

Large-Scale Characterization of Introns in the *Pneumocystis carinii* Genome

BRADLEY E. SLAVEN,^{a,b} ALEKSEY POROLLO,^{a,c} THOMAS SESTERHENN,^{a,b} A. GEORGE SMULIAN,^{a,b}

MELANIE T. CUSHION^{a,b} and JAROSLAW MELLER^{a,c}

^aUniversity of Cincinnati College of Medicine, Cincinnati, Ohio, 46267-0560, and

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

^cCincinnati Children's Hospital Research Foundation Cincinnati, Ohio

PREVIOUS data on *Pneumocystis carinii* introns were limited to 19 genes and 83 introns downloaded from GenBank (Smulian et al. 2001), as well as the sequencing and analysis of a cosmid (15A06). Cosmid 15A06 is a 32,000 base sequence localized to chromosome 1, which contains 15 genes and 55 introns (Slaven et al. 2006). Putative donor and acceptor signals have been previously identified by these studies. However, no branch site patterns have been reported for *P. carinii*.

In the present study, a large-scale characterization of *P. carinii* introns and associated splice site elements was conducted using genomic and transcriptional sequence data obtained from the *Pneumocystis* Genome Project (PGP). Using statistical analysis and motif finding approaches, we identified donor, acceptor, and branch site patterns, as well as intron, exon, and overall A+T base pair (bp) content. This strategy should lead to improvement in the accuracy in identifying *P. carinii* genes and finding novel genes that are not recognized by database homology methods.

MATERIALS AND METHODS

Pneumocystis carinii genomic contigs were assembled using an iterative assembly process, as described elsewhere (Huang and Madan 1999). A total of 7,531 cDNA transcripts were aligned against 4,272 non-telomeric genome contigs ($\sim 6.2 \times 10^6$ bp) using GeneSeqer (Thomas, Leof, and Limper 1999). GeneSeqer uses Hidden Markov Models state transition between exon and intron to determine splice site alignment probabilities and a dynamic programming algorithm to calculate optimal cDNA-to-genomic alignments. Only the highest quality cDNA-to-genomic alignments were selected for inclusion in this study. Alignments required greater than 90% proximal 50-bp exon identity. Alternative and/or conflicting splice alignments were not permitted.

Donor site patterns were determined by excising genomic sequence regions from cDNA-to-genomic alignments 10 bp upstream and downstream of GeneSeqer-determined intron/exon borders. Patterns in the excised regions were determined by Improbizer using its default settings J. Kent (<http://www.cse.ucsc.edu/~kent/improbizer/improbizer.html>). Improbizer uses a variation of the expected maximizer (EM) algorithm to predict statistically significant likely motif patterns. Acceptor site patterns were determined in a similar manner.

Initially, branch site patterns were not observed using complete intron sequences removed from genomic-to-cDNA alignments. However, intron alignments beginning 30 bp upstream of their acceptor sites (i.e. trimming them to 30 bp and ‘right justifying’ them by their 3' donor sites) revealed the branch site patterns. After alignment, sequence patterns were determined by subjecting them to Improbizer analysis.

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

RESULTS AND DISCUSSION

In the database of 7,531 cDNA transcripts, 1,063 unique putative genes were identified using the Cap3 assembler and BLAST(N/X), KEGG/KAAS and Blast2Go homology-based systems (Altschul et al. 1990; Conesa et al. 2005; Cushion et al. 2006) (Kanehisa, M., 2006, Kanehisa Laboratory, Bioinformatics Center, Institute for Chemical Research, Kyoto University. <http://www.genome.jp/kegg/kaas>). Putative unique genes (989) exhibited BLASTX homology to proteins in the NR and UniProt/TrEMBL protein databases (including hypothetical protein hits) with expected e values $< 10^{-6}$ (Usuka, Zhu, and Brendel 2000; National Institutes of Health, National Library of Medicine, National Center for Biotechnology Information, 2006; <ftp://ftp.ncbi.nlm.nih.gov/blast/db/blastdb.html>; Cushion et al. 2006, Wu et al. 2006). After using strict cDNA-to-genomic sequence alignment criteria and after removing alternative and conflicting splice site alignments (Slaven et al. 2006), 1,781 introns (86,694 bp) and 3,593 exons (705,555 bp) were identified in the *P. carinii* genome.

