Wielinga, C., Ryan, U., Thompson, R.C. and Monis, P. (2011) ***Multi-locus analysis of Giardia duodenalis intra-Assemblage B substitution patterns in cloned culture isolates suggests sub-Assemblage B analyses will require multi-locus genotyping with conserved and variable genes.*** **International Journal for Parasitology, 41 (5). pp. 495-503.**

Multi-locus analysis of *Giardia duodenalis* intra-Assemblage B substitution patterns in cloned culture isolates suggests sub-Assemblage B analyses will require multi-locus genotyping with conserved and variable genes

Caroline Wielinga, Una Ryan, R.C. Andrew Thompson, Paul Monis

PII: S0020-7519(10)00385-1
DOI: 10.1016/j.ijpara.2010.11.007
Reference: PARA 3228

To appear in: *International Journal for Parasitology*

Received Date: 29 September 2010
Revised Date: 23 November 2010
Accepted Date: 30 November 2010

Please cite this article as: Wielinga, C., Ryan, U., Thompson, R.C.A., Monis, P., Multi-locus analysis of *Giardia duodenalis* intra-Assemblage B substitution patterns in cloned culture isolates suggests sub-Assemblage B analyses will require multi-locus genotyping with conserved and variable genes, *International Journal for Parasitology* (2010), doi: 10.1016/j.ijpara.2010.11.007

Multi-locus analysis of *Giardia duodenalis* intra-Assemblage B substitution patterns in cloned culture isolates suggests sub-Assemblage B analyses will require multi-locus genotyping with conserved and variable genes★

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

[a] *Division of Veterinary and Biomedical Sciences, Murdoch University, Murdoch, WA 6150, Australia.*

[b] *WHO Collaborating Centre for the Molecular Epidemiology of Parasitic Infections, Division of Veterinary and Biomedical Sciences, Murdoch University, Murdoch, WA 6150, Australia.*

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

*Corresponding author.

Division of Veterinary and Biomedical Sciences, Murdoch University, South Street, Murdoch, WA 6150, Australia.

Tel.: +61 08 9360 2691; fax: +61 08 9360 6628. *E-mail address*: c.wielinga@murdoch.edu.au

★Note: Nucleotide sequence data reported in this paper are available in the GenBank databases under the accession numbers HQ179581 to HQ179653.

Note: Supplementary data associated with this article.

Abstract

Recent research concerning *Giardia duodenalis* has focused on resolving possible sub-assemblages within Assemblages A and B to better understand host-specific and zoonotic relationships. In the present study nine cloned, cultured, Assemblage B isolates were used to investigate the intra-Assemblage B substitution patterns of conserved (ssrDNA, *ef*, *h2b*, *h4*) and variable (*tpi*, *gdh*, *bg*) genes to assess their suitability for further application to sub-assemblage analyses. The resolution of each gene was found to be proportional to its substitution rate and for the genetically narrow sample set examined, the variable genes best represented the consensus phylogeny while the conserved genes only established fractions. However it was demonstrated that the spectra of conserved and variable genes were required to ensure accuracy of inferred phylogeny and it was therefore concluded that further research into sub-Assemblage B groups would require a mixture of conserved and variable genes for the multi-locus analyses of this genetically broad assemblage.

Keywords: *Giardia*, Assemblage B, Sub-assemblages, Genotyping, Phylogeny, MLG, ASH

1. Introduction

Giardia duodenalis is a common intestinal parasite of mammals including humans. Within the *G. duodenalis* species complex, there are currently eight described assemblages (Monis et al., 1999; Gaydos et al., 2008; Lasek-Nesselquist et al., 2010). The majority of these assemblages are host-specific but two, Assemblage A and Assemblage B, are considered zoonotic and are the only two assemblages commonly accepted as being infectious to humans (Sprong et al., 2009).

Within Assemblage A, an apparent host-specific sub-assemblage, AIII from wild ungulates, has been discovered (van der Giessen et al., 2006; Lalle et al., 2007; Langkjaer et al., 2007; Robertson et al., 2007; Caccio et al., 2008). This finding has lead to the possibility that further genetic investigations may identify other host-specific sub-assemblages.

Recent research has therefore focused on resolving possible sub-assemblages within A and B but has been hampered by the lack of genetic tools to allow consistent or sufficient sub-genotyping, particularly within Assemblage B (Caccio et al., 2008; Sprong et al., 2009; Lebbad et al., 2010; Plutzer et al., 2010).

It is possible that the genes currently used for genotyping, such as triose phosphate isomerase (*tpi*), glutamate dehydrogenase (*gdh*) and beta giardin (*bg*) will be unable to consistently define the sub-assemblages within Assemblage B (single or multiple loci) because Assemblage B is too diverse and the high substitution rate of these genes will produce excessive noise for congruent phylogenetic analyses. Therefore the core sub-assemblages of Assemblage B may be better defined by substitution patterns at conserved genes similar to those that differentiate AI/AII at the 3' ssrDNA (Weiss et al., 1992). The aim of the present study was therefore to examine genes with low substitution rates to assess their suitability

for further application in the identification of sub-Assemblages within Assemblage B. Nine cloned, cultured *G. duodenalis* Assemblage B isolates were used to investigate the intra-Assemblage B substitution patterns of several conserved and variable genes.

2. Materials and methods

2.1. Isolates

A total of 11 cloned cultured isolates sourced from humans in Western Australia were analysed (Table 1 and Fig. 1). Two Assemblage A isolates (AI and AII) and nine Assemblage B isolates selected from as broad a genetic range as possible based on previous isozyme data were chosen (Thompson and Meloni, 1993). No clinical information was available for the isolates, only their collection locations. Culture stocks were previously cloned from individual trophozoites (Binz et al., 1991). Cryo-preserved cloned stock isolates were re-cultured in BS-I-33 *Giardia* culture medium as previously described (Steuart et al., 2008). Trophozoites were washed in PBS and DNA extracted using a QIAamp extraction kit (Qiagen, Australia).

2.2. Genes

The genes examined included the genes routinely used for genotyping, *tpi*, *gdh*, *bg* and ssrDNA, as well as the less common typing genes coding for elongation factor 1 α (*ef*), histone 2b (*h2b*) and histone 4 (*h4*). These genes do not appear to be linked, based on the available *G. duodenalis* genome data (*Giardia* DB; http://*Giardia*db.org). For Assemblage B (isolate GS), all genes were on different scaffolds and for Assemblage A (isolate WB); all genes were on different contigs (*bg* AACB02000035, *tpi* AACB02000019, *ef* AACB02000053, *h2b* AACB02000019, *h4* AACB02000048 and AACB02000073) except *tpi* and *h2b*, which were separated by 80 kb. In the Assemblage A reference genome sequence, WB, the ssrDNA has

multiple copies, the *ef* and *h2b* genes have two copies and the *h4* gene three copies, while the remainder are single copy genes. Preliminary data for the Assemblage B reference genome sequence, GS, indicates that all of the genes in the current study are single copy, except the *h4* gene with two copies and the multi-copy ssrDNA (*Giardia* DB; http://*Giardia*db.org).

Primers derived from Teodorovic et al. (2007) and WB were also trialed for ferredoxin (*fd*), ribosomal protein L7a (*rpl7a*) and chaperonin 60 (*cpn60*). Primers are listed in Supplementary Table S1.

2.3. PCR and sequencing

PCRs were performed one isolate at a time to eliminate cross contamination by PCR products. Cycling was conducted using touch down PCRs (96 °C for 5 min; 96 °C for 30 s, 65 °C (-1.0 °C per cycle) for 45 sand 72 °C for 1min for 15 cycles followed by 96 °C for 30 s, 50 °C for 45 s, 72 °C for 1min for 30 cycles with a final extension of 72 °C for 7min). The reaction mixture consisted of 1 x reaction buffer, 1.5-2.5 mM MgCl2, 200 uM of each dNTP, 500 nM of each primer, 1.0-1.5 U Tth^{+} (Biotech International, Australia), 5% DMSO (primary only), 1-2 uL template and H2O.

PCR products were visualised and separated on 1% agarose gels. Products were excised and purified using Wizard columns (Promega, Australia). Purified PCR products were sequenced in both directions with the PCR primers using an ABI Prism™ Dye Terminator cycle sequencing kit (Applied Biosystems, Foster City, California, USA).

