AMERICAN SOCIETY FOR MICROBIOLOGY | Applied and Environmental Microbiology®

ENVIRONMENTAL MICROBIOLOGY

Hidden Diversity within Common Protozoan Parasites as Revealed by a Novel Genomotyping Scheme

Matthew H. Seabolt,[a,b,d] Konstantinos T. Konstantinidis,[b,c] Dawn M. Roellig[a]

[a]Division of Foodborne, Waterborne, and Environmental Diseases, National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA
[b]School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia, USA
[c]School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA
[d]CFD Research Corporation, Huntsville, Alabama, USA

ABSTRACT *Giardia duodenalis* (syn. *Giardia lamblia*, *Giardia intestinalis*) is the causative agent of giardiasis, one of the most common diarrheal infections in humans. Evolutionary relationships among *G. duodenalis* genotypes (or subtypes) of assemblage B, one of two genetic assemblages causing the majority of human infections, remain unclear due to poor phylogenetic resolution of current typing methods. In this study, we devised a methodology to identify new markers for a streamlined multilocus sequence typing (MLST) scheme based on comparisons of all core genes against the phylogeny of whole-genome sequences (WGS). Our analysis identified three markers with resolution comparable to that of WGS data. Using newly designed PCR primers for our novel MLST loci, we typed an additional 68 strains of assemblage B. Analyses of these strains and previously determined genome sequences showed that genomes of this assemblage can be assigned to 16 clonal complexes, each with unique gene content that is apparently tuned to differential virulence and ecology. Obtaining new genomes of *Giardia* spp. and other eukaryotic microbial pathogens remains challenging due to difficulties in culturing the parasites in the laboratory. Hence, the methods described here are expected to be widely applicable to other pathogens of interest and advance our understanding of their ecology and evolution.

IMPORTANCE *Giardia duodenalis* assemblage B is a major waterborne pathogen and the most commonly identified genotype causing human giardiasis worldwide. The lack of morphological characters for classification requires the use of molecular techniques for strain differentiation; however, the absence of scalable and affordable next-generation sequencing (NGS)-based typing methods has prevented meaningful advancements in high-resolution molecular typing for further understanding of the evolution and epidemiology of assemblage B. Prior studies have reported high sequence diversity but low phylogenetic resolution at standard loci in assemblage B, highlighting the necessity of identifying new markers for accurate and robust molecular typing. Data from comparative analyses of available genomes in this study identified three loci that together form a novel high-resolution typing scheme with high concordance to whole-genome-based phylogenomics and which should aid in future public health endeavors related to this parasite. In addition, data from newly characterized strains suggest evidence of biogeographic and ecologic endemism.

KEYWORDS *Giardia duodenalis*, bioinformatics, molecular sequence typing, parasites, waterborne disease

Over the past decade, molecular techniques have brought new insights into the global diversity of microbial eukaryotic pathogens such as *Giardia* and *Cryptosporidium* protozoa. These molecular surveys have been traditionally limited to the sequence or

Citation Seabolt MH, Konstantinidis KT, Roellig DM. 2021. Hidden diversity within common protozoan parasites as revealed by a novel genomotyping scheme. Appl Environ Microbiol 87:e02275-20. https://doi.org/10.1128/AEM.02275-20.

Editor Johanna Björkroth, University of Helsinki

This is a work of the U.S. Government and is not subject to copyright protection in the United States. Foreign copyrights may apply.

Address correspondence to Konstantinos T. Konstantinidis, kostas@ce.gatech.edu.

Received 15 September 2020
Accepted 30 November 2020
Accepted manuscript posted online 4 January 2021
Published 26 February 2021

Downloaded from https://journals.asm.org/journal/aem on 17 July 2021 by 2a02:587:d449:5600:9805:7ecd:78fc:2edf.

restriction fragment polymorphism analysis of short fragments of a few ($n = 2$ to 6) marker genes (1–6). While this approach has served public health risk assessment effectively, it has also become clear that the low resolving power of these markers at the subspecies level hinders determining the epidemiologic importance of different genotypes (7–10). Typing eukaryotic pathogens using genomic approaches has not advanced as much as for several of their bacterial counterparts, such as *Escherichia coli* and *Bacillus anthracis* (11–13). In addition to the traditional reliance on low-resolution markers mentioned above, widespread adoption of whole-genome sequencing (WGS) for pathogens like *Giardia* and *Cryptosporidium* is hindered by difficulties with *in vitro* culturing and/or generating sufficient quantities of DNA for sequencing from clinical and environmental samples (14). Accordingly, there is high need for the development of new, culture-independent methods to accurately capture and describe diversity for epidemiologic and epizootic investigations. Here, we provide a scalable, systematic approach to molecular typing of microbial eukaryotes using *Giardia* as the case study, which is applicable to any organism of interest, particularly those for which obtaining isolates or whole-genome sequencing is challenging.

Giardia duodenalis is a species complex of globally ubiquitous protozoan parasites and the causative agent(s) of giardiasis, a diarrheal disease in humans, domestic animals, and wildlife (15). Giardiasis is a major global public health concern, responsible for an estimated 280 million infections per year and recognized by the WHO as a neglected disease since 2004 (16, 17). Infection is acquired via direct or indirect contact with infected individuals or by consuming water or food contaminated with cysts, which release motile trophozoites that attach to the surface of the gut epithelium, causing acute watery diarrhea. Trophozoites encyst in the gastrointestinal tract before being excreted into the environment in the host's feces, completing the parasite's life cycle. The cysts are resilient and can persist in the environment for up to 7 weeks under suitable conditions (18).

