

The Giardia Genome Project: Production Pipeline and Assembly of LI-COR Bi-Directional Sequence Data

Hilary G. Morrison, Andrew G. McArthur, Julie E. J. Nixon, Nora Q. E. Passamaneck, Ulandt Kim, Melissa K. Crocker, Gregory Hinkle, Michael E. Holder, Rebecca Farr, Claudia I. Reich, Gary J. Olsen, Lorena A. Fierro, Stephen B. Aley, Rodney D. Adam, Frances D. Gillin and Mitchell L. Sogin

The Josephine Bay Paul Center for Comparative Molecular Biology and Evolution
The Marine Biological Laboratory, Woods Hole, MA 02543
E-mail: morrison@mbi.edu, sogin@mbi.edu

ABSTRACT

We have undertaken complete sequencing and annotation of the 12 MB genome of the eukaryotic parasite, *Giardia lamblia* (Figure 1). Our genome project relies on LI-COR bi-directional reads for primary shotgun sequence data. We have achieved four-fold coverage of the genome in three years (>99% of coding capacity) and have assembled the data into approximately 1400 contigs. We are now using directed plasmid and BAC sequencing to join contigs and map contigs to BACs and to *Giardia's* five chromosomes. We have recently begun annotation of the genome. Our production pipeline begins with individual reads (SMP or SAMP files, after basecalling and minimal editing) and ends with annotated contig files. We utilize several pre-existing tools for sequence analysis, including modules from the SEALS, BLAST, GCG, and PHRED/PHRAP/CONSED packages. Additionally, we have created a number of UNIX® and perl scripts specific to LI-COR data and this genome project.

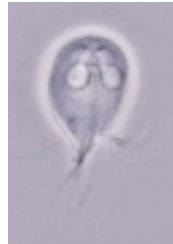


Figure 1. *Giardia lamblia* parasite.

INTRODUCTION

Giardia lamblia is an environmentally transmitted, waterborne, human pathogen. We have selected *G. lamblia* as a model organism for genome analysis because of its well-recognized impact on human health, its relatively small genome containing approximately 12 million base pairs, and the insights it will provide about the origins of nuclear genome organization. Previous comparisons between several gene families have demonstrated *Giardia's* basal position in molecular phylogenies. Since the divergence of *Giardia* lies close to the transition between eukaryotes and prokaryotes in universal ribosomal RNA phylogenies, it is a valuable model for gaining basic insights into the genetic innovations that led to the formation of eukaryotic cells.

In evolutionary terms, the divergence of this organism is at least twice as ancient as the common ancestor for yeast and man. Sequence analyses of the *Giardia lamblia* genome will address several important questions related to human health, including the number, gene organization and regulation of variant-specific surface protein coding regions. The *Giardia* genome project at the Marine Biological Laboratory in Woods Hole is part of an NIH Investigator-Initiated Interactive Research Project Grant (IRPG). The genome sequencing component is a collaborative effort between the laboratories of Mitchell L. Sogin (Josephine Bay Paul Center for Comparative Molecular Biology and Evolution at the MBL), Stephen Aley (University of Texas at El Paso), Rodney Adam (University of Arizona at Tucson), and Gary Olsen (University of Illinois at Urbana-Champaign). The *Giardia* sequencing effort is complemented by its IRPG functional genomics unit, directed by Frances D. Gillin at the University of California at San Diego.

METHODS

Single pass reads

Giardia lamblia genomic libraries were prepared in the plasmid vector pUC18, a 2.7 kb vector with primer sites for M13 forward (universal) and M13 reverse primers. We use a shotgun sequencing approach (Fig. 2).

