# CHAPTER 6

# POSTGENOMIC TAXONOMY OF HUMAN UREAPLASMAS

Fanrong Kong, Gwendolyn L. Gilbert

Centre for Infectious Diseases and Microbiology (CIDM), Institute of Clinical Pathology and Medical Research (ICPMR), Westmead, New South Wales 2145, Australia

This chapter is in press in *International Journal of Systematic and Evolutionary Microbiology .*

## Statement of Joint Authorship

**Kong, F.** (candidate)

Did all the molecular work, interpreted the data and wrote the manuscript.

**Gilbert G. L.** (supervisor)

Supervised the overall project, assisted in research design, analysis and interpretation of data, and made a significant contribution to the manuscript.

## 6.1. SUMMARY

In 2000, the full genome sequence of *Ureaplasma parvum* (previously known as *Ureaplasma urealyticum*) serovar 3 was released. In 2002, after prolonged debate, it was agreed that the former *U. urealyticum*, should be divided into two species **–** *U. parvum* and *U. urealyticum*. To provide additional support for this decision and improve our understanding of the relationship between these two species, we studied four "core" genes or gene clusters in ATCC reference strains of all 14 serovars of *U. parvum* and *U. urealyticum.* These core regions were the rRNA gene clusters, the elongation factor Tu genes (*tuf*), urease gene complexes and multiple banded antigen genes (*mba*). The known *U. parvum* genome sequences (GenBank accession number: NC_002162) were used as reference. DNA insertions and deletions (indels) were found in all of the gene regions studied, except *tuf*, but they were found only between, not within the two species. An incidental finding was that there was inter-copy heterogeneity for rRNA gene complex sequences. Sequence analysis (sequence heterogeneity and especially indels) of all four selected targets consistently support the separation of human ureaplasmas into two species. Except for MBA, there was less heterogeneity in amino acid sequences of proteins, between species, than in the nucleic acid seqeuences of the corresponding genes. The level of heterogeneity in amino acid and base sequences at the 5'end of the species-specific regions of MBA were almost identical. Analysis of our results provided an interesting case study to help resolve some common problems in the use of sequence data to infer phylogenetic relationships and support taxonomic changes. We recommend that, to avoid confusion, the new nomenclature be used for human ureaplasmas in future publications.

## 6.2. INTRODUCTION

There have been two recent major developments that affect our understanding of human ureaplasmas. Firstly, the full genome sequence of *U. parvum* (previous

*U. urealyticum*) serovar 3 was released in 2000 (GenBank accession number: NC_002162) (Glass *et al.*, 2000). Secondly, the taxonomy of human ureaplasmas changed in 2002 (Robertson *et al.*, 2002), when the two former *U. urealyticum* biovars were given full species status, as *U. parvum* (previously biovar parvo or biovar 1) and *U. urealyticum* (previously biovar T-960 or biovar 2) (Robertson *et al.*, 2002).

It is now accepted that a decision to create a new species should be based on many independent phenotypic and genotypic characteristics – the theory of "polyphasic taxonomy" (Vandamme *et al.*, 1996). However, molecular methods and genome-based criteria have become more accessible and attractive (Gurtler & Mayall, 2001; Stackebrandt *et al.*, 2002). DNA-DNA hybridization showing <70% homology between whole genomes is accepted as the most definitive criterion or "gold standard" for separate prokaryote species (Murray & Schleifer, 1994). The traditional 70% DNA-DNA hybridization value used to delineate genomic species was found to correspond to genome mispairings in the range 13-13.6% or 0.097-0.104 nucleotide substitutions per site (Mougel *et al.*, 2002). Similarity of <97% in the 16S rRNA gene is the most widely used practical alternative (Murray & Schleifer, 1994), but these criteria may conflict and additional alternative targets are required (Pettersson *et al.*, 2000; Dellaglio, *et al.*, 2004).

