Optical map of the Genotype A1 WB C6 Giardia lamblia genome isolate

Alexander Perry[1], Hilary G. Morrison[2], and Rodney D. Adam[3]

[1]University of Arizona College of Medicine
Infectious Disease Section
1501 N. Campbell
Tucson, AZ 85724-5039
USA

[2]Josephine Bay Paul Center, MBL
7 MBL Street
Woods Hole, MA 02543-1015
USA

[3](corresponding author) University of Arizona College of Medicine
Infectious Disease Section
1501 N. Campbell
Tucson, AZ 85724-5039
USA

Phone: (520) 626-6887
Fax (520) 626-5183
adamr@u.arizona.edu

Abstract

The Giardia lamblia genome consists of 12 Mb divided among 5 chromosomes ranging in size from approximately 1 to 4 Mb. The assembled contigs of the genotype A1 isolate, WB, were previously mapped along the 5 chromosomes on the basis of hybridization of plasmid clones representing the contigs to chromosomes separated by PFGE. In the current report, we have generated an MluI optical map of the WB genome to improve the accuracy of the physical map. This has allowed us to correct several assembly errors and to better define the extent of the subtelomeric regions that are not included in the genome assembly.

Key words: optical map, genome, pulsed field gel electrophoresis, subtelomeric variation

Introduction

The published sequence of the Giardia lamblia genotype A1 isolate, WB, consists of 11.7 Mb divided among 306 contigs. Some of these contigs were joined into larger scaffolds, primarily by “contig-joining” clones that linked these contigs even in the absence of continuous sequence [1]. The results were supplemented by the use of multiple BAC clones that were end-sequenced and physically mapped to specific chromosomes using pulsed field gel electrophoresis (PFGE). Subsequent physical mapping studies using NotI-digested chromosomes of the genotype A1 isolate, BRIS/83/HEP/106, [2, 3] have made additional contributions to a complete physical map. The current manuscript describes the use of optical mapping to refine and extend the physical map of the WB isolate.

Methods

WB-C6 Giardia trophozoites were used to generate the optical map. The WB isolate was originally axenized from a patient who most likely acquired his giardiasis in Afghanistan [4] and subsequently has been cloned a number of times. The C6 clone from the laboratory of Dr. Fran Gillin, UC San Diego, was used for the genome project and was also used for the optical mapping described here. However, the WB isolate has been subjected to multiple rounds of replication in the laboratory, so any changes that occur rapidly, such as changes in the subtelomeric regions (STRs) may have resulted in differences between the organisms used for the genome project and those used for the optical mapping.

Trophozoites were grown to confluence, pelleted and embedded in soft agarose as previously described [5], followed by digestion with proteinase K in the presence of 1% Sarkosyl. The optical mapping performed by OpGen (Gaithersburg, Maryland) [6, 7] consisted of melting the agarose blocks followed by digestion with B-agarase. The DNA was mounted on a glass optical mapping surface and digested in situ with MluI so that the order of the individual restriction fragments was maintained. The DNA was labeled with fluorescent YOYO-1 and imaged by fluorescent microscopy, allowing the sizes of the fragments to be estimated by the intensity of the fluorescent labeling. OpGen software was used to generate an MluI restriction map and then to compare that map with the available genomic sequence data. The map generated 150-fold coverage. An algorithm that incorporates the length of the alignment and the quality of the individual restriction fragments was used to overlay the sequence contigs (and secondarily the scaffolds) onto the optical map. Individual sequence contigs could be flagged as problematic if regions of match were followed by complete mismatch, suggesting an assembly error in the individual sequence contigs.

Contigs that matched the optical map over their entire sequence were left intact. Those that matched the optical map for only a portion of the map were split at the point of discrepancy (c13, c27 and c29; Table 1). Conversely, if two contigs overlapped on the contig map and had areas of sequence identity consistent with their positions on the optical map, these contigs were joined. (17a and 53, 61 and 29a; Table 1).

Results and Discussion

The MluI optical map yielded a genome size of 12.1 Mb divided among five chromosomes ranging in size from 1.46 to 4.43 Mb. There were 1463 MluI sites with an average restriction fragment size of 8.29 kb. These chromosome sizes compare with PFGE estimates of 1.6 Mb to 3.5 Mb (Table 1). The total genome size estimated by the optical map is remarkably similar to the 12 Mb estimated by PFGE [5] and 11.7 Mb by the published genome, which did not include the rDNA repeats [1]. Although the total size was nearly identical to that estimated by PFGE, the sizes of the individual chromosome estimates differed in that the chromosome 5 size had been underestimated by PFGE (assuming that the optical map is indeed more accurate) and the estimates for other chromosomes were smaller for the optical map than for PFGE.

