Variation in the Major Surface Glycoprotein Genes in *Pneumocystis jirovecii*

Geetha Kutty,1 Frank Maldarelli,2 Guillaume Achaz,3 and Joseph A. Kovacs1

1Critical Care Medicine Department, National Institutes of Health Clinical Center, National Institutes of Health, Bethesda, and 2HIV Drug Resistance Program, National Cancer Institute, National Institutes of Health, Frederick, Maryland; 3UMR 7138 and Atelier de Bioinformatique and Université Pierre & Marie Curie, Paris, France

The genome of *Pneumocystis*, which causes life-threatening pneumonia in immunosuppressed patients, contains a multicopy gene family that encodes the major surface glycoprotein (Msg). *Pneumocystis* can vary the expressed Msg, presumably as a mechanism to avoid host immune responses. Analysis of 24 msg-gene sequences obtained from a single human isolate of *Pneumocystis* demonstrated that the sequences segregate into 2 branches. Results of a number of analyses suggest that recombination between msg genes is an important mechanism for generating msg diversity. Intra-branch recombination occurred more frequently than inter-branch recombination. Restriction-fragment length polymorphism analysis of human isolates of *Pneumocystis* demonstrated substantial variation in the repertoire of the msg-gene family, variation that was not observed in laboratory isolates of *Pneumocystis* in rats or mice; this may be the result of examining outbred versus captive populations. Increased diversity in the Msg repertoire, generated in part by recombination, increases the potential for antigenic variation in this abundant surface protein.

*Pneumocystis* causes life-threatening pneumonia in immunosuppressed hosts. The most abundant surface protein of *Pneumocystis* is the major surface glycoprotein (Msg), which is encoded by a multicopy gene family with ~50–100 copies (of ~3000 kb each) per genome that are clustered in tandem arrays near the telomeres of each chromosome [1]. These genes encode an incomplete protein that lacks an N-terminal peptide and are not expressed unless they are translocated downstream of a unique subtelomeric expression site that encodes the upstream conserved sequence (UCS) [2–7]. Only the variant present at the expression site is translated in a given organism. However, the ~50–100 variant msg genes provide great potential for antigenic variation [8–12]. Variation of the expressed Msg presumably facilitates evasion of immune responses in hosts. Because *Pneumocystis* species are haploid [13], a single organism can express only a single Msg variant. However, multiple variants can be expressed in a single infected lung from an immunosuppressed host [14, 15]. A similar msg-gene organization has been identified in *Pneumocystis* in humans, rats, mice, and ferrets [4, 6, 7, 12]. Recombination may play a role in generating multiple msg variants [3, 16, 17].

Antigenic variability in other species, such as African trypanosomes and *Borrelia*, is associated with evasion of host immune—primarily antibody—responses. The potential for antigenic variation in these organisms is increased not only by the presence, in each organism’s genome, of multiple unique copies of genes encoding their surface proteins, but also by variation in the repertoire of these multicopy genes (i.e., different isolates have unique sets of these genes), which likely further contributes to successful immune evasion [18–20]. To better characterize the family of genes that compose the msg repertoire of *Pneumocystis* (*P. jirovecii*) in humans, we undertook to sequence individual msg variants in a patient with *Pneumocystis* pneumonia and to determine their relationships to each other, specifically focusing on possible recombinant between msg variants. To see whether, like other organisms, *Pneumocystis* species have variable repertoires of the msg-gene family, we used
restriction-fragment length polymorphism (RFLP) analysis to examine the msg repertoire in Pneumocystis in humans, rats, and mice.

