De Novo Assembly of the *Pneumocystis jirovecii* Genome from a Single Bronchoalveolar Lavage Fluid Specimen from a Patient

Ousmane H. Cissé, Marco Pagni, and Philippe M. Hauser

Institute of Microbiology, Centre Hospitalier Universitaire Vaudois and University of Lausanne, Lausanne, Switzerland; and Vital-IT Group, SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

ABSTRACT *Pneumocystis jirovecii* is a fungus that causes severe pneumonia in immunocompromised patients. However, its study is hindered by the lack of an *in vitro* culture method. We report here the genome of *P. jirovecii* that was obtained from a single bronchoalveolar lavage fluid specimen from a patient. The major challenge was the *in silico* sorting of the reads from a mixture representing the different organisms of the lung microbiome. This genome lacks virulence factors and most amino acid biosynthesis enzymes and presents reduced GC content and size. Together with epidemiological observations, these features suggest that *P. jirovecii* is an obligate parasite specialized in the colonization of human lungs, which causes disease only in immune-deficient individuals. This genome sequence will boost research on this deadly pathogen.

IMPORTANCE *Pneumocystis* pneumonia is a major cause of mortality in patients with impaired immune systems. The availability of the *P. jirovecii* genome sequence allows new analyses to be performed which open avenues to solve critical issues for this deadly human disease. The most important ones are (i) identification of nutritional supplements for development of culture *in vitro*, which is still lacking 100 years after discovery of the pathogen; (ii) identification of new targets for development of new drugs, given the paucity of present treatments and emerging resistance; and (iii) identification of targets for development of vaccines.
silico sorting and assembly of P. jiroveci reads out of a mixture representing many organisms. We iteratively built stringent assembly of the reads to detect homology with the available genome of P. carinii, as well as with other fungal genomic data. Ideally, reads should be attributed to a single organism before an assembly of its genome is attempted. However, this simple strategy could not be applied here, as most raw reads were too short to be attributed to a taxonomic group through sequence comparison. Moreover, some Pneumocystis-specific reads would certainly be lost or misidentified, as P. carinii assembly is known to be incomplete. Hence, we have established an overall strategy for read filtration and de novo assembly that allowed us to progressively refine and complete the P. jiroveci genome (see Text S1, section A9, and Fig. S1 in the supplemental material). The rRNA unit and the mitochondrial genome were recovered, assembled, and annotated separately. The 18S nuclear rRNA subunit displayed 99% nucleotide identity with the P. jiroveci public sequence, providing a solid assessment of the organism identity. The repeated telomere regions that contain the major surface glycoproteins were also kept apart. The repeated regions kept apart included probably the centromeres, which were not investigated further in the absence of a genetic or physical map. A posteriori, 25% of the reads could be attributed to P. jiroveci (see Table S3 in the supplemental material). In addition, we deep sequenced total RNA isolated from a nonenriched BALF of a patient with B-cell lymphoma, which provided 2.7% of reads attributable to the fungus. The de novo assembly of RNAseq data yielded 2,667 transcripts, which were used for genome annotation and provided a glimpse into the fulminating infection process in the human host. As previously observed for P. carinii (3), the most abundant category of the transcripts was annotated as major surface glycoproteins (7.2%).

The 8.1-Mb genome assembly is made of 356 contigs (Table 1). A total of 3,878 coding sequences were identified and annotated, with a gene density of 481 genes per Mb. Seventy-seven percent of them were supported by the transcriptome data. We identified 458 single-copy orthologs that were shared by P. jiroveci and several other fungal species, including distantly related ascomycetes and basidiomycetes. A phylogenetic tree was computed from an alignment of these orthologous proteins (Fig. 1). It further documents the taxonomic position of P. jiroveci close to P. carinii and Taphrina deformans.

