Expression and complexity of the \textit{PRT1} multigene family of \textit{Pneumocystis carinii}


\textbf{Pneumocystis carinii} has a multigene family, \textit{PRT1}, that encodes proteins with homology to KEX2-like proteases. \textit{PRT1} genes cluster with \textit{MSG} genes near the telomeres and, like \textit{MSG}, \textit{PRT1} proteins seem to be surface-expressed. The clustering of \textit{PRT1} and \textit{MSG} genes suggested that expression of the two multigene families might be coordinated. Studying gene expression in \textit{P. carinii} has been hampered by the lack of a culture system, and by lack of clonality in \textit{P. carinii} populations in naturally infected rats, the host of this fungus. Heterogeneity can be reduced, however, by low-dose intratracheal inoculation, which can produce \textit{P. carinii} populations dominated by organisms derived from a single progenitor. To study \textit{PRT1} expression, nude rats were inoculated with approximately 10 \textit{P. carinii} each. The clonality of the \textit{P. carinii} populations from inoculated rats was assessed by analysis of the UCS locus, a site in the genome that is known to be very heterogeneous in naturally infected rats, but nearly homogeneous in rats infected by low-dose intratracheal inoculation. Each of the populations had the same \textit{MSG} gene at the UCS locus in at least 80\% of the organisms. To investigate \textit{PRT1} gene expression, RNA was amplified using primers that amplify numerous \textit{PRT1} genes. Seventy-four cloned cDNAs were sequenced, including at least 12 clones from each population of \textit{P. carinii}. Many differently expressed \textit{PRT1} sequences were identified in each population, and a total of 45 different sequences were detected. However, the same \textit{PRT1} sequence was present in 15 of 74 plasmids and was found in 3 of the 5 \textit{P. carinii} populations, suggesting that some \textit{PRT1} genes may be either more commonly expressed or expressed at a higher level. These data show that many members of the \textit{PRT1} gene family can be expressed in populations of \textit{P. carinii} derived from few progenitors and suggest that the regulation of this family is different from that governing expression of the \textit{MSG} gene family.

\textbf{INTRODUCTION}

\textit{Pneumocystis carinii} is a fungus that causes severe pneumonia in rats that lack an intact immune system (Stringer, 2002). While best known for this pathogenic activity, the following features of \textit{P. carinii} suggest that it may be an obligate parasite of the immunocompetent rat. \textit{P. carinii} organisms are difficult to find in the environment (Wakefield, 1994), do not grow well in culture (Cushion & Walzer, 1984a; Cushion, 1989; Cushion & Ebbets, 1990) and appear to be able to colonize immunocompetent rats, but are not able to proliferate in other host species, including mice, even when the host lacks immune function (Gigliotti \textit{et al}., 1993; Sidman & Roths, 1994; Wakefield \textit{et al}., 1998; Mazars & Dei-Cas, 1998; Demanche \textit{et al}., 2001).