Intron characterizations. *Pneumocystis carinii* introns have a strong Adenine and Thymine (A+T) nucleotide bias (Table 1). Over 96% of the introns (1,710/1,781) had canonical (5'GU ... AG3') splice sites. Another 1.12% of the introns (20/1,781) contained non-canonical (5'GC ... AG3') splice sites. Exons had an average length of 196.4 bp with an A+T content of 66.7%. Intron lengths were within a narrow range; >75% (1,344/1,781) consisted of 40–50 bp. The average intron length was 48.7 bps (Table 1). The shortest intron was 36 bp. We therefore hypothesize that 36 bps is the minimum intron length that permits removal by the *P. carinii* spliceosome.

Donor site pattern. Conserved donor and acceptor site patterns extended beyond 2 bp nucleotide signals. *Pneumocystis carinii* donor sites consisted of an overall 8-bp signal. The strongest nucleotide bp preservation was at the ‘GT’ dinucleotide donor site. The guanine and thymine ‘GT’ dinucleotide has 98.9% and 97.9% bp conservation levels, respectively (see Fig. 1A). The overall observed pattern is ‘AGGTATTT,’ with the donor site occurring in the third and fourth nucleotide positions. The pattern may be more broadly represented as

Table 1. Characteristics of *Pneumocystis carinii* introns and exons and compared with the overall G+C and A+T nucleotide content of the entire genome.

Genome regions	Number of putative identified regions	Number of nucleotides (bp)	Average length (bp)	G+C (%)	A+T (%)
Introns	1,781	86,694	48.7	20.7	79.1
Exons	3,593	705,555	196.4	33.3	66.7
Whole genome		$\sim 6.2 \times 10^6$		30.1	69.9

bp, base pair.

