Molecular & Biochemical Parasitology 201 (2015) 1–4

Contents lists available at ScienceDirect

Molecular & Biochemical Parasitology

Short communication

Identification of polymorphic genes for use in assemblage B genotyping assays through comparative genomics of multiple assemblage B *Giardia duodenalis* isolates

Caroline Wielinga [a,*], R.C. Andrew Thompson [b], Paul Monis [c], Una Ryan [a]

[a] *School of Veterinary and Life Sciences, Murdoch University, South Street, Murdoch,Western Australia, 6150, Australia*
[b] *WHO Collaborating Centre for the Molecular Epidemiology of Parasitic Infections, School of Veterinary and Life Sciences, Murdoch University, South Street, Murdoch, Western Australia, 6150, Australia*
[c] *South Australian Water Corporation Adelaide, SA 5000, Australia*

ARTICLE INFO

Article history:
Received 27 February 2015
Received in revised form 6 May 2015
Accepted 8 May 2015
Available online 21 May 2015

Keywords:
Giardia
Assemblage B
Genome
Assembly
Annotation
Genotyping

ABSTRACT

Giardia duodenalis assemblage B is potentially a zoonotic parasite. The characterisation and investigation of isolates has been hampered by greater genetic diversity of assemblage B, limiting the application and utility of current genotyping loci. Since whole genome sequencing is the optimal high-throughput method for gene identification, the present study sequenced assemblage B isolate BAH15c1 and compared the sequence to the draft GS references to identify polymorphic genes for potential use in genotyping assays. The majority of the genome sequence was conserved between the two isolates, producing 508 contigs of 10.4 Mb with 4968 genes. Seventy polymorphic genes for potential use in genotyping assays were identified ranging in variation from elongation factor 1 α, which was the most conserved, through to triose phosphate isomerase, which was the most variable.

Giardia duodenalis (Giardia intestinalis, Giardia lamblia) is a common intestinal parasite of humans and mammals worldwide. Genetic analyses to date segregate what is hypothesised to be a species complex into predominantly host specific assemblages – A, B, C, D, E, F, G, H [1–3]. Assemblages A and B differ from the other assemblages in that they can be zoonotic [4].

The analyses of assemblage A have successfully progressed toward reproducible multiloci identification and characterisation of isolates into sub-assemblages AI, AII and AIII, but analyses of assemblage B have been hampered by the greater diversity encountered between and within these isolates [5–12]. Assemblage B is reported to have 50 times more allelic sequence heterozygosity (ASH) than assemblage A [13], which complicates analyses of the tetraploid organism [14].

Due to the greater genetic diversity of assemblage B, different, more conserved, loci have been sought than the relatively variable loci applied to the analyses of the less divergent assemblage A [15]. It was hypothesized that genes with lower substitution rates may provide a clearer understanding of the potential subgroups within assemblage B. Whole genome sequencing technology enables the entire genome to be examined for more suitable genotyping loci. To date there have been two *G. duodenalis* assemblage B genome assemblies published, both were the GS isolate [13,16]. Here we compare the two draft GS reference assemblies with the assembly of a cloned cultured assemblage B isolate (BAH15c1) and identify polymorphic genes for potential new intra-assemblage B genotyping.

Assemblage B isolate BAH15c1, obtained from a human in Australia, was cultured and DNA extracted as previously described [15,17]. Preparation of non-paired end libraries, template DNA capture beads and sequencing of enriched DNA capture beads with titanium chemistry on a 454 Life Sciences' sequencer were as per the manufacturer's protocols (454 Life Sciences GS Junior System – Rapid Library Preparation Method Manual, emPCR Amplification Method Manual Lib-L and Sequencing Method Manual; March 2012, Roche Applied Science, Mannheim Germany) at the State Agricultural Biotechnology center, Murdoch University. Sequencing generated 250,000 reads (average 430 bp), totalling 109 Mb, equal to 9x coverage.