2.4. Sequence and phylogenetic analysis

Sequences were checked using Finch TV 1.4.0 (Geospiza, Inc, USA; http://www.geospiza.com) and aligned using ClustalX 2.0.11 (Larkin et al., 2007). Maximum likelihood (ML) phylogenetic analyses were performed using PhyML (Dereeper et al., 2008)

and the reliability of the inferred trees was assessed by the approximate likelihood ratio test (aLRT) (Anisimova and Gascuel, 2006). Branches below 60% bootstrap value were collapsed in TreeDyn (Dereeper et al., 2008) and Newick tree files were presented in MEGA 4.0 (Molecular Evolutionary Genetics Analysis software, Arizona State University, Tempe, Arizona, USA).

Models and parameters used for the phylogenetic analyses were computed using the statistical J Model Test programme, 0.1.1 (Posada, 2008).

Concatenated sequences were also compiled and processed through J Model Test prior to analyses. Complete concatenation of all sequences across all loci was not possible due to unavailable data (isolate 49c11 missing from the *h2b* gene and 54c14 from the *h4* gene) hence four (ssrDNA, *ef*, *tpi*, *bg*) and five gene concatenations (ssrDNA, *h2b*, *ef*, *tpi*, *bg* and ssrDNA, *h4*, *ef*, *tpi*, *bg*) were produced.

In an effort to include the degenerate bases (Fig. 2) in phylogenetic analyses (without sub-cloning product variants), two sets of phylogenetic analyses, 'original' and 'divergent', were conducted for comparison purposes. In the 'original' set of sequences, degenerate bases were left as degenerate bases when the signal strength was equal, or converted to the nucleotide of the greater signal where they were unequal, as is standard practice. In the 'divergent' set of sequences, the degenerate bases for an isolate (deemed allelic variants when present in these cloned culture isolates) were altered to a specific nucleotide when the substitution site was shared among other isolates (shown boxed in Fig. 2). Only degenerate bases at substitution sites shared among isolates were altered for the comparative phylogenetic analyses because only the shared substitutions affect grouping-topology. Specific alterations are shown in brackets in Fig. 2 next to the original degenerate base. The alterations changed the degenerate base to the nucleotide least prevalent of the two in the

sample population and hence generated sequences equivalent to the most divergent alleles. For example, if there was a degenerate base Y (equal peaks of C and T), this was changed to a C if the minority of remaining isolates had a C and the majority a T. Alterations in the reverse direction, to the most prevalent base, were not included as an additional sequence set because this results in the same grouping-topology as already seen in the 'original' set because degenerate bases are excluded from phylogenetic calculations. Degenerate bases were noted as two types in Fig. 2, one in apparently equal ratios and represented by Y or R and the other in apparently unequal ratios (1:2/1:3) and represented by the lower case a/t/g/c of the base present in the highest amount (with the transition pair being the lower amount, except in *h4*, where the substitution pair was a transversion). Hence 'original' trees are those where all substitutions shared among isolates are represented at all alleles (all copies of the gene) and 'divergent' trees are those where some substitutions shared among the isolates are represented at only some of the alleles (some copies of the gene).

3. Results

3.1. Multi-locus sequence typing

Complete gene sequences were successfully generated for nine isolates (two Assemblage A, seven Assemblage B) at six loci (ssrDNA, *h2b*, *h4*, *ef*, *tpi*, *bg*) and partial sequences at one locus (*gdh*). Isolates 49c11 and 54c14 were problematic, apparently due to DNA quality and neither the partial *gdh* sequence was obtained, nor *h2b* for 49c11 or *h4* for 54c14 and only partial sequences for the *bg* and *tpi* genes were obtained for 54c14.

More gene sequences were generated for the 3' ends of the *ef* and *tpi* genes where limited sequences previously existed for Assemblage B. Additional gene sequence data for the

centre of the ssrDNA gene verified a previously reported deletion, but failed to validate several reported substitutions from the AMC-4 (human) isolate (**U09491**, van Keulen et al., 1995). The sequences generated in the present study are available on GenBank under accession numbers **HQ179581** to **HQ179653**.

The present study identified several new intra-Assemblage B substitution sites (Fig. 2) and validated many existing sites as seen in alignments of recently collated GenBank sequences (data not shown). Many of the intra-Assemblage B substitutions were represented by variation at an allele within an isolate and are shown in Fig. 2 as degenerate bases (Y/R) where the PCR product distribution appeared equal (1:1) and as a lowercase letter (a/t/g/c) where it appeared unequal (1:2/1:3), with the lesser base being the transition pair. Hence 'degenerate substitutions' are defined as those where a substitution has occurred at some of the alleles for a given gene and 'regular substitutions' are defined as those where a substitution has occurred at all of the alleles for a given gene. This rarely occurred in Assemblage A isolates except in the *ef* gene where all of the intra-Assemblage A sites were also degenerate.

Primer sets for ferredoxin (*fd*), ribosomal protein L7a (*rpl7a*) and chaperonin 60 (*cpn60*) failed to amplify Assemblage B isolates under any of the optimisation conditions trialed. Primers had originally been designed from sequences from a combination of Assemblage A and B isolates (WB; Teodorovic et al., 2007) however subsequent review of the GS genome sequence indicated significantly greater inter-assemblage divergence than initially suspected for all of these genes. The *rpl7a* Assemblage A and B sequences (WB, XM_001706269; GS, ACGJ01000263, 27247-28028) had the least divergence of 0.13 substitutions per nucleotide (comparable with the *gdh* and *tpi* genes, Supplementary Table S2). Although the forward primer from Teodorovic et al. (2007) at the start codon was

homologous, the reverse primer from the present study at the stop codon contained a 6 bp deletion in the Assemblage B sequence. The *fd* (WB, AF393829; GS, ACGJ01002917, 37081-37518) and *cpn60* genes (WB, AF029695; GS, ACGJ01002917, 37596-39236) had greater Assemblage A/B divergence than any of the other genes examined in the present study at 0.22 and 0.24 substitutions per nucleotide, respectively. The original primers, those from Teodorovic et al. (2007) at the start and stop codons on the *fd* gene and those from the present study at the start and near-stop codons and the centre of the *cpn60* gene, had on average only approximately 70% homology to the Assemblage B GS genome sequence and were unsuccessful. New primers were not designed for the present study because the substitution rates of the genes were higher than preferred. A similar problem was encountered for the 5' end section of the *gdh* gene.

The histone genes had multi-priming problems due to their repeating motifs. The *h4* reverse primer was ineffective in sequencing reactions due to excessive stuttering.

3.2. Substitution statistics

In all genes, the majority of the substitutions were inter-assemblage substitutions (between Assemblages A and B), usually followed by intra-Assemblage B substitutions (except in the *gdh* 3' section and the *h4* gene, which had similar amounts of intra-A and intra-B substitutions) and intra-Assemblage A substitutions (Supplementary Table S2). All of the inter-assemblage substitution sites had been previously reported except for the new section of the *ef* gene.

As previously reported, the majority of substitutions were in the *tpi*, *bg* and *gdh* genes, followed by the *ef* gene and then the *h4*, *h2b* and ssrDNA genes (Supplementary Table S2).

Many of the intra-Assemblage B substitutions were degenerate, where the variation was represented by only a proportion of the alleles within an isolate (Fig. 2 and Supplementary Table S2). Allelic sequence heterozygosity (ASH) varied among the isolates and loci. Calculated over the 5.5 kb coding region sequenced, isolate 7c3 had the highest ASH at 0.5% (ranging from 1.2% in *bg* to 0% in ssrDNA) and 15c1 and 42c5 the lowest at 0.02% (from 0.07% in ssrDNA to 0% in the other genes). Intermediate isolates were 39c10, 0.3% (0.9% *tpi* – 0% ssrDNA), 49c11, 0.3% (0.8% *tpi* - 0% ssrDNA), 54c14, 0.2% (1.1% *tpi* - 0% ssrDNA), 33c2, 0.1% (0.3% *gdh* – 0% *h4*), 34c8, 0.05% (0.2% *gdh* – 0% *h4*) and 30c7, 0.05% (0.1% ssrDNA – 0% *h4*). ASH per gene (averaged across the isolates) was approximately proportional to the substitution rates of the genes; *tpi* (0.4%, range 1.1%-0%) > *bg* (0.3%, 1.2%-0%) > *gdh* (0.2%, 0.9%-0%) > *ef* (0.15%, 0.4%-0%) > *h2b* (0.1%, 0.5%-0%) > *h4*/ssrDNA (0.05%, 0.4%-0%, Supplementary Table S2).

In the genes most commonly used, *tpi*, *gdh* and *bg*, there was a relatively even proportion of known intra-Assemblage B substitution sites, new substitution sites and those previously reported but not seen in this set. However, for the less studied genes (ssrDNA, *ef*, *h2b* and *h4*), there was a much higher proportion of new intra-Assemblage B substitution sites detected (Fig. 2 and Supplementary Table S2).

