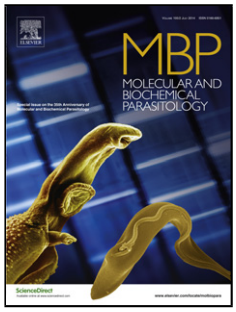


Title: Identification of polymorphic genes for use in assemblage B genotyping assays through comparative genomics of multiple assemblage B *Giardia duodenalis* isolates.

Author: Caroline Wielinga RCAndrew Thompson Paul Monis Una Ryan

PII: S0166-6851(15)30001-3
DOI: http://dx.doi.org/doi:10.1016/j.molbiopara.2015.05.002
Reference: MOLBIO 10898

To appear in: *Molecular & Biochemical Parasitology*

Received date: 27-2-2015
Revised date: 6-5-2015
Accepted date: 8-5-2015

Please cite this article as: Wielinga Caroline, Thompson RCAndrew, Monis Paul, Ryan Una.Identification of polymorphic genes for use in assemblage B genotyping assays through comparative genomics of multiple assemblage B Giardia duodenalis isolates.*Molecular and Biochemical Parasitology* http://dx.doi.org/10.1016/j.molbiopara.2015.05.002

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

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

[a]*School of Veterinary and Life Sciences, Murdoch University, South Street, Murdoch, Western Australia, 6150*

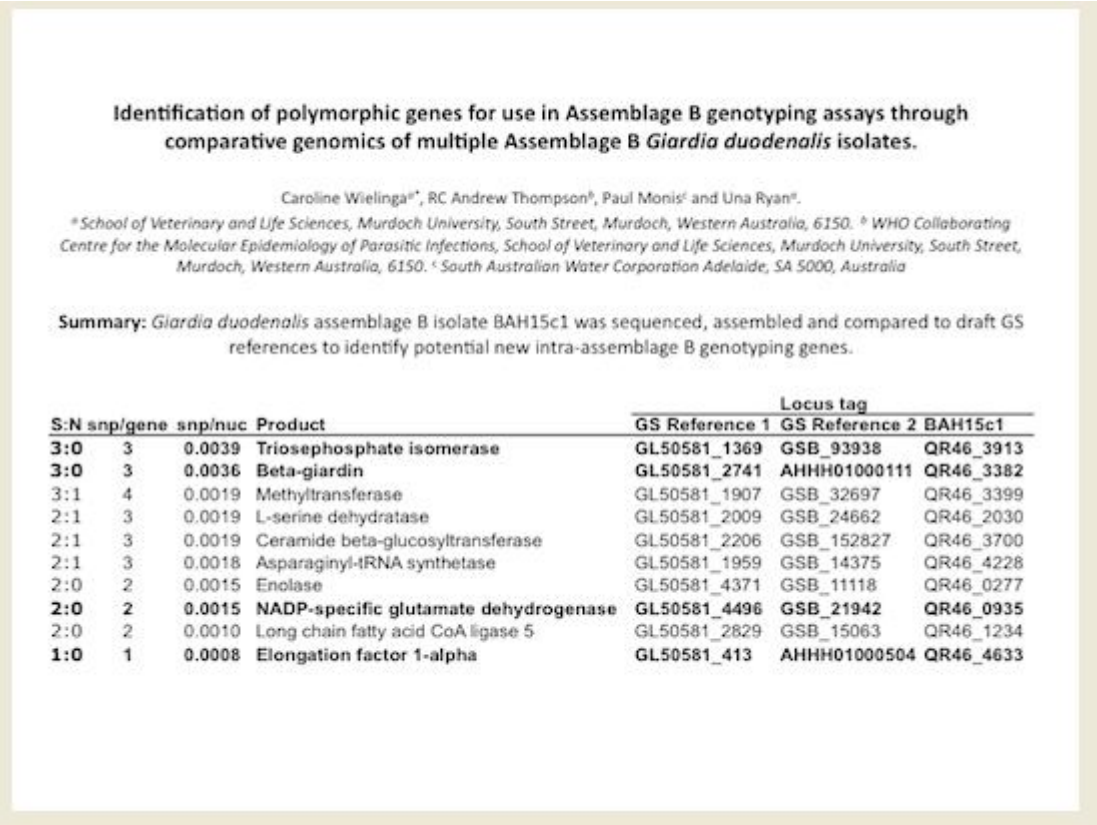
[b] *WHO Collaborating Centre for the Molecular Epidemiology of Parasitic Infections, School of Veterinary and Life Sciences, Murdoch University, South Street, Murdoch, Western Australia, 6150*