The two ureaplasma species exhibit many distinct phenotypic and genotypic properties (including DNA-DNA hybridization showing <70% homology), which support the change in taxonomy and fulfil the requirements of the polyphasic theory (Chrisstiansen *et al.*, 1981; Harasawa *et al.*, 1991; Vandamme *et al.*, 1996). However, continued use of the old single-species nomenclature in some recent publications (Baier *et al.*, 2003; Daxboeck *et al.*, 2003) is potentially confusing. To strengthen the case for acceptance and exclusive use of the new ureaplasma taxonomy (Robertson *et al.*, 2002), we studied four "core" genes/gene clusters – the rRNA gene complex, elongation factor Tu gene (*tuf*), urease gene cluster and multiple banded antigen gene (*mba*) – of all 14 human ureasplasma serovars. These four regions were chosen because previous studies have shown that they were promising targets for study of the phylogeny of ureaplasmas and mycoplasmas (Kamla, *et al.*, 1996; Kong, *et al.*, 1999b, 2000b). In addition, we used this as a case study to help resolve some common problems with the use of sequence data to infer phylogeny and to support the establishment of new taxonomy (Ludwig *et al.*, 1998).

## 6.3. MATERIALS AND METHODS

### 6.3.1. Bacterial strains.

Reference strains of four *U. parvum* and ten *U. urealyticum* serovars, obtained directly from the American Type Culture Collection (ATCC), were the same as those used in our previous studies (Kong *et al.*, 1999a) and are listed in Chapter 1.

### 6.3.2. Value of *U. parvum* serovar 3 genome in oligonucleotide primers design.

The full genome sequence of *U. parvum* serovar 3 (Glass *et al.*, 2000) greatly facilitated sequencing of the selected genes and gene clusters of the other three *U. parvum* and ten *U. urealyticum* serovars. For this study, we used the following steps:  Firstly, to identify conserved regions: we compared known sequences of genes corresponding to our selected genes and gene clusters – rRNA gene complex, *tuf* and urease gene clusters – in other *Mycoplasma* spp. and ureaplasma serovars (Fraser *et al.*, 1995; Himmelreich *et al.*, 1996; Neyrolles *et al.*, 1996), in addition to *mba*, which we and others have studied in detail previously (Zheng, *et al.*, 1995; Kong *et al.*, 1999a). Based on the results, we designed primers and amplified target regions for sequencing. The target regions sequenced were:

a) the whole rRNA cluster, including a short region upstream of 16S rRNA gene (for *U. parvum* serovars only)-16S rRNA gene-16S/23 rRNA intergenic spacer-23S rRNA gene-23S/5S rRNA intergenic spacer-5S rRNA gene-and a short region downstream of 5S rRNA gene;

b) almost the full length of *tuf*;
c) the whole urease gene cluster, including short regions upstream of *ureA*-*ureA*-*ureB*-*ureC*-*ureE*-*ureF*-*ureG*-*ureD*-and downstream of *ureD*.

The amplification and sequencing primers used in the study are shown in Table 6.1. Most primers were used for both amplification and sequencing but some (as inner sequencing primers) were used only for sequencing.

### 6.3.3. DNA preparation and PCR.

DNA preparation and PCR were performed as previously described (Kong *et al.*, 1999b).

### 6.3.4. Sequencing and sequence analysis.

The PCR products were sequenced with Applied Biosystems (ABI) *Taq* DyeDexoy terminator cycle-sequencing kits according to standard protocols. All sites showing unexpected heterogeneity, such as those indicating rRNA gene intercopy sequence variation and the unique heterogeneity site in serovar 13 *tuf* (see below), were sequenced at least twice, to confirm the results. When necessary, different PCR amplicons and/or inner sequencing primers were used for sequencing.

The initial sequencing results were analysed with *Bestfit* program in *Comparison* program group and then joined together to determine sequences of whole genes/gene clusters. The multiple sequence alignments were performed with *Pileup* and *Pretty* programs from the *Multiple Sequence Analysis* program group. All of the programs/program groups are available in WebANGIS (http://www1.angis.org.au/WebANGIS/), ANGIS (Australian National Genomic Information Service).