Abbreviation: MSG, major surface glycoprotein.
The *P. carinii* genome contains a multigene family, PRT1, which encodes different isoforms of a protein that resembles subtilisin, and are also known as ‘dibasic processing proteases’ because these enzymes cleave polypeptides at sites that have two consecutive basic amino acid residues, such as lysine-arginine (Lugli et al., 1997, 1999; Russian et al., 1999). Genes encoding dibasic processing proteases have been observed in many other species. Not surprisingly, *P. carinii* PRT1 proteins are most similar to dibasic processing proteases in other fungi, e.g. *Saccharomyces cerevisiae* (Goffeau et al., 1996), *Schizosaccharomyces pombe* (Wood et al., 2002) and *Candida albicans* (Newport et al., 2003). Unlike *P. carinii*, each of these species has just a few genes encoding dibasic processing proteases. These genes are often named KEX2, after a gene in *S. cerevisiae*. PRT1 genes have also been called KEX (Russian et al., 1999). A dibasic processing protease gene has been identified in two other forms of the *P. carinii* organism displays only one isoform of MSG. Based on studies using an antibody that recognizes multiple PRT1 proteins, the surface of *P. carinii* appears to contain one or more PRT1 proteins, implying that degradation of exogenous proteins is a likely function of PRT1 proteins (Russian et al., 1999; Wada & Nakamura, 1999; Lugli et al., 1999). Yet, surface PRT1 proteins might seem to be a disadvantage as the parasites attempt to colonize the lung of an immunologically intact rat. One way in which other microbes circumvent this problem is by switching surface antigens frequently (Barbour et al., 1991; Borst, 1991; Restrepo & Barbour, 1994; Zhang et al., 1997; Rudenko et al., 1998). Indeed, differential expression of a different surface antigen gene family is known to occur in *P. carinii*. Expression of the multigene family MSG, encoding various forms of the major surface glycoprotein, involves a single-copy expression site, the UCS locus, which can be occupied by only one MSG gene at a time (Stringer & Keely, 2001). MSG genes at the UCS locus give rise to mRNAs, but MSG genes not at this site appear to be transcriptionally silent (Wada et al., 1995; Edman et al., 1996; Sunkin & Stringer, 1996, 1997; Keely et al., 2003). These observations imply that the surface of a given *P. carinii* organism displays only one isoform of MSG.

It is possible that PRT1 genes may also be involved in antigenic variation in *P. carinii*. Interestingly, PRT1 genes are linked to MSG genes. *P. carinii* chromosomes end with gene arrays comprised of unidirectional copies of PRT1 and MSG genes (Sunkin & Stringer, 1996; Underwood et al., 1996; Keely et al., 2001). Such an arrangement could allow both a PRT1 gene and an MSG gene to be expressed from a single promoter via transcriptional read-through, generating a polycistronic message. In addition, processing of *P. carinii* pre-MSG molecules may require a dibasic processing protease peptide (Sunkin et al., 1998).

The information described above suggested that transcription of the PRT1 multigene family might be regulated and coupled to transcription of MSG genes, such that *P. carinii* expressing a specific MSG isoform would express a specific PRT1 gene. Testing this hypothesis was not possible until recently because of the difficulties of working with *P. carinii*, which does not proliferate very well in culture. Cultures can be established by seeding with millions of organisms, but proliferation is too limited to allow the microbe to be cloned in culture (Cushion & Walzer, 1984a, b; Cushion et al., 1985a, b, 1988; Cushion & Ebbets, 1990; Merali et al., 1999). Large numbers of *P. carinii* can be obtained from immunocompromised rats, but populations of the fungus in naturally infected rats have been very heterogeneous with respect to MSG gene expression (Keely et al., 2003). However, it has been shown recently that inoculation of rats with a low dose of *P. carinii* produced populations in which 80–100% of the organisms had the same MSG gene at the UCS locus (Keely et al., 2003). Herein, we describe PRT1 gene expression in such populations of *P. carinii*. Contrary to the prediction of coordinated expression of PRT1 and MSG genes, mRNAs encoding many different PRT1 proteins were present.

**METHODS**

**Provocation and assessment of *P. carinii* infection.** Fisher nu/nu rats (nude rats) 12–17 weeks old, from a *Pneumocystis*-free colony housed under HEPA (high efficiency particulate air)-filtered air conditions (Lille Pasteur Institute, France) were used in this study. The animals were given dexamethasone in the drinking water (1 mg l⁻¹) throughout the study. After 5 weeks of dexamethasone treatment, rats were anaesthetized and infected by non-surgical intratracheal inoculation with 1–10 *Pneumocystis* organisms per rat. *P. carinii* to be introduced into nude rats were derived from two dexamethasone-treated Wistar rats from Harlan (Aliouat et al., 1995). The same preparation of pooled organisms was used to inoculate each nude rat. After inoculation, rats were housed in separate HEPA-filtered air isolators. Nude rats that were not inoculated (sentinel rats) were used to monitor for external infection. Sentinel rats were housed in HEPA-filtered air isolators. Eight weeks post-inoculation, the animals were killed and organism burdens were enumerated by light microscopy. The mean number of organisms per inoculated animal was 8 × 10⁶ ± 0.7 × 10⁶. No *P. carinii* DNA was found in sentinel rats by using PCR amplification of the mitochondrial LSU rRNA locus (Wakefield et al., 1990).

