

Draft Assembly and Annotation of the *Pneumocystis carinii* Genome

BRADLEY E. SLAVEN,^{a,b} JAROSLAW MELLER,^{a,c} ALEKSEY POROLLO,^{a,c} THOMAS SESTERHENN,^{a,b}

A. GEORGE SMULIAN^{a,b} and MELANIE T. CUSHION^{a,b}

^aUniversity of Cincinnati College of Medicine, Cincinnati, Ohio, and

^bCincinnati Veterans Administration Medical Center, Cincinnati, Ohio, and

^cCincinnati Children's Hospital Research Foundation, Cincinnati, Ohio

SEQUENCING and assembly of the *Pneumocystis carinii* genome presented many challenges to currently available molecular and bioinformatic technologies. The unculturable and non-clonal nature of the organism necessitated the in vivo harvesting and purification of the organisms from the rat lung. This resulted in a multi-population organism DNA mix that required the implementation of additional informatics purging processes for contaminating sequences. Sequencing libraries were largely generated from amplified DNA extracted from chromosome bands separated by contour-clamped homogeneous electric fields (CHEF) gels. The A+T richness of the genome (67%) caused cloning inefficiencies resulting in sequence coverage gaps in the assembly (Slaven et al. 2006). Arachne, Phrap and Cap3 assemblers used individually were unable to merge overlapping contigs on a genome-wide basis. We implemented an alternative strategy which took advantage of all three genomic assembly systems.

Before the *Pneumocystis* genome project, limited genomic and transcriptional message sequence was available for analysis. A ~32 kb Cosmid (15A06; Smulian et al. 2001), which included 15 genes and 55 introns, was sequenced by us and incorporated in this first version draft assembly (v1). In addition, 24 *P. carinii* genes containing ~94 kb of sequence were downloaded from GenBank and used in the assembly. Seven telomeric *P. carinii* contigs containing ~225 kb of telomeric end sequences were previously sequenced as a joint project by the University of Cincinnati and the Sanger Center (Keely et al. 2005). These data are not included in this draft. All data from our genomic and transcript sequencing and subsequent annotations are available at the PGP (*Pneumocystis* Genome Project) web site, <http://pgp.cchmc.org>.

MATERIALS AND METHODS

Libraries used for sequencing. The vast majority of our assembly-qualified small (1,000–4,000 bp) insert reads (AQSIR) (50,485 of 58,090) were created using PCR amplification of CHEF-derived chromosomes cloned into the Lucigen pSMART-HCKan cloning vector (Mead, D., Patterson, M., Schoenfeld, T., Smulian, A. G., Cushion, M.T., Baric, R.S. & Godiska, R. 2006. High Stability Vectors for Cloning Unstable DNA, Lucigen Corporation, http://www.lucigen.com/catalog/images/pdfs/GSAC_Poster.pdf). An additional 7.5% (4,368 of 58,090) of the AQSIR were created by blunt-ended restriction digestion (Sesterhenn et al. 2003). The remainder (3,439 of 58,090) were created using traditional sonication, followed by enzymatic blunt ending of the fragments and shotgun insertion techniques. Additional AQSIR (1,463) were obtained by sequencing large pWEB cosmid end inserts. These cosmid end sequences were used primarily to provide scaffolding bridges between small insert contig assemblies. The small insert sequence reads had a mean length of

533 bp ± 129 SD. The lengths of large insert reads averaged 271 bp ± 158 SD.

Assembly strategy. The *P. carinii* genomic assembly process used a hybrid of the genomic shotgun assembly procedure pioneered at The Institute for Genomic Research (TIGR) (Fleischmann et al. 1995) with an added cosmid end sequencing component to create a scaffold with which to anchor the assembled shotgun sequences. A schematic of the overall assembly process is shown in Fig. 1. Input sequences consisted of 134,423 small insert shotgun and cosmid end reads. However, over 55% of these reads (74, 668 sequences) were not included in the assembly process because they were identified as host (rat) DNA, bacterial or viral contaminants, vector sequence, telomeric end sequences, or they had low sequence quality.

Processing pipeline. A plate processing pipeline was used to identify input sequences from a specific plate and insert vector reads by using forward and reverse sequencing names. These identifications were used later in the assembly process to verify genomic contig assemblies and to identify putative super contigs. A super contig is defined as a collection of two or more continuous (contigs) sequences with forward and reverse vector reads in different contigs. They are separated by a sequencing gap.