[c]*South Australian Water Corporation Adelaide, SA 5000, Australia*

*Corresponding author: Tel.: +61 8 9360 2691

E-mail address:c.wielinga@murdoch.edu.au

Graphical abstract



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

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

[a] *School of Veterinary and Life Sciences, Murdoch University, South Street, Murdoch, Western Australia, 6150.* [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.* [c] *South Australian Water Corporation Adelaide, SA 5000, Australia*

Summary: *Giardia duodenalis* assemblage B isolate BAH15c1 was sequenced, assembled and compared to draft GS references to identify potential new intra-assemblage B genotyping genes.

				Locus tag		
S:N	snp/gene	snp/nuc	Product	GS Reference 1	GS Reference 2	BAH15c1
3:0	**3**	**0.0039**	**Triosephosphate isomerase**	**GL50581_1369**	**GSB_93938**	**QR46_3913**
3:0	**3**	**0.0036**	**Beta-giardin**	**GL50581_2741**	**AHHH01000111**	**QR46_3382**
3:1	4	0.0019	Methyltransferase	GL50581_1907	GSB_32697	QR46_3399
2:1	3	0.0019	L-serine dehydratase	GL50581_2009	GSB_24662	QR46_2030
2:1	3	0.0019	Ceramide beta-glucosyltransferase	GL50581_2206	GSB_152827	QR46_3700
2:1	3	0.0018	Asparaginyl-tRNA synthetase	GL50581_1959	GSB_14375	QR46_4228
2:0	2	0.0015	Enolase	GL50581_4371	GSB_11118	QR46_0277
2:0	**2**	**0.0015**	**NADP-specific glutamate dehydrogenase**	**GL50581_4496**	**GSB_21942**	**QR46_0935**
2:0	2	0.0010	Long chain fatty acid CoA ligase 5	GL50581_2829	GSB_15063	QR46_1234
1:0	**1**	**0.0008**	**Elongation factor 1-alpha**	**GL50581_413**	**AHHH01000504**	**QR46_4633**

16 ***Highlights***

17 • An additional *Giardia duodenalis* assemblage B genome sequence for *Giardia* analyses.
18 • *De novo* assembly comparison of Newbler and de Bruijn methods.
19 • *De novo* assembly comparison of sample and existing draft GS references.
20 • Annotation comparison of sample and existing draft GS references.
21 • Seventy potential intra-assemblage B genotyping genes identified.

22

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 10.4Mb 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.

Keywords: Giardia, assemblage B, genome, assembly, annotation, genotyping

*Giardia duodenalis (G. intestinalis, G. 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 multilociidentification and characterisation of isolatesinto sub-assemblages AI, AII and AIII, but analyses of assemblage B have been hampered by the greater diversity encountered between and within theseisolates[5-12]. Assemblage B is reported to have 50 times moreallelic 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 therelatively 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 withinassemblage 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 genomeassemblies published, both were theGS isolate[13, 16]. Here we compare the two draft GS reference assemblies with the assembly ofa cloned culturedassemblage 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'sequencerwere as perthe manufacturer's protocols (454 Life Sciences *GS Junior System*-Rapid Library Preparation Method Manual, emPCR Amplification Method Manual Lib-L and Sequencing Method Manual;March2012, Roche Applied Science,Mannheim Germany) at the

State Agricultural Biotechnology Centre, Murdoch University.Sequencing generated 250,000 reads (average 430bp), totalling 109Mb, equal to 9x coverage.

