Mapping by Sequencing the Pneumocystis Genome Using the Ordering DNA Sequences V3 Tool

Zheng Xu,* Britton Lance,† Claudia Vargas,† Budak Arpinar,* Suchendra Bhandarkar,* Eileen Kraemer,* Krys J. Kochut,* John A. Miller,* Jeff R. Wagner,‡ Michael J. Weise,§ John K. Wunderlich,‡ James Stringer, George Smulian,†† Melanie T. Cushion†† and Jonathan Arnold†,1**

*†Department of Genetics, *Department of Computer Science and ‡Molecular Genetics Instrumentation Facility, University of Georgia, Athens, Georgia 30602, §Accelrys, Madison, Wisconsin 53711-1060, **Department of Molecular Genetics, Biochemistry and Microbiology, University of Cincinnati College of Medicine, Cincinnati, Ohio 45267 and ††Department of Internal Medicine and the Cincinnati VAMC, University of Cincinnati College of Medicine, Cincinnati, Ohio 45220*

Manuscript received June 14, 2002
Accepted for publication December 19, 2002

ABSTRACT

A bioinformatics tool called ODS3 has been created for mapping by sequencing. The tool allows the creation of integrated genomic maps from genetic, physical mapping, and sequencing data and permits an integrated genome map to be stored, retrieved, viewed, and queried in a stand-alone capacity, in a client/server relationship with the Fungal Genome Database (FGDB), and as a web-browsing tool for the FGDB. In that ODS3 is programmed in Java, the tool promotes platform independence and supports export of integrated genome-mapping data in the extensible markup language (XML) for data interchange with other genome information systems. The tool ODS3 is used to create an initial integrated genome map of the AIDS-related fungal pathogen, *Pneumocystis carinii*. Contig dynamics would indicate that this physical map is ~50% complete with ~200 contigs. A total of 10 putative multigene families were found. Two of these putative families were previously characterized in *P. carinii*, namely the major surface glycoproteins (MSGs) and HSP70 proteins; three of these putative families (not previously characterized in *P. carinii*) were found to be similar to families encoding the HSP60 in *Schizosaccharomyces pombe*, the heat-shock ψ protein in *S. pombe*, and the RNA synthetase family (*i.e.*, MES1) in *Saccharomyces cerevisiae*. Physical mapping data are consistent with the 16S, 5.8S, and 26S rDNA genes being single copy in *P. carinii*. No other fungus outside this genus is known to have the rDNA genes in single copy.

IN the past 12 years, genomics have provided scientists a fundamental, comprehensive, and systematic way to understand life. Fungi as simple eukaryotes played a central role in the development of genomics (Arnold 1997, 2001). Their compact genomes have been exploited to develop new genomic tools, such as yeast artificial chromosomes (Burke *et al.* 1987) and novel physical mapping algorithms (Cuticchia *et al.* 1992). The ability to carry out site-directed transformation places fungi in a unique position relative to plants and animals for functional studies of many genes at once (Bennett and Arnold 2001). Objectives of many genome projects include: (1) constructing high-resolution genetic maps; (2) developing physical maps with an emphasis on those that allow investigators access to mapped DNA; (3) determining the complete genomic sequence of a target organism (Goffeau *et al.*, 1996); (4) annotating all of these sequences with special reference to the genes present (Kraemer *et al.* 2001); and (5) developing a capability for collecting, storing, distributing, and analyzing genome data (Kochut *et al.* 1993). In >15 fungal genome projects, high priority is being given to constructing an integrated genome map including all of the information in 1–4, ultimately yielding the total nucleotide sequence of a particular species (Prade *et al.* 1997; Bennett and Arnold 2001; Kelkar *et al.* 2001). The focus here is on developing a bioinformatics tool that facilitates the process in 1–4 (Hall *et al.* 2001) and enables storing, integrating, and distributing genetic, physical, and sequencing information in 5.

Building such an integrated genome map is a required step toward the rational understanding of the general structure, function, and evolution of fungal genomes.

Fungal chromosomes are on the order of 0.2–15 Mb in size and can be separated by pulsed-field gel electrophoresis (Bennett and Arnold 2001). As a consequence, a common starting point in fungal genomics is the bottom-up contig mapping approach to generating physical maps (Olson *et al.* 1986; Hoheisel *et al.* 1993; Mizukami *et al.* 1993; Prade *et al.* 1997). In constructing a physical map researchers first take each chromosome and break it up into small (40- to 150-kbp inserts) overlapping clones. The original ordering of these clones

[1]*Corresponding author:* Genetics Department, University of Georgia, Athens, GA 30602. E-mail: arnold@uga.edu

is then inferred by using a combination of experimental and computational approaches (Prade *et al.* 1997; Bhandarkar *et al.* 2001). One widely used method of contig mapping is clone-probe hybridization-based mapping (*i.e.*, Prade *et al.* 1997; Chibana *et al.* 1998; Enkerli *et al.* 2000; Aign *et al.* 2001). Each clone is tested for the presence of a probe sequence. If two clones hybridize, then they may contain the same sequences and are inferred to overlap. Clone-probe hybridization is an efficient approach to recovering the order of clones on each chromosome and is being used in several fungal genome projects (*i.e.*, Arnold and Cushion 1997; Cushion and Arnold 1997; Aign *et al.* 2001). When such a physical map is coupled with a genetic map, long-range continuity is achieved (Hall *et al.* 2001). Other experimental methods of ordering clones include radiation-hybrid mapping (Slonim *et al.* 1997), restriction mapping (Coulson *et al.* 1995), and optical mapping (Lin *et al.* 1999). The ordering DNA sequences V3 (ODS3) software is designed to support a particular new bottom-up strategy called mapping by sequencing (Mahairas *et al.* 1999).

The mapping-by-sequencing strategy (Venter *et al.* 1996; Mahairas *et al.* 1999) begins by generating end sequences on a deep-coverage large insert (>20×) library. As an example, in the Pneumocystis Genome Project end sequences have been generated on cosmid and small-insert cDNA libraries (Smulian *et al.* 2001). These end sequences are called sequence-tagged connectors (STCs). Seed clones are chosen and sequenced by a shotgun sequencing approach. The resulting cosmid sequence is BLASTed against all STCs, and a cosmid with the least overlap with the sequenced cosmid is selected for sequencing. In this approach, sequence extension is achieved with walks initiated from the seed clones. The progress completes when no more STCs are found to extend the walks. By reiterating this "sequence-then-map" by computer analysis against the STC database strategy, a minimum tiling path of clones can be sequenced at a rate that is primarily limited by the sequencing throughput of individual genome centers. Because the STC resource permits the easy integration of genetic, physical, and sequence maps for chromosomes, it has been a powerful tool for the initial analysis of the human genome and other complex genomes (International Human Genome Sequencing Consortium 2001; Venter *et al.* 2001).

Currently, several computational tools are available for integrating information within genome projects to complement the integrative experimental approach of mapping by sequencing. They are used as (1) a supporting tool to the genome database (Cuticchia 1994; Talbot and Cuticchia 1998), (2) a web-based genome map-browsing tool (Thomas *et al.* 1995), or (3) a general purpose stand-alone visualization system (Durbin and Thierry-Mieg 1994).

Our ODS3 software is a genome visualization tool for fungal genome projects (Bennett and Arnold 2001) and has been developed from its earlier versions, ODS (Cuticchia *et al.* 1993) and ODS2 (Hall *et al.* 2001). The ODS3 tool is a comprehensive software package designed to provide unique functions to construct an integrated genome map at multiple levels and in different representations. New features of ODS3 are: (1) support for integrated genome mapping and sequencing, (2) new algorithms for map building and map integration, (3) hierarchical views for a single map domain, (4) support for mapping and sequencing access with web browsers, (5) support for a mapping-by-sequencing strategy, (6) object-oriented design, fully coded in Java to promote platform independence, (7) extensible markup language (XML)-based mapping data files with document type definitions (DTDs) for easy data distribution and manipulation across genome projects, (8) additional hybridization matrix view of the physical map, and (9) support for GIF file graphics export.

In this "sequence-then-map" approach, the genome map can be mapped and scanned with the use of STCs. The software ODS3 supports an implementation of the mapping-by-sequencing approach and was developed as a Java-based genome mapping visualization system for the Fungal Genome Database (FGDB) (Kochut *et al.* 1993, 2003). The ODS3 tool can be used in three modes. The tool can extract information from the FGDB and generate integrated genome maps at multiple levels (chromosome, contig, or sequence) using automated mapping methods and advanced computer technologies, either locally or on the server. Alternatively, the tool can be used to save the retrieved information locally in XML format and operate on the data in a stand-alone capacity independently of FGDB. The ODS3 tool optimizes the experiments in mapping by sequencing by providing estimates of overlap between clones and will enhance the efficiency and cost effectiveness of genome projects by providing an easily used web-accessible tool for rapid assembly of integrated genome maps. The data in the project are stored in an XML format that will facilitate its integration with other genome projects. Finally, all of these features will dramatically improve the speed of fungal genome mapping by automating a time-consuming step of creating integrated genome maps.

The organization of this article is as follows: in methods we describe the methods for construction of an initial physical map of the *Pneumocystis carinii* genome using a dual strategy of mapping by sequencing and hybridization-based physical mapping, the algorithms used to reconstruct the native order of clones on a chromosome, and the design and implementation of ODS3 to carry out an integrated genome-mapping strategy. Under results we describe an integrated genomic map of *P. carinii* created using the new tool ODS3. In the discussion we frame findings about the Pneumocystis Genome Project and suggest some limitations and needed

extensions to ODS3 and other integrated genome-mapping tools.

