Genetic diversity of *Pneumocystis jirovecii* strains based on sequence variation of different DNA region

MOHAMED ALI JARBOUI, FATMA MSEDDI, HAYET SELLAMI, AMIRA SELLAMI, FATMA MAKNI & ALI AYADI
Fungal and Parasitic Molecular Biology Laboratory, School of Medicine, University of Sfax, Sfax, Tunisia

*Pneumocystis jirovecii* is an important opportunistic pathogen that causes severe pneumonia in immunocompromised patients. The aim of the present study was to investigate the genetic diversity of *P. jirovecii* strains by direct sequencing and analysis of the Upstream Conserved Sequence (UCS) region, mitochondrial large-subunit (mt LSU) rRNA and dihydrofolate reductase (DHFR) genes. We identified the polymorphisms in *P. jirovecii* strains of 15 immunocompromised patients, as well as detecting a new tandem repeat of 5 nucleotides in UCS region. The following three different types of repeat unit were found: type a GCCCA; type b GCCCT; and type c GCCTT. In addition, we identified the repeat unit which consisted of 10 nucleotides and three different patterns of UCS repeats with 3 and 4 repeats, i.e., 1, 2, 3 (86.7%), 1, 2, 3, 3 (6.6%) and a new genotype 2, 2, 3, 3 (6.6%). The polymorphism in the mtLSUrRNA gene was seen primarily at position 85 where we detected three different genotypes. Genotype 3 and genotype 2 were the most abundant with frequencies of 53.3% and 40%, respectively. With regard to the DHFR gene, only two (20%) patients had nucleotide substitution in position 312. In conclusion, the multilocus analysis facilitated the typing of *P. jirovecii* strains and proved the important genetic biodiversity of this fungus.

Keywords genetic diversity, multilocus genotyping, PCR, *Pneumocystis jirovecii*, sequencing

Introduction

*Pneumocystis jirovecii* is a fungus that causes pneumonia in immunocompromised patients, especially those with AIDS [1]. The study of *Pneumocystis* is hampered by the lack of a culture-based system, and therefore, molecular methods are employed for typing and determining strain genotypes. Although numerous methods have been described, direct DNA sequencing remains the most commonly employed method for studies of *Pneumocystis* biodiversity and molecular types [2–4]. Sequence analysis of the thymidylate synthase (TS) and superoxide dismutase (SODA) gene loci, and the mitochondrial small subunit ribosomal RNA (mt SSU rRNA) locus have been used to distinguish *Pneumocystis* species among diversity of mammalian hosts [5–7]. Because of the generally low sequence divergence among *P. jirovecii* strains at these loci, they are not very useful in discriminating *P. jirovecii* types. Several additional loci have proved helpful for molecular epidemiologic applications, including the internal transcribed spacer (ITS) regions of the nuclear rRNA operon [8–11], the mitochondrial large subunit ribosomal RNA locus (mt LSU rRNA) [5], the dihydropteroate synthase (DHPS) gene [12], and dihydrofolate reductase (DHFR) gene [4].

In addition, the major surface glycoprotein (MSG) is the most abundantly expressed protein on the surface of *Pneumocystis*. It appears to play a critical role in the pathogenesis of pneumocystosis. MSG is encoded by a multigene family, of which only one is transcribed per organism at any given time because they come from a unique locus called the upstream conserved sequence (UCS) [3,13–20]. In *P. carinii*, the single copy UCS region probably encodes a signal peptide that is necessary for entry of the MSG into
the endoplasmic reticulum for further processing and export to surface [19].

The presence of variable numbers of repeats in the UCS region of *P. jirovecii* has been reported and quantification of them provides a new method for its rapid typing [3,13]. Genetic typing is of use in the investigation of *Pneumocystis* biodiversity [5], person-to-person transmission [8,9], and analysis of the resistance genes [4,12].

The aim of the present study was to look for genetic diversity in *P. jirovecii* strains by analysis of the Upstream Conserved Sequence (UCS) region, mitochondrial large-subunit (mtLSU) rRNA and dihydrofolate reductase (DHFR) genes.