**Table 6.1. Primers used for sequencing three ureaplasma different genes/gene clusters.**

| **Primer names**[a] | **Target genes/regions**[a] | ***Tm* °C**[b] | **GenBank numbers**[c] | **Primer sequences**[d] |
|---|---|---|---|---|
| 113S1 | UU113 | 68.4 | AE002111 | **9814**GAA GAA CCC ACC AAA TAC GAG CAG**9837** |
| 113S2[e] | UU113 | 62.7 | AE002111 | **9862**TTG TTG GTG AAC AAA AAT ACA TCA**9885** |
| 303S1 | UU303 | 64.9 | AE002127 | **5713**TTG ATG CAA AAA GAT CAG GTT GTA G**5737** |
| 303S2[e] | UU303 | 61.7 | AE002127 | **5800**GAG AAA CAA GCT GAA CAT AAT GAT C**5824** |
| 16S10A[e] | 16S rRNA | 66.4 | AE002127 | **6144**AAT CCT GAG CCA GGA TCA AAC TC**6122** |
| 16S23A | 16S rRNA | 74.3 | AE002127 | **6156**GCC GCC AGC GTT AAT CCT GAG C**6135** |
| SP1623S1 | 16S-23S rRNA spacer | 65.3 | AE002127 | **7830**CTT TCT AAT CAT TGA CAT TAA GTT GTC AGT G**7860** |
| SP1623S2[e] | 16S-23S rRNA spacer | 62.8 | AE002127 | **7843**GAC ATT AAG TTG TCA GTG AAC AGA AAC**7869** |
| 23S32S | 23S rRNA | 63.6 | AE002127 | **7933**CTA AGA GCT TAT GGT GA/GA TGC CTT G**7957** |
| 23S523S | 23S rRNA | 69.4 | AE002127 | **8424**GAA CGG TGA AAA GAA CCC AGA GAT G**8448** |
| 23S503A[e] | 23S rRNA | 65.7 | AE002127 | **8449**CCA TCT CTG GGT TCT TTT CAC C**8428** |
| 23S516A | 23S rRNA | 63.7 | AE002127 | **8464**GGT TCT ATT TCA CTC CCA TCT CTG**8441** |
| 23S1104S | 23S rRNA | 68.7 | AE002127 | **9006**GCA AGG ATG TTG GCT TAG AAG CAG**9029** |
| 23S1082A[e] | 23S rRNA | 68.7 | AE002127 | **9030**GCT GCT TCT AAG CCA ACA TCC TTG**9007** |
| 23S1134S[e] | 23S rRNA | 67.5 | AE002127 | **9034**CGT TTA AAG AGT GCG TAA CAG CTC AC**9059** |
| 23S1117A | 23S rRNA | 70.8 | AE002127 | **9067**CTC GAC AAG TGA GCT GTT ACG CAC TC**9042** |
| 23S1721S | 23S rRNA | 68.9 | AE002127 | **9622**AAG GAA CTC TGC AAA TTA ACC CCG T**9646** |

| | | | | |
|---|---|---|---|---|
| 23S1698A[e] | 23S rRNA | 68.0 | AE002127 | **9647**TAC GGG GTT AAT TTG CAG AGT TCC T**9623** |
| 23S 1729A | 23S rRNA | 68.5 | AE002127 | **9677**TTT TAC AGC GAG CAC CCC TTA TTG**9654** |
| 23S2257S | 23S rRNA | 68.5 | AE002127 | **10158**GAC AGT GTT AGG TGG GCA GTT TGA C**10182** |
| 23S2237A[e] | 23S rRNA | 71.9 | AE002127 | **10186**CCC AGT CAA ACT GCC CAC CTA ACA C**10162** |
| 23S2251A | 23S rRNA | 78.6 | AE002127 | **10198**GGA GGC GAC CGC CCC AGT CAA AC**10176** |
| 23S2578S | 23S rRNA | 74.6 | AE002127 | **10478**GGT TCG GCT GTT CGC CGA TTA AAG AG**10503** |
| 23S2603A | 23S rRNA | 55.8 | AE002127 | **10551**AGA TAG GGA CCA ACC TGT CTC ACG**10528** |
| 23S2772S | 23S rRNA | 65.2 | AE002127 | **10674**AAA CGC TGA AAG CAT CTA AGT GTG**10697** |
| 5S41S | 5S rRNA | 67.7 | AE002128 | **86**GAA ATA CCT GTT CCC ATC CCG A**107** |
| 5S60S[e] | 5S rRNA | 70.0 | AE002128 | **103**CCC GAA CAC AGA AGT CAA GCA CTC**126** |
| 5S45A | 5S rRNA | 67.4 | AE002128 | **134**CGG CTC TAG AGT GCT TGA CTT CTG**111** |
| UU304A1[e] | UU304 | 63.0 | AE002128 | **330**TTC TAA TTG CAA TTC TTC AAG ACG**307** |
| UU304A2 | UU304 | 71.8 | AE002128 | **400**CAC CTT GTT CGC GTG CAT CTT G**379** |
| UU114A1[e] | UU114 | 60.1 | AE002112 | **5246**A/GTT TAT TGT TTT TGG ATA TAC CAC C**5222** |
| UU114A2 | UU114 | 66.2 | AE002112 | **5405**CGT CTT CTG GTG TTT GCA TAA TTG**5382** |
| TUF5S | *tuf* | 61.3 | AE002151 | **1284**TTA ATT TTT AAG GAG ATT TAA AAT GGC**1258** |
| TUF32S[e] | *tuf* | 62.4 | AE002151 | **1256**AAA GCT AAA TTT GAA AGA ACA AAA CC**1231** |
| TUF92S | *tuf* | 60.5 | AE002151 | **1195**ATG GTA AAA CTA CTT TAA CAG CTG C**1171** |
| TUF163A | *tuf* | 61.2 | AE002151 | **1076**TTG TAA TAC CAC GTT CTC TTT CTT C**1100** |
| TUF590S | *tuf* | 67.9 | AE002151 | **697**TTG ATG AAT TAA TGG ACG CAG TTG A**673** |