METHODS

Libraries: *P. carinii*, karyotype form 1 (Cushion *et al.* 1993), obtained from the lungs of individual immunosuppressed rats during fulminate infection, were used to make the cDNA and cosmid libraries. Library constructions were described in Smulian *et al.* (2001) and at http://pneumocystis.uc.edu. The cDNA library is in pSKII and available from the American Type Culture Collection (ATCC) at http://www.atcc.org. For the cosmid library 26 plates are in pWEB (Epicentre Technologies, Madison, WI), and 9 plates are in pLorist6Xh (Kelkar *et al.* 2001) for a total of 35 plates.

DNA preparations: Cosmid DNAs were isolated from overnight cultures grown in Luria broth (LB) media-carbenicillin (50 μg/ml), using the Concert Rapid Plasmid Miniprep system (GIBCO BRL, Gaithersburg, MD).

Probe preparation: Totals of 10 μl of plasmid minipreparations and 5 μl of random oligo primers from the Prime-it II kit (Stratagene, La Jolla, CA) were heated in 100° water for 5 min. Then 5 μl of 5× dCTP buffer, 4 μl ^{33}P label (>10^8 cpm), and 1 μl of Exo(−) Klenow (50 units/μl) were added to the mixture. The solution was incubated at 37° for 1 hr to create a ^{33}P-labeled probe. To stop the reaction, 25 μl of 0.25 M EDTA was added to the probe solution.

DNA hybridization: cDNA libraries were stamped onto nylon membranes (Hybond-XL; Amersham Pharmacia Biotech) in an 8 × 8 array of 55 microtiter plates per membrane with a *Bio*Grid high-density stamping robot (*Bio*Robotics). Inocula were allowed to grow overnight at 37° on LB agar plates containing carbenicillin (50 μg/ml) for the cDNA library. Membranes were treated to lyse the colonies with: (1) 10% SDS for 5 min; (2) 1.5 M NaCl, 0.5 M NaOH for 5 min; (3) 0.5 M Tris-HCl, 1.5 M NaCl, pH 7.2, for 5 min; and (4) 0.3 M NaCl, 0.3 M Na citrate, pH 7.0 (2× SSC), for 5 min. The treated membranes were air dried for 30 min, and the DNA was crosslinked to the nylon membrane by UV radiation (3–12 min) with a UV crosslinker (Stratagene) or alternatively by baking at 80° for 2 hr.

One membrane representing the complete cDNA library was prehybridized at 65° for 2 hr with 9 ml of modified hybridization buffer containing casein hydrolysate instead of bovine serum albumin (Church and Gilbert 1984) in a hybridization oven (Hybaid). After prehybridization, a mixture of 3 ml of Church buffer and 50 μl of radioactive probe (>10^8 cpm) was added to 9 ml of fresh hybridization buffer. The membranes were hybridized with a cosmid probe for 12–18 hr at 65°. The membranes were then washed twice in 2× SSC, 0.1% SDS at 65° for 20 min. Two subsequent washes were done in 0.5× SSC at 65° for 20 min. Membranes were then removed, blotted dry, and imaged by direct β-emission counting on a Packard (Meriden, CT) instant imager. They were also exposed to X-ray film (BIOMAX MR) at −80° with intensifying screens.

DNA sequencing: Cosmids were sequenced using a shotgun subcloning method (Roe *et al.* 1996). A starting culture of 20 ml LB-carbenicillin (50 ug/ml) containing the cosmid was grown at 37° overnight (300 rpm) in 50-ml conical tubes. A 1-ml seed from the starting culture was grown in a 500-ml Ehrlenmeyer flask containing 250 ml LB media-carbenicillin (50 μg/ml) under the same conditions. At least 75 μg of cosmid DNA was recovered using a QIAGEN (Valencia, CA) large construct kit for each cosmid, according to manufacturer directions. The recovered DNA was nebulized (physically sheared) in 500 μl of glycerol and 200 μl of 10× TM buffer, using an IPI Medical Products nebulizer 4207 or Salter Labs REF 8901 nebulizer at 10 psi for 2.5 min in a −20° ethanol/salt/water bath. The nebulized DNA was ethanol precipitated first with a 100% ethanol wash followed by a 70% ethanol wash. Next, DNA was end repaired with Klenow DNA polymerase, phosphorylated with T4 polynucleotide kinase, and size selected using electrophoresis for 1 hr on a 30-ml low melting point agarose gel (0.7%) in 1× TAE (2 wells each loaded with 15 μl of DNA for each cosmid). The band fragment of 1.0–2.0 kb was gel isolated with a High Pure PCR kit (Boehringer-Mannheim, Indianapolis) with changes from manufacturer directions (Vargas 2002) and ligated into Ready-to-Go *Sma*I-digested pUC18 (Amersham, Arlington Heights, IL) for subcloning. Electrocompetent *Escherichia coli* XL1-Blue (Stratagene) cells were transformed with the ligated DNA and plated on LB-carbenicillin (25 μg/ml) agar plates incorporated with isopropyl thiogalactoside (25 mg/ml) and X-GAL (20 mg/ml). After incubation plates were stored at 4° for at least 1 hr. White colonies (196) were robotically picked with a *Bio*Pick (*Bio*Robotics) and inoculated into two 96-well microtiter plates. This sublibrary was replicated into deep well blocks (Marsh Biomedical, Rochester, NY) containing 1.2 ml of LB-carbenicillin (50 μg/ml) media per well and incubated overnight at 37° on a rotary shaker at 250 rpm.

DNA sequencing templates were generated using a Robbins Scientific hydra-based double-stranded DNA-sequencing template isolation procedure (Roe *et al.* 1996). Fifty cycles of *Taq*-polymerized cycle sequencing were performed using BigDye terminator chemistry according to the manufacturer's specification (PE Applied Biosystems, Foster City, CA). Unincorporated dye terminator was removed with ethanol precipitation followed by isopropanol precipitation, and gel electrophoresis was performed on a 96-lane ABI 3700 sequenator.

Cosmid end sequencing: Cosmid end sequencing was challenging because the *E. coli* strain included with the pWEB vector (Epicentre Technologies) was an *endA1* strain (Schoenfeld *et al.* 1995). Sequencing of pWEB ends followed the same protocol as sequencing the

pUC18 clones with two exceptions. At the neutralization step in the PE Applied Biosystems sequencing protocol the solution of lysed, neutralized cells was transferred to −80° for 30 min instead of −20° for 30 min. The number of *Taq*-polymerized cycles was reduced from 50 to 25 cycles to give *endA1* product less time to degrade cosmid DNA.

Automated analysis of all sequencing data: Sequencing data were processed using an automated workflow (Hall *et al.* 2003). Trace files were processed and sequences were assembled by the software Phred/Crossmatch/Phrap running on a Unix server (Ewing and Green 1998; Ewing *et al.* 1998). BLASTX and BLASTN searches were done in 600-bp sliding windows (with an overlap of 60 bp) along the cosmids against GenPept and GenBank (Altschul *et al.* 1990) between 8/02/00 and 4/04/01. Contigs with an *E*-value of *E*-40 or more significance from a BLASTN search against the *E. coli* genome or vector were removed, and then the remaining sequences were reassembled in the workflow.

Tree building: The 186 expressed sequence tags (ESTs) of the HSP-70 genes from EST sequencing can be found at http://www.uky.edu/Pc/ and were used to build a gene genealogy. Accessions aab58248 and aad00455 were used in a Framesearch (Accelrys) against all 186 ESTs. The resulting inferred polypeptides were aligned with PILEUP (Accelrys) and arranged in an UPGMA tree. One representative from each of 11 clades was selected to build a consensus parsimony tree with PAUP (Accelrys).

To cross-validate this analysis, a separate analysis was performed. The ESTs were binned into clusters with the program Fragment Assembly System (Accelrys). The resulting consensus sequence of each bin was translated and aligned with PILEUP (Accelrys). The resulting alignment was used to build a consensus parsimony tree with PAUP (Accelrys).

Physical mapping algorithms: Cuticchia *et al.* (1992) and Mott *et al.* (1993) formulate the probe-ordering problem as the solution to a traveling salesman problem (with the distance between probes measured by a Hamming distance) and solve this optimization problem by simulated annealing. Xiong *et al.* (1996) have shown this method to be consistent. The data on which the choice of probe order is based are the clone/probe hybridization data, which are summarized in a matrix A. A "1" in this matrix indicates hybridization or sequence overlap between a clone and probe; a "0" indicates no hybridization or sequence overlap. For a given hybridization matrix, the physical mapping problem can be simplified to finding a permutation of the probes such that the reordered matrix has the consecutive ones property (*i.e.*, every clone has at most one block of consecutive overlapping probes). The matrix entry $A_{i,j}$ is 1 if the ith clone is believed to contain (overlap) the jth probe on the basis of hybridization results; otherwise, $A_{i,j}$ is 0. Unfortunately, hybridization-based physical mapping is influenced by errors and ambiguities. There can be false-positive and false-negative hybridization signals and inconsistent hybridization signals caused by repetitive sequences, chimeric clones, or clones containing deletions. Also sequence-based physical mapping is influenced by errors and ambiguities. There can be false positives due to repeats and contaminating DNA.

The algorithm used by ODS2 seeks to minimize the sum of the Hamming distances between adjacent probes with respect to probe order by simulated annealing as described by Hall *et al.* (2001) in ODS2. This objective function is referred to as the total linking distance. The following modifications were made over the original algorithm to enhance the computing performance of ODS3.