**Materials and methods**

**PCR amplification**

The study was performed with samples from 15 immunocompromised patients including 12 HIV-positive patients, two with kidney transplants and one with leukaemia. They all had *Pneumocystis* pneumonia with typical symptoms (dyspnea, dry cough and fever) and chest X-rays with interstitial infiltrations. These patients can be considered to be unrelated in time and place but the *P. jirovecii* genotyping was necessary for identification of major, minor and new genotypes in our population. Clinical specimens (bronchoalveolar lavage and sputum) recovered from these patients were centrifuged at 1500 g for 5 min and the pellet was digested overnight by proteinase K at 56°C. DNA was extracted using a commercial kit (QIAamp DNA Minikit; Qiagen GmbH, Hilden, Germany) and was conducted in 50 μl containing: 2.5 mM MgCl2; 0.25 mM each of dATP, dTTP, dCTP and dGTP; 0.25 mM of each primer; and 2.5U of Taq DNA polymerase (Promega, Madison, WI, USA). The touchdown thermal cycling protocol consisted of 10 cycles of 45 s at 95°C and 1 min at 65°C, with a decrease by 1°C every cycle to reach 55°C in the last cycle of the period, and 1 min at 72°C, followed by 25 cycles of 45 s at 95°C, 1 min at 55°C, and 1 min at 72°C.

The same extracted DNA and PCR mixture were used to amplify a fragment of the DHFR gene in samples from 10 patients by using the primers FR208 (5' -GCAG AAAGTAGGTACATTATTACGAGA-3') and FR1038 (5' -AACCAGTTACCTAATCAA ACTATATGTC-3') [23]. All samples collected from all 15 patients were subjected to PCR, and only 10 produced a product. This DNA fragment does not represent the whole gene, but the coding sequence is interrupted by a small intron (42 bp) located between nucleotide positions 267 and 308. We have amplified this segment of the DHFR gene because it contained the nucleotide substitution sites [23,24]. The thermal cycling protocol consisted on 2 min at 94°C, 35 cycles of denaturation (94°C, 1 min), annealing (52°C, 1 min) and extension (72°C, 2 min). The amplification product of 850 bp was detected by electrophoresis on a 1% agarose.

To avoid contamination, each step (DNA extraction, master mix preparation and amplification) was performed in separate rooms with different sets of micropipettes and using barrier tips. PCR mixtures were prepared in a laminar-flow cabinet and several controls were included in the PCR experiments. All amplifications were performed in parallel with a negative control (ultrapure distilled water) and a positive control (BAL samples of patients with definite *Pneumocystis* pneumonia and *P. jirovecii* positive detection by toluidine blue O, Giemsa and immunofluorescent assay).

**DNA sequencing and sequence analysis**

The PCR amplification products were purified by using the MiniElute PCR purification kit (Qiagen GmbH, Hilden, Germany) and were sequenced using the primers: pAZ102-E (5' -GGGTCTGTTTCTGACGGCCA-3') and pAZ102-H (5' -GTGTACGTTGCAGAGTA-3') [22]. The same DNA was also employed for all other PCR studies.

For amplification of the UCS genomic DNA sequence of the *P. jirovecii* UCS region, we used two primers ML40 (5' -TTACGAGGAGTTGGAATG-3') (corresponding to nt 1353–1370) and ML664 (5' -CCAGGCTCCCAACAAATG-3') (complementary to nt 1961–1977) [13]. The primers were designed on the basis of the UCS genomic DNA sequence of the *P. jirovecii* MSG gene (GenBank accession no. AF367050) [14]. The association of the UCS with the transcribed MSG variant was identified previously and the size was estimated to be 3.6 kb. An intron of 480 bp was detected in the UCS by comparison of the genomic and cDNA sequences but varied in size among *Pneumocystis* species. The part of amplified region constituted the intron of UCS and contained a region of tandem repeats that varied in number in different isolates [13,14].

The PCR reaction was carried out in 50 μl containing: 2.5 mM MgCl2, 0.25 mM each of dATP, dTTP, dCTP and dGTP, 0.25 mM of each primer, and 2.5U of Taq DNA polymerase (Promega, Madison, WI, USA). The touchdown thermal cycling protocol consisted of 10 cycles of 45 s at 95°C and 1 min at 65°C, with a decrease by 1°C every cycle to reach 55°C in the last cycle of the period, and 1 min at 72°C, followed by 25 cycles of 45 s at 95°C, 1 min at 55°C, and 1 min at 72°C.

