Infection, Genetics and Evolution 68 (2019) 68–76

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

A novel approach for the identification and phylogenetic delineation of human *Mycoplasma* species and strains using genomic segment sequence analysis

Orville St.E. Roachford[a,*], Karen E. Nelson[b], Bidyut R. Mohapatra[a]

[a] Department of Biological and Chemical Sciences, The University of the West Indies, Cave Hill Campus, Bridgetown, BB 11000, Barbados
[b] J. Craig Venter Institute, 9714 Medical Center Drive, Rockville, MD 20850, USA

ARTICLE INFO

Keywords:
Mycoplasma
Phylogeny
Biomarker
Comparative genomics
Species delineation

ABSTRACT

Human *Mycoplasma* are opportunistic, facultative pathogens that are site-specific in their colonization of mucosal surfaces. They are responsible for significant annual morbidity in humans by causing acute illnesses and chronic auto-inflammatory diseases via modulation of the host's immune system. Accurate and reliable identification of *Mycoplasma* species and their strains are thus of upmost importance. This study, analysed for the first time, the effectiveness of a short (50 kb) genome fragment (termed as R-segment), which includes the complete rRNA operon and the flanking region up to 50 kb, as a single phylogenetic marker for assessing the molecular taxonomy and determining the identity of human *Mycoplasma* species and their strains. The R-segments of human mycoplasmas were shown to have inherent genetic properties [average nucleotide identity (ANI), codon bias index (CBI), genome-to-genome distances (GGD) and % G + C] similar to their whole genome counterparts. Based on the results of our R segment analysis, a species of human *Mycoplasma* can simply be defined as a group of strains that share R-segments with ANIs ≥97%. Additionally, R-segments offered superiority to 16S rRNA gene sequences and multilocus sequences for the delineation of the human *Mycoplasma* species and their strains. The overall comparative genomic results suggest that R-segment analysis can be considered as a promising cost-effective tool for the epidemiological surveillance and differentiation of the closely related species and/or strains of human mycoplasmas.

1. Introduction

Mycoplasmas, the smallest self-replicating wall-less bacteria with redundant minimal genomes, emerged by degenerative evolution from low G + C, Gram positive eubacteria approximately 600 million years ago. They are ubiquitous and phylogenetically diverse. In view of their inability to synthesize essential nutrients, mycoplasmas parasitize both animal (humans, mammals, fish, reptiles, insects, arthropods) and plant hosts (Rogers et al., 1985; Razin, 2006). The survival of mycoplasmas in various environmental niches has been primarily ascribed to an effective arrangement of their genomes that allows for stochastic genophenotypic variations via frequent DNA recombination and other genetic mechanisms (Iverson-Cabral et al., 2006; Roachford et al., 2017). Size and phase variation of their antigenic surface lipoproteins, along with immunomodulation of the host's immune defense system allow them to persist within their hosts (Iverson-Cabral et al., 2007; Ma et al., 2008; Citti et al., 2010).

Mycoplasma species reside within humans, animals and plants either as commensals or pathogens. Significant morbidity can result as a consequence of chronic infection (Bove, 1981; Nicolson et al., 2000; Pereira et al., 2017; Scolten et al., 2017). In humans, they have a tendency to colonise the mucosal surfaces of the respiratory and urogenital systems (Zarei et al., 2013; Guillermo et al., 2014; Dehon and McGowin, 2017). For example, P1 type I and P1 type II *Mycoplasma pneumoniae* species (ca. 816 kb genome) tend to infect the respiratory system while its smaller descendent, *M. genitalium* (ca. 580 kb genome) infects the genitourinary system, though the latter has also been identified in the upper respiratory tract (Baseman et al., 1988; Stein and Baseman, 2006).

Like many other mycoplasmas, *M. pneumoniae* and *M. genitalium* share adhesins with cross-reactive epitopes, while *M. penetrans* and *M. hominis* display ON/OFF switching (phase variation) of their major immunodominant lipoproteins P35 and Vaa, respectively (Neyrolles et al., 1999; Boesen et al., 2001). These features create difficulty not

* Corresponding author.
E-mail address: orville.roachford@mycavehill.uwi.edu (O.S.E. Roachford).

https://doi.org/10.1016/j.meegid.2018.12.002
Received 27 July 2018; Received in revised form 21 November 2018; Accepted 3 December 2018
Available online 04 December 2018
1567-1348/

only to detect them via species-specific PCR and immunologic serological (ELISA) analyses, but also to genotype the closely related *Mycoplasma* species/strains. Inexpensive, accurate and reliable methods for detection and identification of mycoplasmas is paramount for the diagnosis and successful treatment of diseases caused by human mycoplasmas.

To date there are 31 strains (7 species) of human *Mycoplasma* inhabiting the oral (2 species), respiratory (1 species) and genitourinary (4 species) niches (Razin, 2006; Citti and Blanchard, 2013; NCBI Resource, 2016). The availability of whole-genome sequences has led to the construction of supertrees and the generation of a genome-based taxonomy of mycoplasmas (Oren and Ventosa, 2010; Thompson et al., 2013). The correlation between DNA-DNA Hybridization (DDH) and digital Genome-to-Genome Distance (GGD) along with other phylogenetic information [including average amino acid identity (AAI), average nucleotide identity (ANI), dinucleotide bias, codon usage bias (Nc), percentage G + C, Karlin genomic signatures that is inherent in genomic sequences] has led to a re-definition and delineation of *Mycoplasma* species. Generally, a species is defined as a group of bacterial strains with the following features: > 70% DDH similarity, < 5% mol G + C difference of total genome, < 5 °C ΔT_m and > 98% 16S rRNA gene sequences similarity (Stackebrandt and Ebers, 2006). Currently, the genome-based classification defines *Mycoplasma* species as a group of strains sharing ≥97% DNA identity in MLSA, ≥93.9% AAI and ≤ 8 in Karlin genomic signature (Thompson et al., 2011, 2013).

Based on 16S rRNA, MLSA and whole genome sequences, the group of species arising from the same genus form monophyletic clusters whose phylogeny can be determined based on the tree topology. However, these phylogenetic trees often show discrepancies when compared to each other. Furthermore, 16S rRNA gene sequences are inadequate in differentiating closely related species and strains (as for *M. genitalium* and *M. pneumoniae*) since these can have almost 100% identical 16S rRNA gene similarity (Thompson et al., 2011). Supertrees can be formed from whole genome sequences as have been constructed in the determination of *Bifidobacterium* taxonomy (Lugli et al., 2017). These supertrees may serve as a gold standard but the drawbacks of whole genome sequence analyses include the requirement of supercomputers and/or the generation of statistical artifacts during the in silico analysis by decreasing the number of whole genome entries. Supertrees based on specific concatenated genes (Thompson et al., 2009) and MLSA approaches have been used as alternatives to whole genome analyses but the caveat is that not all core genes are shared across all the genera or species.

An alternative to 16S rRNA, MLSA and whole-genome analyses is the possible use of short genome segments at a defined locus whose inherent genetic properties are co-linear with those of the whole genome with respect to % G + C difference, GGD, digital DND-DNA Hybridization (dDDH), ANI and codon usage bias. These short genome segments (called R-segments henceforth), are defined here by us as 50 kb genomic sequences which includes the first entire rRNA operon and the immediate flanking region. The genes downstream to 16S rRNA are highly conserved fundamental housekeeping genes that play a very important role in cell function. In fact, this 50 kb genomic fragment has already been identified as the potential target for the ribotyping to discriminate the species of *Haemophilus influenza*, *Streptococcus pneumoniae*, *Burkholderia cepacia* and *Pseudomonas aeruginosa* because of the presence of additional hypervariable regions (reviewed in Bouchet et al., 2008). Consequently, these 50 kb R-segments are expected to be useful fragments for phylogenetic analyses and in instances where whole-genomes are unavailable or incomplete, analyses could be carried out with the appropriate R-segments. The human mycoplasmas which consist of very closely related species (*M. pneumoniae* and *M. genitalium*) and type-specific strains (*M. pneumoniae* P1 types I & II) allows for testing the use of R-segments for taxonomic relationships. R-segments also possess unique signatures or fingerprints that once identified can be used to accurately determine the identity and/or differentiate *Mycoplasma* species and their strains. The advantages of R-segment analysis are the saving in both cost and time compared to the whole genome and MLSA approaches. The R-segment analysis can be performed by directly sequencing the 50 kb fragments via the targeted sequencing approaches of any next-generation sequencing platform including the SMRT PacBio and Illumina MiSeq systems.