The majority of substitutions were transition substitutions, with transversions usually in a low proportion of the inter-assemblage substitutions. The two main exceptions were one intra-Assemblage B transversion substitution in the *h4* gene and predominantly all transversions for the inter-assemblage substitutions of the *ef* gene (Fig. 2 and Supplementary Table S2).

Only the *tpi*, *gdh*, *ef* and *h2b* genes had non-synonymous substitutions (Fig. 2 and Supplementary Table S2). In all but one case in the *tpi* gene, these substitutions resulted in

amino acid changes that were within recognisable groups (for example polarity, size, etc), with high BLOSUM scores (Henikoff and Henikoff, 1992, data not shown), indicating high probability of substitution. The absence of first and second codon position substitutions in the *bg* gene increased its relative substitution rate. Whereas the intra-Assemblage B substitutions per nucleotide for the *tpi* gene and *bg* gene were comparable at 0.039 and 0.029, respectively, 28% of the intra-Assemblage B substitutions at the *tpi* gene were spread over the first and second codon positions, however all of the substitutions in the *bg* gene were concentrated in the third codon position (Fig. 2 and Supplementary Table S2). The proportion of shared sites was highest in the histone genes (Supplementary Table S2).

3.3. Shared substitutions and phylogenetic analyses

Shared substitution sites are shown boxed in Fig. 2, in the Euler diagram Fig. 3 and in phylogenetic analyses in Figs. 4 and 5.

The majority of shared substitutions grouped the isolates into approximate northern Western Australia and southern Western Australia groups. North - 15c1 (Kununurra), 30c7 (Derby), 33c3 (Perth), 34c8 (Kununurra), 42c5 (Karratha) and South - 7c3 (Katanning), 39c10 (Perth), 49c11 (Northam), 54c14 (Kununurra). There were two exceptions, isolate 33c3 (Perth) in the North group and isolate 54c14 (Kununurra) in the South group (Figs. 2 - 4).

The conserved genes *h2b*, *h4* and *ef* grouped southern isolates 7c3/39c10 (non-synonymous), 7c3/49c11 (synonymous) and 39c10/49c11 (non-synonymous). The ssrDNA gene grouped northern isolates 30c7/42c5 and 33c3/34c8 (Figs. 2 - 4). The variable genes, *tpi* and *bg*, had shared substitutions among the southern isolates, 7c3/49c11, 39c10/49c11 (non-synonymous), 7c3/39c10, 7c3/39c10/49c11, 7c3/39c10/49c11/54c14 (synonymous) and 7c3/54c14, 7c3/49c11/54c14, 7c3/39c10/49c11/54c14 (synonymous), respectively (Figs. 2

- 4). The only shared substitution site in the 3' end of *gdh* gene grouped 7c3-30c7-(34c8)-42c5 (Katanning, Derby, (Kununurra), Karratha) and 15c1-33c2-(34c8)-39c10 (Kununurra, Perth, (Kununurra), Perth) (Fig. 2).

The phylogenetic analyses (Figs. 4 and 5) reflected the shared substitutions and demonstrated that the resolution of each gene was proportional to its substitution rate. The northern isolates formed a cluster in the *tpi* and *bg* genes and the southern isolates grouped into different pairs in each of the conserved genes, *h2b*, *h4* and *ef*, as predicted by their shared substitution patterns (Figs.2 and 3). The southern isolates were unable to form a cluster in either of the variable genes because the shared substitutions within the variable genes grouped different southern isolates at different substitution sites (Figs. 2 and 3), producing a contradictory phylogenetic signal resulting in no grouping (Fig. 4). Combining the remaining significant substitution data (*h2b*, *h4*, *ef*, *tpi*, *bg*), isolates 7c3, 39c10 and 49c11 had the highest substitution activity (total substitutions, unshared substitutions, shared substitutions, non-synonymous substitutions and all of the shared non-synonymous substitutions), with the exception of 54c14, which in the variable genes (*bg*, *tpi*) also had moderate numbers of shared and individual substitutions and in the case of *tpi* non-synonymous substitutions (Figs. 2 and 4).

Phylogenetic analyses using the concatenated merged sequences (Fig. 5) resulted in consensus trees clearly defining a cluster of northern isolates (North group) with high bootstrap support. This was the resultant influence of the uninterrupted, unified phylogenetic signal from the variable genes (Fig. 4). The relationships of the southern isolates were more complex. The strong relationships evident in the substitution patterns in the individual gene trees, grouping different pairs of isolates in each of the conserved genes and within the variable genes (Figs. 2 and 3), could not be supported in the consensus trees. This was the same situation as was

found in the variable genes where the interrelated contradictory signals canceled each other out, instead placing isolate 7c3 at the base followed by 39c10, 49c11 and 54c14 together (Fig. 5). The *gdh* gene was not included in the concatenated analyses due to missing isolate data and negligible contribution to tree topology (data not shown).

The regular shared substitutions and the degenerate shared substitutions predominantly grouped the same isolates (Fig. 2). The inclusion of the degenerate shared substitutions into the phylogenetic analyses led to an increase in phylogenetic resolution (Fig. 4). The increase in resolution was inversely proportional to the substitution rate of the gene where it was most prominent in the conserved genes (*h2b*, *h4*), evident in the *tpi* gene and unnoticeable in the *bg* gene.

4. Discussion

The aim of the present study was to examine genes with low substitution rates to assess their suitability for the identification of sub-assemblages within Assemblage B. The rationale was that the genes with high substitution rates currently used for genotyping are unable to consistently define the sub-Assemblage B groups because Assemblage B is genetically diverse and high substitution rates obscure the true sub-assemblage patterns.

Genes with high substitution rates have a limited number of sites that can be changed (usually due to functional constraints) and hence older substitutions become overwritten by newer ones, obscuring any phylogenetic signal. This overwriting phenomenon can cause homoplasy, i.e. two sequences can have the same base substitution due to independent events rather than via common ancestry. Homoplasies in DNA sequencing data can obscure phylogenetic relationships and contribute to noise in the data. This has been clearly

296 demonstrated in the *tpi* gene at the inter-assemblage level by comparison of the nucleotide
297 and amino acid sequences (Wielinga and Thompson, 2007).

298 The hypothesis that analysis of low substitution rate genes should provide unobscured
299 substitution patterns that better define the sub-Assemblage B groups was based on previous
300 multiple sequence analysis where proposed sub-Assemblage B groups were found to be
301 inconsistent across loci and BIII and BIV were not validated (Wielinga and Thompson, 2007).
302 This has since gained support from extensive multi-locus-genotyping (MLG) studies that
303 repeatedly conclude that congruent sub-typing across the loci for Assemblage B could not be
304 established (Caccio et al., 2008; Lebbad et al., 2008, 2010; Geurden et al., 2009; Levecke et al.,
305 2009; Sprong et al., 2009).

306 Low substitution rate genes also have the advantage of making it easier to interpret
307 mixed intra-assemblage infections and reduce ASH (present findings), factors that have
308 previously prevented the concatenation of MLG data (Caccio et al., 2008; Lebbad et al., 2010).
309 The disadvantage of analyses with low substitution rate genes is the loss of resolution among
310 similar groups, necessitating the use of MLGs with a range of genes of different substitution
311 rates to encompass the as yet undefined extent of divergence in Assemblage B.

312 As expected, the substitution rates of the genes determined their resolution (Figs. 2
313 and 4) and the substitution patterns demonstrated different relationships where the
314 relationships have changed over time and the substitutions have occurred at different times
315 (Figs. 3 and 4).

316 There was geographical sub-grouping detected (Figs. 4 and 5) with results from most
317 genes indicating a cluster of isolates from the North and clustering of some isolates from the
318 South (Figs. 3 and 4).