DNA sequence reads were assembled*de novo*twice,once withNewbler v2.5 (454 Life Sciences)and once with de Bruijn(CLC bio, Qiagen)software andthen the two contig sets were combined,aligned and assembledin 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 weremanually checked to ensure uniform coverage of high identity with both Newbler and de Bruijncontigs (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 500bp; 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 40bp (longest min. overlap possible) and min. output 500bp (>average read length); Geneiousv6.1.5,non-standard parameters (to allow for gaps and possible partial alignments for manual assessment): allow gaps (max. per read 20%, max. size 200bp), min. overlap 40bp, 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 werecombined into asecond orderconsensuscontig set was preferred to compare theassemblies. The Newbler and de Bruijn *de novo* assemblies had similar metrics (2,124 contigs, 10.5Mb, N50=10kb, max. contig 51kb and 2,089 contigs, 10.3Mb, N50=10kb, max. contig 51kb respectively) and when alignedto generate the second order consensus contigs, mostof the alignments (90%),hadpair-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=27kb, max. contig 108kb, illustrating a furtherbenefit of the combined

method.Although the agreement between the Newbler and de Bruijnassembly methods was good, there were variations observed. On 40 occasions,small deletions (5-150bp) and sequence reversals (50-100bp) at the end of a contig were observed with the de Bruijn method, and there were 7 large (1.5kb-14kb)and 14 small (average 230bp) alignment chimeras.Many of the chimeras were in or near genes of multiple copies. Of thelarge 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 Geneiousv6.1.5.Second order consensus contigs consecutively aligning alonga reference contigwere joined wherepair-wise alignment identitywas >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 200bp), 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 2,931 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 25bp), min. overlap 100bp, 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 >20kb respectively). The resultant second order consensus contig set, had further improved metrics of 508 contigs, 10.4Mb, N50=50kb, max. contig 184kb. Most of the assembled genome,9.5Mb (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 0.5-76kb(median=2.5kb), totalling 3Mb. 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 (16x coverage andSanger sequencing and50x 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 8kb gap between ACGJ01000948 and ACGJ01002392 relative to draft GS reference 2 (AHHH1000111) and BAH15c1 contig107; and draft GS reference 2 had a 40kb gap next to AHHH01000146 relative to draft GS reference 1 (ACGJ01002915) and BAH15c1 contig041; draft GS reference 2 on AHHH01000016, had a 6.5kb 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 fromboth 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 4,470 protein coding ORF's across 454 contigs and draft GS reference 2, 6,098 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 proteins (vsps) (41 to 638) and the remaining difference from draft GS reference 2 having more ORF's with multiple copies and those with copies having more replicates.Of the draft GS reference protein coding ORF's,95% (4,253/4,470) of the draft GS reference 1 and 79% (4,807/6,098) of the draft GS reference 2 ORF's, were detected in BAH15c1. Of the ORF's not detected, the majority for draft GS reference 1 were hypothetical ORF's (49%, 107/217) and vsp/protein 21.1/kinase NEK (25%, 55/217) and the majority for draft GS reference 2

were vsp/ankyrin/serine-threonine kinases (53%, 679/1,291) and hypothetical ORF's (23%, 284/1,291).The amount of sequencing coverage was proportional to the vsp/protein 21.1/kinase family gene representatives detected due to their location in difficult to assemble gene regions where it is hypothesized a greater frequency of recombination maintains their variation[19]. There werealso 37 new proposed hypothetical genes with ORF's >1kb identified in BAH15c1 that were not annotated in draft GS references(Supplementary Table 1). The majority of these were located on the ends of contigs and therefore the difference in annotation among the isolates could be due to fragmentation or assembly variations of the repeating regions.The Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under accession JXTI00000000.

