

**Draft Assembly and Annotation of the *Pneumocystis carinii* genome.**

**BRADLEY E SLAVEN<sup>1</sup>, J. MELLER<sup>2</sup>, A. POROLLO<sup>2</sup>, T. SESTERHENN<sup>1</sup>, A. G. SMULIAN<sup>1</sup>, M.T. CUSHION<sup>1</sup>, <sup>1</sup>University of Cincinnati College of Medicine and the Cincinnati Veterans Administration Medical Center, Cincinnati, OH; <sup>2</sup>Childrens Hospital Research Foundation Cincinnati, Ohio**

*Pneumocystis carinii* (Pc) presents many challenges to genomic sequencing and assembly technologies, highlighting current limitations. The genome A+T richness (67%), biased library construction, and subsequent genomic coverage necessitated a customized strategy. Lack of a viable culture system required purification of organisms from the lungs of immunologically suppressed rats during fulminate infection. Sequencing libraries were generated primarily from amplified DNA extracted from chromosome bands separated by CHEF. Cosmid end sequencing provided some scaffolding. Crossmatch (Green et. al., 1996), an implementation of the Needleman-Wunsch algorithm, was used to compare sequences against host, bacterial and viral genomes to remove contaminants. The non-clonal nature of the Pc populations further complicated the assembly process since even slight genetic population drift inhibited genomic assembly merge operations. Approximately 10% of the Pc genome is composed of 3 telomerically located gene (PRT1/MSR/MSG) paralogs (Keely et.al., 2006). Therefore, this draft assembly focused on the less repetitive portion of the genome.

These Pc- genome specific challenges limited the effectiveness of heuristically based Arachne (Jaffe, 2003), Phrap (Green and Ewing, 2002) and Cap3 (Huang, 1999) assembly systems and an alternative assembly strategy was implemented that took

advantage of all 3 programs. Arachne was used to construct backbone contig based scaffolds. Phrap was used to merge the contig based scaffolds. Contigs were compared using Crossmatch to identify non-merged contig overlapping areas. Overlapping contigs were binned and re-assembled in individual bins using Cap3. We tested various overall and localized assembly system results. Assembly system parameters were optimized, and merged contig constructions were computationally validated.

The resulting iterative assembly process produced a draft of the Pc genome containing ~6.2 million sub-telomeric base pairs contained within 4272 contigs. Nearly half of the contigs (2010 of 4272) were merged into 878 directionally oriented supercontigs using sister read names and cDNA alignments. Annotation of the genome is underway. A homology based annotation was conducted on the contigs using the KEGG-KASS<sup>6</sup> annotation server.