1. Calculate the Hamming distances between probes, order the probes, and then fit the clones to the probe order. We write $P = \{p_1, p_2, \ldots, p_m\}$ for the set of m probes, $C = \{c_1, c_2, \ldots, c_n\}$ for the set of n clones, and $P^\pi = \{p_1^\pi, p_2^\pi, \ldots, p_m^\pi\}$ for an ordering (or permutation) of probes. Let $D = |C| \times |P|$ denote a binary matrix, where $D_{i,j}$ is 1 if clone c_i overlaps probe p_j on the basis of the experimental data; otherwise it is 0:

$$D_{i,j} = \begin{cases} 1, & \text{if clone } c_i \text{ has hits of probe } p_j \\ 0, & \text{otherwise.} \end{cases}$$

   The Hamming distance objective function is given by $F(P^\pi) = \sum_{i=1}^{m-1} \sum_{j=1}^{n} D_{j,i}^\pi \oplus D_{j,i+1}^\pi$, where D^π is a matrix derived from D by permuting the columns to the corresponding probe ordering P^π, and $\oplus$ is the Boolean exclusive or operation. After ordering the probes, the longest existing contiguous sequence of probes that hybridized with the given clone is found. The clone is placed in the map such that it spans ordered probes. If more than two such pairs of places are found, the clone is randomly placed in the map in one of the possible positions. If a clone is without hybridization to any probe, it cannot be placed in the map. For those clones that have an ambiguous position in the map, we can adjust or correct their position manually. This modification of the algorithm can decrease the run time of the calculation when the probe is a subset of the set of clones.
2. Use a microcanonical annealing search algorithm (Creutz 1983) instead of simulated annealing for ordering clones. Microcanonical annealing was found to achieve levels of optimization as good as simulated annealing and to do so an order of magnitude faster (Bhandarkar and Machaka 1997).
3. A weighted penalty value ϕ related to the number of misplaced anchored clones is added to the sum of Hamming distances $F(P^\pi)$. When computing a given permutation of probes by the algorithm, the subset of clones that are anchored to the genetic map is placed within the probe ordering. We write $A^\pi =$

$\langle a_1^\pi, a_2^\pi, \ldots, a_k^\pi \rangle$ for a permutation of anchored clones. Let $\text{pos}(a_i^\pi)$ denote the function that returns the position of a marker in the genetic map, where a_i^k is the permutation anchored to that marker. The penalty value is calculated by $\phi = \alpha \Sigma_{i=0}^{k-1} \text{pos}(a_i^\pi) - \text{pos}(a_{i+1}^\pi)$, where α is a scaling factor variable that can be set by the user when creating the genetic map. A higher value of α places more emphasis on genetic map data. This penalty is minimum for a given α when the order of markers implied by A^π is the same as the order of markers in the genetic map.

Implementation of ODS3: The ODS3 tool available at http://gene.genetics.uga.edu/pub is implemented in Java and allows access to the University of Georgia FGDB (Kochut *et al.* 1993) to provide capabilities for generating genome maps and for data mining. The ODS3 tool can be used either as a stand-alone application or as a web-based mapping viewer of data in the FGDB. The Java code for ODS3 has been successfully executed on Sun Solaris 7, Linux, and Windows 98/NT/2000 platforms. A more detailed description of ODS3 is given on the Genetics web site at http://www.genetics.org as supplementary data.

Design of the FGDB: The FGDB (Kochut *et al.* 1993, 2003) is implemented using the Oracle 8*i* Enterprise edition, which is an object-relational database management system (DBMS). With the FGDB, the ODS3 can generate various genomic maps at different levels (chromosome, contig, or sequence). The design of the FGDB schema is given in a UML diagram in Kochut *et al.* (2003) and in the supplementary data at http://www.genetics.org.

Supporting the HTBLAST workflow application: A critical requirement for a large genome laboratory is software to control laboratory workflow while managing the data produced in the laboratory (Rozen *et al.* 1995; Goodman *et al.* 1998). Workflow applications automate the execution of workflows. A workflow application consists of a network of different individual tasks that are performed by a human being or machine (*e.g.*, a computer; Miller *et al.* 1998). The tasks are carried out on data objects that "flow" through the system. The tasks may be existing legacy applications, which may or may not require interaction with the database (Hall *et al.* 2003), or a person carrying out a specialized experimental task like a step in sequencing. As a supporting tool for ODS3, a high-throughput BLAST (HTBLAST) workflow application was developed in Perl to generate and parse BLAST reports on sequences in the project for gene identification.

In our setting a collection of query sequences are to be BLASTed. The BLAST tool helps locate hot spots by dividing a query sequence into all possible subsequences of a given length, which depend on the type of the subsequence involved (Altschul *et al.* 1990). Silicon Graphics Inc. (SGI) undertook a project to parallelize BLAST with HTBLAST (Camp *et al.* 1998) to speed up the search process by making more efficient use of larger numbers of processors. The architecture for BLAST and HTBLAST can be found at http://www.sgi.com/solutions/sciences/chembio/resources/.

The prototype of this workflow is based on a series of tasks for generating an HTBLAST search for sequences. The first task loads sequences that have not been searched by HTBLAST from the database. After the creation of the input file for HTBLAST, the sequence file (describing the collection of query sequences) is sent to an SGI Origin 2000 computer, which has 24, 300-MHz MIPS R12000 processors with 4 MB cache memory and 8 GB of system memory. The HTBLAST search is remotely invoked at a particular time by a local machine. When the execution of HTBLAST is finished, the results file is sent back to the database server via ftp. At this point, ODS3 has modules to parse the HTBLAST report and to update sequence data in the database (*i.e.*, FGDB) with the new BLAST information.

ODS data files: There is a need to develop a standardized but flexible format for representing different types of biological data (Robbins 1996). The XML is a way of sharing, exchanging, and organizing the vast amount of genetic data cooperatively. The XML is a meta-language to produce documents that convey content with semantic structure (Elenko and Reinertsen 2000) and is widely used for data presentation. An XML-based genomic data file from one system can be restructured and presented to another system through an associated DTD file without any change. To take advantage of XML files and to distribute our mapping and sequencing data more efficiently and seamlessly among the fungal genome community, the ODS3 package now supports several classes of input/output operation. For example, ODS3 can either generate the mapping and sequencing data for storage in FGDB or export the results of ODS3 as an XML file.

RESULTS

Results in all figures and tables including Pneumocystis genome data (with the exception of the gene genealogy in Figure 3) were generated with ODS3. The tool was and is used to (1) estimate overlap between cosmids to guide the selection of cosmid probes, (2) generate an integrated map from the available sequence and mapping data as they are collected, (3) calculate statistics about the map, (4) examine the physical map for repeats, and (5) correlate features of the map with Pneumocystis sequence. In mapping by sequencing ODS3 estimates an overlap between each Pneumocystis cosmid clone with every other cosmid with an STC to guide the selection of the next cosmid for sequencing.

Coverage of the genome: A total of 5280 *P. carinii* cDNA clones contain ~2000 distinct genes, listed at http://gene.genetics.uga.edu, being used to link slightly

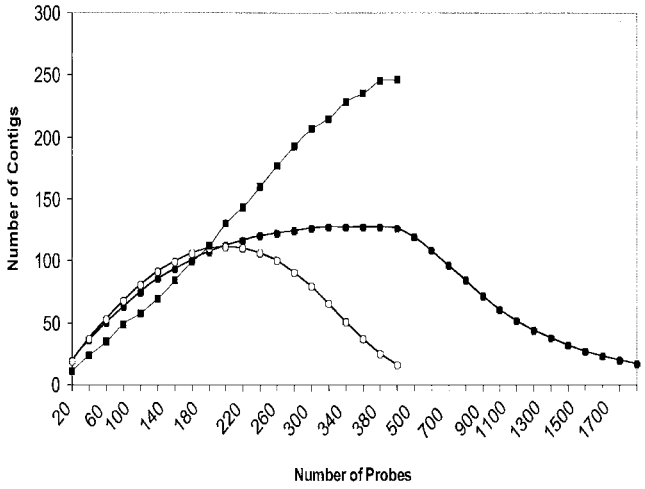

FIGURE 1.—The chart for contig assembly dynamics. The predicted numbers of contigs for sampling with (●) and without replacements (○) are compared with the observed values (■) during genome reconstruction (ZHANG and MARR 1993). Each observed value is the average of 15 map assemblies with 15 distinct random-number seeds.

$< \sim$2500 cosmids (SMULIAN *et al.* 2001). A total of 3720 cosmid end sequences were generated and made available at the same web site with an average read length of 221 bp, implying an STC approximately every 2 kbp (7700 kbp/3720 cosmid ends). A sampling-with-replacement strategy is being followed by linking cosmids with respect to shared ESTs and cosmid sequences (KELKAR *et al.* 2001). While this strategy is slower than sampling without replacement (PRADE *et al.* 1997), using the ESTs or shared cosmid sequences as the linking information sidesteps the problem of the rat host contaminating DNA (estimated to be ~10% in the cosmid library and 14% in the cDNA library; see http://gene.genetics.uga.edu for BLAST reports). The linking EST and cosmid sequences in the physical map can be classified as fungal or rat in origin by BLAST searches of these sequences against public databases generated by the workflow in ODS3 (ALTSCHUL *et al.* 1990).

A chart of the current contig assembly dynamics for the whole genome as 384 cosmid probes are added into the physical map is shown in Figure 1. The project is expected to be complete by 800 probings. Cosmid probes were classified as either sequence probes (113 in number) if they were sequenced or as hybridization probes (271 in number) if they were hybridized to the arrayed cDNA library. The physical map currently contains 1045 cDNAs or $100 \times (1045/5280) = 21\%$ of the redundant cDNA library. The dynamics would indicate that we are about halfway through completing the physical map, but there are significantly more contigs than expected as explained in the next subsection (ZHANG and MARR 1993). The observed number of contigs can be reduced ~18% by manual editing. For example, if the maps with 344 or 384 probes are manually edited after automated assembly, then the contig numbers are reduced to 188 and 203, respectively. This number of 203 contigs is still too high. Theoretically, the total amount of generated contigs for 500 probings should be 119 (Figure 1).