| | | | | |
|---|---|---|---|---|
| TUF646A | *tuf* | 65.3 | AE002151 | **592**CGT CCT GAA ATT GTG AAT ACA TCT TC**617** |
| TUF874S[e] | *tuf* | 66.9 | AE002151 | **414**AAA AGA AGA TGT TGA ACG TGG TCA AG**389** |
| TUF985A | *tuf* | 61.1 | AE002151 | **253**TCT GTT GTT CTA AAA TAG AAT TGT GG**278** |
| TUF1132A[e] | *tuf* | 65.0 | AE002151 | **107**GAC CTA CAG TTT TAC CAC CTT CAC G**131** |
| TUF1159A | *tuf* | 59.0 | AE002151 | **77**ATT AAT TAC TTG TTT TAA TTA CGC TAC C**104** |
| UC424S | *ureC* | 65.6 | AE002140 | **2676**CAG CTG GTG GTT TAG ATA CTC ACG**2653** |
| UC429S[e] | *ureC* | 64.8 | AE002140 | **2672**TGG TGG TTT AGA TAC TCA CGT TCA C**2648** |
| UC910S | *ureC* | 64.9 | AE002140 | **2181**TTT TAC CAG CTT CTA CAA ACC CAA C**2157** |
| UC947A[e] | *ureC* | 60.6 | AE002140 | **2094**AAG TGG TGA CAT ACC ATT AAC ATA TC**2119** |
| UC959A | *ureC* | 60.5 | AE002140 | **2083**CCT TAG GAT TTA AGT GGT GAC ATA C**2107** |
| UC1418S | *ureC* | 68.0 | AE002140 | **1683**GG/TG ATC CAA AC/TG CTT CAA TTC CAA C**1659** |
| UC1450A | *ureC* | 61.0 | AE002140 | **1601**A/GC/TT AGT TAA C/TA/GA ACG TCC ATA A/TGT TCC**1624** |
| UE93S | *ureE* | 59.5 | AE002140 | **1147**GAA CAT TCA TTT AAC AAG CGA CGA C**1126** |
| UE119A | *ureE* | 62.9 | AE002140 | **1073**CCA TAT TCA ACA TTT TGA TCT GAT G**1097** |
| UE143S[e] | *ureE* | 62.9 | AE002140 | **1097**CAT CAG ATC AAA ATG TTG AAT ATG G**1073** |
| UF217S | *ureE* | 59.5 | AE002140 | **632**GCG TTA CTT CTG TAT ATG AAT GAA C**608** |
| UF222A[e] | *ureF* | 60.0 | AE002140 | **578**AAA TTG CC/TA ATA AAT CAC CAT GTA AC**603** |
| UF320A | *ureF* | 63.1 | AE002140 | **484**CGA GTC TCT CTT GCT AAA CCT TG**506** |
| UF721S | *ureF* | 64.7 | AE002140 | **129**C/TCT TGA AAT TGC ACA AAT GGA AC**107** |
| UG3A | *ureG* | 62.4 | AE002139 | **11152**CCT ACA CCA ATA ATT AAT GGT CTT TTC**11178** |
| UG451A[e] | *ureG* | 63.4 | AE002139 | **10706**CAC CAA CAT AAG GAG CTA AAT CAA C**10730** |