**MATERIALS AND METHODS**

**Preparation of Pneumocystis DNA.** At autopsy, samples of P. jiroveci–infected lungs from 6 patients with Pneumocystis pneumonia (5 of whom were infected with HIV) were collected and used for DNA extraction; for P. murina, infected lung samples from scid mice were used; for P. carinii, infected lung samples were obtained from immunosuppressed rats maintained in 2 facilities (Biocon and the Indiana University campus in Indianapolis) and were partially purified by Ficoll-Hypaque density-gradient centrifugations [21]. Genomic DNA was isolated either by use of a QIAamp DNA mini kit (Qiagen) or by proteinase K treatment [22]. Human- and animal-experimentation guidelines of the National Institutes of Health were followed in the conduct of these studies.

**Polymerase chain reaction (PCR) amplification.** Pneumocystis DNA was amplified by use of TaqPlus Long (Stratagene) and primers, from conserved regions (based on alignment of available msg sequences), designed to amplify the entire ~3.3 kb of the msg variable region (table 1 and figure 1, which are available only in the electronic version of the Journal). PCR conditions for both P. carinii (primers GK521 and GK527) and P. murina (primers GK257 and GK261) were as follows: 2 min at 94°C; 35 cycles of 30 s at 94°C, 30 s at 56°C, and 4 min at 72°C; and a final extension for 10 min at 72°C. PCR conditions for P. jiroveci (primers GK126 and GK452) were as follows: 2 min at 95°C; 10 cycles of 30 s at 94°C, 1 min at 68°C with 1°C decremental steps in each cycle, and 4 min at 72°C; and 35 cycles of 30 s at 94°C, 30 s at 58°C, and 4 min at 72°C.

**Sequencing of msg variants.** Genomic DNA from a single infected patient was analyzed by nested PCR using HotstarTaq (Qiagen) after limiting dilution was performed [23]. The first round was as described above but with a 15-min initial denaturation. The second-round conditions (primers GK508 and GK506) were as follows: 15 min at 95°C; 35 cycles of 30 s at 94°C, 30 s at 58°C, and 4 min at 72°C; and a final extension for 10 min at 72°C. For the limiting dilution, DNA was serially diluted (3–10-fold per dilution) in preliminary studies, and 10 replicate nested-PCR reactions were performed at each dilution. The dilution at which only ~3 PCR reactions yielded a product as determined by agarose-gel electrophoresis was used for subsequent subcloning [23]. Amplification products were subcloned into TOPO TA cloning PCR 2.1 (Invitrogen), and clones showing distinct sequences (>1% variability) after initial sequencing of the 3' and 5' ends were completely sequenced. To examine the likelihood of PCR-mediated recombination between 2 msg genes on a single DNA fragment, we used the identical procedure to amplify and sequence a genomic clone with 2 full-length msg genes in tandem repeats (GenBank accession number AF038556). We found no recombinants in 52 fully sequenced clones that were generated during 6 PCR reactions.

**RFLP and Southern blot analysis.** PCR products amplified as above were separated on an agarose gel, excised, purified by use of a Qiaquick gel-extraction kit (Qiagen), digested with the indicated restriction enzymes, analyzed on 1% agarose gels in 1× Tris-borate EDTA buffer, and visualized by use of SYBR green staining (Molecular Probes). Preliminary studies demonstrated that the RFLP pattern was stable during repeat PCR reactions and was not dependent on DNA concentration. DNA blotted onto Nytran membranes (Schleicher&Schuell) was probed with oligonucleotides labeled by use of either a DIG Oligonucleotide Tailing Kit (Roche) or DIG-labeled DNA probes (PCR DIG Probe Synthesis Kit; Roche); Southern blot analysis was performed as described elsewhere [24]. Before rehybridization, blots were stripped at 37°C in buffer containing NaOH at a concentration of 0.2 mol/L and 0.1% SDS.