In this paper, the effectiveness of R-segments for phylogenetic classification and delineation of the human *Mycoplasma* species/strains was evaluated for the first time by directly comparing the R-segment phylogenetic analysis with the 16S rRNA and MLSA phylogenetic approaches using 31 human *Mycoplasma* genomes.

2. Materials and methods

2.1. Mycoplasma genome sequence data

Thirty-one human *Mycoplasma* genome sequences were downloaded from the National Center for Biotechnology Information (NCBI) GenBank for in the in silico comparative phylogenetic analyses. There is a greater number of genome sequences available in GenBank for the strains of human *M. pneumoniae* (16) and *M. genitalium* (5) compared to other species, *M. fermentans* (3), *M. hominis* (4), *M. orale* (1), *M. penetrans* (1) and *M. salivarium* (1). Additionally, 1 genome of the genus *Ureaplasma* (*Ureaplasma urealyticum* serovar 10 ATCC 33699) and 1 genome of an animal *Mycoplasma* (*Mycoplasma arthritidis* 158L3--1) were obtained from GenBank and used in the analyses. The accession numbers and genomic features along with the locus tag and position of their 16S rRNA genes are highlighted in Table 1.

2.2. 16S rRNA gene, multilocus sequence analysis (MLSA) and genome segment sequences

RNAmmer v1.2 prediction server (Lagesen et al., 2007) was utilized to confirm or predict the location of the 16S rRNA gene sequence in the whole and partial genomes of the mycoplasmas. Five highly conserved core genes, *gapA* (encodes glyceraldehyde-3-phosphate dehydrogenase), *tktA* (encodes transketolase), *ligA* (encodes NAD dependent DNA ligase), *gyrB* (encodes DNA gyrase Subunit B) and *adk* (encodes adenylate kinase) were also obtained from NCBI GenBank for all the mycoplasmas. These conserved genes, which participate in energy production and nuclear processes, have previously been used in comparative phylogenetic studies (Thompson et al., 2011; Adeolu et al., 2016; Roachford et al., 2017). The concatenated gene product sequences were aligned using CLUSTALW for the MLSA. The dinucleotide biases (dinucleotide relative abundance values) for consecutive 50 kb sequences are relatively constant along the genome of a given organism but varies between different organisms. Dinucleotide bias thus serves as a genomic signature that allows for genome discrimination among prokaryotes (Karlin and Burge, 1995). The dinucleotide bias is determined by the formula $\rho_{xy} = f_{xy}/f_x f_y$, where f_{xy} is the frequency of the dinucleotide XY and f_x is the frequency of the nucleotide X (Karlin, 1998). The R-segments were aligned using MUSCLE in MEGA 7 (Kumar et al., 2012) and visualized with progressiveMauve in Mauve version 20150226 (Darling et al., 2010). 16S rRNA, MLSA and R-segment phylogenetic trees were generated using the maximum-likelihood algorithm in MEGA 7 with default settings. The statistical reliability of the constructed phylogenetic trees was assessed with the bootstrap method using 1000 replications.

2.3. Comparison of R-segments with the whole genome of human mycoplasmas

The inherent genetic properties [GGD, dDDH, % G + C, codon bias (CBI and Nc) of R-segments were compared to the whole or partial genomes of the equivalent *Mycoplasma* strains. The values of GGD, DDH and the % G + C difference were calculated using the Genome-to-

Table 1
Genomic features of human *Mycoplasma*[b].

Organism	Accession no.	Genome size (bp)	%GC	First 16S rRNA locus tag	First 16S rRNA locus position
Mycoplasma fermentans JER[a]	NC_014552.1	977,524	26.95	MFE_RS02665	622,199 > 623,745
Mycoplasma fermentans M64[a]	NC_014921.1	1,118,751	26.86	MFEM64YM_RS03100	698,997 > 700,543
Mycoplasma fermentans PG18[a]	NC_021002.1	1,004,014	26.87	MBIO_RS00420	78,008 > 79,141
Mycoplasma genitalium G37	NC_000908.2	580,076	31.69	MG_RS00775	170,000 > 171,538
Mycoplasma genitalium M2288	NC_018498.1	579,558	31.67	CM5_r02796	169,843 > 171,366
Mycoplasma genitalium M2321	NC_018495.1	579,977	31.67	CM9_RS00775	169,959 > 171,497
Mycoplasma genitalium M6282	NC_018496.1	579,504	31.67	CM3_RS00780	169,673 > 171,211
Mycoplasma genitalium M6320	NC_018497.1	579,796	31.68	CM1_RS00780	169,844 > 171,382
Mycoplasma hominis PG21[a]	NC_013511.1	665,445	27.12	MHO_RS01540	331,796 > 330,267
Mycoplasma hominis LBD-4[a]	NZ_CP009652.1	715,165	26.90	MLBD4_RS01570	353,551 > 355,080
Mycoplasma hominis AF1	NZ_CP009677.1	700,851	26.91	VE27_RS02290	527,893 > 529,420
Mycoplasma hominis Sprott[a]	NZ_CP011538.1	695,214	27.40	MHOMSp_RS01535	336,346 > 337,873
Mycoplasma orale ATCC 23714	ATUH01000003.1	710,549	24.75	AB353273.1	N/A
Mycoplasma penetrans HF-2	NC_004432.1	1,358,633	25.72	MYPE_RS04745	1,236,488 > 1,238,013
Mycoplasma pneumoniae 309	NC_016807.1	817,176	39.98	MPNA_RS00530	118,306 > 119,841
Mycoplasma pneumoniae 19294	NZ_CP010539.1	817,313	39.98	B434_RS02065	439,427 > 440,962
Mycoplasma pneumoniae 39443	NZ_CP010540.1	817,184	39.98	F536_RS00535	118,263 > 119,798
Mycoplasma pneumoniae 51494	NZ_CP010541.1	816,404	40.01	F533_RS00545	118,275 > 119,810
Mycoplasma pneumoniae 54089	NZ_CP010542.1	816,565	40.01	F530_RS00540	118,263 > 119,798
Mycoplasma pneumoniae 54524	NZ_CP010543.1	816,583	40.01	F531_RS00545	118,275 > 119,810
Mycoplasma pneumoniae 85084	NZ_CP010544.1	816,404	40.01	F537_RS00540	118,311 > 119,846
Mycoplasma pneumoniae 85138	NZ_CP010545.1	816,402	40.01	F535_RS00545	118,299 > 119,834
Mycoplasma pneumoniae FH	NC_017504.1	811,088	40.00	MPNE_RS00540	118,283 > 119,818
Mycoplasma pneumoniae M1139	NZ_CP010547.1	817,045	39.98	F538_RS00545	118,272 > 119,807
Mycoplasma pneumoniae M129	NC_000912.1	816,394	39.86	MPNr01	118,312 > 119,824
Mycoplasma pneumoniae M129-B7	NC_020076.2	816,451	40.01	C985_RS00535	118,278 > 119,813
Mycoplasma pneumoniae M2592	NZ_CP010549.1	817,198	40.01	G667_RS00540	118,246 > 119,775
Mycoplasma pneumoniae M29	NZ_CP008895.1	857,799	39.86	Y923_RS00930	195,779 > 197,314
Mycoplasma pneumoniae MAC	NZ_CP010550.1	817,156	39.98	F529_RS00545	118,267 > 119,802
Mycoplasma pneumoniae PO1	NZ_CP010551.1	817,216	39.98	C680_RS00540	118,260 > 119,795
Mycoplasma salivarium ATCC 23064	NZ_AXZE01000009.1	713,526	26.71	F800_RS0100680	108,004 > 109,535
Mycoplasma arthritidis 158L3-1	NC_011025.1	820,453	30.71	MARTH_RS00785	191,346 > 192,941
Ureaplasma urealyticum ATCC 33699[a]	NC_011374.1	874,478	25.77	UUR10_RS00605	149,403 > 150,935

N/A = Not available.
[a] Possesses two copies of rRNA operon.
[b] All strains are cultured isolates.