The conserved genes (ssrDNA, *h2b*, *h4*, *ef*) were not capable of properly defining the geographical sub-groups in this set because in this sample set their total substitution numbers were too low (Fig. 4). Although the rate of substitution within a gene is constant, the number of substitutions between isolates depends on the divergence of the isolates and in these samples, with relatively recent divergence, there were few substitutions in these genes. Some of the older relationships were demonstrated (Figs. 3 and 4), but the information was incomplete or absent. The information, however, was complementary and presents an example of the benefit of MLG analyses where multiple genes can provide a greater amount of phylogenetic information. The complementary rather than identical nature of the phylogenetic signals may have been due to an old rapid expansion event where the separate divergences occurred in a short time-frame, causing interrelated signals. Hence in summary, it was demonstrated that the low substitution rates of the conserved genes could lead to low resolution among related isolates, where they detected only the older divergence events and not recent events. At the opposite end of the spectrum, the variable genes (*tpi*, *bg*) did differentiate geographical sub-grouping, detecting the apparently recent collective divergence of the northern group (Fig. 4). The most variable gene however, *bg*, did not distinguish the more divergent isolates (7c3, 39c10, 49c11) as such, contradicting the results from the other genes (Fig. 4) where the analyses at conserved genes and first and second codon sites (Fig. 2), suggested they were more divergent than 54c14. This apparent loss of resolution in a variable gene for older divergence events represents the opposite end of the spectrum of a gene's application where variable genes may progress toward misrepresentation of older events because high substitution rates increase the likelihood of homoplastic substitutions. Hence the high substitution rates of the variable genes can also lead to reduced resolution in divergent isolates, where they detect recent divergence events more reliably than older events. This presents another example of the benefit of MLG analyses where using multiple

genes of different substitution rates can ensure an accurate phylogenetic assessment of the whole sample population, both the recent and the older divergence events.

Hence it can be concluded that mixed-substitution-rate MLG will also be required for accurate phylogenetic analysis of the whole of Assemblage B and it can also be extrapolated that the conserved genes will have increasing substitution rates, resolution and application with the increasing divergence of the population examined while the variable genes could lose resolution for older events.

The present study confirmed the high rate of ASH in Assemblage B as reported by Franzen et al. (2009) (Fig. 2). The degenerate substitution rate was higher in variable genes (0.4%) and lower in conserved genes (0.05%) and varied among the isolates (0.5-0.02%) over the 5.5 kb of coding genes examined. The patterns of degenerate substitutions concurred with the patterns of regular substitutions, where similar isolates were grouped together and similar isolates showed divergence (Fig. 2).

Since the trend of the degenerate shared substitutions matched that of the regular shared substitutions, the increase in resolution in tree topology (Fig. 4) was inversely proportional to the substitution rate of the gene. This was because the existing resolution could only be increased where it was not already at maximum, as is more likely the case for high substitution rate genes. Hence in the most variable gene, *bg*, the degenerate shared substitutions provided no new phylogenetic information that was not already presented in the regular shared substitutions, leading to no change in the tree grouping-topology (Fig. 4). In contrast, in the conserved genes (*h2b*, *h4*, *ef*), where the total substitution rate was low, the degenerate substitution patterns contributed significantly to the total information (Fig. 4).

Therefore, although the degenerate substitutions created an extra level of complexity in the analyses, they were found to match the regular substitutions (in rate and pattern) and not to disrupt the evolutionary signals. This aspect may be useful in future analyses for deciphering ASH from mixed intra-assemblage infections where a mixed sample may be conspicuous by its mixture of two divergent 'parent' signals. This would be similar to the way in which mixed inter-assemblage infections are currently suspected, however for intra-assemblage mixtures there would be limitations where mixtures of closely related isolates may appear the same as ASH and this would need to be considered. Further analyses into the specific allelic variants, the distribution of the substitutions at the different alleles, could provide valuable insights into the mechanisms of allele evolution and modes of recombination in *Giardia*.

The present study did not detect inter-assemblage recombination as previously reported in Assemblage B isolates where Assemblage A sequences were retrieved from Assemblage B isolates (Teodorovic et al., 2007; Lasek-Nesselquist et al., 2009). Although the present study did not clone PCR products it is believed that the minimum allelic sequence variation for a single copy gene of one in four was detected. Hence, for genes with a higher copy number, marginal allelic variation may have gone undetected. The present study also did not use assemblage-specific primers which, for the variable genes, could have affected the detection of minority alleles due to primer specificity bias. It is notable, however, that the only two coding genes for which Assemblage A sequences were obtained from Assemblage B isolates in the study by Teodorovic et al. (2007), the *fd* and *cpn60* genes, were also included in this study and in both instances failed to amplify Assemblage B haplotypes. In both of these studies the primers had originally been designed prior to the GS genome sequence availability and can now be shown to be only 70% homologous to the Assemblage B GS genome sequence.

In the present study the primers failed to amplify any products from any of the Assemblage B isolates, indicating they were not suitable for analysis of Assemblage B isolates. The lack of Assemblage A sequences detected from Assemblage B isolates is in agreement with the genome sequencing results from Franzen et al. (2009). Some sequence degeneracy of Assemblage B isolates matching Assemblage A substitutions was found in isolates 7c3 and 54c14 in the *tpi* and *bg* genes (Fig. 2), but this could also be attributed to homoplasy due to variation at sites with high substitution rates.

All of the genes analysed have potential application for phylogenetic analyses on different sample sets with different amounts of divergence. The present study has demonstrated that a gene's resolution was dependent on its rate of substitution which, among isolates, is dependant on their divergence. It was shown that the conserved genes only detected older divergence events while the variable genes reliably detected recent events, therefore necessitating the use of mixed-substitution-rate MLG.

For the current study, among the conserved genes the ssrDNA was of no value on the genetically narrow sample set. No regular intra-Assemblage B substitutions were detected, only degenerate variation in the 5' and 3' ends. Therefore it does not suit routine use at this stage, but may be useful in the future if targetable sub-assemblage variation is found. Alternatively sub-assemblage variation has been found in the adjacent internal transcribed spacer (ITS) regions (Caccio et al., 2010) and may provide a sensitive tool for intra-assemblage analyses. The small histone genes did show intra-Assemblage B variation (regular substitution sites, non-synonymous sites and transversion sites) and could therefore prove useful in analyses of the older intra-Assemblage B groups where their resolution is expected to increase. Although there were some technical PCR difficulties that should be resolved with new or nested primers, they are still desirable targets because they are very small genes (300-

400 bp). The *ef* gene also showed intra-Assemblage B variation, including regular and non-synonymous sites and although it was much larger (1.3 kb) it amplified readily. Some of the conserved genes are known to have multiple copies, which may increase PCR sensitivity, as is the case with the ssrDNA, but may also increase ASH. In the Assemblage A isolates there were no degenerate substitutions in the multi-copy ssrDNA or double-copy *h2b* gene, but there were degenerate substitutions in the double-copy and triple-copy *ef* and *h4* genes (Fig. 2). In the Assemblage B isolates, the multi-copy ssrDNA and double-copy *h4* genes contained no more degenerate substitutions than the other genes examined which were of similar size and substitution rate. Hence it does not appear that degeneracy is directly proportional to copy number, but rather directly proportional to substitution rate and therefore the effect of copy number may be considerable in high substitution rate genes.

The remaining variable genes, although well established for assemblage-level genotyping (Monis et al., 1999; Wielinga and Thompson, 2007), do not often demonstrate congruent phylogenetic analyses. Notably the *tpi* and *bg* genes have a predisposition to homoplastic substitution patterns in divergent samples sets (Wielinga and Thompson, 2007), and their subgroups rarely concur with each other (Caccio et al., 2008; Geurden et al., 2009; Abe et al., 2010) or the *gdh* gene (Gelanew et al., 2007; Lalle et al., 2009; Levecke et al., 2009). However in the present study examining a genetically narrow sample set, the *bg* and *tpi* genes best represented the consensus phylogeny (Figs. 4 and 5) and it is expected their resolution will still be relevant for resolving the phylogeny of similar closely related sub-assemblage groups within Assemblage B. Indeed the main finding of the present study was to demonstrate that different genes of different substitution rates are all required to establish the different levels of relationships, old and new, with confidence because no single gene could encompass such a range in resolution. The *bg* and *tpi* genes are also small and easy to

amplify with a significant sequence database, and are therefore convenient and practical for future use in combination with genes of lower substitution rates.

In contrast, the large *gdh* gene has been shown to infer incongruent phylogeny with different gene sections employed (Souza et al., 2007; Lasek-Nesselquist et al., 2009) and in the present study the partial sequence (636 bp) was shown to be inadequate for phylogenetic analyses. This served to demonstrate the ineffectiveness of utilising partial sequences in phylogenetic analyses (as distinct from typing analyses) prior to the locations of the relevant substitutions being known. Therefore the *gdh* gene was found to be less suitable for sub-Assemblage B phylogenetic analyses, except possibly when analysed in total. Conversely the *gdh* gene may be an ideal gene to apply to Assemblage A analyses as it contained the highest rate of intra-Assemblage A substitutions (Fig. 2, Supplementary Table S2) which may be useful for increasing resolution in this relatively homogeneous assemblage. Since Assemblages A and B have distinct differences in the timing and extent of their divergences, the same genes applied to each assemblage will result in different resolutions, and may therefore necessitate some differences in their gene repertoires for comprehensive analysis.