The assembly of the published WBC6 genome consisted of 306 contigs in descending sizes by increasing ID number. These contigs are identified in Genbank and in the Giardia genome database as AACB0200000**1**-AACB02000**306**. Many of the contigs were joined into scaffolds, most frequently because of longer contig-joining clones. Contigs 1 through 70 with the exception of 66 were placed onto the optical map (Fig 1). (A more detailed demonstration of the placing of the contigs can be seen in Supplementary Figure 1). However, contigs 13, 27 and 29 were each split into two fragments. Contig 13a was placed onto chromosome 5, but contig 13b was not placed. The two fragments of contig 27 were placed on chromosomes 5 and 1, respectively. Contig 29a was also placed onto chromosome 5, but contig 29b (39.8 kb) was not placed on the map. There were nine places on the optical map with “negative gaps”, meaning that there was an overlap between two contigs. In each case, we used BLAST comparisons of the adjacent contig sequences to look for regions of sequence identity near the contig ends that would allow them to be joined. We identified regions of sequence overlap for two of the nine contig pairs. The two pairs of contigs with overlapping sequences were joined and then analyzed using the OpGen software. This analysis confirmed that the joined contigs were compatible with the optical map. For the remaining seven pairs of overlapping but unjoined contigs, it is possible that misassembled sequences are present at one or both of the adjacent ends; this remains to be determined.

The contigs smaller than contig 70 (34.2 kb) had too few MluI sites to allow direct placement onto the optical map. However, several were contained in a scaffold of the published genome. These were left in the same positions if they did not contradict the optical map.

With the exception of the end gaps, 95% of the genome is covered by the optical map. The genome assembly omitted the STRs entirely. A subsequent report [8] described the sequences at most of the STRs, but the optical map provides the first accurate assessment of the extent of the STRs not covered by the sequence assembly. The 10 end gaps ranged in size from 2 to 819 kb, with all but one being less than 45 kb in size. The exception is the 819 kb gap from one end of chromosome 5, much of which consists of a repetitive region with MluI fragments 4400-4600 bp in size. We believe this most likely represents the rDNA repeat region. Although the rDNA repeat is 5566 bp in length and has only one MluI site, this is the only repeat region in the optical map compatible with prior data regarding the location of the rDNA sequence in subtelomeric repeats. Prior data indicated that three genotype A1 isolates (Portland, ISR, and CAT) varied greatly in the locations of the rDNA repeats [9]. These repeats are located in the STRs of different chromosomes in different isolates. Even within different cloned lines of the ISR isolate, the sizes of the rDNA-containing subtelomeric regions varied substantially [10]. This is particularly remarkable since the chromosome-internal regions demonstrate very little sequence variability.

A map placing the contigs and supercontigs onto a physical map that was derived by PFGE hybridization studies has recently been published [11]. Many sections of the map in the current study are identical to those obtained by PFGE, but there are a few notable differences. Some of these differences resulted from the fact that the optical map split some of the contigs and supercontigs (Sc) or allowed the placement of additional Sc between two existing Sc. For example, Sc 1764 and 1761 were adjacent to each other at the right end of chromosome 4 on the PFGE-based map, while Sc 1801 was placed between them on the optical map. The differences not explained by splitting the contigs or supercontigs are found primarily in the subtelomeric regions. For example, Sc 1769 and 1767 were located at the left ends of chromosomes 1 and 2, respectively, in the PFGE-based map, but in the optical map, Sc 1769 was at the end of chromosome 2, while Sc 1767 was at the end of chromosome 1. We suggest that these subtelomeric differences may be the result of using different isolates in the two studies.

The optical map has provided independent verification for the majority of the contigs and supercontigs of the WB Giardia genome as originally published [1]. In addition, it has corrected several errors that resulted from misassembly. We believe the increased accuracy of the current map will facilitate improved analysis of recombination and of gene expression that depends on the local context.