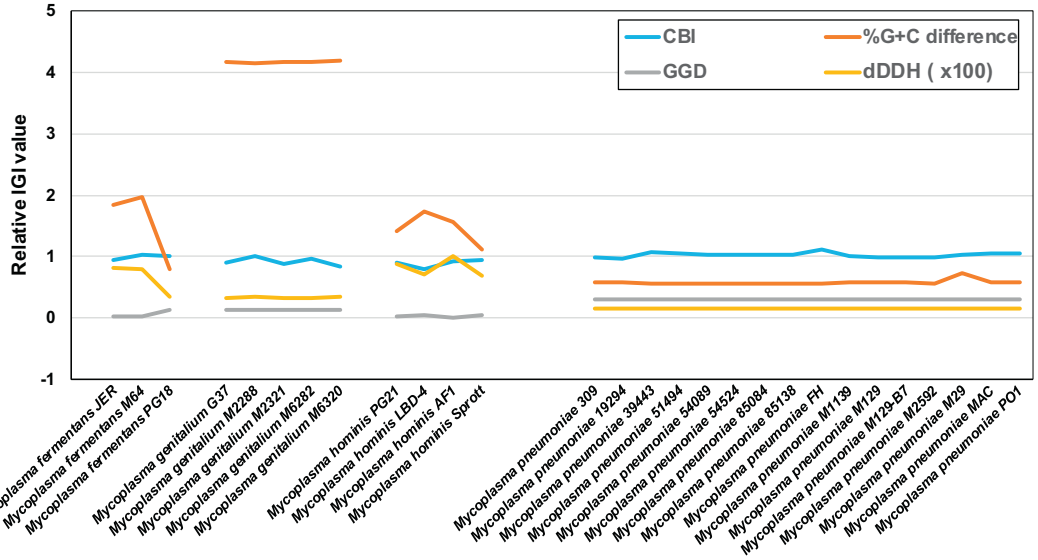

Fig. 1. Comparison of inherent genetic indices (IGIs) CBI, %G + C, GGD and dDDH of R-segments of human *Mycoplasma* against their whole genome counterparts. The X-axis denotes the value of the relative inherent genetic index (relative-IBI). The relative CBI was calculated by dividing the value of a given CBI of an R-segment by the value of the CBI for the respective whole genome. The other relative IBIs were generated by the relevant software used. Each inherent genetic index is color coded; CBI (Blue), %G + C difference (Orange), GGD (Grey) and dDDH (Yellow). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

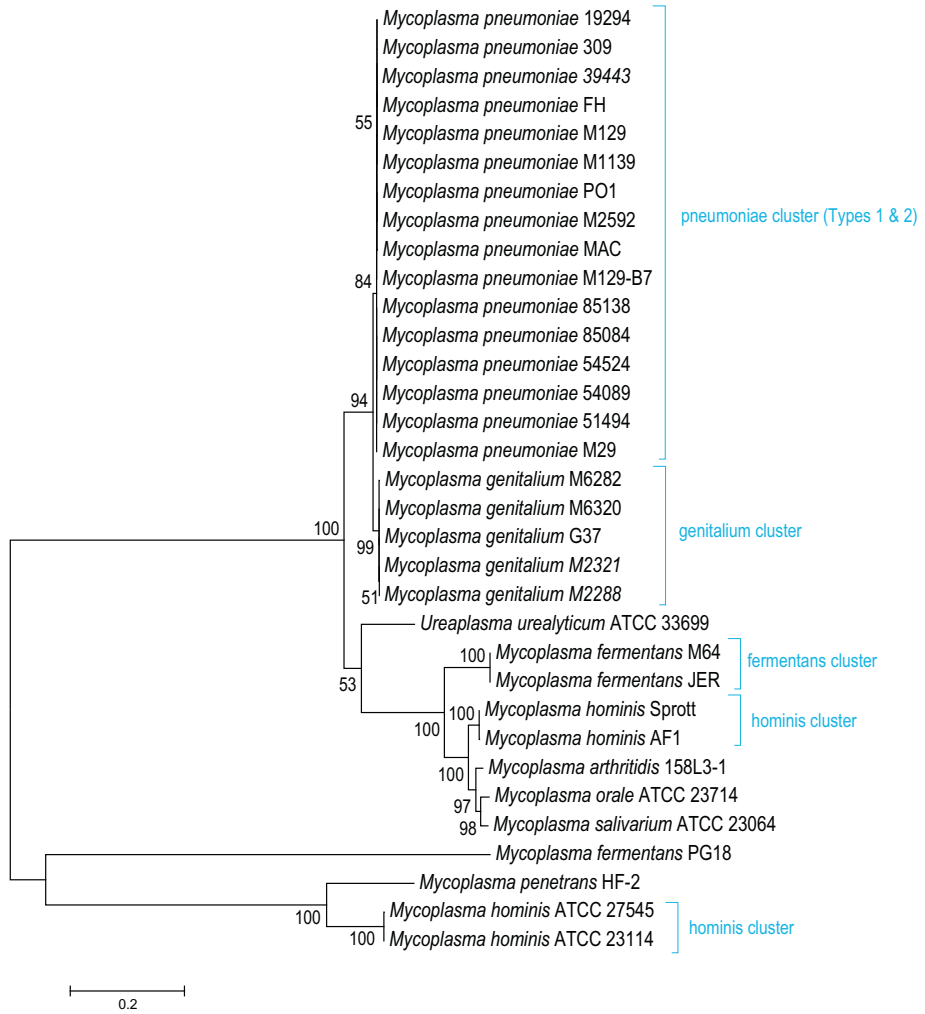

Fig. 2. Phylogenetic relationship among human *Mycoplasma* species and their strains based on 16S rRNA gene sequences. Phylogenetic analysis was performed with MEGA 7.0 using the maximum likelihood algorithm. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Evolutionary distances were calculated using the *p*-distance method. The bootstrap values (1000 replicates) are shown at the node with a cut-off value of 50%.

Genome Distance Calculator version 2.1 (Meier-Kolthoff et al., 2013, 2014). The formula, Identities/High Scoring-Segment Pair (HSP) lengths, within the software was used to infer distances from a set of HSPs obtained by pairwise comparisons of the R-segment with one whole genome. Using a generalized linear model inferred from an empirical reference dataset comprising of real DDH values and genome sequences, the distances were converted to values corresponding to DDH (Auch et al., 2010). The dDDH values for R-segments were corrected using Eqs. 1 and 2 (Auch et al., 2010).

$$dDDH = (1 - d_2)\ X\ 100 \quad (1)$$

$$d_2(x, y) = 1 - (I_{xy} + I_{yx}/H_{xy} + H_{yx}) \quad (2)$$

where d_2 = distance between genomes; x = R-segment; y = whole genome; H_{xy} = total lengths of all HSPs; I_{xy} = sum of the number of all identical base pairs over all HSPs recorded by performing a BLAST search of x against y. H_{yx} and I_{yx} are obtained by using y as the query and x as the subject sequence. dDDH = similarity between genomes $= 1–d_2$. Therefore, % dDDH = $(1–d_2)$ x 100.