| | | | | |
|---|---|---|---|---|
| UG475S | *ureG* | 63.4 | AE002139 | **10730**GTT GAT TTA GCT CCT TAT GTT GGT G**10706** |
| UG499A | *ureG* | 60.6 | AE002139 | **10655**CTT TAT TAC CAC GTG ATT TTA ATG TAT C**10682** |
| UD332S | *ureD* | 59.0 | AE002139 | **10246**CA/TG AA/GC AAC AC/TA CAA ATA TC/TA CA/GT TAG G**10219** |
| UD379A | *ureD* | 62.0 | AE002139 | **10147**TTA AAT TGG/T GCA/G AAC/T TTT CCA TCT TC**10172** |
| UD841A | *ureD* | 62.3 | AE002139 | **9684**CTT C/TTA TGG TTT TCG TAA AAT TAA A/TGG**9710** |
| UU427A1[e] | UU427 | 60.4 | AE002139 | **9446**AAT AAA TTT TGC TAA AAA AGG CAT AC**9471** |
| UU427A2 | UU427 | 56.5 | AE002139 | **9330**GT/CA/T GGT/C TTA AAA T/CTA ACA TCT ACA C**9354** |
| UU427A3 | UU427 | 63.1 | AE002139 | **9175**CAT CAT CAA AAT CTT TAA TAC CAT CAT C**9199** |

*Notes.*

a. Primers were named according to their target genes/regions, the 3'-end locations (the distance from the beginning of correspondent genes/regions), and directions of primers (sense or antisense).

b. The melting temperature (*T*m) values were provided by the primer synthesiser (Sigma-Aldrich, Castle Hill, NSW, Australia).

c. All the GenBank sequences were from the related sections of *U. parvum* serovar three full genome sequences.

d. Numbers represent the numbered base positions at which primer sequences start and finish (numbering start point "1" refer to the start points "1" of correspondent gene GenBank accession numbers). Letters behind "/" indicate alternative nucleotides in different species/serovars, which were based on the comparison with the other related GenBank sequences.

e. Primers used only for sequencing; all the other primers were used for both PCR and sequencing.

### 6.3.5. Nucleotide sequence accession numbers.

The sequence data were deposited into the GenBank Nucleotide Sequence databases with the following accession numbers: rRNA gene complex: AF272599-AF272604, AF073446-AF073459, AF059322-AF059335 and AF272605-AF272630; *tuf*: AF270758-AF270770; urease gene clusters: AF085720-AF085733; *mba*: AF055358-AF055367 and AF056982-AF056984.

## 6.4. RESULTS AND DISCUSSION

### 6.4.1. Advantages of sequencing multiple strains.

In this study, as in our previous study of *mba* (Kong, *et al.*, 2000b), we sequenced the three target genes/gene clusters for 13 ureaplasma serovars (excluding serovar 3 sequences from the full genome, which were used as reference). The advantages of this approach are:

i) since the genes/gene clusters are relatively conserved between serovars within each species, the results for different serovars help to confirm the accuracy of sequencing results (Clayton *et al.*, 1995);

ii) the results help to differentiate interspecies from intraspecies heterogeneity (Mygind *et al.*, 1998).

A previous study showed that there is some intra-serovar, as well as inter-serovar/intra-species and inter-species heterogeneity in *mba* (Knox, *et al.*, 1998). However, there is limited intra-species heterogeneity in the other three gene regions studied and, even in *mba,* intraspecies heterogeneity is much less than between species (Kong, *et al.*, 1999b). Therefore, a single reference strain of each serovar provides enough examples of each species to demonstrate inter-species heterogenity, which was the focus of this study.