Hence in summary, for mixed-substitution-rate MLG a combination of histones, *ef*, *bg* and *tpi* would start to provide the necessary intra-Assemblage B substitution pattern data to decipher sub-Assemblage B groups should they exist. Utilising genes with non-synonymous variation (*h2b*, *ef*, *tpi*) would also provide an extra layer of definition to the analyses, a useful aspect when the expected divergence is unknown, and for increased sensitivity the ITS rDNA region may also be useful.

The main findings of the present study were to demonstrate the effect of a gene's substitution rate on its ability to resolve relationships for a given sample set and the

461 importance of using mixed-substitution-rate MLG to ensure accurate phylogenetic inference
462 in a sample set of unknown divergence.

463 The present findings only partially supported the hypothesis that conserved genes
464 would better define the sub-Assemblage B groups because both conserved and variable genes
465 were required to construct the consensus phylogeny and it was concluded that this would
466 also be the case for the remainder of Assemblage B. However, as Assemblage B in total is
467 likely to be genetically much more diverse than the preliminary study set, the resolution of
468 the conserved genes is likely to provide the most reliable data for defining the core sub-
469 Assemblage B groups.

470

471 **Acknowledgements**

472 The present study utilised culture isolates originating from the work of Adjunct
473 A/Prof. B. Meloni, Murdoch University, Australia and isolate clones originating from the work
474 of Dr. N. Binz, Lions Eye Institute, Australia.

475

476 **References**

477 Abe, N., Tanoue, T., Noguchi, E., Ohta, G., Sakai, H., 2010. Molecular characterization of *Giardia*
478 *duodenalis* isolates from domestic ferrets. Parasitol. Res. 106, 733-736.

479 Anisimova, M., Gascuel, O., 2006. Approximate likelihood-ratio test for branches: A fast,
480 accurate, and powerful alternative. Syst. Biol. 55, 539-552.

481 Binz, N., Thompson, R.C.A., Meloni, B.P., Lymbery, A.J., 1991. A Simple Method for Cloning
482 *Giardia duodenalis* from Cultures and Fecal Samples. J. Parasitol. 77, 627-631.

483 Caccio, S.M., Beck, R., Almeida, A., Bajer, A., Pozio, E., 2010. Identification of *Giardia* species
484 and *Giardia duodenalis* assemblages by sequence analysis of the 5.8S rDNA gene and
485 internal transcribed spacers. Parasitology 137, 919-925.

486 Caccio, S.M., Beck, R., Lalle, M., Marinculic, A., Pozio, E., 2008. Multilocus genotyping of *Giardia*
487 *duodenalis* reveals striking differences between assemblages A and B. Int. J. Parasitol.
488 38, 1523-1531.

489 Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.F., Guindon, S.,
490 Lefort, V., Lescot, M., Claverie, J.M., Gascuel, O., 2008. Phylogeny.fr: robust phylogenetic
491 analysis for the non-specialist. Nucleic Acids Res. 36, W465-W469.

492 Franzen, O., Jerlstrom-Hultqvist, J., Castro, E., Sherwood, E., Ankarklev, J., Reiner, D.S., Palm, D.,
493 Andersson, J.O., Andersson, B., Svard, S.G., 2009. Draft genome sequencing of *Giardia*
494 intestinalis assemblage B isolate GS: is human Giardiasis caused by two different
495 species? PLoS Pathog. 5, e1000560.

496 Gaydos, J.K., Miller, W.A., Johnson, C., Zornetzer, H., Melli, A., Packham, A., Jeffries, S.J., Lance,
497 M.M., Conrad, P.A., 2008. Novel and canine genotypes of *Giardia duodenalis* in harbor
498 seals (Phoca vitulina richardsi). J. Parasitol. 94, 1264-1268.

499 Gelanew, T., Lalle, M., Hailu, A., Pozio, E., Caccio, S.A., 2007. Molecular characterization of
500 human isolates of *Giardia duodenalis* from Ethiopia. Acta Trop. 102, 92-99.

501 Geurden, T., Levecke, B., Caccio, S.M., Visser, A., De Groote, G., Casaert, S., Vercruysse, J.,
502 Claerebout, E., 2009. Multilocus genotyping of Cryptosporidium and *Giardia* in non-
503 outbreak related cases of diarrhoea in human patients in Belgium. Parasitology 136,
504 1161-1168.

505 Henikoff, S., Henikoff, J.G., 1992. Amino-Acid Substitution Matrices from Protein Blocks. P.
506 Natl. Acad. Sci. USA 89, 10915-10919.

507 Lalle, M., Bruschi, F., Castagna, B., Campa, M., Pozio, E., Caccio, S.M., 2009. High genetic
508 polymorphism among *Giardia duodenalis* isolates from Sahrawi children. Trans. R. Soc.
509 Trop. Med. Hyg. 103, 834-838.

510 Lalle, M., di Regallbono, A.F., Poppi, L., Nobili, G., Tonanzi, D., Pozio, E., Caccio, S.M., 2007. A
511 novel *Giardia duodenalis* assemblage a subtype in fallow deer. J. Parasitol. 93, 426-428.

512 Langkjaer, R.B., Vigre, H., Enemark, H.L., Maddox-Hyttel, C., 2007. Molecular and phylogenetic
513 characterization of Cryptosporidium and *Giardia* from pigs and cattle in Denmark.
514 Parasitology 134, 339-350.

515 Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H.,
516 Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G.,
517 2007. Clustal W and clustal X version 2.0. Bioinformatics 23, 2947-2948.

518 Lasek-Nesselquist, E., Welch, D.M., Sogin, M.L., 2010. The identification of a new *Giardia*
519 *duodenalis* assemblage in marine vertebrates and a preliminary analysis of G.
520 *duodenalis* population biology in marine systems. Int. J. Parasitol. 40, 1063-1074.

521 Lasek-Nesselquist, E., Welch, D.M., Thompson, R.C., Steuart, R.F., Sogin, M.L., 2009. Genetic
522 exchange within and between assemblages of *Giardia duodenalis*. J. Eukaryot.
523 Microbiol. 56, 504-518.

524 Lebbad, M., Ankarklev, J., Tellez, A., Leiva, B., Andersson, J.O., Svard, S., 2008. Dominance of
525 *Giardia* assemblage B in Leon, Nicaragua. Acta Trop. 106, 44-53.

526 Lebbad, M., Mattsson, J.G., Christensson, B., Ljungstrom, B., Backhans, A., Andersson, J.O.,
527 Svard, S.G., 2010. From mouse to moose: multilocus genotyping of *Giardia* isolates from
528 various animal species. Vet. Parasitol. 168, 231-239.

529 Levecke, B., Geldhof, P., Claerebout, E., Dorny, P., Vercammen, F., Caccio, S.M., Vercruysse, J.,
530 Geurden, T., 2009. Molecular characterisation of *Giardia duodenalis* in captive non-
531 human primates reveals mixed assemblage A and B infections and novel
532 polymorphisms. Int. J. Parasitol. 39, 1595-1601.

533 Monis, P.T., Andrews, R.H., Mayrhofer, G., Ey, P.L., 1999. Molecular systematics of the parasitic
534 protozoan *Giardia* intestinalis. Mol. Biol. Evol. 16, 1135-1144.

535 Plutzer, J., Ongerth, J., Karanis, P., 2010. *Giardia* taxonomy, phylogeny and epidemiology: Facts
536 and open questions. Int. J. Hyg. Environ. Health. in press.

537 Posada, D., 2008. jModelTest: Phylogenetic model averaging. Mol. Biol. Evol. 25, 1253-1256.

538 Robertson, L.J., Forberg, T., Hermansen, L., Hamnes, I.S., Gjerde, B., 2007. *Giardia duodenalis*
539 cysts isolated from wild moose and reindeer in Norway: Genetic characterization by
540 PCR-RFLP and sequence analysis at two genes. J. Wildl. Dis. 43, 576-585.

541 Souza, S.L., Gennari, S.M., Richtzenhain, L.J., Pena, H.F., Funada, M.R., Cortez, A., Gregori, F.,
542 Soares, R.M., 2007. Molecular identification of *Giardia duodenalis* isolates from humans,
543 dogs, cats and cattle from the state of Sao Paulo, Brazil, by sequence analysis of
544 fragments of glutamate dehydrogenase (*gdh*) coding gene. Vet. Parasitol. 149, 258-264.