The majority of single pass read data comes from one library, which has an average insert size of 2.35 kbp and a range of insert sizes from 1.6 to 2.7 kbp. Plasmid templates are prepared following a standard Qiagen REAL96 protocol on the BioRobot 9600. Template quality is checked by agarose gel electrophoresis. Each end of the cloned insert is sequenced using LI-COR's simultaneous bi-directional sequencing protocol (SBS, Roemer *et al.*, 1997). The M13 universal (forward) primer is labeled with IRDye™ 700 and the M13 reverse primer with IRDye™ 800. We use Epicentre's Excel II® cycle sequencing protocol. Sequencing reactions (7 µl volume) are assembled using the Tecan Miniprep 75. The samples are loaded onto 3.75% KB^{Plus}™ polyacrylamide gels and run for 12 hours on a LI-COR 4200 sequencing machine.



Fig. 2

A single SBS reaction generates 900-1100 base reads from each primer, which are given unique and informative names (Fig. 3). We use LI-COR automatic base-calling software, then manually edit each sample. This results in single pass reads with an extremely high degree of accuracy (>99%), which are made available to the scientific community (www.mbl.edu/Giardia and www.ncbi.nlm.nih.gov).

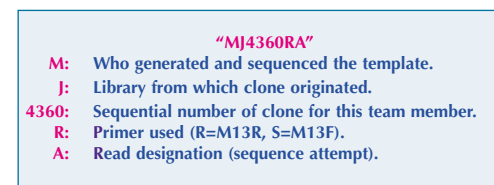


Fig. 3

The data pipeline bypasses normal ABI-type PHRED base calling because of the superior performance of the LI-COR base caller when presented with LI-COR generated TIFF images.

Data analysis (Figures 4 A-F)

After data acquisition, base calling and editing (A, B), SCF and phd (quality value) files are generated on an OS/2 platform using the bulk SCF/phd executable "makeallscfphd," or on a PC using eSeq™. Copies of the SCF and phd files are moved into the assembly and analysis pipeline using a UNIX shell script called "Harvest" and all the original files associated with a sequencing run (text files, TIFFs, sample files, etc.) are archived (C). The phd files are converted to fasta format and trimmed of vector sequence using the PHRED routines "phd2seqfasta" and "crossmatch" (D). As new data are harvested, the file of trimmed fasta format sequences is converted to a BLAST-searchable database using the NCBI program "formatdb" (E). Each fasta file is used as a query sequence to search the nucleotide and protein sequence databases for homologous sequences using the BLAST algorithm (Altschul *et al.*, 1990). BLAST results are parsed and converted to HTML format using a suite of UNIX scripts (PreviewBlast and WebRelease). Included in these scripts are a number of error-checking steps to ensure, for example, that all sequences contain unique identifiers. The BLASTX results are the only "annotation" available for the first pass data (F).

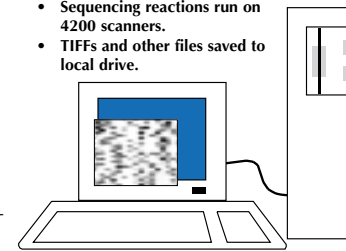


Fig. 4 A.

Data Assembly

The primary reads are assembled into sequence contigs using PHRAP and CONSED (Gordon *et al.*, 1998). As described previously, phd files are trimmed of contaminating vector sequence and converted to fasta format. The merged fasta sequence file and associated quality file are used by PHRAP to generate contigs of overlapping reads, with parameters such as mismatch and minimum overlap set by the user (D). Currently, we set the overlap to 50 bases and allow no more than 2% mismatch. Contig sequences are used to construct a BLAST searchable database (E) and to identify target clones for sequence closure.

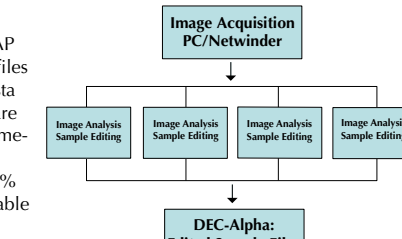


Fig. 4 B.