**Statistical analysis.** msg sequences were aligned by Clustal W (Megaalign module of Lasergene; DNASTAR), with a gap penalty of 15 and a gap-length penalty of 6.66 [25]. Neighbor-joining trees of gap–stripped msg sequences were constructed by PAUP* (version 4.0; Sinauer Associates), with the outgroup being a P. carinii msg sequence from rat. Bootstrap values were calculated on the basis of 1000 resampling replicates. Average pairwise differences between groups were calculated, and population-structure tests were performed on gap-stripped sequences, by the Hudson method [see 24] (for online version of this analysis, see http://wwwabi.snv.jussieu.fr/~achaz/hudsontest.html). On the basis of permutation, this test provides the probability that 2 a priori–defined groups have more similarity within than between groups. Homogeneity was defined as an absence of structure. Simplot and bootscanning analyses of the alignments were performed to identify regions where recombination occurred. SimPlot compares the similarity between a short (~200 nt) window of an individual sequence and the corresponding window for the entire complement of sequences in the alignment. The degree of relatedness is calculated as the window is moved in steps (200 nt) across the alignment. Marked changes in similarity indicate the presence of recombination [26]. Bootscanning analyzes phylogenetic relationships.
The figure is available in its entirety in the online edition of the Journal of Infectious Diseases.
wide analysis (e.g., see sequences Cl-1, Cl-7, Cl-14, and Cl-25). Branch switching that, as a function of nucleotide position, has strong bootstrap support, strongly suggests the presence of recombination. (We use the term “recombination” to refer to both reciprocal and nonreciprocal genetic exchange, although the latter is properly called “conversion”).

Because msg alignments revealed numerous insertions and/or deletions (indels), we undertook specific analysis of indels. This analysis demonstrated that a number of msg indels segregate primarily according to phylogenetic grouping. As shown in figure 4, an insertion of 3 nt after nucleotide position 1339 was found almost exclusively in sequences in principal branch A of the phylogenetic tree, but it was also found in 1 of the group B sequences. Similarly, a deletion of 6 nt was uniformly present after nucleotide position 2908 in group A and in 3 of the 13 sequences in group B; and an insertion of 3 nt after nucleotide position 2925 was uniformly present in group B and was present in 1 sequence in group A. The finding of an indel uniformly present in 1 branch and appearing in the other branch strongly supports the presence of recombination between these msg genes.

To further investigate potential recombination events, we used 2 sliding-window approaches: (1) SimPlot, to measure region-specific similarity, and (2) bootscanning, to quantitate bootstrap values for phylogenetic trees. SimPlot analysis identified a divergence in sequence similarity, with a segregation of sequences after ~1000–1200 nt, that corresponded exactly to branches A and B in the phylogenetic analysis (figure 5A); a recombination event can explain this sharp divergence. Bootscanning analysis of tree structure was used to identify potential areas of recombination more precisely. Phylogenetic trees constructed by use of a sliding window (200 nt) of sequences from candidate recombinants and potential parental sequences are compared by bootstrap analyses; switching of bootstrap support from one parental sequence to another is consistent with recombination. We analyzed sets of sequences that had strong bootstrap support in full-length msg sequence analysis but that switched support in partial-sequence analyses. In the bootscanning analyses (figure 5B) of Cl-14 and Cl-21 and of Cl-14 and Cl-25, clear evidence of 2 recombination events was detected (at nucleotide positions ~700 and ~1520). Exhaustive bootscanning comparisons identified numerous additional areas where switching (presumably secondary to recombination) occurred, within and between groups A and B (data not shown).