The codon bias [Codon Bias Index (CBI] and effective number of codons (Nc)] used in each sequence) were calculated using DnaSP v5 software (Librado and Rozas, 2009). Formulae and tabulated results were listed in Supplementary Table S1.

2.4. Average nucleotide identity (ANI)

Average nucleotide identity (ANI) was determined for all *Mycoplasma* sequences (16S rRNA, MLSA and R-segment) using the ANI calculator available in the EZ BioCloud server (Yoon et al., 2017). The ANI calculator uses the OrthoANIu algorithm which employs USEARCH instead of BLAST. The genomic sequence of each *Mycoplasma* strain was compared against its sister strain and in some instances against strains of distant species. The ANI values were then used to create intraspecies and interspecies similarity matrices (Supplementary Table S2).

3. Results

3.1. Genomic features of human Mycoplasma

The six species of human *Mycoplasma* (Table 1) consist of minimal genomes ranging in size from 580 kb (*M. genitalium* G37) to approximately 1360 kb (*M. penetrans* HF-2). This variation in size is due to the presence of insertion sequences (tandem repeats, phage-like protein coding genes), pseudogenes and redundant copies of virulent genes (paralogs), all of different sizes (Razin and Hayflick, 2010; Roachford et al., 2017). As descendants of low G + C Gram positive bacteria, the mycoplasmas have maintained this low G + C content (averaging 25–40%), a property that allows for their high genomic plasticity. The 16S rRNA gene is located at approximately the same locus position

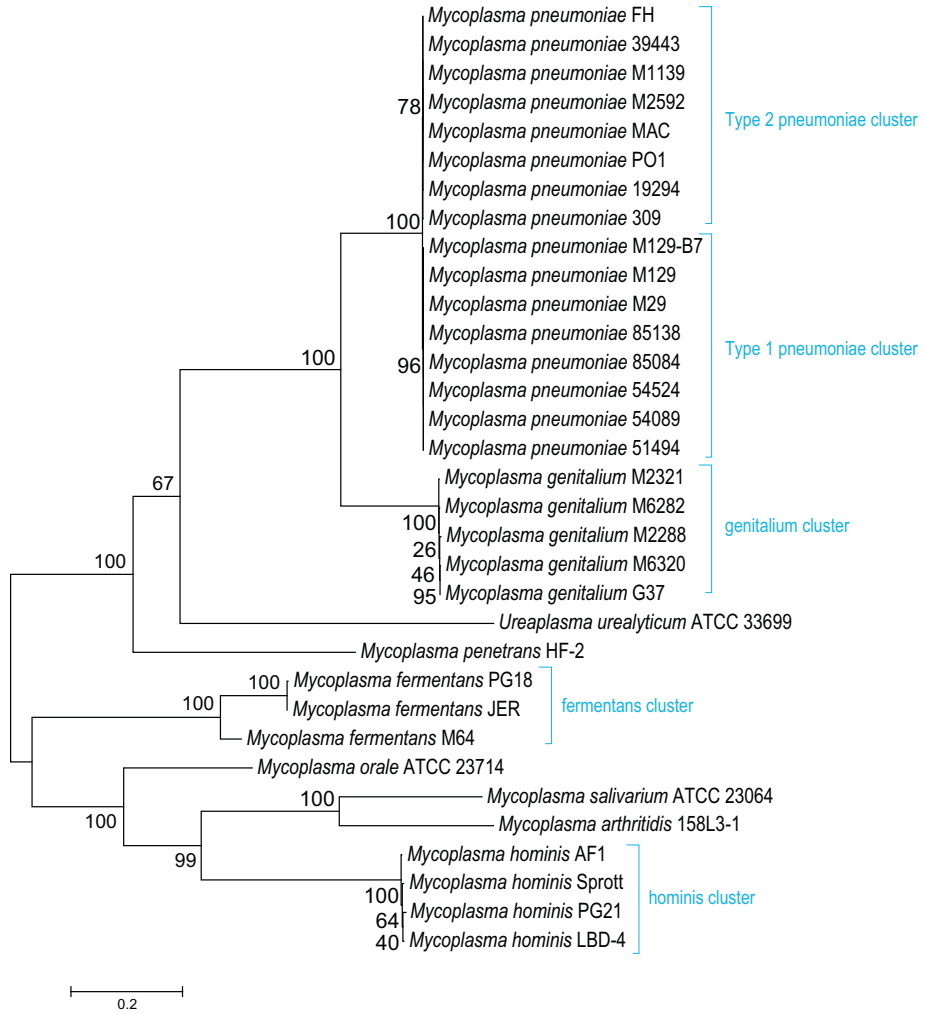

Fig. 3. Phylogenetic relationship among strains of human *Mycoplasma* species, *U. urealyticum* ATCC 33699 and an animal mycoplasma, *M. arthritidis* 158L31, based on concatenated gene sequences (*gapA*, *tktA, ligA, gyrB* and *adk*). The phylogenetic tree was constructed with MEGA 7.0 using the maximum likelihood method and is drawn to scale. Branch lengths are measured in the number of amino acid substitutions per site. Evolutionary distances were calculated using the *p*-distance method. The bootstrap values (1000 replicates) are shown at the node with a cut-off value of 50%.

within the genomes of the human *Mycoplasma* species and their strains except for four anomalies; *M. fermentans* PG18, *M. hominis* AF1, *M. pneumoniae* 19294 and *M. pneumoniae* M29.

3.2. Comparison of R-segments with whole-genome sequences of human Mycoplasma

The genome properties (represented by the indices GGD, dDDH, %G + C, CBI) that are inherent in R-segments (derived from strains of *M. fermentans*, *M. genitalium, M. hominis* and *M. pneumoniae*) in general showed a linear relationship with their whole genome counterparts as evidenced in Fig. 1. The relative GGDs (between R-segments and whole genomes) ranged from 0 to 0.308. A score of zero indicates identical segment to genome properties and is equivalent to an estimated DDH (an in silico similarity value analogous to a real DDH value) of 100. Generally, a DDH value ≥70 suggests that the genomes/sequences are of the same species. Across all strains, the dDDH determined by the GGD Calculator gave ranges from 35 to 81; 32 to 34; and 68 to 100 for strains of *M. fermentans*, *M. genitalium* and *M. hominis*, respectively. For *M. pneumoniae*, the DDH value was ca. 14 for all the strains. Corrected dDDH values ranged from 70% to 100% for all the mycoplasma strains. Intraspecies % G + C differences varied by < 1% across all species/strains except for *M. fermentans* PG18 (ca.1.2%). The relative CBI varied slightly (< 10%) between the *Mycoplasma* species and their strains: 0.05–0.09 (*M. fermentans*), 0.03–0.17 (*M. genitalium*), 0.08–0.12 (*M. hominis*) and 0.01–0.10 (*M. pneumoniae*).