As an independent control on the whole-genome assembly, >14 clones were assigned by pulsed-field gel electrophoresis (PFGE) Southern to chromosome 7. These 14 anchors were then used to retrieve all other clones linked by hybridization or sequence to the anchors for an independent assembly. Six contigs were assembled from 250 clones assigned to chromosome 7, and eight fragments of these six contigs from the independent assembly could be found in the whole-genome assembly. A total of 63 of the clones on chromosome 7 were removed by the filtering in the genome-wide assembly. The conclusion is that acquiring a small number of anchors (~15) per chromosome is very useful in validating the genome-wide assembly.

Examining chromosome 7 also allowed an assessment of coverage. Chromosome 7 is measured by PFGE at 500 kbp (CUSHION *et al.* 1993) and is thought to be a doublet because of staining intensity. A total of six contigs have been generated with ~53 nonredundant ESTs assigned currently to the map or ~1 EST assigned per 20 kbp. There were 48 cosmids in the physical map of chromosome 7, and 104 cosmids were sized by *Bam*HI restriction digests, yielding an estimated insert size of 26 kbp (VARGAS 2002). The proportion of shared sequence between 25 pairs of sequenced cosmids provided an estimate of overlap of 7 kbp or 27% overlap. This estimate is consistent with estimated overlaps in the *Aspergillus nidulans* map (PRADE *et al.* 1997). Links that involved a pair with one or more unsequenced cosmids were conservatively estimated to overlap by 90%. The resulting doublet, designated chromosome 7, was estimated to be 905 kbp in size as compared with the PFGE-sized doublet of $\sim 500 \times 2$ kbp. Subsequently, C. P. VIVARES (unpublished results) confirmed that chromosome 7 is a doublet by two-dimensional PFGE. As a consequence, the completeness of the map with respect to this mid-sized chromosome is estimated to be 90% ($100 \times 905/1000$). The larger chromosomes are likely to have smaller coverage.

Filtering the clone-probe hybridization matrix: Reconstructing an integrated map with the physical mapping algorithm described under METHODS depends on the clone-probe hybridization matrix generated from physical mapping data and sequence data (including the STCs, cDNAs, and genomic sequence). The physical mapping algorithm utilizes inferred overlaps detected on the basis of DNA/DNA hybridization and sequence similarity from *all* available sequence data. Not all of the inferred overlaps, such as those overlaps based on sequences representing *E. coli* contamination, should be relied on. The filtering rule is quite conservative; any sequence contig containing a significant hit (*E*-5 or

TABLE 1

Data sets used to test the integrated mapping tool ODS3

Data group	Data set	Description of the data set	Threshold value
I	1	*E. coli* sequences removed	N/A
	2	*E. coli* sequences removed	*E*-6
	3	*E. coli* sequences removed	*E*-10
	4	*E. coli* sequences removed	*E*-30
II	5	*E. coli* and vector sequences removed	N/A
	6	*E. coli* and vector sequences removed	*E*-6
	7	*E. coli* and vector sequences removed	*E*-10
	8	*E. coli* and vector sequences removed	*E*-30
III	9	*E. coli*, vector, Pseudomonas, and Adenovirus sequences removed	N/A
	10	*E. coli*, vector, Pseudomonas, and Adenovirus sequences removed	*E*-6
	11	*E. coli*, vector, Pseudomonas, and Adenovirus sequences removed	*E*-10
	12	*E. coli*, vector, Pseudomonas, and Adenovirus sequences removed	*E*-30

smaller) with the filter's keyword is removed entirely. For example, on average 9% of the sequence per cosmid was removed as *E. coli* or vector with *E*-40 (see METHODS). To evaluate the map of the entire genome with respect to filtering, 12 filtered data sets were generated and classified into three groups (Table 1) according to removed keyword(s). Each data set contains 384 probes and was prefiltered for the known *E. coli* and vector contaminants with an *E*-40 as described in METHODS.

For evaluation purposes, we used two different methods of filtering the clone/probe hybridization matrix that are based on BLAST reports to discover clone-probe overlaps from sequence data (see Figure 2). Each data set was run 15 times using a scaling factor $\alpha = 1.0$ in the physical mapping algorithm (see METHODS). In the first probe-based approach, we carried out the following operations:

1. Determine clone/probe hybridization from DNA/DNA hybridization data and compute "sequence hits." Sequence probes are defined as all cosmids that were sequenced by shotgun subcloning (see METHODS). A sequence hit between a sequence probe and clone is computed with a BLASTN search against *all* sequence from cosmids (including cosmids with only STCs). An overlap is declared if the *E*-value on the hit is below that chosen by the researcher (*i.e.*, *E*-6 in this article).
2. Filter each probe and its overlapping clone(s). Another BLASTN and BLASTX search of all sequence associated with each probe and each clone that hit a probe was performed against GenBank and GenPept within an automated workflow (see METHODS). Those contigs with the selected *E*-value and keyword(s) in the second set of BLAST reports were filtered out.

The results of this probe-based approach are reported in Table 2.

In the second clone-based approach, we carried out the following operations:

1. Filter the sequence of every clone in FGDB. A BLASTN and BLASTX search was performed for all sequence with an automated workflow (see METHODS). Those contigs with the selected *E*-value and keyword(s) were filtered out.
2. Determine clone/probe hybridization matrix as in step 1 of the probe-based approach.

The results of this clone-based approach are reported in Table 3.

The two approaches can differ in the resulting filtered clone/probe hybridization matrix because every clone has end sequence. When Tables 2 and 3 are compared, we find that there are fewer clone-probe hits and clones in the map with the second approach. This method was more effective in eliminating problematic sequences in the Pneumocystis genome project such as particular clones that are enriched for rat sequences, but took on average 20% longer to execute (during ordering of clones and probes with the filtered matrix). The running time for each data set in ODS3 includes the time used to construct the overlap matrix and to compute with the ordering algorithm (see METHODS). In general, the run time is reasonably short, and an average running time of 252 sec was obtained for each of the 12 data sets using the second approach. As a consequence, the clone-based approach was selected and used in all other analyses in this article.

As the stringency of the filter is increased, we expect that removing contaminating DNA will reduce false joins (and possibly weak true joins) and hence increase the number of contigs and reduce the average size of contigs. From Table 2, we can see that the number of contigs is increased and that the average size of contigs is decreased when the stringency of the filter (*i.e.*, the

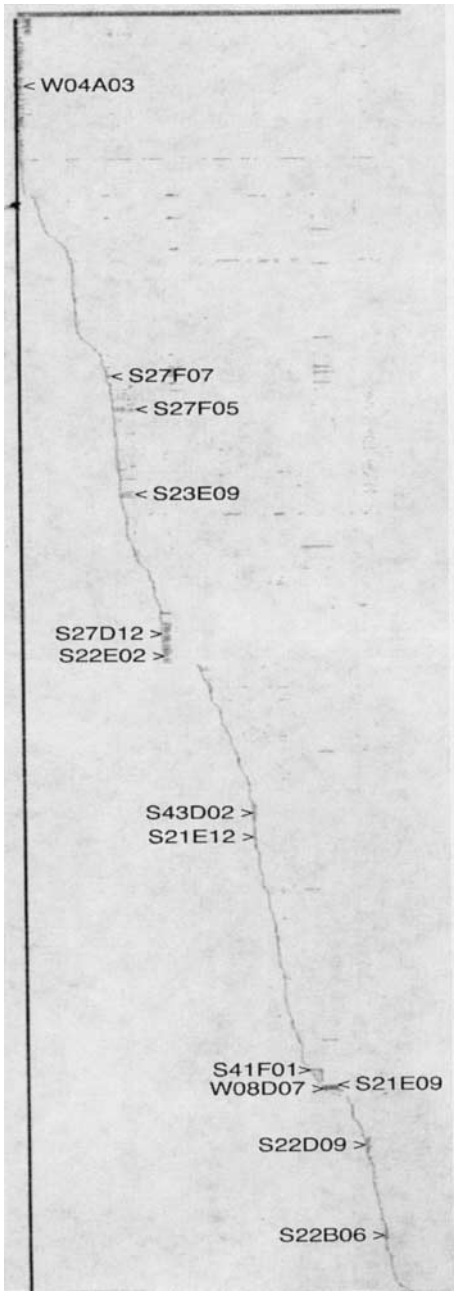

FIGURE 2.—The Pneumocystis physical map. Each putative gene family is indexed by a clone number from Table 4. The location of MSG is indicated by W04A03. HSP70 families are located at S21E09, S21E12, and S22D09 (BiP). No tandem array of the rDNA family is visible on the physical map, which further supports its single-copy nature in Pneumocystis.

filter threshold is decreased) applied to the BLAST search is increased, as expected. The average contig number of these 12 data sets in Table 3 is ~273 (and in Table 2, 267 contigs). The primary pWEB library used to reconstruct the *P. carinii* genome contained ~2500 clones (only ~2200 clones grew) with an average insert size of 26 kb. The estimated genome size is 7.7 Mb (CUSHION *et al.* 1993) by PFGE. The observed and predicted numbers of anchored contigs (hybridized at least once to a probe) as a function of nonoverlapping probes during the mapping experiment were not in agreement with those predicted under sampling with replacement, in which nonoverlapping probes are chosen randomly (Figure 1). The filtering techniques currently in place generate about twice as many contigs as expected. The observed number of contigs can be reduced by ~18% after manual curation in ODS3 at probings 344 and 384. An additional explanation is required for the departure of 203 contigs from the expected of 119 contigs in Figure 1.