The same extracted DNA and PCR mixture were used to amplify a fragment of the DHFR gene in samples from 10 patients by using the primers FR208 (5' -GCAG AAAGTAGGTACATTATTACGAGA-3') and FR1038 (5' -AACCAGTTACCTAATCAA ACTATATGTC-3') [23]. All samples collected from all 15 patients were subjected to PCR, and only 10 produced a product. This DNA fragment does not represent the whole gene, but the coding sequence is interrupted by a small intron (42 bp) located between nucleotide positions 267 and 308. We have amplified this segment of the DHFR gene because it contained the nucleotide substitution sites [23,24]. The thermal cycling protocol consisted on 2 min at 94°C, 35 cycles of denaturation (94°C, 1 min), annealing (52°C, 1 min) and extension (72°C, 2 min). The amplification product of 850 bp was detected by electrophoresis on a 1% agarose.

To avoid contamination, each step (DNA extraction, master mix preparation and amplification) was performed in separate rooms with different sets of micropipettes and using barrier tips. PCR mixtures were prepared in a laminar-flow cabinet and several controls were included in the PCR experiments. All amplifications were performed in parallel with a negative control (ultrapure distilled water) and a positive control (BAL samples of patients with definite *Pneumocystis* pneumonia and *P. jirovecii* positive detection by toluidine blue O, Giemsa and immunofluorescent assay).

**DNA sequencing and sequence analysis**

The PCR amplification products were purified by using the MiniElute PCR purification kit (Qiagen GmbH, Hilden, Germany) and were sequenced using the primers: pAZ102-E and pAZ102-H for the fragment of mtLSUrRNA gene, ML40 and ML664 for UCS region, and the primers FR208 and FR1038 for DHFR gene. Sequencing reactions were performed using the BigDye Terminator technology according to the manufacturer’s protocol (Applied Biosystems), and products were analyzed in an ABI 3100 automated DNA sequencer (Applied Biosystems). Data obtained with forward and reverse sequencing primers
were combined, and the alignments were obtained using BioEdit version 7.

Results

Analysis of UCS region by direct sequencing

The UCS region was amplified and sequenced successfully using the ML40 and ML664 primers for the samples from the 15 patients studied. At the UCS locus, we observed three different variations in a previously described 10 nucleotides repeat unit [3,13,14], i.e., ACAGGCAGCAT (type 1), ACATGCAGCAT (type 2), and GCAGGCAGCAT (type 3).

By direct sequencing of PCR products from samples of the 15 patients, we identified three different patterns of UCS repeats with 3 and 4 repeats, that is, 13 of 15 samples had the same genotype (1, 2, 3), one sample had genotype 1, 2, 3, 3 and one sample contained a new genotype (2, 2, 3, 3) (Fig. 1).

We have detected new repeat units of five nucleotides in the UCS region of *P. jirovecii* strains localized in position number 1778 for strains with genotype (1, 2, 3) and in position 1788 for strains with genotype (1, 2, 3, 3) or (2, 2, 3, 3). Three different types of repeat units were identified; type a GCCCA, type b GCCCT, and type c GCCTT (Fig. 2). These repeat units were present in varying numbers in specimens from different patients. Thirteen patients had four repeat units corresponding to sequence GCCCTGCCCCCTGCT (genotype a, b, b, c) but samples of two patients had only three repeat units corresponding to sequence GCCCTGCCCTGCCT (genotype b, b, c).

We have detected the substitution of ‘TT’ by ‘AA’ immediately before the region of tandem repeats in these two patients. The sequencing of PCR product was repeated several times.