DNA sequence reads were assembled *de novo* twice, once with Newbler v2.5 (454 Life Sciences) and once with de Bruijn (CLC bio, Qiagen) software and then the two contig sets were combined,

* Corresponding author. Tel.: +61 8 9360 2691.
E-mail address: c.wielinga@murdoch.edu.au (C. Wielinga).

http://dx.doi.org/10.1016/j.molbiopara.2015.05.002
0166-6851/

aligned and assembled in Geneious (v6.1.5) to generate a single second order consensus contig set as recommended by Kumar and Blaxter [18]. The second order consensus contigs were manually checked to ensure uniform coverage of high identity with both Newbler and de Bruijn contigs (no internal regions >1kb with only one first order type, pair-wise alignment identity >97%, >96% in overlaps). [Software parameters – de Bruijn CLC bio, standard parameters, min. output 500 bp; Newbler v2.5 non-standard parameters, min. overlap identity 97% (max. before the number of reads assembled markedly reduced), CPU = 0 (used all CPUs), seed step = 1 (max. sensitivity), seed length = 16 (max. selectivity), seed count = 1 (default), min. overlap 40 bp (longest min. overlap possible) and min. output 500 bp (>average read length); Geneious v6.1.5, non-standard parameters (to allow for gaps and possible partial alignments for manual assessment): allow gaps (max. per read 20%, max. size 200 bp), min. overlap 40 bp, min. overlap identity 90%, max. mismatches per read 40%].

The *de novo* assembly was used in preference to comparative assembly (reference guided assembly) so that structural variations between assemblage assemblies could be identified. The use of two distinct *de novo* assembly methods that were combined into a second order consensus contig set was preferred to compare the assemblies. The Newbler and de Bruijn *de novo* assemblies had similar metrics (2124 contigs, 10.5 Mb, N50 = 10 kb, max. contig 51 kb and 2089 contigs, 10.3 Mb, N50 = 10 kb, max. contig 51 kb respectively) and when aligned to generate the second order consensus contigs, most of the alignments (90%), had pair-wise alignment identity >99%, demonstrating the similarity of the assemblies. The second order consensus contig set had improved metrics of 840 contigs, 11.3Mb, N50 = 27 kb, max. contig 108 kb, illustrating a further benefit of the combined method. Although the agreement between the Newbler and de Bruijn assembly methods was good, there were variations observed. On 40 occasions, small deletions (5–150 bp) and sequence reversals (50–100 bp) at the end of a contig were observed with the de Bruijn method, and there were 7 large (1.5–14 kb) and 14 small (average 230 bp) alignment chimeras. Many of the chimeras were in or near genes of multiple copies. Of the large Newbler/de Bruijn alignment chimeras, all aligned with the Newbler-assembled draft GS references in the Newbler format. Six percent of the Newbler and de Bruijn *de novo* contigs were not incorporated into the second order consensus contigs (mostly Newbler, 83%).