3.3. Comparison of phylogenetic trees based on 16S rRNA, MLSA and R-segments

Phylogenetic trees constructed for 16S rRNA gene sequences, concatenated genes sequences of *gapA*, *tktA, ligA, gyrB* and *adk* (for the MLSA approach) and R-segments using maximum-likelihood algorithm were shown to have similar topological configurations of branches as indicated by high bootstrap confidence levels. Generally, in MLSA (Fig. 3) and R-segment (Fig. 4) trees, the two main clusters (*genitalium* and *pneumoniae*) were well separated with type 1 and 2 *M. pneumoniae* strains being well divided. *M. arthritidis and M. salivarium* clustered with *M. hominis*. The pairs of species, *M. arthritidis* and *M. hominis, M. salivarium* and *M. hominis,* and *M. pneumoniae* and *M. genitalium* had 16S rRNA gene similarities of 96%, 78% and 98%, respectively (Fig. 2). The MLSA similarities were 50%, 50%, and 74%, respectively. Phylogenetically, *Ureaplasma urealyticum* was found to be more related to *M. pneumoniae* and *M. genitalium* (Fig. 4) with 16S rRNA gene (and MLSA) similarity values of 84% (49%) and 84% (48%), respectively. More specifically, *M. salivarium* ATCC 23064 was found to be more closely related to *M. arthritidis* 158 L3–1 (62% similarity) than with *M. hominis* Sprott (50% similarity) using the MLSA approach. However, based on

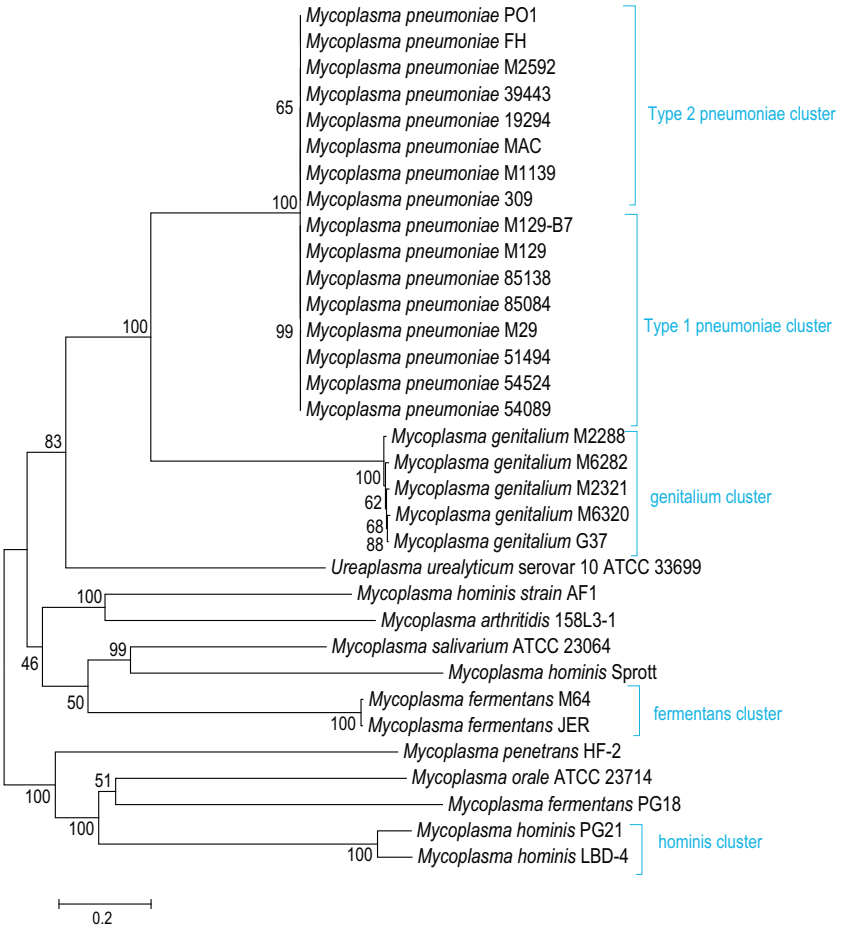

Fig. 4. Phylogenetic relationship among human *Mycoplasma* species-strains, *M. arthritidis* 158L3-1 and *U. urealyticum* ATCC 33699 based on R-segment sequences. The phylogenetic tree was constructed with MEGA 7.0 using the maximum likelihood method and is drawn to scale. Branch lengths were measured as the number of nucleotide substitutions per site. Evolutionary distances were calculated using the *p*-distance method. The bootstrap values (1000 replicates) are shown at the node with a cut-off value of 50%.

analysis of the R-segments, *M. salivarium* ATCC 23064 was noted to be more closely related phylogenetically to *M. hominis* Sprott. Based on the MLSA, *M. orale* ATCC 23714 was more closely related to *M. hominis* AF1 (58% similarity) than to *M. fermentans* PG18 (49% similarity). Again, R-segment analysis contradicted the latter showing *M. orale* ATCC 23714 to be more related to *M. fermentans* PG18 (Fig. 4). The genome similarities for *M. salivarium* ATCC 23064 against *M. arthritidis* 158 L3–1 and *M. salivarium* ATCC 23064 against *M. hominis* Sprott were 90% and 92%, respectively. For *M. orale* ATCC 23714 against *M. hominis* AF1 and *M. fermentans* PG18 the genome similarities were recorded as 71% and 72%, respectively.

3.4. Codon bias (CBI & Nc), %G + C and average nucleotide identity (ANI) for R-segments

The codon bias for the human mycoplasmas was assessed from the values of two parameters; the codon bias index (CBI) and the effective number of codons (Nc). These two parameters varied significantly among the *Mycoplasma* species but the intraspecies CBI and Nc values were relatively consistent. The CBI values ranged from 0.19 (*M. pneumoniae* 19294) to 0.49 (*M. penetrans* HF-2), whereas the Nc values ranged from 40 (*M. penetrans* HF-2) to 56 (*M. pneumoniae* 19294). Intraspecies CBI values for *M. fermentans*, *M. genitalium*, *M. hominis* and *M. pneumoniae* were 0.38–0.41, 0.28–0.34, 0.33–0.38 and 0.18–0.20, respectively. The variation of the intraspecies Nc values for the species *M. fermentans*, *M. genitalium*, *M. hominis* and *M. pneumoniae* were 46–47, 49–51, 48–50 and 56–57, respectively (Table 2).

The interspecies %G + C among the human *Mycoplasma* ranged from 26.1 to 40.6. For the species *M. fermentans*, *M. genitalium*, *M. hominis* and *M. pneumoniae*, the average % G + C values were 26.1–28.8, 35.8–35.9, 28.5–28.7 and 40.5–40.6, respectively. *M. orale* ATCC 23714, *M. penetrans* HF-2 and *M. salivarium* ATCC 23064 had % G + C values of 24.8, 25.2 and 30.6, respectively.

The intraspecies % G + C differences for *M. fermentans*, *M. genitalium* and *M. pneumoniae* were consistently $\leq$0.04% in the absence of *M. fermentans* PG18 whose incomplete genome had an R-segment % G + C value of ~26% and a % G + C difference that ranged from 2.7 to 2.8% relative to the other *M. fermentans* strains. For *M. hominis* strains, the % G + C difference was $\leq$0.21% (Table 2).

The intraspecies ANI for all human *Mycoplasma* strains was $\geq$97.5%. For the species *M. fermentans*, *M. genitalium*, *M. hominis* and *M. pneumoniae,* the intraspecies ANI differences were maximal at the values of 0.60, 0.93, 0.24 and 0.36, respectively. The assessment of interspecies ANI for the selected *Mycoplasma* species indicated that the species of human mycoplasmas were well differentiated based on the ANI values. ANI values varied from 67.3% for the pair of species, *M. genitalium* and *M. hominis* to 95% for the more closely phylogenetically –related species pair, *M. genitalium* and *M. pneumoniae* (Supplementary Table S2).