### 6.4.2. Key characteristics of the "core genes".

#### 6.4.2.1. Inter-copy polymorphisms of the rRNA gene complex.

There was relatively little interspecies heterogeneity in 16S and 23S rRNA genes but considerably more in the two copies of the 5S rRNA genes and the corresponding intergenic spacer regions (Table 6.2.). In common with other targets studied, the intraspecies heterogeneities in these genes were minor and some were assumed to be due to intercopy differences between duplicate copies (Table 6.2.). Previous studies have shown sequence variation in duplicate copies of rRNA genes of other mollicutes (Pettersson *et al.*, 1996). Analysis of the two copies of the rRNA gene complex in *U. parvum* serovar 3 genome (GenBank accession number: AE002111+AE002112 and AE002127+AE002128) showed inter-copy heterogeneity between 16S rRNA genes (one site), 16S-23S rRNA intergenic spacer regions (two sites), 23S rRNA genes (four sites) and 5S rRNA genes (one site) but none in the 23S-5S rRNA intergenic spacer regions. In sequences of the corresponding genes of the other 13 serovars the result was "N" (i.e. unknown or unidentifiable nucleotide) rather than "A, T, C or G", at several sites, even after repeat sequencing or use of different amplication and sequencing primers. We assumed that most, if not all, of these were due to inter-copy polymorphisms (Table 6.2.) (Ueda *et al.*, 1999). If these inter-copy polymorphisms were ignored, the intraspecies heterogeneity in rRNA gene complexes between the two human ureaplasma species was very low. In future, the design of primers or probes or study the phylogenetic relationships should take account of polymorphisms between multi-copy rRNA gene complexes (Gurtler, 1999).

#### 6.4.2.2. *U. urealyticum* serovar 13 EF-Tu gene (*tuf*).

Previous studies have shown that differences in *tuf* can distinguish species and may reflect some phenotypic relationships better than 16S rRNA gene (Kamla *et al.*, 1996). EF-TU gene (*tuf*) DNA sequences were the same in serovars within each species, except for that of serovar 13 of *U. urealyticum*. It contains two base differences (but the same amino acid sequences) compared with the other nine *U. urealyticum* serovars. This difference is of interest in view of another reported difference between serovar 13, which gives an intermediate response in the $Mn^{2+}$ (manganese)-inhibition test, and all other *U. urealyticum* serovars, which are fully inhibited (Robertson & Chen, 1984).

#### 6.4.2.3. Intraspecies and interspecies heterogeneity of urease gene clusters.

There were 21, nine and one intraspecies heterogeneity sites in the urease gene cluster DNA sequences for *U. parvum* serovars 1, 6, and 14 (compared with serovar 3), respectively. There were seven heterogeneity sites (or 12 base pairs – one site has 6 bp difference) in *U. urealyticum* serovar 2, compared with all other *U. urealyticum* serovars. These results show greater heterogeneity between urease gene clusters of *U. parvum* serovars than between those of *U. urealyticum* serovars, as we found previously for MBA genes (Kong *et al.*, 1999a).

Interspecies heterogeneity between urease genes clusters of the two species was greater in the intergenic spacer regions (where it varies from 16.8-31.8%) than in the genes themselves (where the range of heterogeneity is 5.9-9.7%). Variation in amino acid sequences, between species, is less (range 0.97-6.7%) than in nucleic acid sequences of urease genes (Table 6.3.).

#### 6.4.2.4. The molecular clock is different for different MBA gene regions.

As we described previously, different MBA regions apparently evolve at different rates i.e. according to different “molecular clocks” (Bromham & Penny, 2003). The upstream regions are more heterogeneous than MBA gene themselves (Kong *et al.*, 1999a) and the repetitive regions more heterogeneous than the N-terminal regions. This should be taken into account when using different regions to infer the phylogeny (Kong *et al.*, 1999a).

**Table 6.2. Comparison of inter-species, intra-species and inter-copy heterogeneity in rRNA gene complexes of *U. parvum* and *U. urealyticum*.**

| Genes/regions | DNA length | Heterogeneity sites: N (%) | | |
|---|---|---|---|---|
| | | Interspecies[b] | *U. parvum* intraspecies | *U. urealyticum* intraspecies |
| **16S rRNA gene** | 1513[a] | 14 (0.93) | 2 (1[c]) | 2 (1[c]) |
| **16S-23S rRNA gene spacer** | 302 | 13 (4.3) | 2 (2[c]) | 0 |
| **23S rRNA gene** | 2903 | 26 (0.90) | 8 (7[c]) | 10 (9[c]) |
| **23S-5S rRNA gene spacer** | 71 | 3 (4.2) | 0 | 0 |
| **5S rRNA_1 gene** | 115 | 7 (6.1) | 1 (1[c]) | 3 (3[c]) |
| **5S rRNA_2 gene** | 115 | 8 (7.0) | | |
| **5S rRNA_1 gene-UU114 spacer[d]** | 136 | 31 (3.8) | - | - |
| **5S rRNA_2 gene-UU304 spacer[d]** | 105 | 20 (19.0) | - | - |

*Notes.*

a. Length modification was based upon two other mollicutes 16S rRNA genome annotations (Fraser *et al.*, 1995; Himmelreich *et al.*, 1996), especially considering *M. pneumoniae* annotation (Himmelreich *et al.*, 1996).

b. *U. parvum* and *U. urealyticum* interspecies heterogeneity sites were determined independently of intraspecies heterogeneity sites.

c. Numbers in parenthesis were assumed numbers to contain inter-copy heterogeneity (see text for further explanation).

d. The rRNA gene complex (or 5S rRNA gene) downstream external spacer regions.