Another explanation is that true "weak links" are being eliminated by filtering. The evidence for that explanation is that by manual review of the hit matrix for the doublets labeled 7 and 10/11 (CUSHION *et al.* 1993) and by assembling only those clones with an assignment to chromosome 7 or 10/11 a total of 14 contigs were generated, which totals ~1.88 Mb. By extrapolation we would expect 57 contigs for the whole 7.7-Mb genome, which is closer to the expected 119 contigs in Figure 1.

It is natural then to ask what clones are generating the discrepancy. As an example, after filtering these clones with multiple hybridization signals, the average size of contigs decreased on average from 23.2 (data sets 1, 5, and 9) to 7.1 clones (data sets 4, 8, and 12). Most of the hybridizations eliminated by the filtering (see Tables 2 or 3) occur in clones with three or more hybridization signals (61%), *i.e.*, clones with ambiguous placement on the physical map. These additional hybridization signals will lead to an ambiguous position assignment for the clone with these multiple hybridization signals.

We also evaluated the filtering technique on chromosome 7 of *P. carinii*. The filtered matrix was manually curated by BLASTing each cosmid probe and its associated sequences individually and was performed by Dr. George Smulian at the University of Cincinnati. The results were compared with a filtered and reordered matrix (generated by ODS3), in which sequences with BLAST reports containing the keyword *coli*, vector, Pseudomonas, or Adenovirus were removed, and BLAST hit results were filtered by a threshold value *E*-6 (*i.e.*, BLAST hits with an *E*-value less than or equal to *E*-6 were picked). The two hybridization matrices were in substantial agreement. All clones examined either were linked to the same probe or were positioned discordantly because of hybridization to three or more probes (*i.e.*, implying ambiguous placement on the map). It can be seen that the manually prepared map has 6 contigs as opposed to the automatically generated map with 31 contigs.

Multigene families in the physical map: One of the advantages of the hybridization matrix as a representation of the physical map is the straightforward identification of repeat families (KELKAR *et al.* 2001). Putative families are listed in Table 4 and were identified from off-diagonal hits in Figure 2 generated by ODS3 and by

TABLE 2

BLAST filtering on probes

Data group	Data set	No. of clones	No. of contigs	Avg. size of contigs	No. of hits			Avg. running time (sec)
					Sequence	Hybridization	Seq/Hyb	
I	1	6,742	260.3	25.9	52,176	2,940	223	309.8
	2	2,531	265.9	9.6	7,550	2,969	194	215.4
	3	2,484	266.1	9.3	7,197	2,970	193	207.6
	4	1,961	270.3	7.3	3,197	2,981	182	197.3
II	5	5,756	262.2	21.9	44,787	2,949	214	287.6
	6	2,017	266.3	7.6	3,682	2,978	185	202.3
	7	1,969	268.5	7.3	3,395	2,979	184	182.2
	8	1,850	270.3	6.8	2,571	2,987	176	172.4
III	9	5,761	266.5	21.6	43,963	2,951	212	252.9
	10	2,013	269	7.5	2,794	2,980	183	191.3
	11	1,966	270.2	7.3	2,503	2,981	182	167.4
	12	1,847	270.9	6.8	1,675	2,989	174	150.9

Each map statistic is the average of 15 map assemblies (initiated with 15 distinct random-number seeds) after BLAST filtering on probes.

clicking on the entries to yield the associated BLAST reports generated by the HTBLAST workflow application. One of the most prominent families of repeats found on the ends of all Pneumocystis chromosomes is the ~100 members of the Pneumocystis major surface glycoprotein (MSG) family (Stringer 1996). The location of the MSG family can be easily seen in Figure 2 at the location containing cosmid W04A03 near the top of the physical map. In the 10× pWEB library of 2496 clones, we expect 250 clones to contain MSG sequences. Under seven distinct DNA/DNA hybridization experiments, 300 cosmid clones were positive at least once to MSG probes, and 130 cosmids were positive at least twice. Among the 130 strong positives, 40 of these cosmids hybridized both to a conserved recombination junction element (CRJE) and to a conserved 3′ sequence common to all MSGs. The representation of MSG in the library then falls somewhere between 5% ($100 \times 130/2500$) and 17% ($100 \times 430/2500$). As a consequence of the high similarity of MSG family members the tool ODS3 merges all of the MSG sequences. Embedded unique sequences in fully sequenced MSG-containing cosmids will be necessary to resolve the true chromosomal locations of family members.

There are also examples of embedded rat sequence needing removal in the physical map. Examples include the region around cDNA S03F07 and cosmid W08D07. These regions contain sequence similar to human and

TABLE 3

BLAST filtering on clones

Data group	Data set	No. of clones	No. of contigs	Avg. size of contigs	No. of hits			Avg. running time (sec)
					Sequence	Hybridization	Seq/Hyb	
I	1	7,012	266.4	26.3	49,812	2,769	211	321.7
	2	2,632	270.9	9.7	7,202	2,743	237	288.3
	3	2,512	273.2	9.2	6,590	2,761	219	250.5
	4	2,005	275.2	7.3	2,988	2,734	246	203.3
II	5	5,987	270.8	22.1	41,912	2,771	209	309.7
	6	2,212	271.3	8.2	3,200	2,776	204	270.1
	7	2,034	274.8	7.4	2,782	2,726	254	255.3
	8	1,927	277.2	7	2,300	2,762	218	190.4
III	9	5,792	271.2	21.3	39,972	2,773	207	289.4
	10	2,109	272.8	7.7	2,345	2,759	221	240.5
	11	2,054	276.3	7.4	2,134	2,737	243	210.4
	12	1,992	279.6	7.1	1,603	2,749	231	192.3

Each map statistic is the average of 15 map assemblies (initiated with 15 distinct random-number seeds) after BLAST filtering on clones.

TABLE 4

Ten putative gene families are identified from the physical map in Figure 2

Description of putative gene family	Clone landmark of putative repeat family
MSG	W04A03
Rat sequence (homologous to mouse genomic sequence)	S03F07
Unknown	S27F05
Unknown	S23E09
Heat-shock ψ protein	S27D12
Unknown	S22E02
Unknown	S43D02
HSP60	S41F01
Rat tandem array of globins	W08D07
HSP70 and HSP70 (BiP)	S22D09 (BiP), S21E09, and S21E12
tRNA synthetase family (*i.e.*, MES1)	S50F05
Unknown	S22B06

The unknown families show no homology to known vector, *E. coli*, or mammalian sequences and are thus inferred to be Pneumocystis sequences.

rat sequence found by examining BLAST reports stored in the FGDB (from the HTBLAST workflow) and associated with the clone/probe hybridization matrix viewed with ODS3 (Altschul *et al.* 1990).

Several other families were detected including genes encoding homologs to HSP70, ψ heat-shock protein in *Schizosaccharomyces pombe*, HSP60 (Silver and Way 1993), and tRNA synthetase (S50F05; Senger *et al.* 2001). In the case of the HSP70 family there are at least two forms (SA1 and SB1) in the cytoplasm (Stedman *et al.* 1998) with 73% sequence identity at the amino acid level and one form in the endoplasmic reticulum with 62% sequence identity at the amino acid level to the cytoplasmic forms [HSP70 (BiP); Stedman and Buck 1996]. All three forms localize to distinct chromosomes by PFGE (Stedman *et al.* 1998). Only one cytoplasmic form (SA1) has been detected in the physical map at cosmid W01D12 to date (accession U80967) by BLASTN search of the EST collection. Also the HSP70 member (BiP) in the endoplasmic reticulum has been mapped here as well to a distinct region of the physical map (cosmid W04C09). There was a third region of the map (W05B04) with a hit to S21D05 (with a cytoplasmic HSP-70 EST), but this is extremely tentative. The ψ protein family (and its homolog in *Saccharomyces cerevisiae*, *SIS1)* are thought to lend specificity to the function of the HSP70 family by means of a DnaJ motif shared by family members. The ψ protein family is also thought to regulate the HSP70 family members. There are also five additional putative repeat families as yet uncharacterized (see Table 4).

One of the striking features of Pneumocystis is the hypothesis that there is only one rDNA; *i.e.*, the rDNA is not a multigene family in this organism (Giuntoli *et al.* 1994). A cosmid in the physical map was identified with the 16S, 5.8S, and 26S rDNA genes (W03C05) on chromosome 7. Contigs 13, 19, and 21 were 99% identical (*E*-value $<$ *E*-180) with sequence of the rDNA genes of *P. carinii* (in accession PCM26S58R or PC16SRR1). A total of 47,986 bp of cosmid W03C05 were sequenced and found to have no other rDNA sequences other than those in contigs 13, 19, and 21 containing the entire set of rDNA genes. The sequence assembly was no more than five sequence reads deep for contigs 13, 19, and 21, indicating little evidence of excessive stacking due to possible incorrect assembly of a tandem array of rDNA genes. No adjacent cosmids in the physical map appeared to contain rDNA sequence on the basis of hybridization, and there was no indication from Figure 2 that the rDNA sequence was tandemly arrayed. The insert of W03C05 was liberated by a *Not*I digest, gel isolated, and hybridized to a cosmid array in quadruplicate. There were no more than 20 very strong positives to this probe (with an expectation of 10 copies if the rDNA probe were single copy), which is in the realm of variation in representation of a single-copy gene in a genomic library (see Kelkar *et al.* 2001; Table 3). The W03C05 probe was confirmed to hybridize to a rDNA sequence (partly contained in cDNA S23E07) in the cDNA array. The copy number of the rDNA probe in the pWEB library was consistent with the rDNA genes being single copy (Giuntoli *et al.* 1994).