Assembly Annotation

CRITICA (Coding Region Identification Tool Invoking Comparative Analysis; Badger and Olsen, 1999) is used to identify likely protein coding sequences in each contig. In the comparative analysis component of CRITICA, regions of DNA are aligned with related sequences from public databases, and greater than expected amino acid identity indicates a likely coding sequence. Proteins identified by CRITICA are entered into the ERGO database at Integrated Genomics (www.integratedgenomics.com) for functional analysis. In this process, a model of connected metabolic pathways is constructed. Using a model-based approach, rather than looking at isolated coding regions, means that functional annotations are more reliable and "missing" functions are more readily recognized. The initial static model is refined, using a web-based interface, by curators at Integrated Genomics, the MBL, and other institutions.

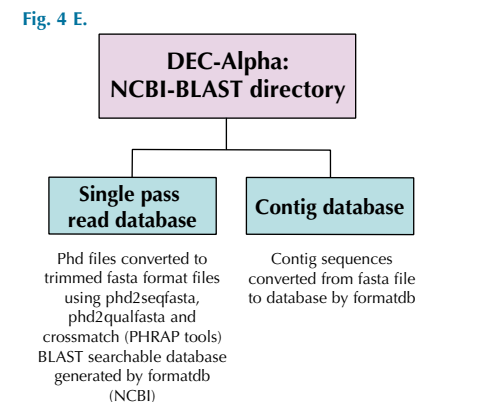
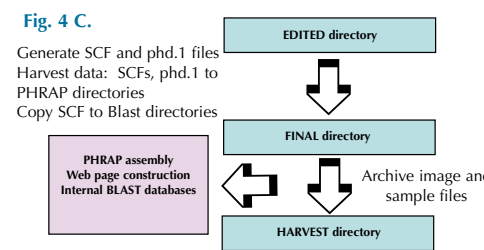


Fig. 4 E.

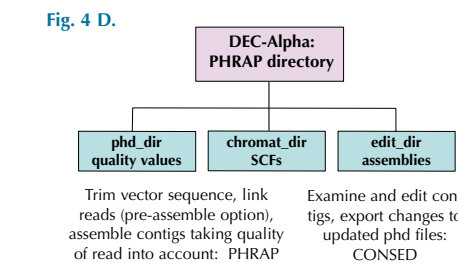


Fig. 4 D.

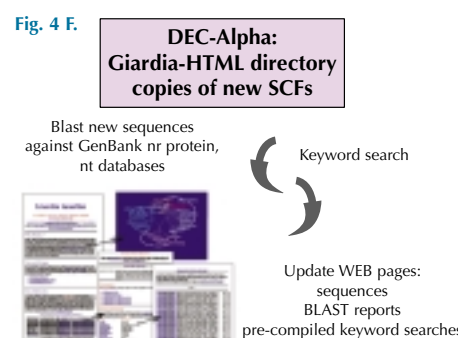


Fig. 4 F.

RESULTS

At the end of three years of our sequencing effort, the *Giardia lamblia* sequence database (www.mbl.edu/Giardia) contains over 50 million bases, representing four-fold genome coverage. Average trimmed read length is 885 nucleotides (Fig. 5). Data quality is very high (phd quality value > 20) out to 950-1000 bases (Figure 6). At the time of the last data release, the reads assembled into approximately 1400 contigs, with a total length of greater than 11 Mbp (Figure 7). Our results demonstrate that a shotgun sequencing approach using bi-directional reactions and LI-COR automated sequencers is well suited to a small genome project.

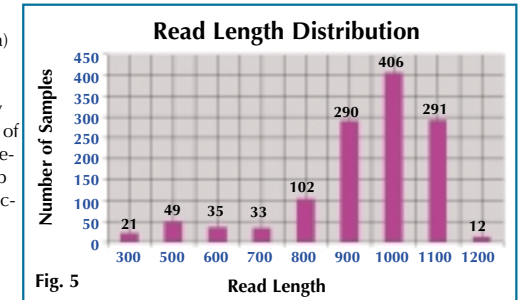


Fig. 5

Our ability to generate long and accurate reads means that reliable first-pass sequence data can be released to the scientific community and used to jump-start specific research. Furthermore, the long, bi-directional reads have allowed us to link several hundred of the contigs into "super-contigs," since the two reads from a single clone sometimes assemble into two different contigs.

