DETERMINATION OF DISCRIMINATORY POWER OF GENETIC MARKERS USED FOR GENOTYPING *GIARDIA DUODENALIS*

Suradej Siripattanapipong[1], Saovanee Leelayoova[2], Mathirut Mungthin[2], RC Andrew Thompson[3], Parima Boontanom[2], Wilai Saksirisamphant[4] and Peerapan Tan-ariya[1]

[1]Department of Microbiology, Faculty of Science, Mahidol University, Bangkok; [2]Department of Parasitology, Phramongkutklao College of Medicine, Bangkok, Thailand; [3]WHO Collaborating Centre for the Molecular Epidemiology of Parasitic Infections, School of Veterinary and Biomedical Sciences, Murdoch University, Murdoch, Western Australia, Australia; [4]Department of Parasitology, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand

Abstract. Small subunit ribosomal DNA (SSU-rDNA), glutamate dehydrogenase (*gdh*), β-giardin, triosephosphate isomerase (*tpi*), and elongation factor 1-alpha (*ef1*-α) genes are useful genetic markers for genotypic analysis of the intestinal protozoan, *Giardia duodenalis* (syn. *G. lamblia*, *G. intestinalis*), the cause of enteric disease in humans. To quantitatively compare the discriminatory power of these loci, 43 fecal samples were collected from central, northern and eastern Thailand and *G. duodenalis* specimens were analyzed using PCR-based genotyping and subcloning methods. Approximately equal prevalence of assemblage A (21) and B (22) were present among these populations. Analysis of Simpson's index and Wallace coefficient values from assemblage B isolates together with the data obtained from GenBank showed that the combination of two loci provides a higher discrimination power for subgenotyping *G. duodenalis* than using any single locus.

Keywords: *Giardia duodenalis*, multilocus typing, genetic diversity, discriminatory power

INTRODUCTION

Giardia duodenalis has increasingly gained attention as a common cause of diarrheal disease in humans in both developed and developing countries (Savioli *et al*, 2006). In developing countries of Asia, Africa, and Latin America, approximately 200 million people are infected with this organism with an average of 500,000 new cases per year (Thompson *et al*, 2000). The high number of infected cases in these countries is mainly due to poor sanitation and personal hygiene (Thompson *et al*, 2000). To achieve effective control and prevention strategies, an indepth understanding of the epidemiology of *Giardia* infection is required particularly of the patterns of spread and routes of transmission.

Genetic characterization has been used to evaluate the genetic variability

Correspondence: Dr Peerapan Tan-ariya, Department of Microbiology, Faculty of Science, Rama VI Road, Ratchathewi, Bangkok 10400 Thailand.
Tel: +66 (0) 2201 5524; Fax: +66 (0) 2644 5411
E-mail: scptn@mahidol.ac.th

within *G. duodenalis*, and to assess the role of the disease epidemiology (Caccio *et al*, 2005; Smith *et al*, 2006). It has been shown that *G. duodenalis* exhibits considerable heterogeneity with genetically and phenotypically distinct strains (Andrews *et al*, 1989; Hopkins *et al*, 1997). Characterization of the glutamate dehydrogenase (*gdh*) gene using PCR sequencing and phylogenetic analysis indicated that *G. duodenalis* can be grouped into eight major genetic groups, termed assemblages A-H (Hopkins *et al*, 1997; Monis *et al*, 2003; Read *et al*, 2004; Lasek-Nesselquist *et al*, 2010). Most of the infections with assemblages A and B are associated with humans (Souza *et al*, 2007; Volotao *et al*, 2007; Sahagun *et al*, 2008). Moreover, molecular studies have also revealed the existence of genetic structures within each of these assemblages. Assemblage A is comprised of two distinct clusters, AI and AII, while assemblage B consists of BIII and BIV. Of these two major assemblages, assemblage B possesses significantly greater nucleotide diversity than assemblage A (Caccio and Sprong, 2010).