The polymorphism in UCS region has been studied in six sites of single nucleotide polymorphisms, which correspond to nucleotide 1501, 1533, 1624, 1696, 1802 and 1889 [3,13]. The nucleotide substitution were observed in position 1501 (C→T), 1533 (A→G), 1696 (G→A) and 1889 (G→A). Every nucleotide substitution has been detected with the frequency of 13.33% (two strains). Nine strains had conserved a ‘C’, ‘A’, ‘G’ and ‘G’ in these positions respectively (Table 1). No substitutions were detected in nucleotide number 1624 and 1802 with conserving a G and C respectively.

Analysis of mtLSUrRNA and DHFR gene

All samples were positive for the mtLSUrRNA gene with a 346 bp fragment on agarose gels. In our study, all strains

![Fig. 1](https://academic.oup.com/mmy/article-abstract/51/6/561/974851)  
Example of UCS sequence for three detected genotypes: (A) genotype 1, 2, 3; (B) genotype 1, 2, 3, 3; (C) genotype 2, 2, 3, 3. This Figure is reproduced in color in the online version of Medical Mycology.

© 2013 ISHAM, Medical Mycology, 51, 561–567
**Fig. 2** UCS sequences of (b, b, c) and (a, b, b, c) genotypes. This Figure is reproduced in color in the online version of Medical Mycology.

<table>
<thead>
<tr>
<th>Patient</th>
<th>Disease</th>
<th>UCS genotype</th>
<th>Nucleotide substitution</th>
<th>mtLSU rRNA genotype</th>
<th>DHFR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HIV</td>
<td>1, 2, 3/a, b, b, c</td>
<td>1501 (C→T) 1533 (A→G) TGGG</td>
<td>Genotype 3</td>
<td>ND</td>
</tr>
<tr>
<td>2</td>
<td>HIV</td>
<td>1, 2, 3/b, b, c</td>
<td>1669 (G→A) CAAG</td>
<td>Genotype 3</td>
<td>Wild type</td>
</tr>
<tr>
<td>3</td>
<td>HIV</td>
<td>1, 2, 3/a, b, b, c</td>
<td>CAGG</td>
<td>Genotype 2</td>
<td>ND</td>
</tr>
<tr>
<td>4</td>
<td>HIV</td>
<td>1, 2, 3/a, b, b, c</td>
<td>CAGG</td>
<td>Genotype 2</td>
<td>Mutant 312 (T to C)</td>
</tr>
<tr>
<td>5</td>
<td>HIV</td>
<td>1, 2, 3/a, b, b, c</td>
<td>1501 (C→T) 1889 (G→A) T A G A</td>
<td>Genotype 2</td>
<td>Wild type</td>
</tr>
<tr>
<td>6</td>
<td>Kidney transplant</td>
<td>1, 2, 3/b, b, c</td>
<td>1669 (G→A) CAAG</td>
<td>Genotype 3</td>
<td>Wild type</td>
</tr>
<tr>
<td>7</td>
<td>HIV</td>
<td>1, 2, 3/a, b, b, c</td>
<td>1501 (C→T) 1889 (G→A) T A G A</td>
<td>Genotype 2</td>
<td>Wild type</td>
</tr>
<tr>
<td>8</td>
<td>Leukaemia</td>
<td>1, 2, 3/a, b, b, c</td>
<td>CAGG</td>
<td>Genotype 3</td>
<td>ND</td>
</tr>
<tr>
<td>9</td>
<td>Kidney transplant</td>
<td>1, 2, 3/a, b, b, c</td>
<td>CAGG</td>
<td>Genotype 2</td>
<td>Wild type</td>
</tr>
<tr>
<td>10</td>
<td>HIV</td>
<td>1, 2, 3/a, b, b, c</td>
<td>1889 (G→A) CGAG</td>
<td>Genotype 3</td>
<td>ND</td>
</tr>
<tr>
<td>11</td>
<td>HIV</td>
<td>1, 2, 3/a, b, b, c</td>
<td>1533 (A→G) CGGG</td>
<td>Genotype 1</td>
<td>Wild type</td>
</tr>
<tr>
<td>12</td>
<td>HIV</td>
<td>1, 2, 3/a, b, b, c</td>
<td>CAGG</td>
<td>Genotype 3</td>
<td>ND</td>
</tr>
<tr>
<td>13</td>
<td>HIV</td>
<td>2, 2, 3/a, b, b, c</td>
<td>CAGG</td>
<td>Genotype 2</td>
<td>Wild type</td>
</tr>
<tr>
<td>14</td>
<td>HIV</td>
<td>1, 2, 3/a, b, b, c</td>
<td>CAGG</td>
<td>Genotype 2</td>
<td>Wild type</td>
</tr>
<tr>
<td>15</td>
<td>HIV</td>
<td>1, 2, 3/a, b, b, c</td>
<td>CAGG</td>
<td>Genotype 3</td>
<td>Wild type</td>
</tr>
</tbody>
</table>