Table 1: Contig changes and chromosome sizes and coverage

Joined Sequences[1]	Nucleotide Sequence	Overlap Region	Sequence Length	Chromosome
53_17a	53: 62321-1 17a: 124018-1	53: 2139-1 17: 124628 - 121880	184,200	2
61_29a	61: 1- 46797 29a: 124018-1	61: 45748 - 46797 29: 124018 - 122599	169,395	5

Split Sequences	Nucleotide Sequence	Sequence Length	Chromosome
13a	1-209,610	209,610	5
13b	209,611-266,103	56,493	N/A
17a	1-124,018	124,018	2
17b	124,019-203,025	85,161	1
27a	1-108,674	108,674	5
27b	108,675-148,504	39,830	1
29a	1-123,648	123,648	5
29b	123649-143,621	19,972	N/A

Chrom	Size by PFGE(Mb)	Size by Optical Map	Total coverage	Total coverage excluding end gaps	Total internal gaps	Total end gaps
1	1.6	1.487	90.21%	93.02%	103,720	41,886
2	1.6	1.504	96.64%	99.29%	10,818	40,123
3	2.3	1.944	94.18%	96.68%	64,935	49,032
4	3.0	2.788	94.73%	95.94%	112,140	33,349
5	3.5	4.429	73.97%	92.84%	319,957	842,792
Total	12.0	12.096	87.05%	94.98%	611,570	1,007,182

The sequence numbers refer to the numbers of the 306 contigs in the assembly, matching the final three digits of the GenBank/GiardiaDB entries. Thus, sequence 17 would be AACB02000017. The a or b letter suffix is used for contigs that were split into a and b sections by the optical map.

Acknowledgments

This work was funded in part by the Woods Hole Center for Oceans and Human Health, jointly funded by the National Science Foundation (OCE-0430724) and the National Institute for Environmental Health Sciences (P50 ES012742).

Figure Legends

Fig 1. Optical maps of the five chromosomes are overlaid with the contigs from the WB genome assembly. The upper number indicates the contig number, which matches the last three digits of the GenBank/GiardiaDB entry. Thus, the GenBank entry for contig 1 is AACB0200000**1** The number in parentheses indicates the supercontig/scaffold (sc) number from the published assembly (www.giardiadb.org). The number shown consists of the last four digits of the GiardiaDB entry (eg. Sc 1767 is identified as CH991767 in the GiardiaDB web site). Chromosome 5 is shown on two lines and the long repeat region on the right represents what may be rDNA repeats. The three scaffolds that were split by the optical map (1763, 1767, 1769) are shown in unique colors to display the new locations.

Supplementary Fig 1: The placement of the contigs along each of the chromosomes is shown. Sequence ID is the full name of the contig sequence in the GiardiaDB. “Contig” is the shortened name which corresponds to the unique last three digits of the full name. The “length” column gives the lengths of the individual contig sequence, while the gap gives the number of bp between contigs. A negative gap indicates overlap between adjacent contigs. “Along the chromosome” indicates the cumulative distance across the chromosome as determined by the optical data.

References

[1] Morrison HG, McArthur AG, Gillin FD, Aley SB, Adam RD, Olsen GJ, et al. Genomic minimalism in the early diverging intestinal parasite Giardia lamblia Science. 2007;317:1921-6.

[2] Upcroft JA, Krauer KG, Burgess AG, Dunn LA, Chen N, Upcroft P. Sequence map of the 3-Mb Giardia duodenalis assemblage A chromosome. Chromosome Res. 2009;17:1001-14.

[3] Krauer KG, Burgess AG, Dunn LA, Upcroft P, Upcroft JA. Sequence map of the 2 Mb Giardia lamblia assemblage A chromosome. Journal of Parasitology. 2010;96:660-2.

[4] Smith PD, Gillin FD, Spira WM, Nash TE. Chronic giardiasis: studies on drug sensitivity, toxin production, and host immune response. Gastroenterology. 1982;83:797-803.

[5] Adam RD, Nash TE, Wellems TE. The Giardia lamblia trophozoite contains sets of closely related chromosomes. NucleicAcidsRes. 1988;16:4555-67.

[6] Reslewic S, Zhou S, Place M, Zhang Y, Briska A, Goldstein S, et al. Whole-genome shotgun optical mapping of Rhodospirillum rubrum. Applied & Environmental Microbiology. 2005;71:5511-22.