**Preparation of *P. carinii* organisms.** *Pneumocystis* organisms were extracted from five separate lung homogenates as described previously (Aviles et al., 2000). Extracted *P. carinii* organisms were either used directly or were frozen at −80°C.

**Processing *P. carinii* for amplification of nucleic acids.** To prepare DNA, approximately 1 × 10⁸ *P. carinii* organisms were treated with proteinase K and genomic DNA was isolated by 2-propanol extraction (Sambrook et al., 1989). The DNA was dissolved in 0.01 M Tris/0.001 M EDTA, pH 8.

To prepare RNA, 1 ml prewarmed Trizol (Gibco) was added to 50 μl *P. carinii* (approx. 1 × 10⁸ organisms). After 5 min at 37°C, the preparation was placed at −80°C overnight. Chloroform (200 μl) was added to the thawed preparation, the tube was agitated with a vortex mixer and then subjected to centrifugation at 13 000 g for 30 min at 4°C to separate the organic and aqueous phases. The aqueous upper layer containing the RNA was removed and placed in a new tube, to
which was added an equivalent amount of 2-propanol (500 μl). After vigorous mixing, the RNA was allowed to precipitate at 4 °C overnight. RNA was collected by centrifugation (13 000 g for 30 min at 4 °C). The RNA pellet was dried in a laminar flow cabinet, then dissolved in 30 μl RNase-free water. The RNA was stored as 10 μl aliquots at −80 °C. Prior to use, RNA was treated with DNase (DNA-free; Ambion). To terminate the reaction, 1 μl DNase inactivation reagent (Ambion) was added. After 2 min at room temperature, the inactivation reagent was removed by centrifugation for 1 min at 13 000 g. The supernatant containing the RNA was used for reverse transcription.

Reverse transcriptase amplification of P. carinii RNA. For first strand synthesis, 1 μl RNA was mixed with 1 μl primer (1 μM final concentration) and 8 μl sterile filtered water. This mixture was incubated at 70 °C for 10 min to denature the RNA. Then 4 μl 5× buffer, 250 mM Tris (pH 8-3), 37-5 mM KCl, 15 mM MgCl₂, (Gibco BRL) 1 μl 10 mM dNTPs and 2 μl 0-1 M DTT were added. After 2 min at 37 °C, 200 U reverse transcriptase (Gibco BRL) was added and the reaction was incubated for 1 h at 37 °C.

Amplification of nucleic acids. To amplify the UCS locus, an upstream primer that binds in the UCS locus (primer –145) was paired with a downstream primer (C2) that binds to a site present in at least 90% of MSG genes (Keely et al., 2003). The amplicons were cloned into a plasmid vector to produce a library of the UCS/MSG junctions. One microlitre of P. carinii genomic DNA was subjected to PCR amplification using primer –145 (5′-TAGACGATGAAAGGGAGAAT-3′), an upstream primer that binds in the UCS, paired with primer C2 (5′-ATACATTTTCTCTCATGTTTT-3′), a downstream primer, under the following conditions: 95 °C for 30 s, 57 °C for 1 min, 72 °C for 1 min 30 s, 35 cycles of incubation at 95 °C for 30 s, 57 °C for 1 min, 72 °C for 1 min 30 s. The reaction mixture volumes were 25 μl and contained dNTPs each at 100 μM, 1-25 U Taq polymerase (Epicerin), 1-5 mM MgCl₂ and 20 ng each primer. PCR products were cloned into plasmid pCRII-TOPO (Invitrogen) which was introduced into the strain of Escherichia coli provided with the vector.