DISCUSSION

Physical mapping strategy: A physical map for the 7.7-Mbp *P. carinii* genome is being constructed using a dual strategy: mapping by sequencing to generate an efficient sequencing resource with high connectivity to other genome resources and hybridizing cosmids to a cDNA collection to generate a gene-rich map (Kuspa and Loomis 1996). In mapping by sequencing, the strategy is to sequence the ends of a cosmid library in the vectors pWEB (Epicentre Technologies) and pLorist-

6Xh (Kelkar *et al.* 2001) and then shotgun sequence ~100 cosmids with an average insert size of 26 kbp. In addition, ~300 cosmids thought to be nonoverlapping with the sequenced cosmids were hybridized to date to the arrayed cDNA collection to generate a gene-rich map. There are ~2000 distinct cDNAs in the cDNA library, and ~(2000/5280) × 1045 = 396 of these genes are currently represented in the physical map. The current integrated map represents at least 21% × 7700 kbp = 1617 kbp of the genome. Gene density on the physical map is then 1617/396 or ~1 gene/4 kbp currently from the cDNA library in the physical map. In that the cDNA library contains about half of the genes in the Pneumocystis genome, this is consistent with the gene density estimate from Smulian *et al.* (2001) of 1 gene/2 kbp. To carry out this dual strategy it was necessary to develop a new tool for building an integrated genome map of Pneumocystis containing both hybridization data and sequence data. The new tool is ODS3.

The new tool presents views of the data that capitalize on two advantages of the dual physical mapping strategy being used. The tool enables the viewer to inspect biologically interesting regions with repeats as identified by Figure 2. The tool also allows the viewer to inspect links in the physical map to validate that they are not based on contaminating rat DNA.

Coverage and gene density of the physical map of Pneumocystis: The progress on the physical map is on target with >55% of the genome covered (Figure 1), but there are too many contigs (~200 contigs after manual curation or ~13 contigs per chromosome). The likely explanation for the greater-than-expected number of contigs (Zhang and Marr 1993) is the extremely conservative filtering rule: remove any sequence contig if it contains a BLAST report with an *E*-value of *E*-40 and a filtered keyword (Tables 2 and 3). This explanation was validated by manual filtering (in which the annotator still makes use of sequence in a contig that is contaminated) and automated assembly of chromosome 7. Currently >345 nonredundant ESTs are assigned to the physical map so that the current state of the physical map is as gene rich as the genetic maps of many classical systems like *Neurospora crassa* (Perkins *et al.* 2001). The physical map also provides an automatic normalization of the cDNA library.

Multigene families in *P. carinii*: In Table 4 there are 10 putative multigene families from Pneumocystis, and we confirmed an earlier hypothesis that the rDNA family found in most organisms is in fact a single-copy gene in *P. carinii*. Of the 12 putative multigene families listed, 2 have been previously identified (MSG and HSP70) in Pneumocystis. The families (heat-shock ψ protein, HSP60, and tRNA synthetase) are multigene families in *S. pombe* or *S. cerevisiae*. Three of the 10 families are heat shock related; the abundance of stress-related messages in cDNA libraries has been a frequent finding (Prade *et al.* 2001). Two of the remaining families are likely contaminating rat host DNA (to be discarded), and the remaining 5 families are unknown.

A striking feature of Pneumocystis biology is the MSGs. The MSGs are the predominant antigenic species found on all *P. carinii* populations and are encoded by a gene family containing ~100 members (Stringer 1996). The MSGs have been implicated as a means of attachment to host cells (Ezekowitz *et al.* 1991; Pottratz *et al.* 1991) or as a mechanism to circumvent immune surveillance via antigenic variation (Stringer 1996). It has been proposed that at least three tandemly arrayed MSG genes are found in subtelomeric positions on each of the 16 *P. carinii* chromosomes (Garbe and Stringer 1994; Wada *et al.* 1995). Their expression appears to be regulated by placement of one MSG gene in a unique expression site located at the end of a single chromosome (Wada *et al.* 1995; Sunkin and Stringer 1996). *P. carinii* has dedicated ~10% of its genome to this family of genes (Stringer 1996), and the size of this family is reflected in Figure 2. The MSG-related cosmids and cDNAs make up 12.2% of the physical map in Figure 2, and the representation in the cosmid library based on DNA/DNA hybridization data appeared to be between 5 and 17%. There is no other known member of the fungal kingdom that possesses such a complex antigenic expression system. The MSG system shares most similarities with the antigenic switching mechanism of the protozoan trypanosomes. Currently the sequence similarity of chromosome ends is sufficiently high not to be able to separate the ends on a hybridization-based approach, but unique sequence embedded around the repeats should enable the resolution of chromosome ends in the physical map. A total of 40 cosmids have been identified by hybridization to contain complete MSG sequences using both a CRJE on the 5′ end and a conserved sequence on the 3′ end of MSG family members that will allow the identification of unique sequence tags for the MSG-containing cosmid clones.

The HSP70 is known to contain at least three family members. We found two of these existing family members. This raises the possibility that there is a third member of the family somewhere in the genome, and a third extremely tentative location was detected at cosmid W05B04. To test this possibility we created a multiway alignment of the 186 ESTs with homology to one of the Pneumocystis HSP70 genes previously reported (Stedman and Buck 1996; Stedman *et al.* 1998). While the lengths of the sequences are not sufficient (97 amino acids) to evaluate the significance of the groupings, all of the ER HSP70s (BIP) were clustered in the parsimony tree on or above EST S22D09 in Figure 3, and all of the cytoplasmic HSP70s were clustered together in the tree on or below S21E09 in Figure 3. The two clusters were separated by ESTs with hits to other HSP sequences like *N. crassa* HSP-88 (S44F12) and the rat HSP70 (S02A10). ESTs S50G10 and S22D09 are localized on or near cosmid W01D12 on the physical map. EST S21E09

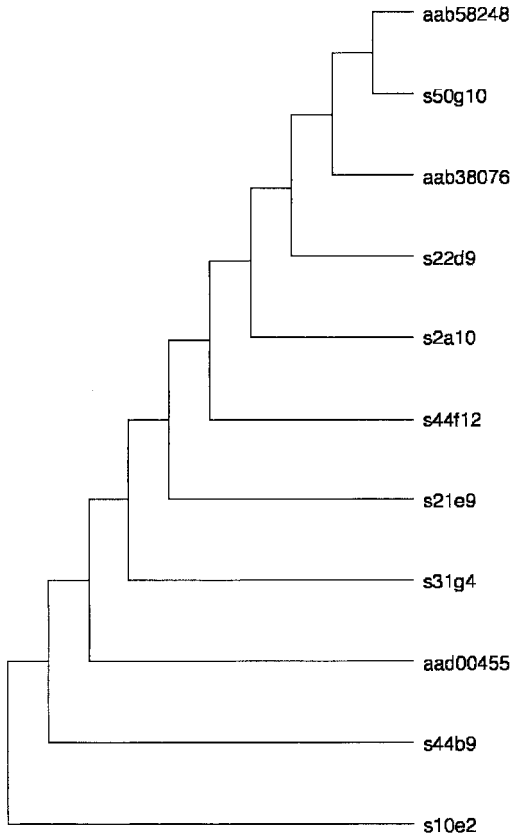

FIGURE 3.—A gene genealogy tree was generated through comparisons of deduced amino acid sequences of HSP70 gene family members and the standard sequences available in GenPept. Accessions numbers are indicated by aa. Each HSP70 family member description is in Table 4.

localizes to a distinct location on cosmid W04C09. The gene genealogy did not separate out the third location at S21D05 on cosmid W05B04. The sequence and physical map concordantly support only two family members at this time. We obtained results completely concordant with those in Figure 3 with an independent analysis beginning with a sequence assembly (see METHODS) applied to the ESTs to sort them into bins and subsequent parsimony analysis (results not shown).

Pneumocystis has an unusual biology in several respects, and another example is the argument that rDNA genes are single copy (GIUNTOLI *et al.* 1994). No other genus of fungi is known to have only a single copy of the rDNA genes. A cosmid was sequenced and found to contain the rDNA genes (W03C05). The rDNA sequence appears to be surrounded on both sides by unique sequence of 1–2 kb in extent, and >20 kbp of cosmid DNA in which the rDNA genes were embedded were found to have no further copies. The spacing between rDNA family members in other fungi appears to be ~1 kbp. Because chromosome 7 contains the rDNA sequence by PFGE Southern and because coverage of chromosome 7 is estimated to be 90%, it is unlikely that we have missed an extension of the rDNA sequences. These data are consistent with a single copy of the rDNA genes.

Limitations to mapping by sequencing: The tool ODS3 enabled the identification of certain families of repeats from the clone/probe hybridization matrix. Once found, the precise placement of the repeats is needed, a function that ODS3 cannot currently provide. Further analysis requires the use of associated unique sequence to provide a context for the repeats, and other assembly tools are needed that successfully separate repeats by the associated unique sequence.

Different filtering strategies using the BLAST reports from the HTBLAST workflow application on all sequences in the integrated genome map vary as to their success in removing contaminating DNA. It was found that a clone-based filtering strategy (as opposed to a probe-based strategy) removed more of the problematic sequences in Tables 2 and 3 when end sequences of all clones were available. Most of the hits (61%) removed were multiple-hybridization signals (beyond two hits) to particular clones. These multiple-hybridization signals will lead to ambiguous placement of clones in the physical map and will complicate the assembly. Moreover, filtering sequences based only on probes might leave sequences in clones (not used as probes) that might mislead researchers using the sequence on the web. Four kinds of sequences were filtered out: *E. coli*, vector, Adenovirus, and Pseudomonas. The first two contaminants likely arose in the cosmid DNA preparation during shotgun sequencing. The Adenovirus and Pseudomonas sequences are hypothesized to arise because the libraries are derived from an individual immunosuppressed rat lung, which can contain other microbes (SMULIAN *et al.* 2001).