The BAH15c1 second order consensus contig set was then aligned to each draft GS reference (draft GS reference 1 and 2, accession numbers ACGJ00000000 and AHGT00000000) [13,16] using Geneious v6.1.5. Second order consensus contigs consecutively aligning along a reference contig were joined where pair-wise alignment identity was >97% (or >97% at the join for alignments with chimeric ends) and gaps were <1kb. Alignments were completed for both draft GS references and configurations were accepted if they were supported by both references, or by one reference if the other reference was not in disagreement (merely fragmented or absent) and not in a region with repeating genes. [Geneious v6.1.5 parameters - non-standard (to allow for gaps and possible partial alignments for manual assessment), iterate 10 times, allow gaps, (max. per read 20%, max. size 200 bp), max. mismatches 20%]. For the draft GS reference 1 (n = 2,931 contigs) a workable subset of contigs was first established by running a *de novo* assembly on the 2931 contigs to determine those contigs that were potentially redundant. Small contigs internal to the larger ones with >96% pair-wise alignment identity, were put aside and the remaining contigs (n = 1,608) were used in further analyses as their original sequence (not as a consensus). [Geneious v6.1.5, non-standard parameters (to increase the alignment identities): allow gaps, (max. per read 10%, max. size 25 bp), min. overlap 100 bp, min. overlap identity 87%, max. mismatches 20%]. The comparative alignment and joining of the second order consensus contigs with both draft GS reference 1 or 2 produced very similar results. Both draft GS references had similar numbers of large contigs (175 and 167 contigs > 20 kb respectively). The resultant second order consensus contig set, had further improved metrics of 508 contigs, 10.4 Mb, N50 = 50 kb, max. contig 184 kb. Most of the assembled genome, 9.5 Mb (91%), was contained within the first 200 contigs. Of the original 840 second order consensus contigs initially aligned to the draft GS references, 473 (56%) could be joined by comparative alignment (to make 141 contigs) and 367 (44%) could not. Those contigs not joined ranged in size from 0.5–76 kb (median = 2.5 kb), totalling 3 Mb. Although both of the draft GS reference alignments produced similar results, there were 15 notable chimeric alignments (between contig pairs ACGJ01000930 and AHHH01000001; ACGJ01002492 and AHHH01000001; ACGJ01002923 and AHHH01000012; ACGJ01002483 and AHHH01000009; ACGJ01002330 and AHHH01000073; ACGJ01002231 and AHHH01000016; ACGJ01002568 and AHHH01000080; ACGJ01002287 and AHHH01000015; ACGJ01002893 and AHHH01000393; ACGJ01002297 and AHHH01000066; ACGJ01002930 and AHHH01000064; ACGJ01002568 and AHHH01000021; ACGJ01002923 and AHHH01000195; ACGJ01000719 and AHHH01000098; ACGJ01001465 and AHHH01000033). Of these, BAH15c1 alignments agreed with more draft GS reference 1 alignments (n = 6) than draft GS reference 2 (n = 4) and some with neither (n = 5) due to gaps. In several instances, the draft GS reference 1 and 2 chimeric swap occurred in copies of genes – such as in the thioredoxin peroxidase gene, (ACGJ01000930 and AHHH01000001, AHHH01000106) and the histone gene (ACGJ01001465 and AHHH01000033). Since both draft GS reference 1 and 2 had sound assembly methodology (16× coverage and Sanger sequencing and 50× coverage with paired end sequencing respectively) these inconsistencies were inconclusive and require further GS analysis. There were also 5 occasions where BAH15c1 and draft GS reference 2 did not align but the draft GS reference 1 was too fragmented for comparison. Other variations included two examples of missing data and a reversed section [draft GS reference 1 had an 8 kb gap between ACGJ01000948 and ACGJ01002392 relative to draft GS reference 2 (AHHH1000111) and BAH15c1 contig107; and draft GS reference 2 had a 40 kb gap next to AHHH01000146 relative to draft GS reference 1 (ACGJ01002915) and BAH15c1 contig041; draft GS reference 2 on AHHH01000016, had a 6.5 kb region in reverse relative to draft GS reference 1 (ACGJ01002231) and BAH15c1 contig169].

The second order BAH15c1 consensus contigs were then annotated by transferring annotations from both references and confirming open reading frames (ORF's) in Geneious [v6.1.5, 65% transfer similarity (to include gaps), standard parameters, ORF finder]. Draft GS reference 1, reported 4470 protein coding ORF's across 454 contigs and draft GS reference 2, 6098 across 492 contigs. In the present study, comparative alignment and annotation with the draft GS references produced 4886 protein coding ORF's on 348 contigs (Supplementary Table 1). The majority of the ORF's (94%) were on the first 200 contigs. Most ORF's (81%) were confirmed by both draft GS references, but 18% were annotated from only one draft GS reference (mostly draft GS reference 2, 68%) (Supplementary Table 1). Comparison of the draft GS reference ORF's together and relative to BAH15c1 was complicated by non-standard nomenclature including 180 ORF's typed by draft GS reference 1 but hypothetical in draft GS reference 2 (Supplementary Table 1). A comparison of the draft GS reference 1 and 2 ORF's showed that the variation was due to the numbers of copies of genes, where half of the difference (806/1,628 ORF's), was due to draft GS reference 2 having increased numbers of kinases (from 291 to 341), ankyrins/protein 21.1 (from 224 to 383) and variant specific surface