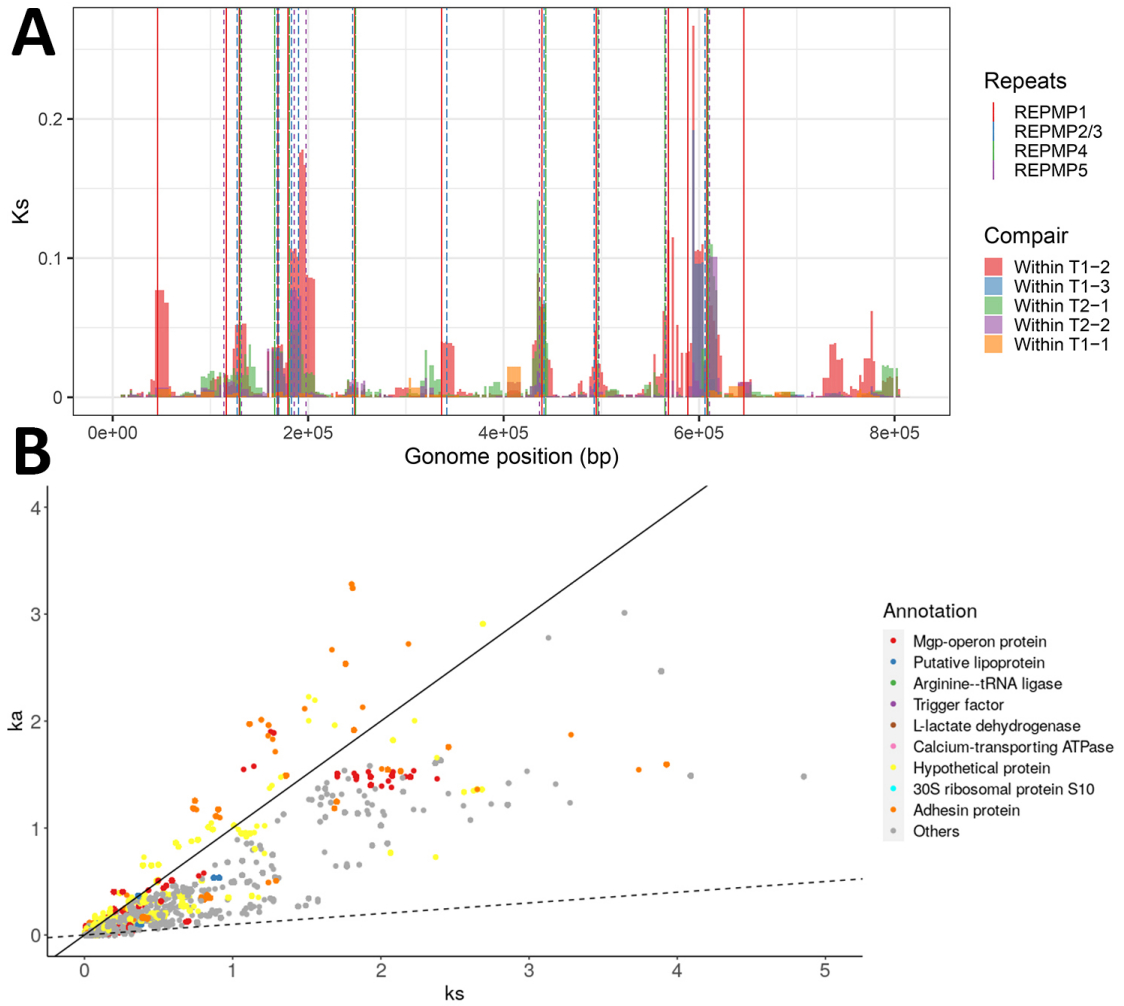


Appendix 2 Figure 1. A) Global recombination predictions for 284 *Mycoplasma pneumoniae* genomes. *M. pneumoniae* phylogeny based on single-nucleotide polymorphism density is shown on the left. Colored blocks represent putative recombination regions identified by using Gubbins software. Coordinates of recombination sites were based on those for strain M129. B) Putative recombination events detected in *M. pneumoniae* subtype 2. Gubbins analysis identified 1 region of predicted recombination (red lines) in clade T2–1 compared with strain FH across the genome.