To locate the new genes for use in intra-assemblage B genotyping, ORF's were sorted based on representation by references, length and percent identity. ORF's were included if they were represented by both draft GS references to allow comparison of the references and as a validation of reproducibility. Degeneracy within and variation between the draft GS references was excluded as potential allelic variation or divergence within an isolate, which would complicate genotyping analyses. The length range selected was between 1,300-2,500bp; greater than the currently used triose phosphate isomerase (*tpi*, 744bp) and beta giardin (*bg*, 822bp) genes, to allow for an increased number of sites for variation in divergent samples, and similar to elongation factor 1 α (*ef1*α, 1329bp) and the currently used glutamate dehydrogenase (*gdh*, 1350bp),but small enough to potentially PCR amplify in part or in segments. Substitutions per nucleotide between the sample and draft GS references were selected to be less than 0.00365 substitutions/nucleotide, which was equal to or less than the already used variable *tpi* and *bg* genes (0.00388 and 0.00365 substitutions/nucleotide respectively). Substitutions per gene were selected to be greater than one substitution/gene to select genes with potential to distinguish samples and the ratio of non-synonymous to synonymous substitutions was restricted to 1:2 so that no more than one third of the total substitutions in a gene were non-synonymous, to select for conserved

genes(dN/dS<1).Potential new genotyping genes for intra-assemblage B analyses are listed in Table 1. Of the total 4886 protein coding ORF's, 4024 were present in both draft GS references, 70% of these were smaller (n=2460) or larger (n=1024) than the designated preferred length range and a further 28% had too high or too low a rate of substitutions between the draft GS references or sample and references, leaving 2% or 70 ORF's in the preferred range. The 70 ORF's consisted of enzymes (46%), hypothetical products (33%), binding proteins (20%) and structural proteins (1%).

In conclusion, since there has been no phylogenetic consensusof the current genotyping genes,more genes have been sought. By starting with a large gene selection it will be possible with additional genome samples to determine the phylogenetic consensus,a subset of new genotyping genes and which of the current genotyping genes (*gdh, tpi, bg*), if any, are informativefor intra-assemblage B analyses.

199 **References**

200

201 1. Monis, P.T., et al., *Molecular systematics of the parasitic protozoan Giardia intestinalis.*
202 Molecular Biology and Evolution, 1999. **16**(9): p. 1135-1144.
203 2. Gaydos, J.K., et al., *Novel and canine genotypes of Giardia duodenalis in harbor seals (*
204 *Phoca vitulina richardsi).* The Journal of Parasitology, 2008. **94**(6): p. 1264-8.
205 3. Lasek-Nesselquist, E., D.M. Welch, and M.L. Sogin, *The identification of a new Giardia*
206 *duodenalis assemblage in marine vertebrates and a preliminary analysis of G. duodenalis*
207 *population biology in marine systems.* International Journal for Parasitology, 2010.
208 **40**(9): p. 1063-74.
209 4. Ryan, U. and S.M. Caccio, *Zoonotic potential of Giardia.* Int J Parasitol, 2013. **43**(12-13):
210 p. 943-56.
211 5. Caccio, S.M., et al., *Multilocus genotyping of Giardia duodenalis reveals striking*
212 *differences between assemblages A and B.* International Journal for Parasitology, 2008.
213 **38**(13): p. 1523-1531.
214 6. Lasek-Nesselquist, E., et al., *Molecular characterization of Giardia intestinalis haplotypes*
215 *in marine animals: variation and zoonotic potential.* Diseases of Aquatic Organisms,
216 2008. **81**(1): p. 39-51.
217 7. Geurden, T., et al., *Multilocus genotyping of Cryptosporidium and Giardia in non-*
218 *outbreak related cases of diarrhoea in human patients in Belgium.* Parasitology, 2009.
219 **136**(10): p. 1161-8.
220 8. Lalle, M., et al., *High genetic polymorphism among Giardia duodenalis isolates from*
221 *Sahrawi children.* Transactions of the Royal Society of Tropical Medicine and Hygiene,
222 2009. **103**(8): p. 834-838.
223 9. Sprong, H., S.M. Caccio, and J.W. van der Giessen, *Identification of zoonotic genotypes of*
224 *Giardia duodenalis.* Public Library of Science Neglected Tropical Diseases, 2009. **3**(12):
225 p. e558.
226 10. Levecke, B., et al., *Molecular characterisation of Giardia duodenalis in captive non-*
227 *human primates reveals mixed assemblage A and B infections and novel polymorphisms.*
228 International Journal for Parasitology, 2009. **39**(14): p. 1595-1601.
229 11. Lebbad, M., et al., *From mouse to moose: multilocus genotyping of Giardia isolates from*
230 *various animal species.* Veterinary Parasitology, 2010. **168**(3-4): p. 231-9.
231 12. Lebbad, M., et al., *Multilocus genotyping of human Giardia isolates suggests limited*
232 *zoonotic transmission and association between assemblage B and flatulence in children.*
233 PLoS Negl Trop Dis, 2011. **5**(8): p. e1262.
234 13. Franzen, O., et al., *Draft genome sequencing of giardia intestinalis assemblage B isolate*
235 *GS: is human giardiasis caused by two different species?* Public Library of Science
236 Pathogens, 2009. **5**(8): p. e1000560.
237 14. Ankarklev, J., S.G. Svard, and M. Lebbad, *Allelic sequence heterozygosity in single Giardia*
238 *parasites.* BMC Microbiol, 2012. **12**: p. 65.
239 15. Wielinga, C., et al., *Multi-locus analysis of Giardia duodenalis intra-Assemblage B*
240 *substitution patterns in cloned culture isolates suggests sub-Assemblage B analyses will*
241 *require multi-locus genotyping with conserved and variable genes.* Int J Parasitol, 2011.
242 **41**(5): p. 495-503.
243 16. Adam, R.D., et al., *Genome sequencing of Giardia lamblia genotypes A2 and B isolates (DH*
244 *and GS) and comparative analysis with the genomes of genotypes A1 and E (WB and Pig).*
245 Genome Biol Evol, 2013. **5**(12): p. 2498-511.