Genomic sequence reads were trimmed using Phred-base quality scores (Ewing and Green 1998; Ewing et al. 1998), which are log transformed measures of sequencing error. The quality trimming program we developed for the *P. carinii* genome found the highest quality read length segment above a minimum size and within a specified log—transformed error rate. *Pneumocystis carinii* sequences were required to have a minimum length of 300 bp and an expected error rate of 5%. Sequence reads not meeting minimum length and quality score levels were removed from the assembly process. This algorithm is similar to one that is used in the Arachne genomic assembly system (Jaffe et al. 2003).

The pipeline executed the Phred/Phrap script to screen for cloning vector contaminants (Sesterhenn et al. 2003). An additional contaminant-specific screening process removed bacterial and viral contaminants as well as rat DNA from the sequence collection. CROSSMATCH (Green, P. & Ewing, B. 2002. CROSSMATCH, <http://www.phrap.org>), an implementation of the Smith-Waterman algorithm was used for this purpose (Temple et al. 1981). All input sequences were screened against the entire genomes of *Rattus norvegicus* (the rat host), *Staphylococcus aureus* subsp. *aureus* MRSA252, *Pseudomonas putida*, *Pasteurella multocida* subsp. *multocida* str. Pm70, *Pseudomonas aeruginosa*, *Bacillus subtilis*, *Haemophilus influenzae* KW20, *Murine adenovirus 1*, *Murine adenovirus A*, *Murid herpesvirus 2* and *Escherichia coli*.

Because of the repetitive nature of the telomeric ends in the *P. carinii* genome, a non-contaminant telomeric screen was conducted to exclude the chromosomal ends from this initial draft assembly process. The BLAST(N/X) algorithm was used to screen reads against MSR/MSG/PRT telomeric gene cassettes. Using this process (Altschul et al. 1990, Smulian et al. 2001; Temple et al. 1981) 9,997 high quality sequence reads were identified. A *P. carinii* telomeric end assembly is planned in subsequent versions of the genomic assembly.

Corresponding Author: B. Slaven, Department of Internal Medicine, Division of Infectious Diseases, University of Cincinnati College of Medicine, 231 Albert Sabin Way, Cincinnati, OH, 46267-0560 USA—Telephone number: +1-513-290-8700; FAX number: +1-513-636-2056; e-mail: bradley.slaven@uc.edu

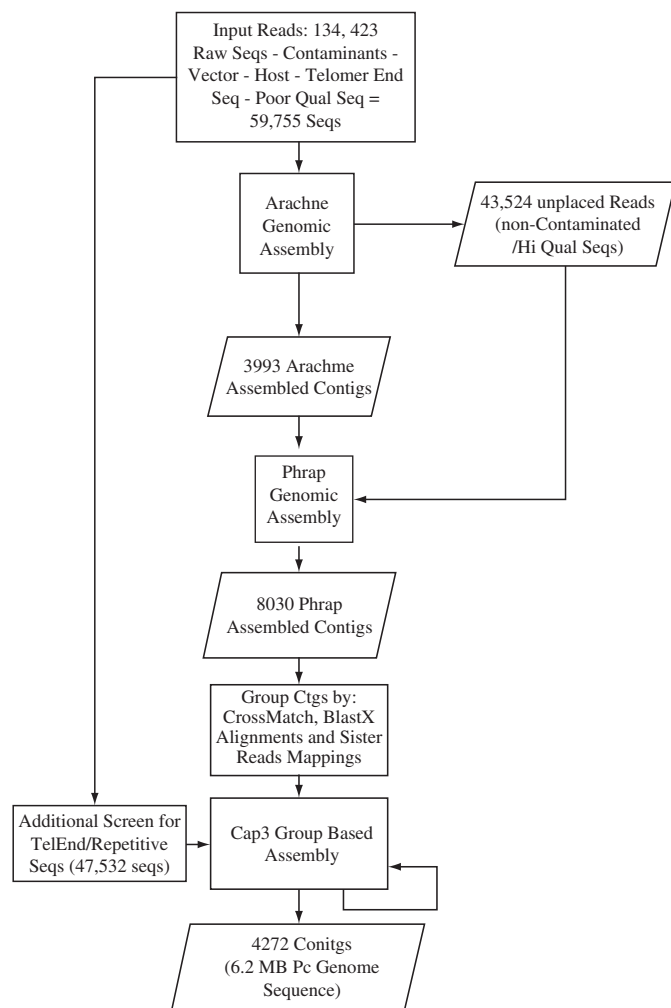


Fig. 1. The multi-staged process flow-chart used to assemble the *Pneumocystis carinii* genome.