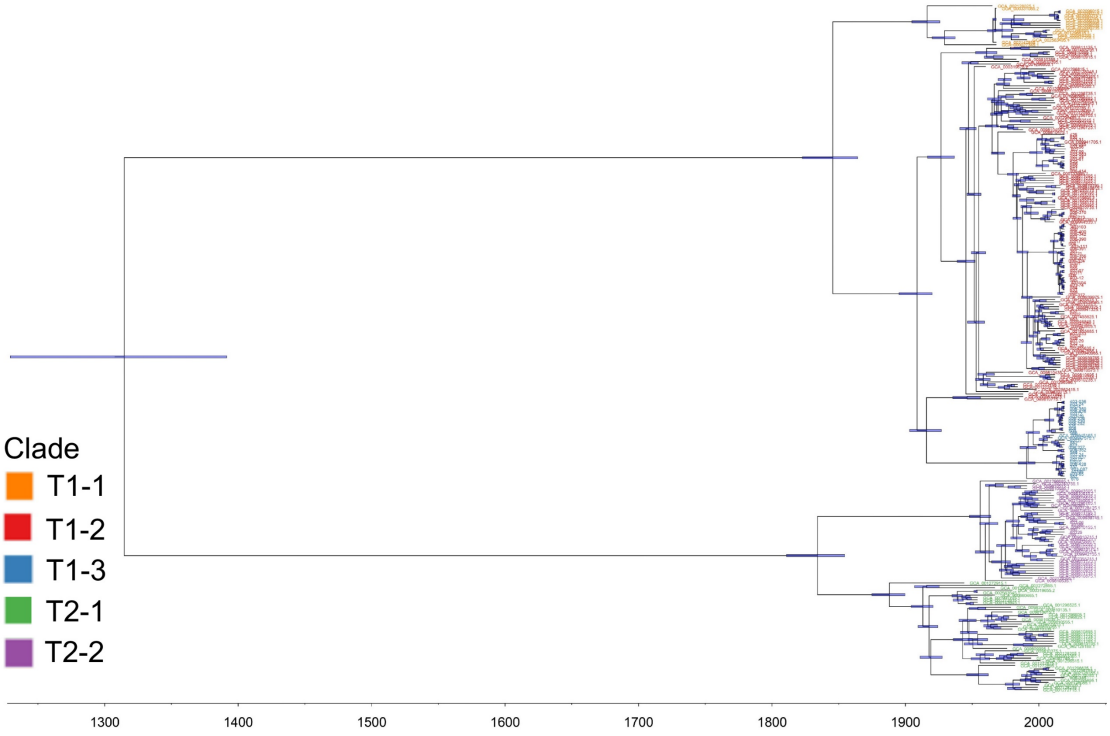

A
Total SNPs
Clades
T1-1
T1-2
T1-3
T2-1
T2-2
B
Number of SNPs inside recombination
C
Number of SNPs outside recombination
D
Number of Recombination Blocks
E
SNPs in recombination events per SNPs outside of recombination events (r/m)
F
Recombination blocks per SNPs outside of recombination events (rho/theta)