To date, studies have been conducted using PCR amplification and sequencing analysis of a number of genetic loci of *G. duodenalis, eg*, *gdh*, *tpi*, SSU-rDNA, and β-giardin genes. The results showed that some gene polymorphisms could reproducibly group isolates into assemblages (Wielinga and Thompson, 2007). *Gdh* and *tpi* can be used to identify all assemblages and subassemblages, whereas SSU-rDNA and β-giardin genes showed uncomparable results wherein the subassemblages of assemblage B could not be identified. However, some multilocus studies have occasionally detected discordance of the assemblage identification between loci, and indicated that they carried different genetic information (Traub *et al*, 2004; Caccio *et al*, 2008; Geurden *et al*, 2009). When using only a single gene marker in an epidemiological study, these phenomena may result in the distortion of subgenotypic identification. Nevertheless, due to the lack of evaluative information regarding the usefulness of these different loci in identifying isolates, these loci could independently become selected for genotypic characterization in each study without any specific protocols. Therefore, a standard protocol for using genetic markers to study the molecular epidemiology of *G. duodenalis* is necessary to be systemically performed.

Thus, this study evaluated the discriminatory power of genotypic markers for identifying nucleotide diversity within subgenotypes of *G. duodenalis* assemblage B.

MATERIALS AND METHODS

Parasite isolates

Forty-three fecal specimens of *G. duodenalis* were obtained from three regions of Thailand as part of a public health survey. Ten isolates were from the north, 19 from the east, and 13 from the central region of Thailand.

This study was approved by the Royal Thai Army Medical Department Ethics Committee, Thailand. Written informed consents were provided by all study participants or their legal guardians.

Specimen processing and DNA extraction

Giardia cysts were concentrated by sodium nitrate flotation technique (O' Grady and Slocombe, 1980). In brief, approximately 2 g of fresh stool were added to 2 ml of 60% sodium nitrate and mixed until homogenized. After filtering through gauze, the filtrate was added with 20 ml of saturated sodium nitrate solution, mixed

and left at room temperature for 20 minutes. One ml aliquot of the mixture surface was sedimented, washed twice with phosphate buffered saline (PBS). Finally, the sediment was suspended in 1.5 ml of PBS and kept at -20°C until used.

Fifteen µl aliquot of each specimen was applied onto a 6 mm-diameter FTA disk (Whatman, Bioscience, Newton, MA) and then air-dried overnight. Disks were washed twice with 200 µl of FTA purification reagent (Whatman, Bioscience) for 5 minutes, then washed twice with 200 µl of TE^{-1} buffer (10 mM Tris-HCl pH 8.0; 0.01 mM EDTA pH 8.0) for 5 minutes and air-dried overnight. FTA disks were used directly as DNA templates in PCR. QIAamp Stool Mini Kit (Qiagen, Hilden, Germany) was additionally used for DNA extraction of specimens that gave negative results with the FTA method.

DNA amplification, cloning and sequencing

Segments of SSU-rDNA (176 bp), β-giardin (511 bp), *gdh* (456 bp), *tpi* (530 bp), and *ef1*-α (191 bp) genes were individually amplified as previously described (Hopkins *et al*, 1997; Monis *et al*, 1999; Sulaiman *et al*, 2003; Lalle *et al*, 2005; Boontanom *et al*, 2010). The amplicons were electrophoresed on a 1.5% agarose gel and purified using QIAquick gel extraction kit (Qiagen, Germany) according to the manufacturer's instructions.

All amplicons were bidirectionally sequenced using their respective secondary primers, and then manually assembled and edited using BioEdit version 7.0.1. When a singleton substitution was found, the sequencing was repeated with the amplicon from another independent PCR amplification. If heterozygosity was identified, subcloning of this amplicon was performed to confirm the existence of multiple templates by inserting and transfecting into pGEM-T Easy vector (Promega, Madison, WI) JM109 competent cells. Recombinant plasmids were purified from three to ten transformed colonies using HiYield Plasmid Mini Kit (RBC bioscience, Taiwan) and sequenced using universal primer SP6. DNA sequencing was conducted by 1st Base, Singapore.

Sequence analysis

Sequences were aligned with SSU-rDNA gene and *gdh* reference sequences obtained from GenBank, using the program ClustalX version 2.0.12 (Larkin *et al*, 2007). The reference sequences used for alignment were as follows: M54878 (SSU-rDNA assemblage A); U09492 (SSU-rDNA assemblage B); L40509 (*gdh* assemblage AI); L40510 (*gdh* assemblage AII); AF069059 (*gdh* assemblage BIII); L40508 (*gdh* assemblage BIV) (Read *et al*, 2004).