545 Sprong, H., Caccio, S.M., van der Giessen, J.W., 2009. Identification of zoonotic genotypes of
546 *Giardia duodenalis*. PLoS Negl. Trop. Dis. 3, e558.

547 Steuart, R.F., O'Handley, R., Lipscombe, R.J., Lock, R.A., Thompson, R.C., 2008. Alpha 2 giardin
548 is an assemblage A-specific protein of human infective *Giardia duodenalis*. Parasitology
549 135, 1621-1627.

550 Teodorovic, S., Bravermanj, J.M., Elmendorf, H.G., 2007. Unusually low levels of genetic
551 variation among *Giardia* lamblia isolates. Eukaryot. Cell 6, 1421-1430.

552 Thompson, R.C.A., Meloni, B.P., 1993. Molecular Variation in *Giardia*. Acta Trop. 53, 167-184.

553 van der Giessen, J.W.B., de Vries, A., Roos, M., Wielinga, P., Kortbeek, L.M., Mank, T.G., 2006.
554 Genotyping of *Giardia* in Dutch patients and animals: A phylogenetic analysis of human
555 and animal isolates. Int. J. Parasitol. 36, 849-858.

556 van Keulen, H., Homan, W.L., Erlandsen, S.L., Jarroll, E.L., 1995. A three nucleotide signature
557 sequence in small subunit rRNA divides human *Giardia* in two different genotypes. J.
558 Eukaryot. Microbiol. 42, 392-394.

559 Weiss, J.B., van keulen, H., Nash, T.E., 1992. Classification of Subgroups of *Giardia*-Lamblia
560 Based Upon Ribosomal-Rna Gene Sequence Using the Polymerase Chain-Reaction. Mol.
561 Biochem. Parasitol. 54, 73-86.

562 Wielinga, C.M., Thompson, R.C.A., 2007. Comparative evaluation of *Giardia duodenalis*
563 sequence data. Parasitology 134, 1795-1821.

564

565

566 **Figure Legends**

567 Fig. 1. Collection locations of Western Australian *Giardia duodenalis* isolates used in the
568 present study.

569 Fig. 2. Compilation of substitution tables for each gene from the study population (*tpi*, triose
570 phosphate isomerase; *bg*, beta giardin; *gdh*, glutamate dehydrogenase ; *ef*, elongation factor;
571 ssrDNA; *H4*, histone 4). All base numbers are from the start codon of each gene. Bold base
572 numbers indicate non-synonymous substitution sites. Bold intra-Assemblage B nucleotides
573 indicate substitutions from the majority of the population; degenerate nucleotide bases
574 represent degenerate substitution sites with even amounts of each nucleotide base detected;
575 lower case nucleotide bases represent degenerate substitution sites with uneven amounts of
576 each base detected (the lower case base is the nucleotide base detected in the majority at that
577 substitution site and its transition base is the base detected in the minority, unless otherwise
578 stated). Assemblage A sequence is the consensus of the two Assemblage A isolates;
579 Assemblage B sequence in the consensus of the Assemblage B isolates.

580 Fig. 3. Euler diagram of shared substitutions amongst Assemblage B isolates in each gene.
581 Numbers represent the isolate numbers (without the prefix BAH or suffix clone number) and
582 circles represent the shared substitutions. The gene names (*tpi,* triose phosphate isomerase;
583 *bg*, beta giardin; *ef*, elongation factor; ssrDNA; *h4*, histone 4, *h2b*, histone 2b) on each circle
584 correspond to the gene in which the shared substitution occurred (and each occurrence
585 thereof), bold gene names indicate non-synonymous shared substitutions and the line weight
586 is proportional to the number of occasions the shared substitution occurred.

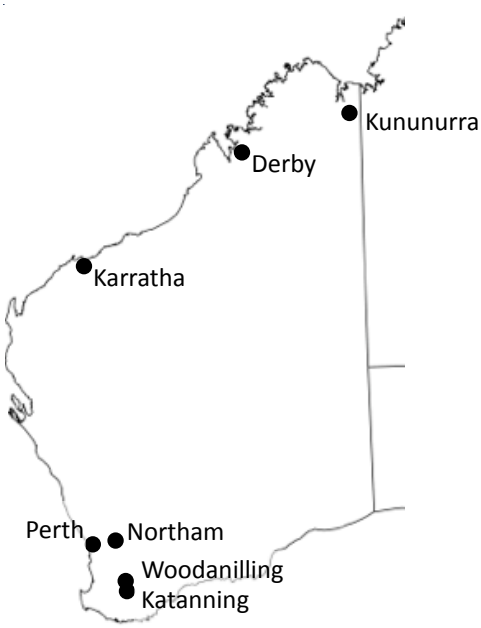
587 Fig. 4. Maximum likelihood phylogenetic analyses of the study population for individual loci.
588 'Original' refers to analyses incorporating shared substitutions present at all alleles (present

589 in all copies of the gene). 'Divergent' refers to analyses incorporating shared substitutions
590 present at some alleles (present in some copies of the gene). *tpi*, triose phosphate isomerase;
591 *bg*, beta giardin; *ef*, elongation factor; ssrDNA; *H4*, histone 4, *h2b*, histone 2b.

592 Fig. 5. Maximum likelihood phylogenetic analyses of the study population for concatenated
593 gene sequences. 'Original' refers to analyses incorporating shared substitutions present at all
594 alleles (present in all copies of the gene). 'Divergent' refers to analyses incorporating shared
595 substitutions present at some alleles (present in some copies of the gene). *tpi*, triose
596 phosphate isomerase; *bg*, beta giardin; *ef*, elongation factor; ssrDNA; *H4*, histone 4, *h2b*,
597 histone 2b.

Research Highlights

- The substitution rate of a gene determines its ability to resolve different relationships formed at different times.
- Mixed-substitution-rate multi-locus-genotyping ensures accurate phylogenetic inference in divergent populations.
- Allelic sequence heterozygosity substitution patterns follow the same trend as regular substitution patterns (those at all alleles).



***tpi* (748 of 774 bp, 17 to 764)**		***Base number from start codon***																																				
Sample	***Differentiation***	39	**91**	126	129	**139**	162	165	168	**182**	210	**263**	399	402	429	**445**	483	492	516	537	552	564	567	**575**	**599**	612	**613**	**625**	630	672	675	681	684	711	714
2c2	Intra-A				T								C			G							C							T					
46c2					C								T			A							A							C					
B					C								T			G							C							T					
A	Intra-B	G	C	C		C	A	C	G	G	G	G		G	G		C	G	G	G	G	G		G	A	G	G	A	T	G		G	T	T	C
15c1/30c7/33c3/34c8/42c5		A	T	T		C	G	T	T	A	A	G		A	A		A	C	G	G	A	A		G	A	A	G	A	C	G		G	C	A	T
7c3		**G**	**C**	**Y**		**Y**	**A**	**C**	**C**	A	**G**	G		A	**G**		A	**c**	**g**	**g**	A	**G**		G	A	**a**	G	A	C	G		**A**	**T**	**G**	**C**
39c10		A	T	T		C	G	T	T	**R**	A	**R(*A*)**		**R**	**G**		**G**	C	G	G	A	**R(*G*)**		G	A	A	**R**	**R**	C	G		G	C	A	**t(*C*)**
49c11		A	**t(*C*)**	T		C	G	T	T	A	A	**g(*A*)**		A	**a(*G*)**		A	C	G	G	**a**	**a(*G*)**		G	**a**	A	G	A	C	G		G	C	A	T
54c14		-	-	-		-	-	-	-	**-**	-	**-**		A	**R(*G*)**		A	C	G	G	A	A		**R**	A	A	G	A	**Y**	**R**		G	C	A	T

***bg* (784 of 819 bp, 22 to 805)**		***Base number from start codon***																												
Sample	***Differentiation***	63	129	147	204	210	228	273	369	438	460	468	495	516	528	564	588	597	606	609	645	648	654	657	660	693	705	729		
2c2	Intra-A										C	T							C									A		
46c2											T	C							T									G		
B											C	C							C									G		
A	Intra-B	C	C	C	A	C	G	G	C	C			C	C	C	C	C	G		T	C	G	C	C	C	C	G			
15c1/30c7/33c3/34c8/42c5		T	C	C	A	T	A	A	T	T			C	C	C	T	C	G		C	C	G	C	C	C	C	G			
7c3		T	C	C	**R(*G*)**	**t(*C*)**	**R**	A	**C**	**Y(*C*)**			C	**Y(*T*)**	C	**Y**	C	G		**Y**	**Y**	G	**Y**	C	C	C	G			
39c10		T	C	C	A	T	A	A	T	**C**			**T**	**T**	**c**	T	C	G		C	C	**g**	C	**c**	C	**c**	**g**			
49c11		T	**T**	C	**G**	T	A	A	**C**	**C**			C	**T**	C	T	**c**	**g**		C	C	G	C	C	**T**	C	G			
54c14		**C**	C	**c**	**G**	**t(*C*)**	A	**a**	**C**	**C**			C	**T**	C	T	-	-		-	-	-	-	-	-	-	-			