3.5. R-segment sequence signatures for Mycoplasma strain identification

The sequences of the R-segments for the human *Mycoplasma* strains were aligned and visualized with progressiveMauve in Mauve version

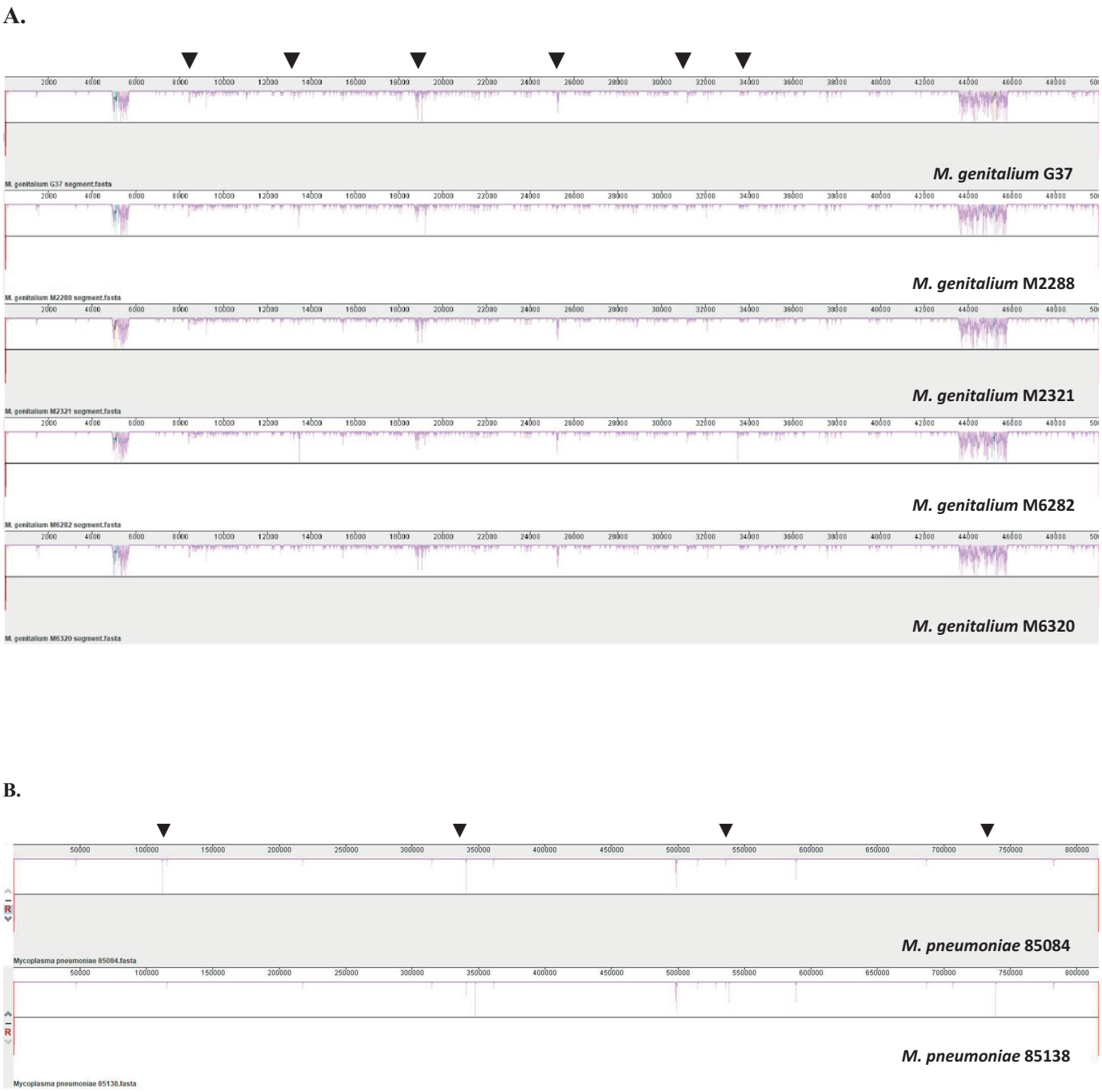

Fig. 5. (A) Visualized R-segment sequence signatures of *M. genitalium* strains. (B) Visualized sequence signatures of *M. pneumoniae* 85084 and *M. pneumoniae* 85138 whole genomes. The black arrow heads (▼) highlight some areas of variation between the signatures of the five *M. genitalium* strains and between those of the two *M. pneumoniae* strains.

Table 2
Inherent genetic properties of R-segments of human *Mycoplasma* species.

	M. fermentans			*M. genitalium*			*M.hominis*			*M. pneumoniae*		
Indices	Min	Max	diff	Min	Max	diff	Min	Max	diff	Min	Max	diff
Relative CBI	0.95	1.04	0.09	0.84	1.02	0.18	0.79	0.94	0.15	0.97	1.11	0.14
Nc	45.50	46.70	1.20	49.10	50.80	1.70	48.30	49.60	1.30	55.60	56.50	0.90
% G + C	28.79[a]	28.83	0.04	35.83	35.87	0.04	28.48	28.69	0.21	40.53	40.58	0.05
dDDH	98.00	98.00	0.00	87.00	88.00	1.00	70.00	100.00	30.00	69.00	70.00	1.00
ANI	99.31	99.61	0.61	98.25	99.18	0.93	97.55	99.35	0.80	99.63	99.99	0.07

[a] minimum value for *M. fermentans* JER, *M. fermentans* PG18 incomplete genome was 26.07.

20,150,226 using default settings except for the color scheme which was set at backbone color. The R-segments produced sequence signatures (Supplementary Figure S1) that were unique for all strains except for *M. pneumoniae* 85,084 and *M. pneumoniae* 85,138 which appeared identical. However, these two strains were differentiated using their whole genomes (Fig. 5B).

4. Discussion

The R-segments (50 kb long genomic sequences whose upstream regions commence with the 5′ end of the first 16S rRNA gene sequence of the genome) were shown to share similar genetic information (codon bias index, genome-to-genome distances, digital DNA-DNA hybridization estimates and % G + C) as the corresponding whole genomes for the four species of human *Mycoplasma* (*M. fermentans*, *M. genitalium*, *M. hominis* and *M. pneumoniae*) and their strains.

The estimates of DNA-DNA Hybridization (DDH), which assessed the similarity between an R-segment and the related whole genome, were inconsistent by displaying the values from 14 to 100%. Two genomes or sequences which share the same identity are expected to have DDH values ≥70% (Goris et al., 2007). The inconsistency in the estimated DDH values might be attributed to the GGD algorithm used to predict DDH, which becomes less effective as the length of the queried sequence falls < 20% of the reference genome (Auch et al., 2010). The length of the R-segments, which is < 10% of the corresponding whole genomes, might therefore have caused the inconsistency in the estimated DDH values. However, the corrected DDH (dDDH) values (calculated from Eq. 0.2) for the mycoplasmas used in this study were ≥ 70% except for *M. pneumoniae* strains which had relative dDDH values of 69.2–69.7%. Overall, the dDDH values of the R-segments had a linear relationship with those of the corresponding whole genomes except for the incomplete genome of *M. fermentans* PG18 (Fig. 1).

The CBI and Nc values generally differentiated the species except for *M. fermentans* JER and *M. hominis* AF1 which shared identical CBI values (0.382) but different Nc values of 46.7 and 48.8, respectively. All strains could be differentiated based on the CBI and Nc values except for *M. pneumoniae* strains 85,084 and 85,138 which had identical CBI (0.190) and Nc (56.177) values for their R-segments. The % G + C values were unable to clearly differentiate the species. None of these parameters could adequately differentiate the *Mycoplasma* strains.