Comparison of discriminatory power

The diversity index and Wallace's coefficients were calculated employing web server http://darwin.phyloviz.net/ComparingPartitions/index.php?link=Tool. Any different allele found was considered as distinct type.

Simpson's index of diversity (Hunter and Gaston, 1988) was applied to measure the discriminatory ability of each locus. This index indicates the probability of two isolates sampled randomly from a population belonging to two different types. Simpson's index (*D*) and confidence intervals (CIs) are calculated from the following equations:

$$D = 1 - \frac{1}{N(N-1)} \sum_{j=1}^{S} n_j (n_j - 1)$$

$$\sigma^2 = \frac{4}{N} \left[\sum \pi_j^3 - \left(\sum \pi_j^2 \right)^2 \right]$$

$$CI = \left[D - 2\sqrt{\sigma^2}, D + 2\sqrt{\sigma^2} \right]$$

Table 1
The number of allele types, discriminatory power and 95% confident interval of each gene in this study.

Locus	No. of allele types	*D*	95% CI
SSU-rDNA gene	1	0	0.00 - 0.00
gdh	8	100	100.00 - 100.00
β-giardin gene	7	96.43	88.77 - 104.08
tpi	8	100	100.00 - 100.00
ef1-α	2	42.86	12.24 - 73.48

where N is the total number of isolates in the sample population, S the total number of types described, n_j the number of isolates belonging to the jth type, and π_j the frequency n_j/N.

The Wallace's coefficient (Wallace, 1983) is the clustering comparison index for evaluation of the strength and the directionality of the correspondences between typing loci. Given two partition schemes of the same data set, P and P', these coefficients are calculated based on the fact that a pair of isolates from the data set will fall into one of the following conditions: (a) the number of point pairs that are in the same cluster in P and P', (b) the number of point pairs that are in the same cluster in P but not in P', (c) the number of point pairs that are in the same cluster in P' but not in P, or (d) the number of point pairs that are in different clusters in P and P'. The coefficients can then be defined as

$$W_{ij}(P, P') = \frac{a}{a+b}$$

$$W_{ji}(P, P') = \frac{a}{a+c}$$

Wallace's coefficients provide an estimate of how much new information is obtained from another typing locus. A low value of Wallace's coefficient indicates that types defined by a given locus could have been predicted from the results of another locus, suggesting that the use of this locus is redundant.

RESULTS

Amplification of target genes

Due to a limited amount of DNA, 4 of 43 isolates were not available for PCR amplifications except for *gdh* gene. PCR reactions were positive for 42 (98%), 38 (97%), 35 (90%), 24 (61%) and 22 (56%) of the isolates for *gdh*, SSU-rDNA, *ef1*-α, β-giardin, and *tpi* gene, respectively.

Assemblage assignment and sequence analysis

Sequence alignments for assemblage assignment using data from the SSU-rDNA and *gdh* genes showed that 21 isolates were assemblage A, while the other 22 isolates were assemblage B.

Direct sequencing of the amplicons from SSU-rDNA, β-giardin, and *ef1*-α genes showed that all of these isolates had homozygous gene at these regions, whereas the two remaining genes, *gdh* and *tpi*, showed heterozygous sequences.

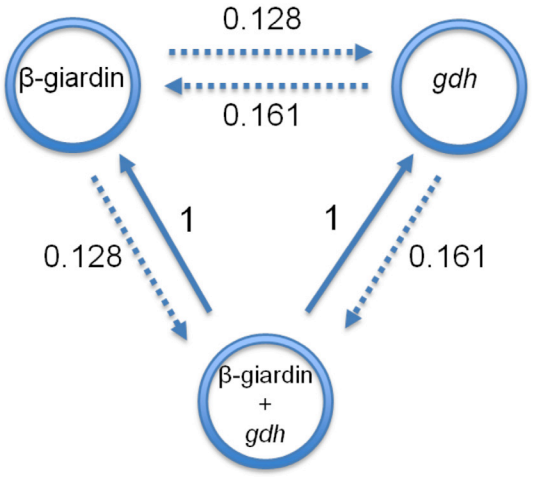

Fig 1–Representation of correspondences between the typing genes used in the 30 isolates, calculated by Wallace's coefficients. Sequence data were from this study and those of Robertson *et al* (2006).