To amplify PRT1 cDNA, 5 μl of a first strand cDNA synthesis reaction was added to a reaction that contained 50 mM KCl, 10 mM Tris (pH 8-0), 0-1% Triton X-100, 3 mM MgCl₂, 0-4 mM of each of the four dNTPs, 1 μM oligonucleotide primers and 0-025 units Taq polymerase ml⁻¹. Primers used were PcProt1 (5′-TTATTCTTGTAGC-TGGGGAC-3′) and PcProt3 (5′-CACGGCATGTAAGATTAGGA-3′). Negative controls where H₂O was used as a template were included and all reagents were handled in a laminar flow cabinet, using sterile tube and pipette tips and aliquoted reagents. A Perkin Elmer Cetus PCR machine was used with cycling conditions as follows: 94 °C for 90 s, 57 °C for 90 s and 72 °C for 120 s (40 cycles). The size of the PCR amplification product was estimated by electrophoresis using a Bio-Rad horizontal electrophoresis system on 1.5 % agarose gels, and 1 x TBE buffer and 0-1 μg ethidium bromide ml⁻¹ were included in the gels. The DNA was visualized with UV light and the gel images were obtained using the Amersham Pharmacia Biotech Imagemaster VDS system.

Prior to insertion into a plasmid vector, PCR products were purified using the SpinPrep PCR Clean-up Kit (Novagen). PCR products were ligated to the plasmid vector pGEM T-Easy (Promega) according to the manufacturer’s protocol. The resulting plasmids were transformed into competent Escherichia coli DH5α cells (Novagen). For a screen by PCR amplification, an initial cycle of 94 °C for 5 min was included to break open the bacteria and release the plasmid DNA. For positive clones either the amplified DNA was sequenced directly or plasmid DNA was extracted from the respective colony using the Wizard Miniprep kit (Promega) and was sequenced.

DNA sequence analysis. UCS/MSG sequences were aligned using DNASN software (Lynnson BioSoft) using the default settings. The alignments were optimized by introducing a limited number of gaps, which were not counted in relatedness calculations. Relatedness of pairs of aligned sequences was calculated using the ‘observed divergence method’, which counts the number of directly unmatched residues and divides this number by the total number of residues compared. The calculated values were used to construct a distance matrix.

PRT1 sequences were analysed using Chromas 1.44 software (C. McCarthy, Griffith University, Australia). Sequence alignments were performed using the University of Wisconsin GCG sequence analysis Version 10.1, (GGW, Wisconsin, USA) and Pregap4 and Gap4 (Staden Package).

Studies on PCR error indicated that Taq polymerase can be expected to commit 1 x 10⁻³ errors per base pair (Cline et al., 1996). Nevertheless, given the greater complexity of amplifying DNA from P. carinii mRNA, it seems prudent to assume that the error rate in the PRT1 analysis could have been as high as 1 x 10⁻² errors per base pair. At this rate, 80 % of PRT1 amplicons will contain at least one error. According to the Poisson distribution, 32 % of mutant amplicons will have one error, 26 % will have two, 14 % will have three, 6 % will have four errors and 2 % will have five errors. These calculations suggest that PRT1 sequences that differ at four or fewer positions may have originated from templates that were the same and that sequences with more than four differences are very unlikely to have come from the same template.

**RESULTS**

**Production of populations of P. carinii that have primarily one MSG gene at the UCS locus**

Previous studies have indicated that more than one PRT1 gene can be represented by mRNA in a given population of P. carinii. Nevertheless, these data do not exclude the possibility of regulation within individual organisms of the population, with different organisms expressing different PRT1 genes, as appears to be the case for MSG genes. The aim of this study was to determine the diversity and complexity of PRT1 gene expression within clonally derived P. carinii populations.

To study PRT1 expression, five P. carinii populations were produced by low-dose inoculation of nude rats. The organisms used to prepare inocula were from a population that had at least 23 different MSG genes at the UCS locus. This information was obtained by sequencing UCS-MSG junctions amplified from genomic DNA. The inserts in 34 plasmids were sequenced. For record-keeping purposes, the 23 different sequences observed were designated MSG-A to MSG-W.