The sequence reads were initially assembled using the Arachne Assembly system. This assembly resulted in 3,993 Arachne contigs containing 4.5 Mb of the *P. carinii* genome (Table 1). However, only 27.4% (16,231 of 59,755) of the non-contaminated high-quality sequence reads were assembled into contigs using this assembly process. The remaining 43,534 sequence reads were

Table 1. The arachne, secondary phrap assembly, and iterative cap3 assembly contig lengths are listed as nucleotide bp for each major stage in the overall *Pneumocystis carinii* genome assembly.

Assembled contig lengths (bp)	(a) Arachne assembly	(b) Secondary phrap assembly	(c) Iterative cap3 assembly
0–1,000	1,596	4,881	2,476
1,001–3,000	2,260	2,681	1,386
3,001–5,000	103	341	246
5,001–7,500	22	81	102
7,501–10,000	5	28	40
10,001–15,000	4	15	21
15,001–20,000	1	1	2
20,001–30,000	1	0	3
30,001–40,000	1	1	1
40,001–50,000	0	1	1
Total contig lengths	4,507	9,585	6,258

left as unassembled sequence fragments. Only 12 Arachne-assembled contigs had lengths greater than 7,500 bp.

In an effort to more extensively utilize existing sequencing information and extend contig lengths, we used Phrap to combine the 3,993 Arachne contigs, the 43,534 unassembled sequence reads, the 25 GenBank downloaded genes, and a 32,000 bp previously assembled contig (15A06) into a second genome-wide assembly (Wu et al. 2006). Simulated uniform quality scores of 25 were used for the 15A06 cosmid and the downloaded GenBank bases. Actual Phred scores were used for all remaining sequences that were available in the GenBank database. This level two Phrap assembly process produced 8,030 contigs which contained ~ 9.6 Mb of sequence. Forty-six of these contigs were greater than 7,500 bp in length. Because of the more lenient assembly requirements used by the Phrap system, this assembly placed a significantly higher percentage of the sequence reads, 89.0% (57,443 of 64,538), into the 8,030 assembled contigs. This process took advantage of the more strictly assembled Arachne contigs as a basis for the genomic structure. Then, Phrap extended the genomic assembly using the additionally layered sequences which were excluded by Arachne.

Though the secondary Phrap assembly used significantly more non-contaminated, high quality reads in the assembly and produced longer genomic contigs, many putative overlapping contigs were still not assembled. The secondary Phrap assembly contained 9.6 Mb of sequence. Considering that the *P. carinii* genome contains ~ 7.0 Mb of non-telomeric chromosomal DNA, we estimated that 36% unmerged nucleotide base pairs resulted from the Phrap assembly [1.0–(9.6/7.0) = 0.37].

We identified overlapping sequence matches between unmerged Phrap-assembled contigs by using CROSSMATCH. The unmerged contigs with significant sequence overlaps were grouped into related clusters, then re-assembled individually (within each group) using the Cap3 assembler (Huang and Madan 1999). This process resulted in the 8,030 Phrap contigs being merged into 4,272 iteratively assembled genomic contigs.

Annotation. Putative genes and biochemical pathways were identified in this draft genome using cDNA sequencing and homology-based search tools. A multi-fasta file was created by splitting the genomic contigs into 2,000-bp segments with 1,000 bp overlapping sections. Homology comparisons were conducted using BLAST(N/X) searches against the GenBank, NR and UniProt/TrEMBL databases (Wu et al. 2006) (National Institutes of Health, National Library of Medicine, National Center for Biotechnology Information, 2006. <http://ftp.ncbi.nlm.nih.gov/blast/db/blastdb.html>). The multi-fasta file was also submitted to the KEGG/KAAS annotation server to obtain *P. carinii*-specific biochemical pathway maps (Kanehisa, M. 2006. <http://www.genome.jp/kegg/kaas>, Kanehisa Laboratory, Bioinformatics Center, Institute for Chemical Research, Kyoto University). In a parallel project, we sequenced ~ 5,000 *P. carinii* cDNA to obtain 1,063 unique transcriptional messages. The transcriptional messages were aligned with the genomic contigs. That process added additional validity to the database-driven gene predictions.

RESULTS AND DISCUSSION

Assembly of the genome. The 4,272 iterative assembled contigs contain ~ 6.3 Mb of distinct *P. carinii* genomic sequence (Table 1). Sixty-eight of these 4,272 contigs have lengths greater than 7,500 bp. Nearly half of the contigs (2,010 of 4,272) were merged into 878 directionally oriented supercontigs using sister read names and cDNA alignments that spanned adjacent contigs.