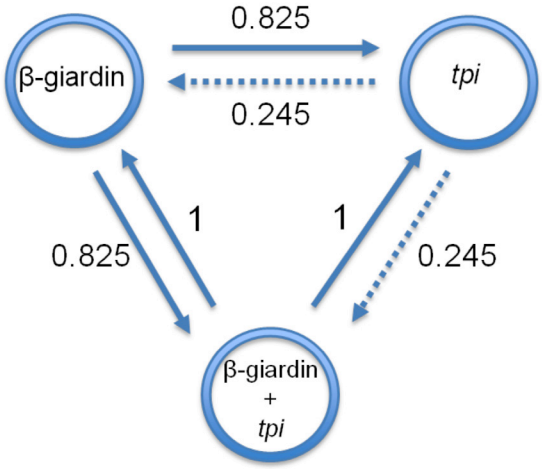

Fig 2–Representation of correspondences between the typing genes used in the 34 isolates, calculated by Wallace's coefficients. Sequence data were from this study and those of Geurden *et al* (2009).

Analysis of all assemblage B sequences from all genes showed that only the SSU-rDNA gene contained no variation. For *gdh*, a total of 66 sequences/clones were selected for sequencing showing that they were 52 different alleles when compared with their reference sequences (data not shown). For *tpi*, from 26 sequences/clones, there were 25 different alleles. In addition, seven and two different alleles were observed in the β-giardin and *ef1*-α gene, respectively.

Discriminatory power comparison

Simpson's index of diversity (*D*) provides a measurement of the discriminatory power of different genes in assemblage B. Only eight isolates, having complete PCR amplifications of the five genes, were eligible for the calculation. Table 1 summarizes the coefficient for each gene used in this study. The discrimination power of *gdh*, *tpi*, β-giardin, *ef1*-α, and SSU-rDNA genes was 100, 100, 96.43, 42.86, and 0%, respectively. The SSU-rDNA gene had no discriminatory power because all sequences were identical. On the other hand, all eight isolates showed totally different alleles in the *gdh* and *tpi* regions. By comparing the *D* values and 95% CI values of the four variable genes, only the 95% CI value of the *ef1*-α did not overlap with the other genes. Thus, *gdh*, β-giardin, and *tpi* genes were likely to have similar discriminatory powers higher than that of *ef1*-α.

To confirm the results of the SSU-rDNA gene and *ef1*-α, the sequence data of these genes retrieved from GenBank were included in the calculation. Overall, 22 GenBank acquired sequences of the SSU-rDNA gene, were identical to the sequences in this study, and the discriminatory power of the SSU-rDNA gene still remained at 0%. Results obtained

for all eight acquired sequences of *ef1*-α reduced the discriminatory power of this gene from 42.86% to 23.55% (1.79-45.31). Thus, studied regions of the SSU-rDNA gene and *ef1*-α were likely to have low discriminatory power and were less useful for typing this organism at the subassemblage level.

For sequence data of *gdh*, β-giardin gene, and *tpi*, in order to increase the reliability of the estimations, additional data were obtained from the GenBank database. Unfortunately, as the sequence data of the *gdh*, β-giardin gene, and *tpi* from GenBank comprised different amplified regions, they could not be used for sequence analysis in this study. Only sequence data of the *gdh* and β-giardin gene from the study of Robertson *et al* (2006) and sequence data of β-giardin and *tpi* gene from that of Geurden *et al* (2009) were matched with the amplified regions of this study. Thus, the estimation was separated into two analyses.

The first analysis was performed using the *gdh*, β-giardin gene, and *gdh*+β-giardin datasets. Amplifications of these loci were successful for 11 isolates in this study. The combination of datasets from this study and those of Robertson *et al* (2006) increased this number to 30 isolates. The number of distinct alleles was 18, 15, and 26 for the *gdh*, β-giardin gene, and *gdh*+β-giardin genes, respectively. The results showed similar index values for the *gdh* and β-giardin gene at 0.161 and 0.128, indicating a strong correlation between these two genes. In contrast, *gdh*+β-giardin showed higher value when compared with the *gdh* or β-giardin gene alone (Fig 1). When using *gdh*+β-giardin as the standard of comparison, the value of Wallace's index was 1. This was the highest value obtained among all three datasets, meaning that if two isolates were identified as the same type by the data from the *gdh* or β-giardin gene, there was a 100% chance to identify a different type by the data from *gdh*+β-giadin gene. This reflects that the *gdh*+β-giardin improves the classification when compared with a single locus, *gdh* or β-giardin gene.