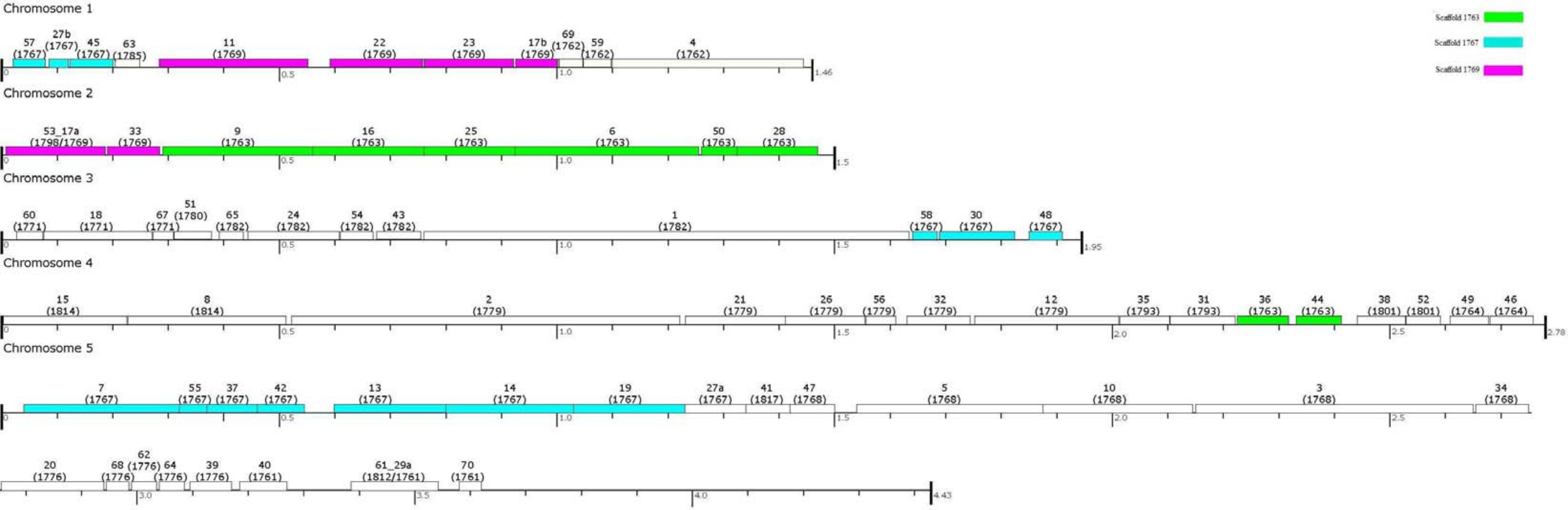
[7] Chen Q, Savarino SJ, Venkatesan MM. Subtractive hybridization and optical mapping of the enterotoxigenic Escherichia coli H10407 chromosome: isolation of unique sequences and demonstration of significant similarity to the chromosome of E. coli K-12. Microbiology. 2006;152:1041-54.

[8] Prabhu A, Morrison HG, Martinez CR, III, Adam RD. Characterisation of the subtelomeric regions of Giardia lamblia genome isolate WBC6. Int J Parasitol. 2007;37:503-13.

[9] Adam RD, Nash TE, Wellems TE. Telomeric location of Giardia rDNA genes. Mol Cell Biol. 1991;11:3326-30.

[10] Adam RD. Chromosome-size variation in Giardia lamblia: the role of rDNA repeats. NucleicAcidsRes. 1992;20:3057-61.

[11] Upcroft JA, Krauer KG, Upcroft P. Chromosome sequence maps of the Giardia lamblia assemblage A isolate WB. Trends in Parasitology. 2010;26:484-91.


Chromosome 1
57 (1767)
27b (1767)
45 (1767)
63 (1785)
11 (1769)
22 (1769)
23 (1769)
17b (1769)
69 (1762)
59 (1762)
4 (1762)
0
0.5
1.0
1.46
Scaffold 1763 🟩
Scaffold 1767 🟦
Scaffold 1769 🟪
Chromosome 2
53_17a (1798/1769)
33 (1769)
9 (1763)
16 (1763)
25 (1763)
6 (1763)
50 (1763)
28 (1763)
0
0.5
1.0
1.5
Chromosome 3
60 (1771)
18 (1771)
67 (1771)
51 (1780)
65 (1782)
24 (1782)
54 (1782)
43 (1782)
1 (1782)
58 (1767)
30 (1767)
48 (1767)
0
0.5
1.0
1.5
1.95
Chromosome 4
15 (1814)
8 (1814)
2 (1779)
21 (1779)
26 (1779)
56 (1779)
32 (1779)
12 (1779)
35 (1793)
31 (1793)
36 (1763)
44 (1763)
38 (1801)
52 (1801)
49 (1764)
46 (1764)
0
0.5
1.0
1.5
2.0
2.5
2.78
Chromosome 5
7 (1767)
55 (1767)
37 (1767)
42 (1767)
13 (1767)
14 (1767)
19 (1767)
27a (1767)
41 (1817)
47 (1768)
5 (1768)
10 (1768)
3 (1768)
34 (1768)
0
0.5
1.0
1.5
2.0
2.5
20 (1776)
68 (1776)
62 (1776)
64 (1776)
39 (1776)
40 (1761)
61_29a (1812/1761)
70 (1761)
3.0
3.5
4.0
4.43