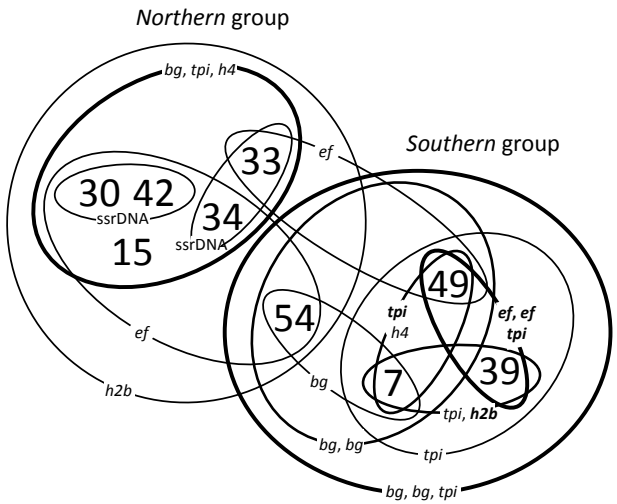
***gdh* (636 of 1350 bp, 689 to 1324)**		***Base number from start codon***																				
Sample	***Differentiation***	699	753	**794**	807	822	831	861	867	870	894	**902**	921	933	963	969	1014	1080	1143	**1178**	1254	1266
2c2	Intra-A	T	C		C		C	T	T	T	C	C						T			G	
46c2		C	T		T		T	C	C	C	T	T						C			A	
B		C	C		T		C	C	C	C	C	T						C			C	
A	Intra-B			T		G							G	C	T	C	C		C	A	C	
7c3				T		**g**							**R**	**c**	**c**	**Y**	C		**c**	A	**T**	
15c1				T		G							A	C	C	C	C		C	A	C	
30c7				T		G							A	C	C	C	C		C	A	**T**	
33c3				**t**		G							A	C	C	C	C		C	**a**	C	
34c8				T		G							A	C	C	C	C		C	A	**Y**	
39c10				T		G							A	C	C	C	**T**		C	A	C	
42c5				T		G							A	C	C	C	C		C	A	**T**	

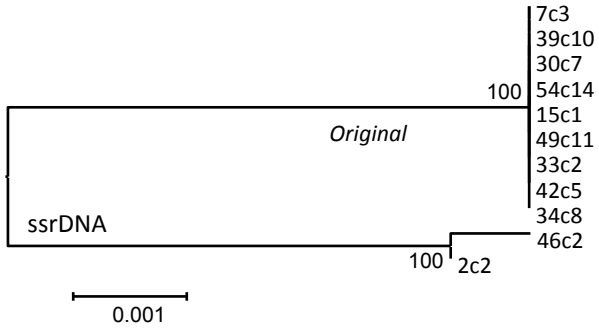
***ef* (1253 of 1329 bp, 21 to 1273)**		***Base number from start codon***																		
Sample	***Differentiation***	177	204	228	240	**256**	336	**448**	**524**	**539**	657	693	**797**	840	912	915	936	939	969	1161
2c2	Intra-A		G	G	Y															Y
46c2			R	K	C															C
B			G	G	C															C
A	Intra-B	C				A	G	A	T	A	C	C	A	C	C	C	C	C	C	
15c1/30c7/34c8/42c5/54c14		C				A	G	A	T	A	C	C	A	C	C	G	C	C	C	
7c3		**c**				A	G	A	T	A	**Y**	C	A	C	**Y**	G	C	**c**	**c**	
39c10		C				**R(*G*)**	G	**a**	**Y(*C*)**	A	C	C	**a**	C	C	**g**	C	C	C	
49c11		C				**g**	**R**	A	**Y(*C*)**	**R**	C	**T**	A	**T**	C	G	**Y**	C	C	
33c3		C				A	G	A	T	A	C	C	A	**T**	C	G	C	C	C	

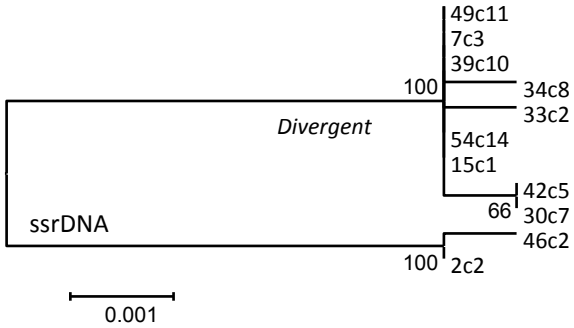
ssrDNA (1432 of 1450 bp, 6 to 1437)		***Base number from start codon***						
Sample	***Differentiation***	25	95	1079	1195	1218	1299	1398
2c2	Intra-A						C	
46c2							T	
B							C	
A	Intra-B	C	C	C	C	C		G
7c3/39c10/54c14		C	C	C	C	C		G
15c1		C	**c**	C	C	C		G
30c7		C	C	C	C	**Y**		**R(*A*)**
33c3		**Y**	C	C	**Y(*T*)**	C		G
34c8		C	C	C	**Y(*T*)**	C		G
42c5		C	C	C	C	C		**R(*A*)**
49c11		C	C	**Y**	C	C		G

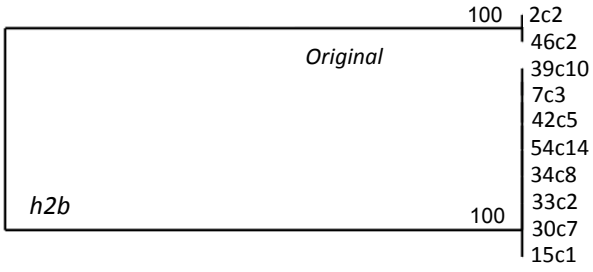
***h4* (265 of 300bp, 29 to 293)**				
Sample	99	177	213	258
2c2	T		A	
46c2	t		G	
B	T		G	
A		C		T
15c1/30c7/33c3/34c8/42c5		C		G
7c3		**c(*A*)**		G
39c10		C		**A**
49c11		**A**		G

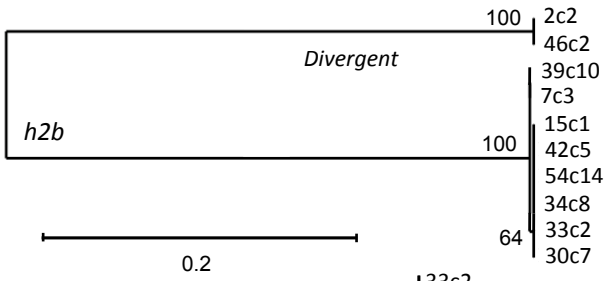
***h2b* (381 of 393bp, 10 to 390)**		
Sample	**37**	303
A	A	C
15c1/30c7/33c3/34c8/42c5/54c14	G	C
7c3	**g(*A*)**	**c**
39c10	**R(*A*)**	C