A DDH value of 70%, which corresponds to an average nucleotide identity (ANI) of 95% or an average amino acid (AAI) of ~95–96% for whole genomes, have been used for the delineation of prokaryotic species (Konstantinidis and Tiedje, 2005; Goris et al., 2007). The taxonomic application of ANI for random sequences of ≥20% of the genome have been used successfully to differentiate bacterial species (Richter and Rossello-Mora, 2009). ANI values determined by different ANI calculators, such as ANI Calculator, JSpecies and EzGenome, have not had any significant impact on the phylogenetic classification of species (Figueras et al., 2014).

Generally, polyphasic approaches have been undertaken to define a species. Generally, a species is defined as a group of bacterial strains having a DDH similarity > 70%, G + C difference < 5% mol of total genome, < 5 °C ΔTm and > 98% 16S rRNA sequence similarity (Stackebrandt and Goebel, 1994). Recently, from a genome-based view point *Mycoplasma* species have been defined as a group of strains sharing ≥97% DNA identity in MLSA (multilocus sequence alignment), ≥93.9% AAI and ≤ 8 in Karlin genomic signature (Thompson et al., 2011, 2013).

In this study, we have taken a monophasic approach for the species definition. Using a single phylogenetic biomarker (R-segment), we have re-defined *Mycoplasma* species and developed a novel approach to differentiate human *Mycoplasma* species and their strains. Thus, a *Mycoplasma* species may simply be defined as a group of strains that share R-segments with ANIs ≥97%. R-segments, which were visualized by an appropriate software (progressiveMauve) produced unique sequence signatures for the human mycoplasmas. Therefore, human *Mycoplasma* species and/or their strains can potentially be identified in silico by determining the R-segment of an isolate and comparing it with R-segment sequence signatures of known *Mycoplasma*. In rare cases where the sequence signature, ANI, CBI and Nc of R-segments for a pair of strains is identical, the strains can be differentiated by comparison of their whole genome signatures as was demonstrated for *M. pneumoniae* 85,084 and *M. pneumoniae* 58,138.

In conclusion, the use of R-segments offers an effective genome-based approach for the delineation and differentiation of *Mycoplasma* species and their strains whilst avoiding the combination of multiple parameters in species definition. With the advent of next-generation sequencing techniques, R-segments can also supplement other molecular tools such as MLSA for *Mycoplasma* genomic phylogeny (Jironkin et al., 2016) because, in some instances, MLSA has failed to represent the bacterial phylogeny (Tsang et al., 2017). This failure of MLSA might be ascribed to the involvement of tedious methodologies, including the search for the availability of annotated whole genomes, identification of 5–10 loci from at least 12–18 potential loci, design and testing of several degenerate PCR primers for the selected genes, spatio-temporal selection of epidemiologically unrelated strains, determination of strain discrimination and analysis of nucleotide sequences for evolutionary neutrality of the alleles at their loci. Furthermore, to date, the only human *Mycoplasma* species with an established MLSA scheme is *M. pneumoniae* (https://pubmlst.org/databases/). Unlike the genes used in MLSA, R-segment analysis can be a potentially cost-effective and robust complementary molecular tool for species/strains delineation of human mycoplasmas via the direct sequencing of the 50 kb fragments using the targeted sequencing approach of a next-generation DNA sequencing equipment. The R-segment analysis, which utilizes short partial sequences that are < 10% of the length of whole genomes, is also a more economical approach compared to the whole genome analysis because of the substantial reduction in the cost associated with sequencing, and computational analyses. Future studies need to be focused in comparing the robustness of R-segments analysis with other phylogenetic tools for the delineation of the unknown isolates of human *Mycoplasma* species and their strains.

Acknowledgments

The authors gratefully acknowledge the Head, Department of Biological and Chemical Sciences, The University of the West Indies, Cave Hill Campus, Barbados for providing necessary facilities. This research was partly funded by the UWI Campus Research Funds (awarded to BRM) and UWI Graduate Student Research Award (awarded to OR).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2018.12.002.

References

Adeolu, M., Alnajar, S., Naushad, S., Gupta, R.S., 2016. Genome-based phylogeny and taxonomy of the 'Enterobacteriales': Proposal for enterobacterales ord. nov. divided into the families Enterobacteriaceae, Erwiniaceae fam. nov., Pectobacteriaceae fam. nov., Yersiniaceae fam. nov., Hafniaceae fam. nov., Morgane. Int. J. Syst. Evol. Microbiol. 66, 5575–5599. https://doi.org/10.1099/ijsem.0.001485.

Auch, A.F., Klenk, H.P., Göker, M., 2010. Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. Stand. Genomic Sci. 2, 142–148. https://doi.org/10.4056/sigs.541628.

Baseman, J.B., Dallo, S.F., Tully, J.G., 1988. Isolation and characterization of *Mycoplasma genitalium* strains from the human respiratory tract. J. Clin. Microbiol. 26, 2266–2269.

Boesen, T., Fedosova, N.U., Kjeldgaard, M., Birkelund, S., Christiansen, G., 2001. Molecular design of *Mycoplasma hominis* Vaa adhesin. Protein Sci. 10, 2577–2586. https://doi.org/10.1110/ps.ps.31901.

Bouchet, V., Huot, H., Goldstein, R., 2008. Molecular genetic basis of ribotyping. Clin.