To assess the clonality of P. carinii in the inoculated rats, the MSG genes present at the UCS locus were examined by PCR amplification followed by DNA sequencing. At least 10 plasmids from each rat were sequenced. Seven different sequences were obtained. One of these matched MSG-A, one of the 23 sequences observed in the population used to prepare inocula. For recording keeping purposes, the other six sequences were designated MSG-a to e and MSG-Y. Of
the 29 different MSG sequences observed, 27 had a complete ORF encoding an MSG peptide with 158 aa.

The five inoculated rats produced *P. carinii* populations that exhibited very little diversity of MSG genes at the UCS locus (Table 1). All ten of the plasmids from the population EA21 library had the same sequence (MSG-A, GenBank accession no. AY387716). Similarly, all but one of the 19 plasmids from population EA15 had the MSG-A sequence. The other three *P. carinii* populations produced plasmid sets in which the MSG sequences were at least 80 % homogeneous (Table 1). This lack of heterogeneity in the set of MSG genes retrieved from the UCS locus contrasted with the situation in the *P. carinii* population that had been used to inoculate each rat (Table 1). In this population, 23 different MSG genes were observed at the UCS locus. Thus the low-dose inoculation procedure yielded the populations of *P. carinii* needed to determine if PRT1 gene expression is regulated by restricting expression to one PRT1 gene per organism.

Multiple PRT1 mRNAs were present in populations in which 90 % or more of the organisms had the same MSG gene at the UCS locus

To estimate the number of different PRT1 mRNA sequences present in the *P. carinii* populations, the mRNA in each population was converted to cDNA by reverse transcriptase amplification using the universal PRT1 primers PcProt1 and PcProt3. These primers were used for several reasons. First, all of the PRT1 genes analysed to date contain the sequences for these primers. Second, the primers amplify a region known to vary in sequence (but not in length), even though this region encodes a part of the conserved enzyme catalytic domain. A BLASTN search using the sequence from one PRT1 gene as the query sequence showed that the NCBI database contained sequences from eight different *P. carinii* PRT1 genes. Nucleotide identities among these sequences ranged between 85 and 95 %. Third, the amplified region includes two introns of approximately 50 bp each. The presence of introns allowed detection of any amplicons that might have come from contaminating genomic DNA within the RNA preparations. Only those sequences that lacked both introns were scored as mRNAs.

Amplified cDNA was inserted into a plasmid vector and a library was made for each *P. carinii* population. From each library, at least 12 plasmids were sequenced. *P. carinii* populations from three rats, EA15, EA17 and EA19, were analysed first. Multiple PRT1 cDNA sequences were detected in all three populations (Table 2). At least six different PRT1 sequences were observed in every case. The degree of divergence among the different sequences from a given population was as high as 40 % (data not shown). The mean divergence among the PRT1 sequences observed in these three rats was 15 %.

While none of the three populations produced a single PRT1 cDNA sequence, some sequences were identified in more than one plasmid (Table 2). Of note, one sequence, referred to henceforth as comPRT (GenBank accession no. AY387750) was prominent in all three rats. Ten exact replicas of the comPRT sequence were detected, three in EA15, four in EA17 and three in EA19. Another five sequences were at least 99 % identical to the comPRT sequence. Sequences that match at 99 % or more positions, i.e. differ at four or fewer nucleotide positions, could be

<table>
<thead>
<tr>
<th>Table 1</th>
<th>MSG gene sequences at the UCS locus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source of P. carinii</strong></td>
<td><strong>No. of plasmids sequenced</strong></td>
</tr>
<tr>
<td>Rat EA15</td>
<td>19</td>
</tr>
<tr>
<td>Rat EA17</td>
<td>10</td>
</tr>
<tr>
<td>Rat EA19</td>
<td>10</td>
</tr>
<tr>
<td>Rat EA21</td>
<td>10</td>
</tr>
<tr>
<td>Rat EA24</td>
<td>10</td>
</tr>
<tr>
<td>Inoculum†</td>
<td>34</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>93</td>
</tr>
</tbody>
</table>

*Plasmids were placed in one of two categories: 1, those that contained a sequence that was present in only one plasmid; and 2, those that contained a sequence that was present in more than one plasmid. The values in this column were calculated by dividing the number of plasmids in category 1 by the total number of plasmids.