#### 6.4.2.5. Indels.

Analysis of insertions and deletions (indels) is a very useful tool with which to study bacterial phylogeny (Britten *et al.*, 2003; Gupta & Griffiths, 2002). We compared the distribution of indels in the four core genes/gene clusters of four *U. parvum* serovars with those of ten *U. urealyticum*. All indels were consistent between serovars within each species, which strongly supports the separation of *Ureaplasma* spp., based on indels. The rRNA gene complex of *U. parvum* differed from that of *U. urealyticum* as follows:

a) a TGTG insertion in 16S rRNA gene;
b) an AT (for operon 1) or A and C (for operon 2) deletion and AT insertion (for operons 1 and 2) in the 16S-23S rDNA intergenic spacer region;
c) a G insertion in 23S-5S rDNA intergenic spacer region;
d) a TTAGG (for operon 1) or AAAAA (for operon 2) deletion in the 5S rRNA gene.

There were no indels in the EF-TU gene (*tuf*). In the urease gene clusters, there were:

a) a TCAAT deletion in the *ureA-ureB* spacer;
b) AAC, T and CTA insertions in *ureB-ureC* spacer;
c) a CA deletion in *ureC-ureE* spacer and
d) an ACATT insertion in the *ureF-ureG* spacer.

Despite these specific differences, the numbers of insertions or deletions, sites and total number of bases in these three genes were not significantly different between the two species. In *mba*, there were no indels in species-specific sites, but there was an AAATT insertion, an AA deletion, a 45 bp deletion and a TC deletion in *U. parvum* upstream of *mba* (Kong, *et al.*, 2000b).

### 6.4.3. Genes, intergenic spacers or gene clusters?

Many studies have shown that intergenic spacer regions are more heterogeneous than the neighbouring genes (Garcia-Martinez *et al.*, 1999; Kong, *et al.*, 1999b). Our study confirmed this by showing greater heterogeneity in the intergenic spacers, especially the external spacer regions, of both the rRNA gene complex and the urease gene clusters compared with the corresponding genes (Tables 6.2. and 6.3.) (Jung *et al.*, 2003). In addition, for urease gene clusters and *mba*, indels only existed in the gene spacer regions. Because the gene cluster as a whole is a functional group, we suggest that there are many advantages in considering them as a unit in basic and applied research.

i) Whole gene clusters contain both conserved and variable sequences and phylogenetic data derived from them are stable and discriminatory (Gurtler, 1999), which is valuable in solving taxonomic problems (Harasawa, 1999).

ii) Species-specific primer pairs based on whole gene clusters are generally more specific and easier to be designed than primers based on any single component (Kong *et al.*, 2000a).

### 6.4.4. DNA or protein sequence? Which protein or gene region?

To fulfil polyphasic theory requirements (Vandamme *et al.*, 1996), DNA sequences and protein amino acid sequences should be considered together (Agosti *et al.*, 1996). However, DNA sequences often reflect the phylogeny more accurately and have greater (about double) discriminatory power (Simmons *et al.*, 2002). Our study showed that, the *mba* species-specific region (5'-end or N-terminal) DNA (67/430= 15.6%) and the corresponding amino acid sequences (24/147=16.3%) have nearly identical levels of heterogeneity (Kong *et al.*, 2000b). However, urease gene subunit (Table 6.3.) and EF-TU gene DNA sequences are more heterogeneous than their corresponding protein amino acid sequences. For example, for the ureaplasma EF-TU gene DNA sequence heterogeniety was 54 of 1185 (4.6%) bases compared with 2 of 394 (0.5%) differences in amino acids between the two species. Presumably, genetic changes that significantly alter the structure, and therefore the function, of proteins such as enzymes are incompatible with survival. On the other hand, genetic