17. Binz, N., et al., *A Simple Method for Cloning Giardia-Duodenalis from Cultures and Fecal Samples.* Journal of Parasitology, 1991. **77**(4): p. 627-631.
18. Kumar, S. and M.L. Blaxter, *Comparing de novo assemblers for 454 transcriptome data.* BMC Genomics, 2010. **11**: p. 571.
19. Manning, G., et al., *The minimal kinome of Giardia lamblia illuminates early kinase evolution and unique parasite biology.* Genome Biol, 2011. **12**(7): p. R66.

S:N	snp/gene	snp/nuc	Product	Locus tag GS Reference 1	GS Reference 2	BAH15c1
3:0	**3**	**0.0039**	**Triosephosphate isomerase**	**GL50581_1369**	**GSB_93938**	**QR46_3913**
3:0	**3**	**0.0036**	**Beta-giardin**	**GL50581_2741**	**AHHH01000111**	**QR46_3382**
5:0	5	0.0036	Protein 21.1	GL50581_1174	GSB_3760	QR46_2743
7:2	9	0.0036	Hypothetical protein	GL50581_4377	GSB_150994	QR46_0284
4:2	6	0.0036	Tubulin binding cofactor C family protein	GL50581_2555	GSB_4789	QR46_2004
7:0	7	0.0036	Protein 21.1	GL50581_2864	GSB_16220	QR46_1933
7:1	8	0.0036	Hypothetical protein	GL50581_2071	GSB_17405	QR46_2297
7:1	8	0.0036	Glucose-6-phosphate 1-dehydrogenase	GL50581_4318	GSB_8682	QR46_0716
7:1	8	0.0035	Hypothetical protein	GL50581_1591	GSB_151824	QR46_4239
4:1	5	0.0035	Transcription factor tfiib/IIIB subunit BRF	GL50581_3049	GSB_4125	QR46_2981
4:2	6	0.0034	Histone acetyltransferase Elp3	GL50581_2520	GSB_16639	QR46_2377
6:2	8	0.0034	Hypothetical protein	GL50581_2685	GSB_16602	QR46_3104
4:1	5	0.0034	Hypothetical protein	GL50581_2842	GSB_15067	QR46_1247
6:0	6	0.0033	Protein tyrosine phosphatase	GL50581_1988	GSB_25035	QR46_4378
6:1	7	0.0032	Oligosaccharyl transferase STT3 subunit	GL50581_675	GSB_137685	QR46_1190
4:1	5	0.0032	Putative SP-RING zinc finger protein	GL50581_3949	GSB_10261	QR46_3749
6:1	7	0.0032	Superfamily I DNA helicase/HCS1	GL50581_3862	GSB_24376	QR46_1829
6:0	6	0.0031	Adenylate cyclase	GL50581_13	GSB_14367	QR46_2534
4:0	4	0.0030	Hypothetical protein	GL50581_1424	GSB_4149	QR46_3284
4:1	5	0.0030	TCP-1 chaperonin subunit epsilon/cpn60	GL50581_1213	GSB_11992	QR46_4181
3:1	4	0.0030	Basal body protein	GL50581_4364	GSB_150986	QR46_0269