The second analysis was performed using β-giardin gene, *tpi*, and β-giardin+*tpi* datasets. Amplification of these loci were successful for the eight isolates in this study. Combination of datasets from this study and those of Geurden *et al* (2009) increased the number of isolates to 34. The number of distinct alleles was 13, 10, and 18 for the β-giardin gene, *tpi*, and β-giardin+*tpi* genes, respectively. When considering the β-giardin gene and β-giardin+*tpi* genes as the standard of comparison, the value of Wallace's index was 1 and 0.825 respectively, whereas the index value from the *tpi* gene was 0.245. The data indicated a similar index value between the information provided by the β-giardin gene and β-giardin+*tpi* genes (Fig 2), showing that β-giardin+*tpi* and β-giardin genes improved the classification when compared to *tpi*. However, the combination of two genes and the β-giardin+*tpi* genes still provided slightly more information than the single β-giardin gene.

DISCUSSION

The main objective of this study was to measure and compare the discriminatory power of five genetic markers most widely used for genotyping of *G. duodenalis*. This study focused on the data from assemblage B isolate because it contained significantly more genetic diversity than assemblage A. Our study provided a framework for the quantitative assessment of correspondence between type assignments obtained from different loci.

The quantification was achieved by the use of Simpson's index of diversity and Wallace's coefficient. An important application of this study was the evaluation of the capability to identify subtypes by using different loci, leading to the selection of the most informative and useful locus to genetically characterize the isolates.

The significantly low discriminatory powers of SSU-rDNA gene and *ef1*-α indicated that they could not be used to genotype subassemblage. In spite of *gdh*, β-giardin, and *tpi* loci showing similar discriminatory powers, due to the dataset restrictions, the comparative analysis among these genes had to be separated into two analyses. For both analyses, when compared with a single locus, the combination of two loci showed significant improvement of the classification of the isolates. These were indicated by the higher coefficient values obtained.

In summary, this study showed that Simpson's index calculation indicated that SSU-rDNA gene and *ef1*-α loci had lower discriminatory power than *gdh*, β-giardin gene, and *tpi* loci, and that the combination of two loci provided more discriminatory power than using a single locus. Although the comprehensive comparison between these analyses implied that the combination of the loci enhanced the classification, unfortunately, these results could not be directly compared since the datasets were from different sources. Further studies are necessary to clarify which of the typing loci would provide a more discriminatory and informative genotyping system.

ACKNOWLEDGEMENTS

This study was supported by the Thailand Research Fund (Grant No. BRG4880003), Phramongkutklao College of Medicine Research Fund, and The Royal Golden Jubilee PhD Program (Grant No. PHD/0236/2548).

REFERENCES

Andrews RH, Adams M, Boreham PF, Mayrhofer G, Meloni BP. *Giardia intestinalis*: electrophoretic evidence for a species complex. *Int J Parasitol* 1989; 19: 183-90.

Boontanom P, Siripattanapipong S, Mungthin M, Tan-Ariya P, Leelayoova S. Improved sensitivity of PCR amplification of glutamate dehydrogenase gene for detection and genotyping of *Giardia duodenalis* in stool specimen. *Southeast Asian J Trop Med Public Health* 2010; 41: 280-4.

Caccio SM, Beck R, Lalle M, Marinculic A, Pozio E. Multilocus genotyping of *Giardia duodenalis* reveals striking differences between assemblages A and B. *Int J Parasitol* 2008; 38: 1523-31.

Caccio SM, Sprong H. *Giardia duodenalis*: genetic recombination and its implications for taxonomy and molecular epidemiology. *Exp Parasitol* 2010; 124: 107-12.

Caccio SM, Thompson RC, Mclauchlin J, Smith HV. Unravelling *Cryptosporidium* and *Giardia* epidemiology. *Trends Parasitol* 2005; 21: 430-7.

Geurden T, Levecke B, Caccio SM, *et al*. Multilocus genotyping of *Cryptosporidium* and *Giardia* in non-outbreak related cases of diarrhoea in human patients in Belgium. *Parasitology* 2009; 136: 1161-8.