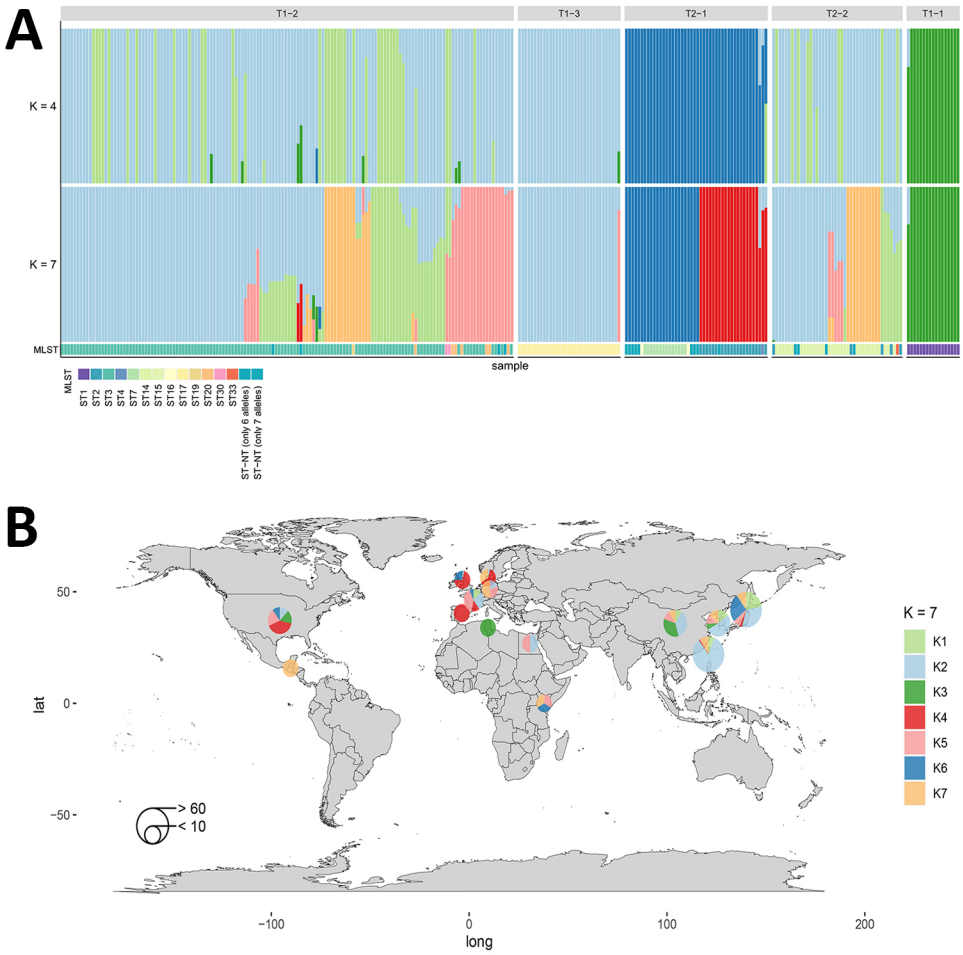
Appendix 2 Figure 3. A) Clade-specific recombination profiles at the gene level for *Mycoplasma pneumoniae*. Each horizontal line represents a gene, with colors corresponding to means of putative recombination events. B) Divergence (proportion of differing nucleotides) of the *MPN141* gene for each clade in 148 *M. pneumoniae* genomes compared with those for strain M129.



Appendix 2 Figure 4. Genome-wide variation among 284 *Mycoplasma pneumoniae* strains. A) Distribution of K_s divergence in M129 genomes with repeat elements. B) K_a/K_s distribution of 3,644,877 pair in 971 paralogs in 284 *M. pneumoniae* strains. Solid line indicates $K_a/K_s = 1$, and dotted line indicates slope = 0.01. The top 10 functions, which showed exhibited higher K_a mean values, are colored. Mgp, *M. genitalium*.



Appendix 2 Figure 5. Maximum-likelihood phylogenetic tree of aligned nucleotide sequences of *Mycoplasma pneumoniae* isolates. Dates along the x-axis are in years (CE), and blue bars indicate 95% posterior probabilities.



Appendix 2 Figure 6. Genetic structure and geographic analysis of 284 *Mycoplasma pneumoniae* isolates. A) Genetic variability in *M. pneumoniae* populations identified by using a STRUCTURE plot (https://omicsspeaks.com/strplot2). Results show admixture proportions between the fourth and seventh (*K* = 4 and *K* = 7) ancestral populations. Each person is indicated by a vertical line, which is partitioned into *K* = 4 and *K* = 7 colored segments with height proportional to estimated ancestry. Two estimated ancestry majors can be observed in T2–1 clade (blue and red), and 1 estimated ancestry major can be observed in T1–1 clade (dark green). B) Admixture proportions by sampling location. World map shows geographic distribution of admixture proportions from STRUCTURE (*K* = 7) in panel A, which suggests that *M. pneumoniae* populations are related to country. Pie charts are placed at the approximate locations where the strains were isolated, and isolate numbers are indicated by pie chart diameter. Color in pie charts are indicated by the mean composition of 7 ancestral populations. lat., latitude; long., longitude; MLST, multilocus sequence typing.