5:0	5	0.0030	Hypothetical protein	GL50581_1966	GSB_151829	QR46_4233
5:1	6	0.0030	Chromosome segregation SMC	GL50581_3867	GSB_14971	QR46_1833
4:0	4	0.0029	Hypothetical protein	GL50581_155	GSB_152343	QR46_0678
6:0	6	0.0029	Kinase/PLK/Ser/thr protein kinase	GL50581_1588	GSB_104150	QR46_0134
4:0	4	0.0028	Kinase/NEK/Ser/thr protein kinase	GL50581_4111	GSB_150268	QR46_0073
3:1	4	0.0028	6-phosphogluconate dehydrogenase	GL50581_4317	GSB_14759	QR46_0715
4:0	4	0.0028	Histone methylacetyltransferase MYST1/NuA3/SAS3	GL50581_2825	GSB_17263	QR46_1230
4:1	5	0.0028	Protein 21.1	GL50581_2341	GSB_150419	QR46_2056
4:1	5	0.0028	Chromosome segregation SMC/spindle pole	GL50581_995	GSB_87149	QR46_1727
5:0	5	0.0028	Hypothetical protein	GL50581_4380	GSB_150996	QR46_0287
5:1	6	0.0027	Nucleolar GTP-binding protein 2	GL50581_4386	GSB_89887	QR46_0293
3:1	4	0.0027	Hypothetical protein	GL50581_2107	GSB_14675	QR46_1093
5:1	6	0.0026	Hypothetical protein	GL50581_783	GSB_5567	QR46_0384
4:1	5	0.0025	Hypothetical protein	GL50581_2035	GSB_15594	QR46_1954
4:2	6	0.0024	Kinesin motor domain protein/Kinesin-16	GL50581_2553	GSB_16161	QR46_2002
5:1	6	0.0024	Hypothetical protein	GL50581_1001	GSB_153374	QR46_1720
3:0	3	0.0023	Elongation IF 5C/EIF4 gamma/eIF2b epsilon	GL50581_996	GSB_7522	QR46_1726
3:0	3	0.0023	UsoAp/Chromosome segregation SMC	GL50581_1453	GSB_17536	QR46_3061
3:0	3	0.0023	Hypothetical protein	GL50581_1146	GSB_94224	QR46_1414
3:0	3	0.0022	Hypothetical protein	GL50581_2675	GSB_151385	QR46_3094
3:1	4	0.0022	Hypothetical protein	GL50581_2111	GSB_14677	QR46_1088
3:0	3	0.0022	UTP-glucose-1-phosphate uridyltransferase	GL50581_2296	GSB_29307	QR46_1447
2:1	3	0.0022	Hypothetical protein	GL50581_2229	GSB_33434	QR46_0628
4:1	5	0.0022	Coiled-coil protein/ATP-binding protein	GL50581_3950	GSB_17508	QR46_3750
2:1	3	0.0021	Lipase	GL50581_109	GSB_152676	QR46_2880
2:1	3	0.0021	Hypothetical protein	GL50581_2775	GSB_152857	QR46_1390
2:1	3	0.0020	Hypothetical protein	GL50581_4347	GSB_92760	QR46_0251
5:0	5	0.0020	Protein phosphatase 2A B'/PP2A Wdb1	GL50581_1938	GSB_16443	QR46_4420
3:1	4	0.0020	Mn-dependent inorganic pyrophosphatase	GL50581_4366	GSB_8163	QR46_0271