Hopkins RM, Meloni BP, Groth DM, Wetherall JD, Reynoldson JA, Thompson RC. Ribosomal RNA sequencing reveals differences between the genotypes of *Giardia* isolates recovered from humans and dogs living in the same locality. *J Parasitol* 1997; 83: 44-51.

Hunter PR, Gaston MA. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. *J Clin Microbiol* 1988; 26: 2465-6.

Lalle M, Jimenez-cardosa E, Caccio SM, Pozio E. Genotyping of *Giardia duodenalis* from

humans and dogs from Mexico using a beta-giardin nested polymerase chain reaction assay. *J Parasitol* 2005; 91: 203-5.

Larkin MA, Blackshields G, Brown NP, *et al*. Clustal W and Clustal X version 2.0. *Bioinformatics* 2007; 23: 2947-8.

Lasek-Nesselquist E, Welch DM, Sogin ML. The identification of a new *Giardia duodenalis* assemblage in marine vertebrates and a preliminary analysis of *G. duodenalis* population biology in marine systems. *Int J Parasitol* 2010; 40: 1063-74.

Monis PT, Andrew RH, Mayrhofer G, Ey PL. Molecular systematics of the parasitic protozoan *Giardia intestinalis*. *Mol Biol Evol* 1999; 16: 1135-44.

Monis PT, Nadrews RH, Mayrhofer G, Ey PL. Genetic diversity within the morphological species *Giardia intestinalis* and its relationship to host origin. *Infect Genet Evol* 2003; 3: 29-38.

O' Grady MR, Slocombe JO. An investigation of variables in a fecal flotation technique. *Can J Comp Med* 1980; 44: 148-57.

Read CM, Monis PT, Thompson RC. Discrimination of all genotypes of *Giardia duodenalis* at the glutamate dehydrogenase locus using PCR-RFLP. *Infect Genet Evol* 2004; 4: 125-30.

Robertson LJ, Hermansen L, GJerde BK, Strand E, Alvsvag JO, Langeland N. Application of genotyping during an extensive outbreak of waterborne giardiasis in Bergen, Norway, during autumn and winter 2004. *Appl Environ Microbiol* 2006; 72: 2212-7.

Sahagun J, Clavel A, Goni P, *et al*. Correlation between the presence of symptoms and the *Giardia duodenalis* genotype. *Eur J Clin Microbiol Infect Dis* 2008; 27: 81-3.

Savioli L, Smith H, Thompson A. *Giardia* and *Cryptosporidium* join the 'Neglected Diseases Initiative'. *Trends Parasitol* 2006; 22: 203-8.

Smith HV, Caccio SM, Tait A, Mclauchlin J, Thompson RC. Tools for investigating the environmental transmission of *Cryptosporidium* and *Giardia* infections in humans. *Trends Parasitol* 2006; 22: 160-7.

Souza SL, Gennari SM, Richtzenhain LJ, *et al*. Molecular identification of *Giardia duodenalis* isolates from humans, dogs, cats and cattle from the state of Sao Paulo, Brazil, by sequence analysis of fragments of glutamate dehydrogenase (gdh) coding gene. *Vet Parasitol* 2007; 149: 258-64.

Sulaiman IM, Fayer R, Bern C, *et al*. Triosephosphate isomerase gene characterization and potential zoonotic transmission of *Giardia duodenalis*. *Emerg Infect Dis* 2003; 9: 1444-52.

Thompson RC, Hopkins RM, Homan WL. Nomenclature and genetic groupings of *Giardia* infecting mammals. *Parasitol Today* 2000; 16: 210-3.

Traub RJ, Monis PT, Robertson I, Irwin P, Mencke N, Thompson RC. Epidemiological and molecular evidence supports the zoonotic transmission of *Giardia* among humans and dogs living in the same community. *Parasitology* 2004; 128: 253-62.

Volotao AC, Costa-macedo LM, Haddad FS, Brandao A, Peralta JM, Fernandes O. Genotyping of *Giardia duodenalis* from human and animal samples from Brazil using beta-giardin gene: a phylogenetic analysis. *Acta Trop* 2007; 102: 10-9.

Wallace DL. A method for comparing two hierarchical clusterings: comment. *J Am Stat Assoc* 1983; 78: 569-76.

Wielinga CM, Thompson RC. Comparative evaluation of *Giardia duodenalis* sequence data. *Parasitology* 2007; 134: 1795-821.