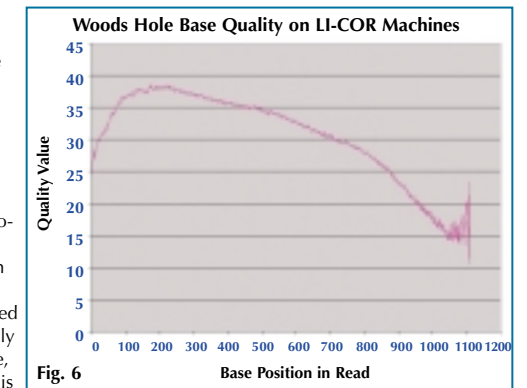


Fig. 6

BLAST results show that over 27% of the reads contain significant similarity to published protein sequences. Among these are proteins involved in intermediary metabolism, nucleic acid processing, cytoskeletal structure, and cell division. Interestingly, although introns have not been reported in *Giardia*, we have identified genes similar to PRP8 and RNP specific proteins, both of which are involved in RNA splicing, and have detected genes which potentially contain introns. And, although *Giardia* is amitochondriate, we have found homologues of mitochondrial proteins. This suggests that at one time in its evolutionary history, *Giardia* harbored a prokaryotic endosymbiont. Another surprising discovery is a protein that displays nearly 95% similarity at the amino acid level with a cDNA that is expressed in embryonic mouse and human placental tissue. The function of this protein is unknown in any system. As expected, we have discovered many open reading frames that do not return any significant BLASTX hits and potentially encode unique or novel proteins.

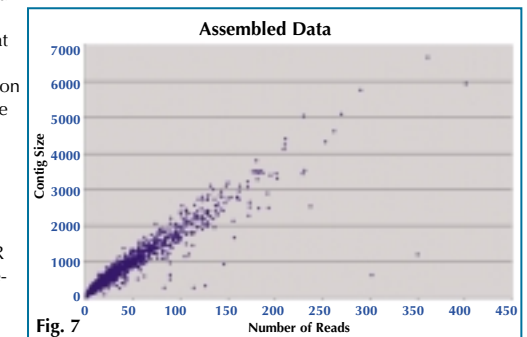


Fig. 7

REFERENCES

- S. Roemer, J. Amen, R. Bruce *et al.*, 1997. A New Near-IR Fluorescence Automated DNA Sequencer. Poster presented at Automation in Mapping and DNA Sequencing, Heidelberg, Germany, March 1997. LI-COR Application Note #484. <http://bio.licor.com>.
- S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215:403-410.
- D. Gordon, C. Abajian, and P. Green.1998. CONSED: A Graphical Tool for Sequence Finishing. *Genome Research* 8:195-202.
- J. H. Badger and G. J. Olsen. 1999. CRITICA: Coding Region Identification Tool Invoking Comparative Analysis. *Molecular Biology Evolution* 16:512-524.

ACKNOWLEDGMENTS

Supported by grant AI43272 to M.L.S. from the National Institutes of Health, LI-COR Biotechnology Division, and the generosity of the G.Unger Vetlesen Foundation. The following have also contributed to this work: Bruce Luders, Scott Bressoud, Elizabeth Duffy, Margaret Bradley, Seth Ament, Dave Gellis, Jeff Kim, John Darga, Alexandria Papa and Martin Foster.

©2001 LI-COR, inc. LI-COR, IRDye and e-Seq are registered trademarks of LI-COR, inc. All other trademarks belong to their respective owners.

Application Note #541

This poster can be viewed online at www.licor.com



4308 Progressive Ave. • P.O. Box 4000 • Lincoln, Nebraska 68504 USA
US & Canada: 800-645-4267 • Int'l: 402-467-0700
Germany: +49 (0) 6172 17 17 771 • UK: +44 (0) 1223 422104
www.licor.com