†Population of *P. carinii* that was used to prepare the dilute suspensions of *P. carinii* that were introduced into nude rats via intratracheal inoculation.

‡MSG-A was in 18 plasmids, MSG-a was in 1 plasmid.

§MSG-Y was in 8 plasmids, MSG-b was in 1 plasmid, MSG-c was in 1 plasmid.

∥MSG-A was in 8 plasmids, MSG-a was in 1 plasmid, MSG-d was in 1 plasmid.

*MSG-Y was in 8 plasmids, MSG-b was in 1 plasmid, MSG-e was in 1 plasmid.

#MSG-A and MSG-I were in four plasmids each; MSG-J was in three plasmids; MSG-C, MSG-N and MSG-T were in two plasmids each.
The presence of a predominant PRT1 cDNA sequence was not due to a PRT1 gene that is highly expressed in all populations

The finding that the comPRT1 cDNA sequence was prominent in these three rats suggested that some PRT1 mRNA sequences may be present at a higher copy number than others. It would be expected that at least one PRT1 gene would be constitutively expressed to produce the dibasic processing protease thought to be needed to cleave preproteins as they traverse the secretory pathway. To test the possibility of constitutive expression of the comPRT1 mRNA, the populations in two additional inoculated rats, EA21 and EA24, were analysed. An exact copy of the comPRT1 sequence was not detected in either population, but one of the 12 EA24 sequences was 99% identical to comPRT1. To determine if the absence of the comPRT1 sequence from EA21 was due to inadequate sampling of the cDNA population, a second RNA preparation was extracted from the P. carinii of population EA21, cDNA produced, amplified and cloned. Twelve plasmids were sequenced. None of these plasmids carried an exact replica of the comPRT1 sequence, yet one sequence was 98% identical to the comPRT1 sequence. Hence, these data showed that the comPRT1 sequence was not highly abundant in every population of P. carinii. These results were divergent from what would be expected if the comPRT1 sequence were to encode the dibasic processing protease that would be required in the secretory pathway.

The presence of a predominant PRT1 cDNA sequence was not correlated with a specific MSG gene

The observation that the comPRT1 sequence was not prominent in all populations raised the possibility that its variable abundance might reflect regulation of PRT1 gene expression. The linkage of PRT1 genes to MSG genes suggests one plausible scenario for regulation, whereby the PRT1 gene that is downstream of the UCS-linked MSG gene is expressed. However, features of the UCS locus in the five populations are difficult to reconcile with the co-expression model. Whereas P. carinii populations EA15 and EA19 had the same predominant MSG sequence at the UCS locus, population EA17 did not feature this MSG sequence at the UCS locus. Hence, co-expression cannot explain the comPRT1 sequence detected in these three populations unless the comPRT1 sequence is linked to both types of MSG sequence.