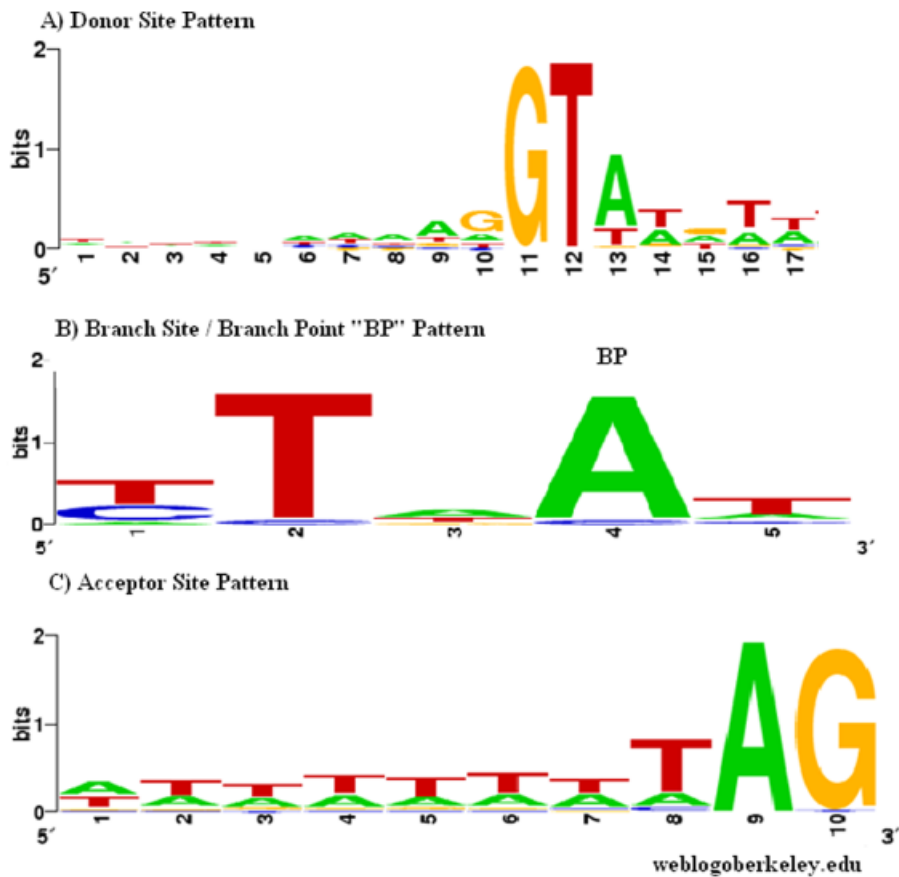


Fig. 1. Graphical representation of the *Pneumocystis carinii* genome. (A) Donor site. (B) Branch site pattern with branch point indicated. (C) Acceptor site patterns. Letter height indicates a proportional shared nucleotide pattern. The graphics for these logos were generated using <http://weblogoberkeley.edu>.

“WRGTWWW,” using the International Union of Pure and Applied Chemistry (IUPAC) nucleotide codes, with “W” representing adenine or thymine and “R” standing for adenine or guanine (International Union of Pure and Applied Chemistry (IUPAC), http://www.iupac.org/dhtml_home.html).

Acceptor site pattern. The acceptor site exhibited an overall 5 bp signal. The adenine and guanine “AG” di-nucleotide acceptor site have 99.0% and 98.4% bp conservation levels, respectively (see Fig. 1C). The overall conserved pattern was “TTTAG,” with the acceptor site occupying the last two nucleotide positions. This pattern may be described as “WWWAG,” using W for the IUPAC code for adenine or thymine.

Branch site pattern. The most frequently occurring *P. carinii* branch site pattern is “TTAAT.” This pattern was found 8–17 bp from the 3′ intron acceptor sites (see Fig. 1B). The branch point was identified as the second adenine nucleotide in the pattern and had the highest base pair conservation (96.4%). The pattern may be more broadly represented as “HTDAH,” using IUPAC codes with “H” representing not guanine and “D” meaning not cytosine.

These specific donor, acceptor, and branch site patterns, combined with tightly limited intron lengths, and intron and exon nucleotide biases will be used to improve the accuracy level of gene prediction in *P. carinii*. For example, we might be able to search known donor, branch site, and acceptor patterns within a 100-bp sliding window to identify putative intron locations. The putative coding regions in the *P. carinii* genome may then be determined

by evaluating the sequences between introns. This strategy should lead to the identification of novel genes in *P. carinii* that currently cannot be done by database homology methods alone. This will provide more complete annotation data for the *P. carinii* genome in the future.

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.
- Conesa, A., Gota, S., Garcia-Gomez, J. M., Perol, J., Talon, M. & Robles, M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **15**:3674–3676.
- Cushion, M. T., Smulian, A. G., Slaven, B. E., Sesterhenn, T., Porollo, A. & Meller, J. 2006. Analysis and functional classification of transcripts expressed by *Pneumocystis carinii* during fulminate infection. Abstract PL82. Ninth International Workshops on Opportunistic Protists.
- Huang, X. & Madan, A. 1999. CAP3: a DNA sequence assembly program. *Genome Res.*, **9**:868–877.
- Slaven, B., Meller, J., Porollo, A., Sesterhenn, T., Smulian, A. & Cushion, M. 2006. Draft assembly and annotation of the *P. carinii* genome. *J. Eukaryot. Microbiol.*, **53**:588–590.

- 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.
- Thomas, C., Leof, E. & Limper, A. 1999. Analysis of *Pneumocystis carinii* introns. *Infect. Immun.*, **67**:6157–6160.
- Usuka, J., Zhu, W. & Brendel, V. 2000. Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics*, **16**:203–211.
- 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.