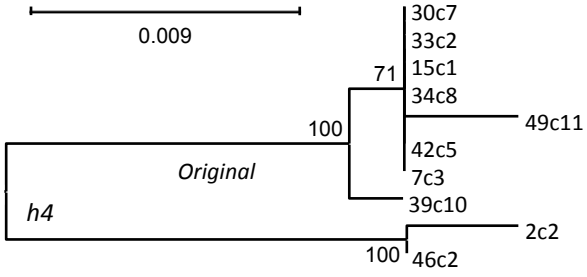


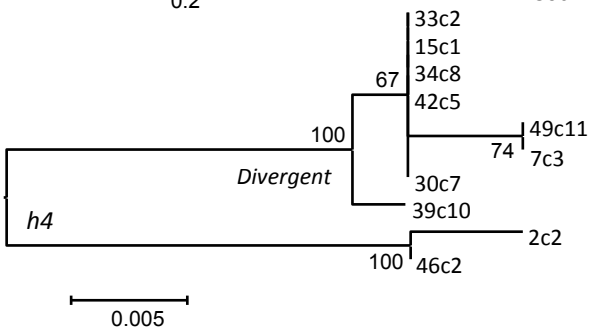


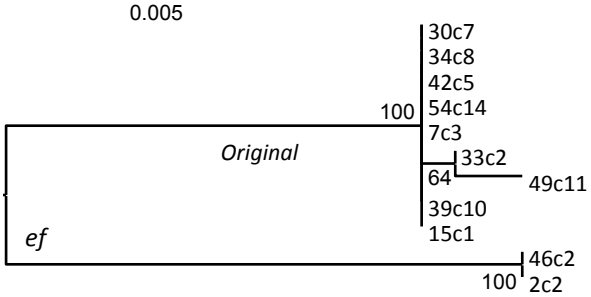


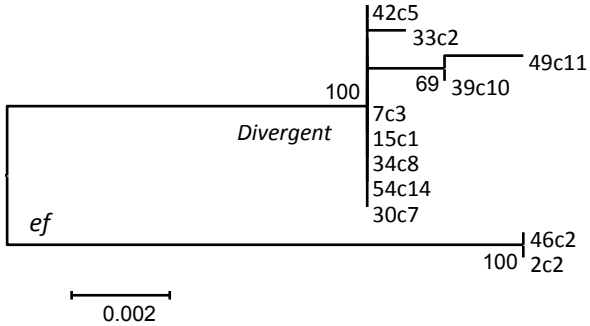


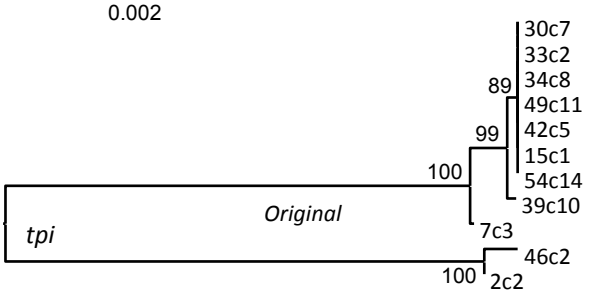


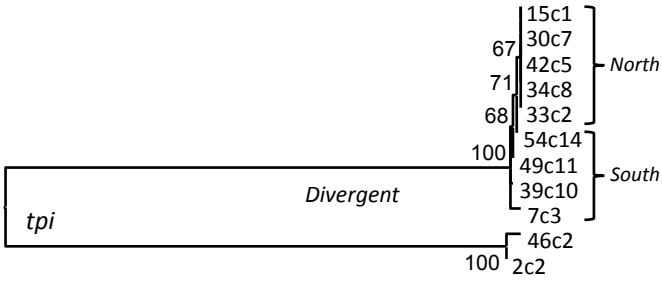


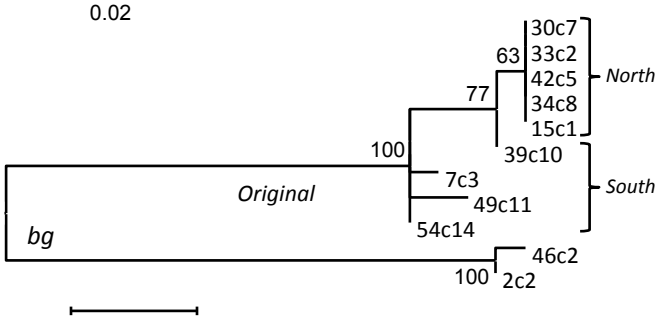


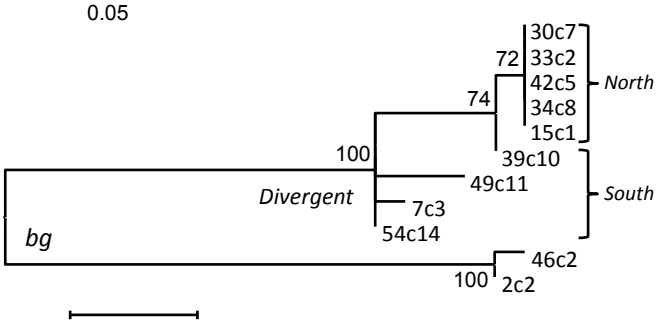


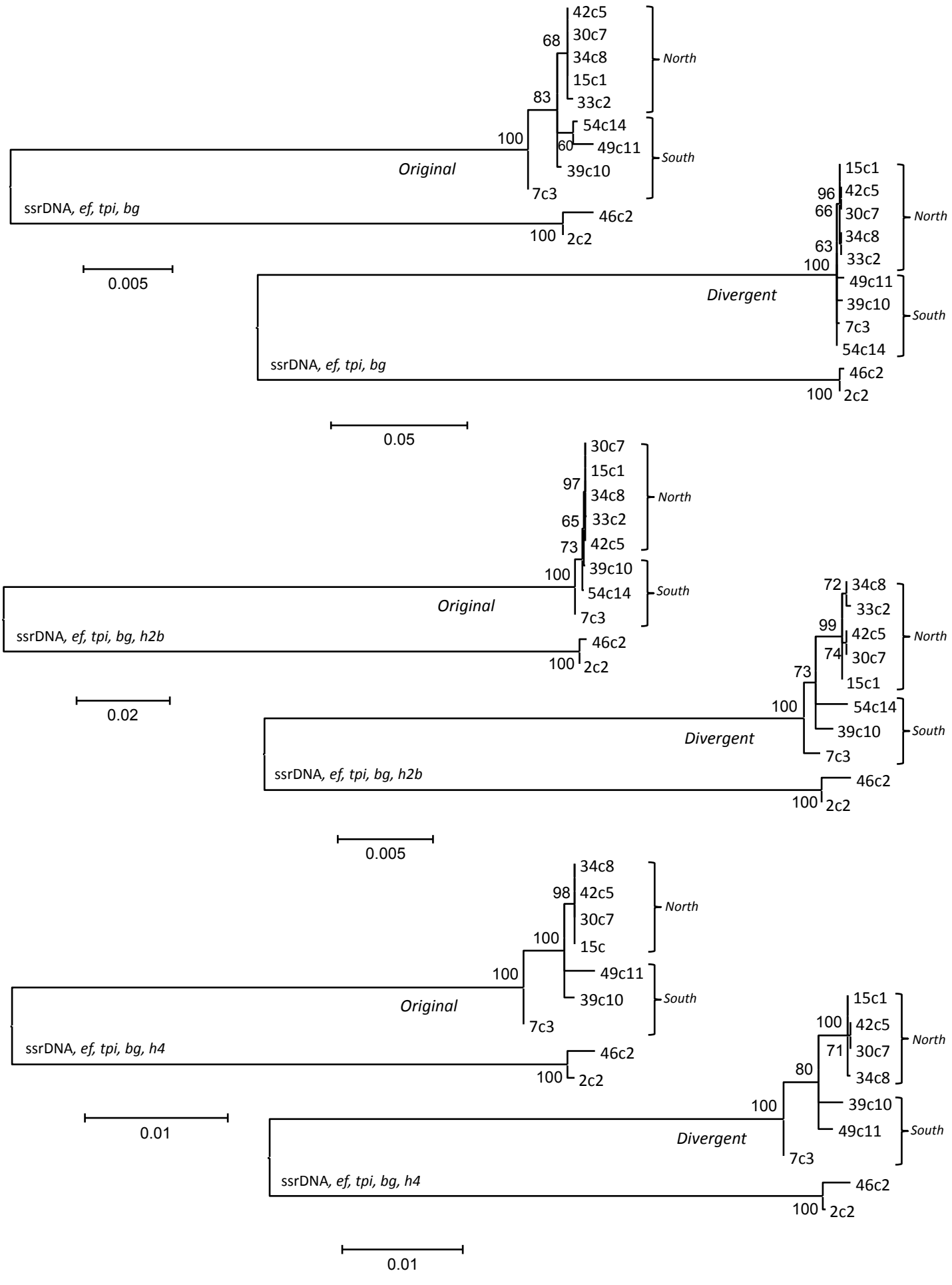












606 Table 1. List of cloned, cultured isolates, collection origin and assemblage.

Isolate	Town	Region	Assemblage
BAH 2c2	Woodanilling	South-Western Australia	AI
BAH 46c2	Perth	South-Western Australia	AII
BAH 7c3	Katanning	South-Western Australia	B
BAH 15c1	Kununurra	North-Western Australia	B
BAH 30c7	Derby	North-Western Australia	B
BAH 33c2	Perth	South-Western Australia	B
BAH 34c8	Kununurra	North-Western Australia	B
BAH 39c10	Perth	South-Western Australia	B
BAH 42c5	Karratha	North-Western Australia	B
BAH 49c11	Northam	South-Western Australia	B
BAH 54c14	Kununurra	North-Western Australia	B

607

608

609

610