**Analysis of msg-gene variation in different Pneumocystis isolates.**

Given the evidence for recombination between different msg genes, we determined whether the msg repertoires (the 50–100 msg genes per genome) are identical in *P. jirovecii* isolates from different patients. Primers based on highly conserved regions (in the known msg variants) at the beginning of the coding region and downstream of the stop codon (figure 1, which is available only in the electronic version of the *Journal*) were used to amplify the whole msg repertoire of all individual genomes from an isolate. The resultant PCR product of ~3.3 kb was subjected to RFLP analysis. If the msg repertoire is highly conserved, then the RFLP pattern among different isolates should be very similar or identical. Among 6 isolates from 6 patients, RFLP patterns were strikingly different when the PCR products were digested with *MboI*, *HindIII*, or *DraI* (figure 6A). Southern blot
analysis using an oligonucleotide designed on the basis of a conserved area near the 3' end of msg, which was selected on the basis of an alignment of known P. jirovecii msg sequences, is shown in figure 6B. The 6 isolates showed distinct patterns, especially when digested with HindIII or DraI. To further examine this difference in patterns, the blot was reprobed with an oligonucleotide specific for msg32, a previously characterized msg variant [22]; strong hybridization was seen only for isolate 1 (figure 6C), which shows that this specific variant is not present in all P. jirovecii, a finding that again corroborates the high diversity of repertoire in this species. To verify that PCR amplification was not introducing an artifact into the RFLP analysis, restriction digestion using genomic DNA from 4 samples was performed, followed by hybridization. Again the pattern variation among isolates was high (figure 6D). These observations demonstrate that the msg repertoires of P. jirovecii are highly variable.

Given (1) that the Pneumocystis species that infects humans is different from those that infect rats and mice and (2) that these Pneumocystis species also have similar multicopy msg-gene families and systems for expressing msg, we performed RFLP analysis to examine the diversity of the msg repertoires in Pneumocystis in rats and mice [6, 10]. When we used the DNA samples from

**Figure 5.** A, Simplot analysis of msg sequences. Sequences were subjected to Simplot analysis comparing all other msg sequences to sequence CI-5 of group A. Each line represents a single clone. All group A sequences (blue lines) are relatively closely related to CI-5 (similarity to CI-5 score, 70%–100%), whereas group B sequences (orange lines) diverge sharply between nucleotide positions ~1200 and ~2300 (similarity to CI-5 score, <60%), suggesting recombination at the divergence site. The similarity of the scores of the group B sequences increases near the end of the msg sequence, suggesting additional recombination events. B, Bootscanning detection of evidence of frequent recombination. msg sequences were subjected to bootscanning analysis using CI-14 as the query sequence and CI-7, CI-21, and CI-25 as the comparison sequences. For simplicity, only the comparisons between CI-14 and CI-21 (black) and between CI-14 and CI-25 (red) are shown. The x-axis depicts the nucleotide position within the msg sequence; the y-axis depicts the percentage of trees grouping with the query sequence. The number of permuted trees grouping CI-21 with CI-14 when a sliding window of 200 nt is used shows sharp switches in frequency, with demarcations suggesting 2 recombination events, at nucleotide positions ~700 and ~1520; in contrast, the number of CI-14 permuted trees grouping with CI-25 shows a reciprocal switch in homology at approximately the same positions. Additional potential switches in homology are present in both comparisons, indicating multiple recombination events.
the lungs of 6 P. murina–infected scid mice housed in a single cage, we observed an identical RFLP pattern when the PCR products were digested with HindIII or DraI (data not shown). When we used lung samples collected at our facility during 1999–2004 (figure 7A), we found that RFLP patterns were again identical in all 6 mice. Southern blot analysis using an oligonucleotide designed on the basis of a conserved region of P. murina msg showed an identical pattern of hybridization for all isolates in each study (results not shown). Because all the mice were from a single colony, we conducted a similar analysis of P. carinii in rats that were obtained from 2 different facilities over a period of years. When the PCR products were digested with HindIII or DraI, RFLP patterns were very similar in all rats (figure 7B). Southern blot analysis using an oligonucleotide from the conserved region of P. carinii msg showed a nearly identical pattern of hybridization in all 7 rats (results not shown). Southern blot analysis of genomic DNA further confirmed that the RFLP pattern was highly conserved among isolates (figure 7C). Thus, Pneumocystis in rats and mice that are bred in captivity did not demonstrate the same degree of variability in msg repertoires as did P. jirovecii.
DISCUSSION

We have demonstrated by RFLP analysis that, among different isolates, there is substantial diversity of msg genes in *P. jirovecii*: no 2 isolates had the same repertoire of 50–100 genes. Sequence analysis of 24 unique msg genes from a single isolate showed that recombination between msg variants plays a major role in generating msg diversity. Because all msg genes appear to be located in clusters near telomeres [3, 15, 32], recombination either upstream of the clusters or within the msg genes can further increase msg diversity [17]. Elsewhere, recombination has been hypothesized to play an important role in generating such haplotype diversity [3, 16, 17].