ND, not done.
had a ‘C’ at position 248 and a ‘T’ at position 288. In the samples from the 15 patients, nucleotide variation was observed at only position 85. On the basis of a single-base polymorphism at position 85, we were able to divide the isolates into three groups, i.e., genotype 1 had a C (6.7%), genotype 2 had an A (40%) and genotype 3 had a T (53.3%). Some strains with identical mtLUrRNA genotype (genotype 1) displayed different UCS repeat number (3 or 4).

DHFR gene amplified by PCR from 10 patients was directly sequenced with the FR208 and FR1038 primers. In the present study, an 850 bp fragment of the *P. jirovecii* was detected (Fig. 3) and sequenced and four substitution sites were studied. Specimens of only two (20%) patients had nucleotide substitution in position 312 and these patients were infected by the mutant type of *P. jirovecii* with a nucleotide change in position 312. All other patients had infections caused by the wild type. No patients had nucleotide substitutions in position 200, 539 or 540.

Twelve different multiple-locus genotypes were observed, i.e., (1, 2, 3 / a, b, b; CAGG; Genotype 3; Wild-type) and (1, 2, 3 / a, b, b; CAGG; Genotype 2; Wild-type) were the most common and were detected in specimens from patients 5, 15 and 9, 14, respectively. The genotype (1, 2, 3 / a, b, b; CAGG; Genotype 2) was detected in patients 3 and 12, but the DHFR genotype of these strains was not determined (Table 1).

**Discussion**

Given that the UCS is present as a single copy per genome, quantifying the number of repeats provides a potential new method for rapidly typing *P. jirovecii*. The originality of our study was the identification of new region of tandem repeats in the UCS intron and sequence variability in the repeat units. This new region contains the following three different repeat units; GCCCA (type a), GCCCT (type b) and GCCTT (type c).

We have also observed three repeat units which consisted on 10 nucleotides. The types 2 and 3 each differ from type 1 by a single nucleotide substitution, which suggests that they arose independently from type 1, perhaps during the duplication process. The repeat type (1, 2, 3) is the major genotype in our population (86.7%) and four repeats were identified in 13.3% of the samples. We have isolated a new sequence corresponding to genotype (2, 2, 3, 3) which was not described previously by other authors. Sequence analysis suggests a hierarchy of evolution of *P. jirovecii* strains with four repeats identified to date probably derived from the three repeats with types 1, 2, and 3, rather than those with types 1, 1, and 2 or 1, 2, and 2. All these strains derived from type (1, 2, 3) by a simple insertion of unit 1 to obtain type (1, 1, 2, 3), unit 2 for type (1, 2, 2, 3) and unit 3 for type (1, 2, 3, 3) [13]. Otherwise, the development of a new type (2, 2, 3, 3) from the three repeats with type (1, 2, 3) would have occurred after insertion of one unit of type 3 and substitution of G by T in first unit, which appears to be unlikely. The new identified genotype (2, 2, 3, 3) was probably derived from the (1, 2, 3, 3) type by a simple substitution of G by T in the first repeat unit.