One filter fits all? One other decision about the filter is what *E*-value in HTBLAST to use as a filter for declaring an overlap. In constructing a physical map in Figure 2 an *E*-value of *E*-5 was used as a cutoff. The impact of the choice can be seen in Tables 2 and 3. This filter allowed the separation of the HSP70 family members to different locations in the genome, but the MSG family members coagulated into one location. Furthermore, a threshold of *E*-40 at a prefiltering step (to remove *E. coli* and vector sequence, see METHODS) entirely removed the rDNA genes because of their high similarity to those of *E. coli*. Whether or not a particular filter "works" will depend on the sequence similarity of family members. One filter is not likely to work for all families with differing levels of sequence similarity between family members within a genome. Instead what may be required is a filter that looks at the context of surrounding repeats in declaring an overlap (KECECIOGLOU and MYERS 1995).

The immediate challenge: In spite of the limitations of the tool, ODS3 does a remarkable job of providing

a first pass at the genome and tracking how a project is unfolding. Normally, physical mapping of individual fungal chromosomes with on the order of 1000 clones and 100 probes requires 1 month to edit one chromosome (Prade *et al.* 1997). With the tool ODS3 one can accomplish the same task in a nearly platform-independent manner within 1 day and make the results available to the community over the web in a format that can be viewed or exchanged with other databases. This accomplishment was achieved by enhancements of existing physical mapping algorithms, utilizing the Oracle 8*i* object-relational database system, Java-based implementation of ODS3 promoting platform independence, utilization of an automated workflow for parsing sequences through a parallelized version of BLAST for annotation, and storage of the data in an XML format for easy export to other resources. The result is a timely presentation of the Pneumocystis genome reconstruction.

Near-term challenge of integrating genome projects: At present, almost all genome information systems are constructed from scratch with little reuse of software developed elsewhere (Goodman *et al.* 1995). This is especially true for genetic data sources, including genome databases (*e.g.*, ACEDB, GDB) and flat file systems (*e.g.*, BLAST, FastA, Staden). With >14 fungal genome projects underway (Bennett and Arnold 2001), the most pressing challenge is providing a framework to unite the diverse ontologies of these fungal genome projects so that information about one system can help to elucidate problems in other fungal systems. Researchers wishing to perform multiple-pass analysis of data, by feeding results of one program to another, encounter the problem of changing data from one format to another. This can be time consuming, frustrating, and error prone. The barriers to information flow include the lack of exchange mechanisms between the diverse array of information systems acting as repositories for these projects and the diversity of nomenclatures used (Bennett and Arnold 2001). It is unlikely that this diversity will go away, so a successful integration tool will need to embrace this diversity rather than try to limit it. There is a need to develop a standardized but flexible language for representing different types of fungal genome data (Robbins 1996).

The XML is a way of sharing, exchanging, and organizing the vast amount of genomics data cooperatively. The XML is a meta-language to produce documents that convey content with semantic structure (Elenko and Reinertsen 2000) and is widely used for data presentation. An XML-based genomic data file from one system can be restructured and presented to another system through an associated DTD file without any change. The DTD file provides control over the XML schema and validation. The receiver system can understand and verify the integrity of received XML files by checking the DTD file.

The extensible markup language is becoming a prevailing document and information exchange standard on the web (see BioML at http://www.bioml.com/BIOML/index.html, GAME at http://bioxml.org/Projects/game, GEML at http://www.geml.orgn, and OpenBSA at http://industry.ebi.ack.uk/openBSA/. As long as a genome information system can provide a means to export data and views of the data in XML, then other systems will have the capability to import this information. We propose that XML be used as a data representation and exchange medium for fungal genome projects. To this end we have added to ODS3 the capability to store an integrated genome map in XML format for reuse in other projects. This is a small but important step toward developing a standard representation method for fungal genomics data. In this way standardized XML DTDs can be used to publish genomics information on the web instead of *ad hoc* representation mechanisms as is currently done. This approach should increase reusability and simplify integration across genome projects (Davidson *et al.* 2001) if one additional step is taken. After embracing XML it will be necessary to encourage data sources like the FGDB to set up and register XML messaging systems (*i.e.*, web services) to enable the integration of disparate data sources in one "bioinformatics nation" (Stein 2002).

We gratefully acknowledge the support from the National Science Foundation in the form of grants MCB-9630910 (J.A.) and BIR-9512887 (J.A.), from the National Institutes of Health (R01 AI44651 to M.C., J.A., G.S., and J.S.), from the U.S. Department of Agriculture (USDA-2002-35300-12475 to S.B. and J.A.), and from the Georgia Research Alliance (J.A.).

LITERATURE CITED

Aign, V., U. Schulte and J. D. Hoheisel, 2001 Hybridization-based mapping of *Neurospora crassa* linkage groups II and V. Genetics **157:** 1015–1020.

Altschul, S. F., W. Gish, W. Miller, E.W. Myers and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. **215:** 403–410.

Arnold, J., 1997 Editorial. Fungal Genet. Biol. **21:** 254–257.

Arnold, J., 2001 Foreword. Genetics **157:** 933.

Arnold, J., and M. T. Cushion, 1997 Constructing a physical map of the *Pneumocystis* genome. J. Eukaryot. Microbiol. **6:** 8S.

Bennett, J., and J. Arnold, 2001 Genomics for fungi, pp. 267–297 in *The Mycota VIII. Biology of the Fungal Cell*, edited by R. J. Howard and N. A. R. Gow. Springer-Verlag, New York.

Bhandarkar, S. M., and S. A. Machaka, 1997 Chromosome reconstruction from physical maps using a cluster of workstations. J. Supercomput. **11:** 61–86.

Bhandarkar, S. M., S. A. Machaka, S. S. Shete and R. N. Kota, 2001 Parallel computation of a maximum likelihood estimator of a physical map. Genetics **157:** 1021–1043.

Burke, D. T., G. F. Carle and M. V. Olson, 1987 Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science **235:** 1046–1049.

Camp, N., H. Cofer and R. Gomperts, 1998 White paper: high throughput BLAST (http://www.sgi.com/solutions/sciences/chembio/resources/papers/HTBlast/HT_Whitepaper.html).

Chibana, H., B. B. Magee, S. Grindle, Y. Ran, S. Scherer *et al.*, 1998 A physical map of chromosome 7 of *Candida albicans*. Genetics **149:** 1739–1752.

Church, G., and W. Gilbert, 1984 Genomic sequencing. Proc. Natl. Acad. Sci. USA **81:** 1991–1995.

Coulson, A., C. Huynh, Y. Kozono and R. Shownkeen, 1995 The physical map of the *Caenorhabditis elegans* genome. Methods Cell Biol. **48:** 533–550.

Creutz, M., 1983 Microcanonical Monte Carlo simulation. Physiol. Rev. Lett. **50:** 1411–1414.

Cushion, M. T., and J. Arnold, 1997 Proposal for a *Pneumocystis* genome project. J. Eukaryot. Microbiol. **44:** 7S.

Cushion, M. T., M. Kaselis, S. L. Stringer and J. R. Stringer, 1993 Genetic stability and diversity of *Pneumocystis carinii* infecting rat colonies. Infect. Immun. **61:** 4801–4813.

Cuticchia, A. J., 1994 A primer for relational databases, pp. 346–349 in *Automated DNA Sequencing and Analysis*, edited by M. D. Adams, C. Fields and J. C. Venter. Academic Press, New York.

Cuticchia, A. J., J. Arnold and W. E. Timberlake, 1992 The use of simulated annealing in chromosome reconstruction experiments based on binary scoring. Genetics **132:** 591–601.

Cuticchia, A. J., J. Arnold and W. E. Timberlake, 1993 ODS (ordering DNA sequences): a physical mapping algorithm based on simulated annealing. Comput. Appl. Biosci. **9:** 215–219.

Davidson, S. B., J. Crabtree, B. Brunk, J. Schug, V. Tannen *et al.*, 2001 K2/Kleisli and GUS: experiments in integrated access to genomic data sources. IBM Syst. J. **40:** 512–530.

Durbin, J., and J. Thierry-Mieg, 1994 The ACEDB genome database, pp. 45–55 in *Computer Methods Genome Research,* edited by S. Suhai. Plenum Press, New York.

Elenko, M., and M. Reinertsen, 2000 XML & CORBA (http://cgi.omg.org/library/adt.html).

Enkerli, J., H. Reed, A. Briley, G. Bhatt and S. F. Covert, 2000 Physical map of a conditionally dispensable chromosome in *Nectria haematococca* mating population VI and location of chromosomal breakpoints. Genetics **155:** 1083–1094.

Ewing, B., and P. Green, 1998 Base-calling of automated sequencer traces using *Phred* II. Error probabilities. Genome Res. **8:** 186–194.

Ewing, B., L. Hillier, M. C. Wendl and P. Green, 1998 Base-calling of automated sequencer traces using *Phred* I. Accuracy assessment. Genome Res. **8:** 175–185.

Ezekowitz, R. A. B., D. J. Williams, H. Koziel, M. Y. K. Armstrong, A. Warner *et al.*, 1991 Uptake of *Pneumocystis carinii* mediated by the macrophage mannose receptor. Nature **351:** 155–158.

Garbe, T. R., and J. R. Stringer, 1994 Molecular characterization of clustered variants of genes encoding major surface antigens of human *Pneumocystis carinii*. Infect. Immun. **62:** 3092–3101.