Table 2. PRT1 cDNA sequences

<table>
<thead>
<tr>
<th>Source of P. carinii</th>
<th>No. of plasmids sequenced</th>
<th>No. of sequences observed in sample</th>
<th>No. of sequences not found in any other sample*</th>
<th>Plasmids with a sequence that was observed once in the group P. carinii population (%)†</th>
<th>Plasmids with a sequence that was observed more than once in the group of plasmids made from this P. carinii population (%)§</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rat EA15</td>
<td>13</td>
<td>9</td>
<td>4</td>
<td>61</td>
<td>39</td>
</tr>
<tr>
<td>Rat EA17</td>
<td>12</td>
<td>6</td>
<td>3</td>
<td>33</td>
<td>67§</td>
</tr>
<tr>
<td>Rat EA19</td>
<td>14</td>
<td>10</td>
<td>6</td>
<td>64</td>
<td>36#</td>
</tr>
<tr>
<td>Rat EA21</td>
<td>24</td>
<td>18</td>
<td>12</td>
<td>62</td>
<td>38**</td>
</tr>
<tr>
<td>Rat EA24</td>
<td>11</td>
<td>9</td>
<td>8</td>
<td>64</td>
<td>36†</td>
</tr>
<tr>
<td>Inoculum‡</td>
<td>14</td>
<td>12</td>
<td>6</td>
<td>71</td>
<td>29‡</td>
</tr>
<tr>
<td>1984§</td>
<td>12</td>
<td>9</td>
<td>6</td>
<td>58</td>
<td>42§§</td>
</tr>
<tr>
<td>Totals</td>
<td>100</td>
<td>72</td>
<td>45</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Sequences from all 100 plasmids were compared and those that did not match any other cDNA in the set were scored as unique. Sequences that were at least 99% identical were not considered to be from different genes.
†Plasmids were placed in one of two categories: A, those that contained a sequence that was present in only one plasmid; and B, those that contained a sequence that was present in more than one plasmid. The values in this column were calculated by dividing the number of plasmids in category A by the total number of plasmids.
‡Population of P. carinii that was used to prepare the dilute suspensions of P. carinii that were introduced into nude rats via intratracheal inoculation.
§An archival population of P. carinii from a naturally infected rat.
||One sequence was in five plasmids.
‖One sequence was six plasmids; another one was in two plasmids.
#One sequence was in five plasmids.
**One sequence was in three plasmids, three sequences were in two plasmids each.
††Two sequences were in two plasmids each.
‡‡Two sequences were in two plasmids each.
§§One sequence was in three plasmids; another one was in two.
Diversity of the PRT1 multigene family

While it is known that the *P. carinii* genome contains multiple sequences capable of producing PRT1 dibasic processing proteases, and that more than one of these PRT1 genes can be transcribed, there has been no systematic study of the PRT1 multigene family in this regard. The analysis of the PRT1 cDNAs described above presented an opportunity to better define the expression capacity of the family. A total of 74 plasmids were sequenced during the analysis of the five populations of *P. carinii* from the low-dose inoculated rats. Two additional PRT1 cDNA libraries were also made, one from the population used to prepare the inocula and the other from an archival population called 1984. Fourteen plasmids from the inoculation population and 12 from population 1984 were sequenced. Thus, in total 100 plasmids were sequenced.

The number of different PRT1 sequences (defined as more than 1% divergent from all other sequences) observed in each population is shown in Table 2. The numbers in the third column of Table 2 add up to 72, but the number of different sequences observed was less than this value because some sequences were detected in at least two populations. The number of sequences that were found only once among all populations was 45. Thus, at least 45 different PRT1 genes can be expressed. Mean nucleotide divergence among the 45 sequences was approximately 15%.

Forty-one of the 45 sequences contained an ORF encoding the expected peptide. The four sequences that lacked an ORF contained either a single base pair change that created a stop codon or a frameshift. Conceptual translation of the 41 complete ORFs showed that there was more divergence among the peptides than among the mRNAs. The mean divergence among peptides was 20% while the mean divergence among nucleotides was 12%. This situation reflects the predominance of non-synonymous nucleotide changes, which at 38 outnumbered synonymous changes approximately fourfold (Fig. 1). As might be expected, given the function of this region which encodes part of the catalytic domain, nearly 70% of the nonsynonymous changes occurred in regions known to be variable in dibasic processing proteases (Fig. 1) (Siezen et al., 1994). These variable regions constitute approximately one half of the protein. Ten of the non-synonymous changes were in the region known to be hypervariable in dibasic processing proteases (HV in Fig. 1). Another 16 of the non-synonymous changes were located between regions that encode peptide motifs with conserved secondary structure. The region between β5 and β6 contained 10 and the region between β7 and αF contained 6. By contrast, the synonymous nucleotide changes were located more randomly, with approximately half in regions featuring conserved secondary structure. However, the trend towards clustering of non-synonymous changes in variable regions was not statistically significant (*P*=0.19 by Fisher’s Exact Test). Only one insertion/deletion was observed and occurred in the HV domain.