**Table 6.3. Comparison of interspecies heterogeneity of DNA and amino acid sequences of the urease gene clusters of *U. parvum* and *U. urealyticum*.**

| **Genes/regions** | **DNA** | | **Amino acid** | |
|---|---|---|---|---|
| | **Length (bases)** | **Heterogeneity: N (%)** | **Length (a.a.)** | **Heterogeneity: N (%)** |
| **Upstream of *ureA***[a] | 149 | 25 (16.8) | | |
| ***ureA*** | 306 | 18 (5.9) | 101 | 5 (5.0) |
| ***ureA*-*ureB* spacer** | 51 | 9 (17.6) | | |
| ***ureB*** | 375 | 31 (8.3) | 124 | 6 (4.8) |
| ***ureB*-*ureC* spacer** | 45 | 11 (24.4) | | |
| ***ureC*** | 1797 | 175 (9.7) | 598 | 26 (4.3) |
| ***ureC*-*ureE* spacer** | 66 | 21 (31.8) | | |
| ***ureE*** | 450 | 37 (8.2) | 149 | 10 (6.7) |
| ***ureF*** | 753 | 64 (8.5) | 250 | 15 (6.0) |
| ***ureF*-*ureG* spacer** | 81 | 18 (22.2) | | |
| ***ureG*** | 621 | 41 (6.6) | 206 | 2 (0.97) |
| ***ureG*-*ureD* spacer** | 10 | 3 (30) | | |
| ***ureD*** | 864 | 82 (9.5) | 287 | 15 (5.2) |
| ***ureD*-UU427spacer**[b] | 122 | 35 (28.7) | | |

*Notes.*

a. The urease complex (*ureA*) up stream external spacer region.

b. The urease complex (*ureD*) down stream external spacer region.

variation that causes antigenic variation in MBA is not only consistent with survival but also an advantage if it helps the organism to evade the host immune immune response.

Many surface protein antigen genes are used to study the phylogeny of different microbes and to develop practical identification and typing schemes (Bush & Everett, 2001; Stackebrandt, *et al.*, 2002). Sometimes, the gene or even gene region selected can significantly affect the results (Bromham & Penny, 2003). For example, ureaplasma MBA genes contain both species-specific (5'-end or N-terminal) and serovar definition sites (repetitive regions or 3'-end). Thus, the 5'-end or N-terminal would be the appropriate region for studying species-level phylogeny, rather than the repetitive regions (Zheng, *et al.*, 1995). If different bacterial species share almost identical protein antigens as a result of lateral gene transfer (Lawrence, 2002) – for example the *Streptococcus agalactiae* Alp3 protein and *Streptococcus pyogenes* R28 protein (Stalhammar-Carlemalm *et al.*, 1999) – the corresponding genes lose their value for studying species-level taxonomy (Thornton, 2002).

### 6.4.5. Why "core" genes?

Of the two ureaplasma species **–** *U. parvum* and *U. urealyticum* (Robertson *et al.*, 2002) – a full genome sequence was available only for *U. parvum* (Glass *et al.*, 2000). In future, our understanding of human ureaplasmas would be significantly improved by availability of the full genome sequence of *U. urealyticum* also. In particular, it would help to elucidate the nature and significance of the >80 kbp size difference between the two human ureaplasma species (Robertson *et al.*, 1990;

Fraser, *et al.*, 2000) in reverse evolution (Rocha & Blanchard, 2002) and pathogenesis (Povlsen *et al.*, 2002). Meanwhile, alternative strategies such as analysis of selected “core” genes or gene clusters (as in this study) can be used to infer the phylogenetic relationship between species (Daubin *et al.*, 2002). The rationale for the choice of these four genes was that the rRNA gene cluster (Stackebrandt, *et al.*, 2002) and *tuf* (Kamla, *et al.*, 1996) have been widely accepted targets for phylogenetic/taxonomic studies and the urease gene cluster and ***mba*** are unique determinants (Stackebrandt, *et al.*, 2002) of ureaplasma metabolism (Neyrolles, *et al.*, 1996) and antigenicity (Zheng, *et al.*, 1995).

## 6.5. CONCLUSION

Analysis of four “core” genes/gene clusters further supported the establishment of two separate human ureaplasma species **–** *U. parvum* and *U. urealyticum.* Significant differences between genes/gene clusters in the degree of heterogeneity between and within species sheds further light on the relationships between them and makes a useful case study to help understand common problems in use of sequence data to infer phylogeny and support taxonomic change (Ludwig *et al.*, 1998).