Giuntoli, D., S. L. Stringer and J. R. Stringer, 1994 Extraordinarily low number of ribosomal RNA genes in *Pneumocystis carinii*. J. Eukaryot. Microbiol. **41:** 88S.

Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon *et al.*, 1996 Life with 6000 genes. Science **274:** 546–567.

Goodman, N., S. Rozen and L. D. Stein, 1995 *The Case for Componentry in Genome Information Systems.* Whitehead/MIT Center for Genome Research (http://www-genome.wi.mit.edu/informatics/componentry.html), Cambridge, MA.

Goodman, N., S. Rozen, L. D. Stein and A. G. Smith, 1998 The LabBase system for data management in large scale biology research laboratories. Bioinformatics **14:** 562–574.

Hall, D., J. Wang and S. M. Bhandarkar, 2001 ODS2: a multiplatform software application for creating integrated physical and genetic maps. Genetics **157:** 1045–1056.

Hall, D., J. A. Miller, J. Arnold, K. J. Kochut, A. P. Sheth *et al.*, 2003 Using workflow to build an information management system for a geographically distributed genome sequencing initiative, pp. 359–371 in *Genomics of Plants and Fungi*, edited by R. A. Prade and H. J. Bohnert. Marcel Dekker, New York.

Hoheisel, J. D., E. Maier, R. Mott, L. McCarthy, A. V. Grigoriev *et al.*, 1993 High resolution cosmid and P1 maps spanning the 14 Mb genome of the fission yeast *S. pombe*. Cell **73:** 109–120.

International Human Genome Sequencing Consortium, 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860–918.

Kececioglou, J. D., and E. W. Myers, 1995 Combinatorial algorithms for DNA sequence assembly. Algorithmica **13:** 7–51.

Kelkar, H. S., J. Griffith, M. E. Case, S. F. Covert, R. D. Hall *et al.*, 2001 The *Neurospora crassa* genome: cosmid libraries sorted by chromosome. Genetics **157:** 979–990.

Kochut, K. J., J. Arnold, J. A. Miller and W. D. Potter, 1993 Design of an object-oriented database for reverse genetics, pp. 234–242 in *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, edited by L. Hunter, D. Searls and J. Shavlik. AAAI Press, Menlo Park, CA.

Kochut, K. J., J. Arnold, A. Sheth, J. A. Miller, E. Kraemer *et al.*, 2003 IntelliGEN: a distributed workflow system for discovering protein-protein interactions. Parallel Distributed Databases **13:** 43–72.

Kraemer, E., J. Wang, J. Guo, S. Hopkins and J. Arnold, 2001 An analysis of gene-finding programs for *Neurospora crassa*. Bioinformatics **17:** 901–912.

Kuspa, A., and W. F. Loomis, 1996 Ordered yeast artificial chromosome clones representing the *Dictyostelium discoideum* genome. Proc. Natl. Acad. Sci. USA **93:** 5562–5566.

Lin, J., R. Qi, C. Aston, J. Jing, T. Anantharaman *et al.*, 1999 Whole-genome shotgun optical mapping of *Deinococcus radiodurans*. Science **285:** 1558–1562.

Mahairas, G. G., J. Wallace, K. Smith, S. Swartzell, T. Holzman *et al.*, 1999 Sequence-tagged connectors: a sequence approach to mapping and scanning the human genome. Proc. Natl. Acad. Sci. USA **96:** 9739–9744.

Miller, J. A., D. Palaniswami, A. P. Sheth, K. J. Kochut and H. Singh, 1998 WebWork: METEOR's web-based workflow management system. J. Intell. Inf. Syst. **10:** 185–215.

Mizukami, T., W. I. Chang, I. Garkavtseve, N. Kaplan, D. Lomardi *et al.*, 1993 A 13 kb resolution cosmid map of the 14 Mb fission yeast genome by nonrandom sequence-tagged site mapping. Cell **73:** 121–132.

Mott, R., A. Grigoriev, E. Maier, J. Hoheisel and H. Lehrach, 1993 Algorithms and software tools for ordering clone libraries: application to the mapping of the genome of *Schizosaccharomyces pombe*. Nucleic Acids Res. **21:** 1965–1974.

Olson, M. V., J. E. Dutchik, M. Y. Graham, G. M. Brodeur, C. Helms *et al.*, 1986 Random-clone strategy for genomic restriction mapping in yeast. Proc. Natl. Acad. Sci. USA **83:** 7826–7830.

Perkins, D. D., M. S. Sachs and A. Radford, 2001 *Chromosomal Loci of Neurospora crassa*. Academic Press, New York.

Pottratz, S. T., J. Paulsrud, J. E. Smith and W. J. Martin, 1991 *Pneumocystis carinii* attachment to cultured lung cells by *Pneumocystis* gp120, a fibronectin binding protein. J. Clin. Invest. **88:** 403–407.

Prade, R. A., J. Griffith, K. J. Kochut, J. Arnold and W. E. Timberlake, 1997 *In vitro* reconstruction of the *Aspergillus* (=*Emericella*) *nidulans* genome. Proc. Natl. Acad. Sci. USA **94:** 14564–14569.

Prade, R. A., P. Ayoubi, S. Krishnan, S. Macwana and H. Russell, 2001 Accumulation of stress and cell wall degrading enzyme associated transcripts during asexual development in *Aspergillus nidulans*. Genetics **157:** 957–967.

Robbins, R. J., 1996 Bioinformatics: essential infrastructure for global biology. J. Comput. Biol. **3:** 465–478.

Roe, B. A., J. S. Crabtree and A. S. Khan, 1996 *DNA Isolation and Sequencing*. John Wiley & Sons, New York.

Rozen, S., L. D. Stein and N. Goodman, 1995 LabBase: a database to manage laboratory data in a large-scale genome-mapping project. IEEE Comput. Med. Biol. **14:** 702–709.

Schoenfeld, T., J. Mendez, D. R. Storts, E. Portman, B. Patterson *et al.*, 1995 Effects of bacterial strains carrying the *endA1* genotype on DNA quality isolated with Wizard™ plasmid purification systems. Promega Notes Mag. **53:** 12–21.

Senger, B., L. Despons, P. Walter, H. Jakubowski and F. Fasiolo, 2001 Yeast cytoplasmic and mitochondrial methionyl-tRNA synthetases: two structural frameworks for identical functions. J. Mol. Biol. **311:** 205–216.

Silver, P. A., and J. C. Way, 1993 Eukaryotic dnaJ homologues and the specificity of HSP70 activity. Cell **74:** 5–6.

Slonim, D., L. Kruglyak, L. Stein and E. Lander, 1997 Building human genome maps with radiation hybrids. J. Comput. Biol. **4:** 487–504.

Smulian, A. G., T. Sesterhenn, R. Tanaka and M. T. Cushion, 2001 The *ste3* pheromone receptor gene of *Pneumocystis carinii* is surrounded by a cluster of signal transduction genes. Genetics **157:** 991–1002.

Stedman, T. T., and G. A. Buck, 1996 Identification, characterization, and expression of the BiP endoplasmic reticulum resident

chaperonins in *Pneumocystis carinii*. Infect. Immun. **64:** 4463–4471.

Stedman, T. T., D. R. Butler and G. A. Buck, 1998 The HSP70 gene family in *Pneumocystis carinii*: molecular and phylogenetic characterization of cytoplasmic members. J. Eukaryot. Microbiol. **45:** 589–599.

Stein, L., 2002 Creating a bioinformatics nation: a web-services model will allow biological data to be fully exploited. Nature **417:** 119–120.

Stringer, J. R., 1996 *Pneumocystis carinii*: What is it, exactly? Clin. Microbiol. Rev. **9:** 489–498.

Sunkin, S. M., and J. R. Stringer, 1996 Translocation of surface antigen genes to a unique telomeric expression site in *Pneumocystis carinii*. Mol. Microbiol. **19:** 283–295.

Talbot, C. C., and A. J. Cuticchia, 1998 Human mapping databases, pp. 1.13.1–1.13.12 in *Current Protocols in Human Genetics*, edited by N. Dracopoli, J. Haines, B. Korf *et al.* John Wiley & Sons, New York.

Thomas, S. W., E. A. Rundensteiner and A. J. Lee, 1995 Visualization and database tools for YAC and cosmid contig construction. Project-Oriented Databases and Knowledge Bases in Genome Research, Biotechonology Computing Track, Twenty-Seventh Hawaii International Conference of System Sciences, HICSS-28, Kihei, HI, No. 128.

Vargas, C., 2002 The genomic study of multigene families of *Pneumocystis carinii* for potential drug targets. Honors Thesis, B.S. Microbiology, University of Georgia, Athens, GA.

Venter, J. C., H. O. Smith, P. W. Li, R. J. Mural and L. Hood, 1996 A new strategy for genome sequencing. Nature **381:** 364–366.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural *et al.*, 2001 The sequence of the human genome. Science **291:** 1304–1351.

Wada, M., S. M. Sunkin, J. R. Stringer and Y. Nakamura, 1995 Antigenic variation by positional control of major surface glycoprotein gene expression in *Pneumocystis carinii*. J. Infect. Dis. **171:** 1563–1568.

Xiong, M., H. J. Chen, R. A. Prade, Y. Wang, J. Griffith *et al.*, 1996 On the consistency of a physical mapping method to reconstruct a chromosome in vitro. Genetics **142:** 267–284.

Zhang, M. Q., and T. G. Marr, 1993 Genome mapping by nonrandom anchoring: a discrete theoretical analysis. Proc. Natl. Acad. Sci. USA **90:** 600–604.

Communicating editor: Z-B. Zeng