The identification of 2 distinct branches of msg genes in *P. jirovecii* obtained from a single infected patient was surprising. It is possible that this patient was infected with 2 unique strains of *Pneumocystis*, although ITS1 as well as UCS typing suggested
infection with a single strain. Alternatively, the 2 branches may represent msg genes in a single isolate that were inherited from 2 parental strains via sexual reproduction and that have not yet had a chance to genetically intermingle extensively. This hypothesis is consistent with the identification, in individual PCR reactions, of msg genes from both branches, a finding that suggests that they are on 1 DNA fragment. It is also possible that, for biological or other reasons, the 2 sets of genes are limited to recombination primarily within a branch rather than across a branch. We do not believe that the 2 branches represent unique families of genes similar to those that we and others have previously identified in P. carinii (e.g., msg and mbr) [17, 33, 132], because the upstream primer region for amplification of P. jirovecii msg genes includes a highly conserved sequence homologous to the conserved recombination-junction element in P. carinii; in P. carinii, a primary characteristic distinguishing between msg genes and variants is the presence of this conserved recombination-junction element in the former but not the latter.

Recombination between genetically distinct organisms, between genetically identical organisms, or between different msg genes within an individual organism could increase the diversity of the msg repertoire. The conservation of the RFLP pattern in Pneumocystis in rats and mice, compared with the diverse patterns seen in Pneumocystis in humans is striking but may simply be due to examination of a captive versus an outbred population; alternatively, it is possible that P. jirovecii has developed a mechanism for increasing msg diversity, a mechanism that is not present in the other Pneumocystis species. RFLP analysis of Pneumocystis obtained from wild animals rather than from colony-bred animals would definitively address this issue.

Although the function of Msg may be to facilitate adherence to host cells or proteins [34, 35], Msg antigenic variation likely confers an immunologic advantage to the organism in its interaction with the host. Similar antigenic variability in other species, such as African trypanosomes and Borrelia species, is associated with evasion of host immune, primarily antibody, responses; repertoire variation in these organisms has been documented and likely contributes to successful immune evasion [18–20]. Given (1) that cell-mediated immune responses—especially CD4+ T lymphocyte responses—appear to be the most critical to control of Pneumocystis infection [36–38] and (2) that antibody responses to Msg are easily detected in humans [39, 40], the primary function of Msg diversity may be to evade cell-mediated responses. Consistent with this hypothesis is our inability to detect in vitro proliferative responses to a recombinant Msg isoform when we have used human peripheral-blood mononuclear cells, despite the fact that antibodies to the same antigen are easily detected (J.A.K., unpublished observations); this result may reflect a low probability that the individuals had been infected with P. jirovecii expressing the cloned msg isoform. Although Pneumocystis infection in healthy hosts does not appear to result in the chronic waxing and waning infection [41] that is seen in infection by other species with antigenic variation (e.g., trypanosome species [42, 43]), msg diversity may facilitate reinfection of healthy hosts by delaying the development of effective cellular immune responses.

The substantial variability that RFLP analysis has demonstrated in the msg repertoire potentially provides a robust method for typing P. jirovecii. Current typing methods rely primarily on examination of variations in one or more single-nucleotide polymorphisms within a single locus or a limited number of loci [44], whereas RFLP analysis of amplified msg genes potentially allows interrogation of 50–100 genes. Such analysis may help us to determine the relationship among isolates from putative outbreaks of Pneumocystis pneumonia [45].

Acknowledgments

We thank Rene Costello and Howard Mostowski for their assistance with the animal studies.

References