The genomic sequencing and analysis of a non culturable A+T-rich organism harvested from the lungs of immunocompromised rats proved to be a significant challenge to current genomic

technologies and genomic assembly systems. DNA amplification of chromosome gel bands and use of the Lucigen pSMART HPKAN vector contributed significantly to the success of this sequencing effort. The non-clonal nature of the organism and the lack of a culture system proved to be significant obstacles to genomic assemblers trained on clonal populations. Given that most organisms in nature do not have established culture systems, this severely limits the scope of traditional shotgun assembly processes. Significant adaptations to traditional sequencing and assembly processes, as we have discussed here, will likely be necessary for their assembly. Because of the heuristic nature of genomic assembly algorithms, as implemented by Arachne and Phrap, many overlapping sequences (contigs) were not assembled together. In particular, Smith–Waterman-based contig comparisons using CROSSMATCH identified a number of overlapping genomic contigs that were not combined by traditional assembly systems.

Putative *P. carinii* open reading frames (3,067) were identified using database-driven homology searches of our genomic contig results, complemented by a cDNA sequencing and alignment project. We expect these open reading frames to represent ~ 75% of the gene complement of the *P. carinii* genome. Addition of the MSG–MSR–PRT genes should provide an additional 10–15%. Microarrays based on these data were printed and assessments of initial hybridizations were presented at this meeting (IWOP-9).

Annotation of the genome. Homology-based annotations using the KEGG/KASS annotation server placed 820 putative *P. carinii* genes into over 100 biochemical pathways. In particular, over half of the elements in the *P. carinii* cell cycle have been putatively identified. Genes in both sexual and asexual modes of replications were identified. Twenty of 32 elements of the proteasome (using the *S. cerevisiae* model) and half of the elements of the ubiquinone biosynthetic pathway were populated by putative *P. carinii* genes. Of note, the ubiquitin proteasome system is essential for many biochemical processes, including cell cycle regulation, cell surface modulation, secretion, DNA repair and transcriptional regulation in fungal and mammalian organisms (Glickman and Ciechanover 2002). This initial genomic analysis has putatively identified 20 of 26 eukaryotic transcriptional factors, which may lead to novel approaches to better understand and disrupt organism growth during infection.

ACKNOWLEDGMENTS

This work was supported by grants R21 AI055338 and AI44651 (NIAID) from the National Institutes of Health.

LITERATURE CITED

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman D. . J. 1990. Basic local alignment search tool. *J. Mol. Biol.*, **215**:403–410.
- Ewing, B. & Green, P. 1998. Basecalling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**: 186–194.
- Ewing, B., Hillier, L., Wendl, M. & Green, P. 1998. Basecalling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**:175–185.
- Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, K. G., FitzHugh, W., Fields, C., Gocayn, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser C. . V., Smith, H. O. & Venter, J. C. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**: 496–512.
- Glickman, M. H. & Ciechanover, A. 2002. The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction. *Physiol. Rev.*, **82**:373–428.
- Huang, X. & Madan, A. 1999. CAP3: a DNA sequence assembly program. *Genome Res.*, **9**:868–877.
- Jaffe, D. B., Butler, J., Gnerre, S., Mauceli, E., Lindblad-Toh, K., Mesirov, J. P., Zody, M. C. & Lander, E. S. 2003. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.*, **13**:91–96.
- Keely, S. P., Renauld, H., Wakefield, A. E., Cushion, M. T., Smulian, A. G., Fosker, N., Fraser, A., Harris, D., Murphy, L., Price, C., Quail, M. A., Seeger, K., Sharp, S., Tindal, C. J., Warren, T., Zuiderwijk, E., Barrell, B. G., Stringer, J. R. & Hall, N. 2005. Gene arrays at *Pneumocystis carinii* telomeres. *Genomics*, **70**:1589–1600.
- Sesterhenn, T., Slaven, B. E., Smulian, A. G. & Cushion, M. T. 2003. Generation of sequencing libraries for the *Pneumocystis* Genome project. *J. Eukaryot. Microbiol.*, **50**:663–665.
- Slaven, B. E., Porollo, A., Sesterhenn, T., Smulian, A. G., Cushion, M. T. & Meller, J. 2006. A large scale characterization of the introns in the *Pneumocystis carinii* genome. *J. Eukaryot. Microbiol.*, **53**, this issue.
- Smulian, A. G., Sesterhenn, T., Tanaka, R. & Cushion, M. T. 2001. The ste3 pheromone receptor gene of *Pneumocystis carinii* is surrounded by a cluster of signal transduction genes. *Genetics*, **157**:991–1002.
- Temple, F., Smith, T. F. & Waterman, M. S. 1981. Identification of common molecular subsequence. *J. Mol. Biol.*, **147**:195–197.
- Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Mazumder, R., O'Donovan, C., Redaschi, N. & Suzek, B. 2006. The universal protein resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**(database issue):D187–D191.