3:1	4	0.0019	Methyltransferase	GL50581_1907	GSB_32697	QR46_3399
2:1	3	0.0019	L-serine dehydratase	GL50581_2009	GSB_24662	QR46_2030
2:1	3	0.0019	Ceramide beta-glucosyltransferase	GL50581_2206	GSB_152827	QR46_3700
2:1	3	0.0018	Asparaginyl-tRNA synthetase	GL50581_1959	GSB_14375	QR46_4228
3:1	4	0.0018	Protein 21.1	GL50581_3875	GSB_16326	QR46_1840
2:1	3	0.0018	Dipeptidyl-peptidase I precursor	GL50581_3606	GSB_8741	QR46_2905
2:1	3	0.0018	Protein required for cell viability	GL50581_4516	GSB_8782	QR46_3033
2:1	3	0.0018	Kinase/STE20/Ser/thre kinase	GL50581_2522	GSB_2796	QR46_2379
3:0	3	0.0017	Chromosome segregation SMC/coiled-coil	GL50581_2523	GSB_151957	QR46_2380
2:1	3	0.0017	Hypothetical protein	GL50581_4350	GSB_4768	QR46_0255
2:1	3	0.0017	Coiled-coil protein	GL50581_1647	GSB_9659	QR46_2795
2:1	3	0.0015	Retinoic acid induced/MIZ/SP-RING Zn finger	GL50581_3324	GSB_150734	QR46_4349
2:0	2	0.0015	Hypothetical protein	GL50581_2592	GSB_152994	QR46_3511
2:0	2	0.0015	Enolase	GL50581_4371	GSB_11118	QR46_0277
2:0	**2**	**0.0015**	**NADP-specific glutamate dehydrogenase**	**GL50581_4496**	**GSB_21942**	**QR46_0935**
2:0	2	0.0014	Kinase/NEK/Ser/thr protein kinase	GL50581_4387	GSB_137719	QR46_0294
3:0	3	0.0014	Hypothetical protein	GL50581_4351	GSB_33672	QR46_0256
2:0	2	0.0013	Vacuolar ATP synthase subunit C	GL50581_2882	GSB_87058	QR46_2961
2:0	2	0.0013	Putative KRI1-like family protein	GL50581_4473	GSB_10569	QR46_0308
2:0	2	0.0012	Kinase/NEK	GL50581_451	GSB_5489	QR46_0878
2:0	2	0.0011	Kinase/NEK/Ser/thr protein kinase	GL50581_4391	GSB_16733	QR46_0298
2:0	2	0.0010	Long chain fatty acid CoA ligase 5	GL50581_2829	GSB_15063	QR46_1234
2:0	2	0.0010	Hypothetical protein	GL50581_4545	GSB_154089	QR46_1577
1:0	**1**	**0.0008**	**Elongation factor 1-alpha**	**GL50581_413**	**AHHH01000504**	**QR46_4633**

Table 1: Potential new intra-assemblage B genotyping genes for *Giardia duodenalis* analyses. Listed in order of decreasing substitution rate [substitutions (single nucleotide polymorphisms, snps) between reference and sample per gene and pernucleotide, with ratio of synonymous (S) and non-synonymous (N) substitutions per gene]. Existing genotyping genes for comparison in bold. Corresponding locus tags listed unless gene unannotated where corresponding contig number is shown.