DISCUSSION

The *P. carinii* populations produced by low-dose intratracheal inoculation of nude rats exhibited the expected lack of complexity with respect to the UCS locus. In each of the five populations, the UCS locus was at least 80% homogeneous. These data suggest that the *P. carinii* in these rats were derived from only a few, perhaps just one, progenitor. Although the UCS locus was not 100% homogeneous, a single progenitor is still a possibility because, in theory, the UCS locus can be changed as the population expands. Despite the uniformity at the UCS locus, many different expressed PRT1 gene sequences were observed in each of these five populations. Therefore, if expression of the PRT1 multigene family is regulated, it is not by a simple system whereby one PRT1 gene is stably expressed per organism. Either more than one PRT1 gene is expressed in an organism, or organisms can switch from one PRT1 to another rapidly enough to generate the diversity observed in these studies.
Even though PRT1 gene expression did not appear to be strictly controlled, the presence of a predominantly expressed gene, comPRT, in three out of the five populations, suggests that some form of regulation for PRT1 gene expression might be occurring. Yet, modulations that involve multiple genes would be difficult to discern at this level of analysis. Alternatively, PRT1 gene expression could be regulated post-transcription. Furthermore, infection of nude immunosuppressed rats with P. carinii may not be representative of the situation that occurs when P. carinii infects an immunocompetent rat in the wild. It is possible that PRT1 gene expression is more tightly regulated when subject to the pressures of a fully functioning rat immune system.

Regulated or not, the PRT1 multigene family encodes a highly diverse family of proteins. A minimum of 45 different PRT1 genes were identified in this study. A high degree of protein diversity implies that the evolution of a gene family may have been influenced by selection in favour of nucleotide changes that cause amino acid changes (non-synonymous nucleotide substitutions). Other genes that exhibit a large number of non-synonymous polymorphisms include those encoding proteins whose function require diversity, such as the immune system genes and antigenically variable proteins of pathogenic microbes and viruses (Woelk et al., 2002; Jansa et al., 2003).

P. carinii is the only pathogenic fungus known to possess a large number of genes encoding different dibasic processing proteases. The reason for the diversity of the PRT1 multigene family is a matter of speculation. One possibility is that the PRT1 multigene family aids survival of P. carinii organisms in the face of the host immune system. P. carinii organisms attempting to colonize a normal rat lung can be presumed to be fully exposed to this system. Gene families that encode various forms of surface proteins can be used to overcome an attack by providing variant cells that do not have the target antigen. The simplest systems of antigenic variation allow for the expression of a single antigenic variant in the presence of the non-variant form of the same antigen (Dei-Cas & Guillot, 2001). In some viruses, the antigenic variants are regulated by the same promoter region (Woelk et al., 2002; Jansa et al., 2003). In P. carinii, the antigenic variants are expressed in a cell type-specific manner (Cushion et al., 1989).

PRT1 is a highly diverse family of proteins. A minimum of 45 different PRT1 genes can be located downstream of the MSG gene (S. M. Sunkin & J. R. Stringer, unpublished observation).

ACKNOWLEDGEMENTS

This work was supported by NIH grant AI-36701 (J. R. S.), the Royal Society and the Wellcome Trust (A. E. W.), the Medical Research Council (H. E. A.) and the Camden Primary Care Trust (R. F. M.). The French part of this work was developed in the framework of the EA3609 group, supported in part by both the French Ministry of Research & Education and the European Commission (the Fifth Framework Programme of the European Commission (grant QLK2-CT-2000-01369).