The most striking feature of P. jiroveci was the lack of certain metabolic capabilities. As we previously found in P. carinii (4), the category of amino acid metabolism pathways was underrepresented in P. jiroveci (see Table S4 in the supplemental material), and a manual analysis revealed that most enzymes specifically dedicated to the synthesis of amino acids were absent (see Table S5 in the supplemental material). The loss of these pathways is a hallmark of obligate parasites (5). This strongly suggests that P. jiroveci scavenges these compounds from human lungs. Accordingly, an important proportion of its genes (22%) corresponded to transporters (e.g., amino acid permeases). Such transporters are believed to be necessary in P. carinii for scavenging amino acids as well as other compounds, such as host cholesterol and S-adenosylmethionine (reviewed in reference 6). The P. jiroveci genome presents a low GC content (29%) and a smaller size than its free-living relatives T. deformans and Schizosaccharomyces pombe (42% and 36%, 13 Mb and 14 Mb, respectively). In obligate bacterial parasites, the reduction of GC content is associated with the reduction of genome, which in turn is generally due to a loss of the genes responsible for the synthesis of compounds that can be scavenged from the host (7). P. jiroveci also lacks the hallmark enzymes of the glyoxylate cycle, a significant virulence factor of fungal pathogens (8), as well as polyketide synthase clusters, responsible for production of secondary metabolites, such as toxins. Together, these features are compatible with the view that P. jiroveci is an obligate parasite specialized in colonization of human lungs, which causes deadly disease only in immunocompromised individuals. This conclusion is further supported by the mechanism of surface antigen variation (6), the coevolution with hosts (9), and the strict host specificity of Pneumocystis spp. (10). Obligate parasitism also fits epidemiological studies, which failed to identify free-living forms as a source of infection. The host and reservoir of P. jiroveci is likely to be only humans, including immunologically impaired individuals, but also infants experiencing primo-infection (11), elderly people (12), pregnant women (13), and healthy transitory carriers (14). Consequently, the entire life cycle of P. jiroveci most probably takes place only within human lungs.

For an organism of clinical importance that cannot be grown in the laboratory as P. jiroveci, the genome sequence represents a

| Characteristic | Result | P. jiroveci | P. carinii |
|---------------|--------|------------|-----------|
| Assembly | Assembly size (Mb) | 8.1 | 6.3 |
| | Mean 544 read depth | 36 | NAa |
| | Mean Illumina read depth | 1,315 | NA |
| | No. of contigs | 358 | 4,278 |
| | Nc (kb) | 41.6 | 2.2 |
| | Mean GC content (%) | 28.4 | 32.5 |
| Annotation | No. of CDSs | 3,898 | 4,591c |
| | Coding regions (%) | 68.9 | 50.0 |
| | No. of KEGG orthologsd | 269 | 252 |
| | No. of tRNA genes | 77e | 36f |
| | Mean gene length (bp) | 1,472 | 891 |
| | Mean exon length (nt) | 211 | 223 |
| | Mean no. of introns per gene | 4.5 | 2.5 |
| | Mean intron length (nt) | 61 | 54 |
| | Repeat density (%) | 9.86 | 5.71 |
| Mitochondrial genome | No. of contigs | 27 | 23g |
| | GC content in whole genome (%) | 29.5 | 31.1 |
| | GC content in coding genes (%) | 32.5 | 30.9 |
| | No. of protein coding genes (CDSs) | 17 | 17 |
| | No. of rRNA genes | 2 | 2 |
| | No. of tRNA genes | 12 | 20 |

a The P. carinii assembly was downloaded from the Pneumocystis genome project website (http://ppg.cchmc.org/) and corresponds to the sequences published by Slaven et al. (15). This assembly is known to be incomplete. b Not applicable. c Some of these peptides were derived from incomplete CDSs located at the extremity of a contig. Hence, the numbers of predicted peptides provide only a rough estimation of the proteome size. d The number of KEGG orthologs was computed as described previously (4). e Not including pseudo tRNAs. f Only complete copies are shown; the total number of tRNAs in the genome may be more important. g P. carinii mitochondrial data were computed from the published genome (16).
wealth of new information for future research. It allows new analyses to be performed, such as RNA sequencing and comparative genomics, which open avenues to solve critical issues such as (i) the identification of nutritional supplements for culture in vitro development; (ii) the identification of new targets for development of new drugs, an important issue because only antifolates are presently efficient and development of drug resistance has been documented; and (iii) the identification of targets for the development of vaccines. The relevant *P. jirovecii* genes can now be used in these studies rather than those of *P. carinii* as models, which is particularly crucial for development of new drugs.

### SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at [http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.00428-12/-/DCSupplemental](http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.00428-12/-/DCSupplemental).

### ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation, grant 310030-124998. Computations were performed at the Vital-IT Center for High-Performance Computing of the Swiss Institute of Bioinformatics (http://www.vital-it.ch).

We are indebted to Melanie Cushion, Georges Smulian, and Alexey Porollo for providing helpful information on *P. carinii* genomic data-
bases. We thank Cécile-Marie Aliouat-Denis and Chabe Magali for providing *P. carinii* ESTs. We thank Laurent Keller, Michel Monod, Jacques Rougemont, Ioannis Xenarios, and Dominique Sanglard for critical reading. We thank Ioannis Xenarios, Thomas Bernard, Laurent Falquet, Sandra Calderon-Copete, Oksana Riba-Grognuz, Lorenzo Cerutti, Sébastien Moretti, Emmanuel Beaudoin, Sylvain Pradervand, and Thomas Junier for support and helpful discussions.

REFERENCES

1. Thomas CF, Jr, limper AH. 2007. Current insights into the biology and pathogenesis of *Pneumocystis* pneumonia. Nat. Rev. Microbiol. 5:298–308.
2. Lu JJ, et al. 1994. Typing of *Pneumocystis carinii* strains that infect humans based on nucleotide sequence variations of internal transcribed spacers of rRNA genes. J. Clin. Microbiol. 32:2904–2912.
3. Cushion MT, et al. 2007. Transcriptome of *Pneumocystis carinii* during fulminate infection: carbohydrate metabolism and the concept of a compatible parasite. PLoS One 2:e423. [http://dx.doi.org/10.1371/journal.pone.0000423](http://dx.doi.org/10.1371/journal.pone.0000423).
4. Hauser PM, et al. 2010. Comparative genomics suggests that the fungal pathogen *pneumocystis* is an obligate parasite scavenging amino acids from its host’s lungs. PLoS One 5:e15152. [http://dx.doi.org/10.1371/journal.pone.0015152](http://dx.doi.org/10.1371/journal.pone.0015152).
5. Payne SH, Loomis WF. 2006. Retention and loss of amino acid biosynthetic pathways based on analysis of whole-genome sequences. Eukaryot. Cell 5:272–276.
6. Cushion MT, Stringer JR. 2010. Stealth and opportunism: alternative lifestyles of species in the fungal genus *Pneumocystis*. Annu. Rev. Microbiol. 64:431–452.
7. Merhej V, Royer-Carenzi M, Pontarotti P, Raoult D. 2009. Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. Biol. Direct 4:13.
8. Lorenz MC, Fink GR. 2001. The glyoxylate cycle is required for fungal virulence. Nature 412:83–86.
9. Demanche C, et al. 2001. Phylogeny of *Pneumocystis carinii* from 18 primate species confirms host specificity and suggests coevolution. J. Clin. Microbiol. 39:2126–2133.
10. Wakefield AE, Stringer JR, Tamburrini E, Dei-Cas E. 1998. Genetics, metabolism and host specificity of *Pneumocystis carinii*. Med. Mycol. 36(Suppl 1):183–193.
11. Vargas SL, et al. 2001. Search for primary infection by *Pneumocystis carinii* in a cohort of normal, healthy infants. Clin. Infect. Dis. 32:855–861.
12. Vargas SL, et al. 2010. *Pneumocystis* colonization in older adults and diagnostic yield of single versus paired noninvasive respiratory sampling. Clin. Infect. Dis. 50:e19–e21.
13. Vargas SL, et al. 2003. Pregnancy and asymptomatic carriage of *Pneumocystis jiroveci*. Emerg. Infect. Dis. 9:605–606.
14. Miller RF, Ambrose HE, Wakefield AE. 2001. *Pneumocystis carinii* f. sp. hominis DNA in immunocompetent health care workers in contact with patients with *P. carinii* pneumonia. J. Clin. Microbiol. 39:589–591.
15. Slaven BE, et al. 2006. Draft assembly and annotation of the *Pneumocystis carinii* genome. J. Eukaryot. Microbiol. 53(Suppl 1):S89–S91.
16. Sesterhenn TM, et al. 2010. Sequence and structure of the linear mitochondrial genome of *Pneumocystis carinii*. Mol. Gen. Genomics 283:63–72.