Microbiol. Rev. 21, 262–273. https://doi.org/10.1128/CMR.00026-07.
Bove, J.M., 1981. *Mycoplasma* infections of plants. Isr. J. Med. Sci. 17, 572–585.
Citti, C., Blanchard, A., 2013. Mycoplasmas and their host: emerging and re-emerging minimal pathogens. Trends Microbiol. https://doi.org/10.1016/j.tim.2013.01.003.
Citti, C., Nouvel, L.-X., Baranowski, E., 2010. Phase and antigenic variation in mycoplasmas. Future Microbiol 5, 1073–1085. https://doi.org/10.2217/fmb.10.71.
Darling, A.E., Mau, B., Perna, N.T., 2010. Progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5. https://doi.org/10.1371/journal.pone.0011147.
Dehon, P., McGowin, C.L., 2017. The Immunopathogenesis of *Mycoplasma genitalium* Infections in Women: a Narrative Review. Sex. Transm. Dis. 44, 428–432.
Figueras, M.J., Beaz-Hidalgo, R., Hossain, M.J., Liles, M.R., 2014. Taxonomic affiliation of new genomes should be verified using average nucleotide identity and multilocus phylogenetic analysis. Genome Announc. 2, e00927. https://doi.org/10.1128/genomeA.00927-14.
Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., Tiedje, J.M., 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int. J. Syst. Evol. Microbiol. 57, 81–91. https://doi.org/10.1099/ijs.0.64483-0.
Guillermo, M., David, P.M., Santos, Y., Antonio, J., Constantino, G.J., Ramírez, C., Lilia, M., 2014. Presence of *Mycoplasma* spp. in patients with asthma or allergic rhinitis. Adv. Microbiol. 4, 720–725.
Iverson-Cabral, S.L., Astete, S.G., Cohen, C.R., Rocha, E.P.C., Totten, P.A., 2006. Intrastrain heterogeneity of the *mgpB* gene in *Mycoplasma genitalium* is extensive in vitro and in vivo and suggests that variation is generated via recombination with repetitive chromosomal sequences. Infect. Immun. 74, 3715–3726. https://doi.org/10.1128/IAI.00239-06.
Iverson-Cabral, S.L., Astete, S.G., Cohen, C.R., Totten, P.A., 2007. *mgpB* and *mgpC* sequence diversity in *Mycoplasma genitalium* is generated by segmental reciprocal recombination with repetitive chromosomal sequences. Mol. Microbiol. 66, 55–73. https://doi.org/10.1111/j.1365-2958.2007.05898.x.
Jironkin, A., Brown, R.J., Underwood, A., Chalker, V.J., Spiller, O.B., 2016. Genomic determination of minimum multi-locus sequence typing schemas to represent the genomic phylogeny of *Mycoplasma hominis*. BMC Genomics 17, 964. https://doi.org/10.1186/s12864-016-3284-z.
Karlin, S., 1998. Global dinucleotide signatures and analysis of genomic heterogeneity. Curr. Opin. Microbiol. 1, 598–610. https://doi.org/10.1016/S1369-5274(98)80095-7.
Karlin, S., Burge, C., 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. 11, 283–290.
Konstantinidis, K.T., Tiedje, J.M., 2005. Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. U. S. A. 102, 2567–2572. https://doi.org/10.1073/pnas.0409727102.
Kumar, S., Stecher, G., Peterson, D., Tamura, K., 2012. MEGA-CC: Computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. Bioinformatics 28, 2685–2686. https://doi.org/10.1093/bioinformatics/bts507.
Lagesen, K., Hallin, P., Rødland, E.A., Stærfeldt, H.H., Rognes, T., Ussery, D.W., 2007. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35, 3100–3108. https://doi.org/10.1093/nar/gkm160.
Librado, P., Rozas, J., 1 June 2009. DnaSP v5.1: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25 (11), 1451–1452.
Lugli, G.A., Milani, C., Duranti, S., Mancabelli, L., Mangifesta, M., Turroni, F., Viappiani, A., van Sinderen, D., Ventura, M., 2017. Tracking the taxonomy of the genus *Bifidobacterium* based on a phylogenomic approach. Appl. Environ. Microbiol 02249. https://doi.org/10.1128/AEM.02249-17.
Ma, L., Taylor, S., Jensen, J.S., Myers, L., Lillis, R., Martin, D.H., 2008. Short tandem repeat sequences in the *Mycoplasma genitalium* genome and their use in a multilocus genotyping system. BMC Microbiol. 8, 130–142. https://doi.org/10.1186/1471-2180-8-130.
Meier-Kolthoff, J.P., Auch, A.F., Klenk, H.-P., Göker, M., 2013. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics 14, 60. https://doi.org/10.1186/1471-2105-14-60.
Meier-Kolthoff, J.P., Klenk, H.P., Göker, M., 2014. Taxonomic use of DNA G + C content and DNA-DNA hybridization in the genomic age. Int. J. Syst. Evol. Microbiol. 64, 352–356. https://doi.org/10.1099/ijs.0.056994-0.
NCBI Resource, Coordinators, 2016. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 44, D7–D19. https://doi.org/10.1093/nar/gks1189.
Neyrolles, O., Chambaud, I., Ferris, S., Prevost, M.C., Sasaki, T., Montagnier, L., Blanchard, A., 1999. Phase variations of the *Mycoplasma penetrans* main surface lipoprotein increase antigenic diversity. Infect. Immun. 67, 1569–1578.
Nicolson, G.L., Nasralla, M.Y., Franco, A. Robert, Nicolson, N.L., Erwin, R., Ngwenya, R., Berns, P.A., 2000. Diagnosis and integrative treatment of intracellular bacterial infections in chronic fatigue and fibromyalgia syndromes, Gulf war illness, rheumatoid arthritis and other chronic illnesses. Clin. Pract. Altern. Med. 1, 92–102.
Oren, A., Ventosa, A., 2010. International committee on systematics of prokaryotes subcommittee on the taxonomy of *Halobacteriaceae* and subcommittee on the taxonomy of *Halomonadaceae*. Int. J. Syst. Evol. Microbiol. 60, 2257–2259. https://doi.org/10.1099/ijs.0.027078-0.
Pereira, C.E.R., Vannucci, F.A., Gabardo, M.D.P., Lucas, F., Mores, N., Guedes, R.M.C., Vannucci, F.A., Gabardo, M.P., Santos, L.F., Mores, N., 2017. *Mycoplasma hyorhinis* infection in early cases of mycoplasmal pneumonia in swine and evaluation of diagnostic assays 1. Pesq. Vet. Brassica 37, 1057–1063.
Razin, S., 2006. Chapter 1.2.29.The Genus *Mycoplasma* and Related Genera (Class Mollicutes), the Prokaryotes. https://doi.org/10.1007/0-387-30744-3.
Razin, S., Hayflick, L., 2010. Highlights of mycoplasma research-an historical perspective. Biologicals. https://doi.org/10.1016/j.biologicals.2009.11.008.
Richter, M., Rossello-Mora, R., 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl. Acad. Sci. 106, 19126–19131. https://doi.org/10.1073/pnas.0906412106.
Roachford, O., Nelson, K.E., Mohapatra, B.R., 2017. Comparative genomics of four *Mycoplasma* species of the human urogenital tract: Analysis of their core genomes and virulence genes. Int. J. Med. Microbiol. 307, 508–520. https://doi.org/10.1016/j.ijmm.2017.09.006.
Rogers, M.J., Simmons, J., Walker, R.T., Weisburg, W.G., Woese, C.R., Tanner, R.S., Robinson, I.M., Stahl, D. a, Olsen, G., Leach, R.H., 1985. Construction of the *Mycoplasma* evolutionary tree from 5S rRNA sequence data. Proc. Natl. Acad. Sci. U. S. A. 82, 1160–1164. https://doi.org/10.1073/pnas.82.4.1160.
Scolten, S.S., Tozon, N., Koprivec, S., Felda, K., Florjančič, M., Bencina, D., Slavec, B., 2017. Molecular detection and seroprevalence of mycoplasmas in clinicaly healthy working dogs. Slov. Vet. Res. 54, 155–161. https://doi.org/10.26873/SVR-377-2017.
Stackebrandt, E., Ebers, J., 2006. Taxanomic Parameters Revisited. tarnished gold standards. Microbiol. Today. Vol. 33. pp. 152–155.
Stackebrandt, E., Goebel, B.M., 1994. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int. J. Syst. Evol. Microbiol. 44, 846–849. https://doi.org/10.1099/00207713-44-4-846.
Stein, M.A., Baseman, J.B., 2006. The evolving saga of *Mycoplasma genitalium*. Clin. Microbiol. Newsl. https://doi.org/10.1016/j.clinmicnews.2006.03.001.
Thompson, C.C., Vicente, A.C.P., Souza, R.C., Vasconcelos, A.T.R., Vesth, T., Alves, N., Ussery, D.W., Iida, T., Thompson, F.L., 2009. Genomic taxonomy of vibrios. BMC Evol. Biol. 9, 1–16. https://doi.org/10.1186/1471-2148-9-258.
Thompson, C.C., Vieira, N.M., Vicente, A.C.P., Thompson, F.L., 2011. Towards a genome based taxonomy of mycoplasmas. Infect. Genet. Evol. 11, 1798–1804. https://doi.org/10.1016/j.meegid.2011.07.020.
Thompson, C.C., Chimetto, L., Edwards, R.A., Swings, J., Stackebrandt, E., Thompson, F.L., 2013. Microbial genomic taxonomy. BMC Genomics 14. https://doi.org/10.1186/1471-2164-14-913.
Tsang, A., Hwei, L., Siu-Ming, Y., Susanna, L., Patrick, W., 2017. Failure of phylogeny inferred from multilocus sequence typing to represent bacterial phylogeny. Sci. Rep. 7, 4536. https://doi.org/10.1038/s41598-017-04707-4.
Yoon, S.H., Ha, S. min, Lim, J., Kwon, S., Chun, J., 2017. A large-scale evaluation of algorithms to calculate average nucleotide identity. Antonie van Leeuwenhoek. Int. J. Gen. Mol. Microbiol. 110, 1281–1286. https://doi.org/10.1007/s10482-017-0844-4.
Zarei, O., Rezania, S., Mousavi, A., 2013. *Mycoplasma genitalium* and cancer : a brief review. Asian Pacific J. Cancer Prev. 14, 3425–3428.