The presence of variable numbers of tandem repeats has been described in other studies [3,13,14]. The tandem repeats (microsatellites or minisatellites) in non-coding sequence tend to vary much more than in coding regions. This characteristic increases the frequency of detection of new genotype and new repeat units and we have used this region in typing of *P. jirovecii* to obtain more genetic variation and differentiate among all strains. What is new about UCS typing is its use in both of length polymorphism (number of repeats) and sequence (repeat type) for the

---

**Fig. 3** PCR amplification of *Pneumocystis jirovecii* DHFR gene in BAL and sputum samples. M: molecular mass marker (100 bp DNA ladder), lane 1: negative control (water), lane 2: positive control (BAL), lane 3 to 10: positive samples.
characterisation of *P. jirovecii* species. The importance of UCS analysis is to facilitate the genotyping of *P. jirovecii* strains and to better understand the impact of this region on expression of MSG gene. In addition, further studies using additional samples are needed to determine whether the sequence variability of UCS region can provide insights into the evolution of *P. jirovecii* strains. The biological role of these tandem repeats in the UCS region is unknown [13,14]. The identification of two distinct tandem repeats region in UCS of *P. jirovecii* could be importance in the diversity and evolution, as well as modulating the expression of MSG of this fungi. A large number of different MSG genes have been observed at the *Pneumocystis* expression site, suggesting that MSG genes can be moved to this locus by DNA recombination [15–17]. However, the multi-copy gene family of MSG provides a large potential for antigenic variation and facilitates evasion of immune responses in hosts [18,25]. In addition, the UCS functions not only to regulate expression of the MSG, but also probably encodes a signal peptide that may be necessary for entry of the MSG into the endoplasmic reticulum for further processing and export to surface [14,19,20].

The genetic diversity of *P. jirovecii* strains has been searched in mtLSU rRNA and DHFR genes. The mitochondrial DNA sequence data have been highly useful in detecting intraspecific differences between populations of diverse organisms [26]. The 346 bp fragment of *Pneumocystis* mtLSU rRNA gene is known to be variable by the existence of three point mutations [5,27–29]. In our study, the analysis of the mtLSU rRNA portion of the gene indicated that genetic diversity in *P. jirovecii* exists, given the nucleotide polymorphism at positions 85. We found three different genotypes among our *P. jirovecii* strains, with genotype 3 (53.3%) appearing to be overrepresented. This result was similar to that obtained by Latouche et al. [5] after analysis of the French and Italian *P. jirovecii* strains. A different allelic distribution pattern was detected in other studies. In Spain and Portugal, genotype 1 was the most frequent [30,31] and similar results were reported in the USA, UK and Austria [2,32,33]. Some differences were seen in these reports in that in the Spanish population, genotype 3 was the second most frequent and genotype 2 was the third most common but in Austria genotype 2 was absent and genotype 3 was the less frequent [2,30,31]. The absence or low frequency of genotype 4 was reported in all studies. These data indicate the existence of geographic variation in allelic frequency detected at the mtLSU rRNA locus. Epidemiological factors inherent to different geographical locations may influence the circulation and transmission of genetically distinct *P. jirovecii* organisms [30,34].

In addition, in the present study we analyzed four substitution sites with two non-synonymous and two synonymous changes in the coding region of DHFR. In our investigation, one synonymous substitution was observed at nucleotide positions 312 (T to C) in two patients (20%) and it is the most abundant mutation in this gene [23,24,35]. This mutation does not induce the modification of the corresponding amino acid and does not cause a structural change of the DHFR enzyme. Costa and co-workers found the wild-type sequence of the *P. jirovecii* DHFR gene in 72.7% of the patients while 27.3% had point substitutions. In these patients, 77% had a mutation at position 312 [4].

In conclusion, the direct sequencing approach showed the genetic variability in the UCS region although it was not frequently used in molecular typing of *P. jirovecii*. Also, the multilocus analysis facilitates the distinguishing of *P. jirovecii* strains differences and proved the important biodiversity of this fungus.

**Declaration of interest:** The authors report no conflicts of interest. The authors alone are responsible for the content and the writing of the paper.

This study was carried out with the financial support of the Ministry of Higher Education and Scientific Research of Tunisia.

**References**

11. Totet A, Pautard JC, Racourt C, Rous P, Nevez G. Genotypes at the internal transcribed spacers of the nuclear rRNA operon of *Pneumo-


20 Schaffzin JK, Stringer JR. Expression of the Pneumocystis carinii major surface glycoprotein epitope is correlated with linkage of the cognate gene to the upstream conserved sequence locus. Microbiology 2004; 150: 677–686.


This paper was first published online on Early Online on 4 December 2012.