Eight genotypic groups (referred to as assemblages A through H) are recognized within *G. duodenalis*, and it has been proposed that these should be elevated to species rank based on molecular and ecological evidence (19, 20). Assemblages A and B have broad host ranges and are considered zoonotic infections in humans (21–23). Assemblage B is responsible for approximately 58% of the human infections worldwide each year (21). Each assemblage can be further divided into subtypes, which have variable host ranges and infectivity to humans. However, lack of morphological characters or diagnostic techniques capable of discriminating between subtypes limits our understanding of the epidemiology and public health significance of clinically relevant variants.

In this study, we have used 37 available whole-genome sequences of *Giardia duodenalis* assemblage B to develop a genome-driven approach to identify the best three markers that precisely delineate the genomic diversity of the species. Using validated PCR assays for each of these three markers, we subsequently doubled the number of typed *Giardia* genomes available. Phylogenetic analysis of these genomes revealed several stable, discrete evolutionary entities, or clonal complexes as defined for prokaryotes (24) and roughly equivalent to near-clades as defined in protozoan literature (25), as opposed to a genetic continuum. These results suggest a signal of ecological and/or geographic differentiation that likely reflects important epidemiologic differences among the clonal complexes and (differential) ongoing speciation.

RESULTS

Whole-genome relatedness of *Giardia* genomes using ML and ANI. To build a robust phylogeny of the available 37 genomes in NCBI's database, we employed maximum likelihood (ML) analysis of a concatenated superalignment of all 4,514 identified core genes shared by all genomes (approximately 8.55 Mbp), corresponding to 93.45% of the coding genes in each genome. Since these genomes are very closely related to one another and members of other assemblages, we expected a limited number of

Downloaded from https://journals.asm.org/journal/aem on 17 July 2021 by 2a02:587:d449:5600:9805:7ecd:78fc:2edf.

variable sites due to insufficient time to accumulate mutations, and we also expected that functional constraints and/or selection on the coding sequence of a gene may result in multiple independent substitutions at some sites. Thus, the ML distances calculated using the best-fitting model represent a powerful estimate of the evolutionary distances across the genomes that is robust with respect to very short evolutionary distances. ML is computationally expensive and not well suited to scaling up to very large data sets; therefore, we also investigated the utility of genome-aggregate average nucleotide identity (ANI) as an alternative method for estimating genetic relatedness. Briefly, ANI represents the average nucleotide identity of all genes shared between any two genomes. The distance matrix generated by ML analysis showed strong linear correlation to the corresponding distance matrix computed using FastANI (Pearson r^2 = 0.961 [Fig. 1a to d]), which is consistent with previous results from bacterial genomes for such short evolutionary scales (26). Therefore, we employed the ANI distances for our purposes of identifying which genes provide the same genetic relationships among the genomes as those revealed by whole-genome ML or ANI analysis, since these are easier to conceptualize than ML distances, unless otherwise mentioned.

All genomes compared were very closely related to one another, showing pairwise ANI values in the range of 98.63% to 99.96%. Note that the closest relative which has a genome available (*G. duodenalis* assemblage E) shares approximately 77% ANI with these assemblage B genomes (27); thus, this genome is divergent and would represent a novel species even based on the prokaryotic standards (28). The ML phylogeny revealed 16 distinct clades, which appear to be predominantly clonal (i.e., the maximum ANI distance between any two genomes in a clade was $<$0.1%). The ANI phylogeny recovered the same clades; however, their relationship to one another differed slightly between the two topologies, which is presumably due to unstable topologies in the ANI and/or ML tree and the short evolutionary distances among the genomes. PAM (partitioning around *k*-medoids) clustering confirmed that the number of groupings we observed manually, 16, was the optimal number of genome clusters (Fig. 1e) (29, 30). Collectively, these results reveal that the 16 clades represent biologically relevant groups found in nature. The high degree of clonality within each group indicated that intergroup genetic recombination is restricted or rare; thus, we here refer to these groups as clonal complexes, which can best be described as "stable phylogenetic clustering clouded by occasional recombination" to discuss similar patterns in prokaryotes and parasitic protozoan taxa (24, 25).

Defining clonal complex boundaries between genomes using MCE. To assess how discrete the 16 recovered clonal complexes were, we attempted to define distance thresholds that delineated them. This was accomplished by transforming the ANI distance matrix into a graph using maximal clique enumeration (MCE), a graph-based clustering technique. Briefly, each genome is represented by a vertex in the graph and there exists an edge (a linkage) between a pair of genomes if they share a pairwise distance less than a given threshold value. This method of clustering is agnostic of existing classifications and thus allowed us to explore genomic relatedness without bias, which was critical to our aim to advance current understanding of subtype (=clonal complex) diversity and the epidemiologic differences among the complexes.

MCE was initially applied with a starting threshold of 98.6% ANI, the maximum distance between any pair of genomes, and produced a single clique containing all 37 genomes. Sequentially, the threshold was increased by 0.1% identity units for considering a pair of genomes linked. At 99.9% identity, we observed cliques (clusters) identically matching the 16 clonal complexes observed on the phylogeny and PAM, recovering 1 clique containing 9 genomes, 1 clique containing 4 genomes, 4 cliques of 3 genomes each, 2 cliques which both contain a pair of genomes, and 8 groups represented by a single genome (Fig. 1f). In the context of evolutionary distance, this threshold corresponded to up to approximately 12,000 single-nucleotide polymorphism (SNP) differences between genomes classified in the same clique, confirming our prior observation that the genomes within a given complex are highly clonal. Using the

Downloaded from https://journals.asm.org/journal/aem on 17 July 2021 by 2a02:587:d449:5600:9805:7ecd:78fc:2edf.

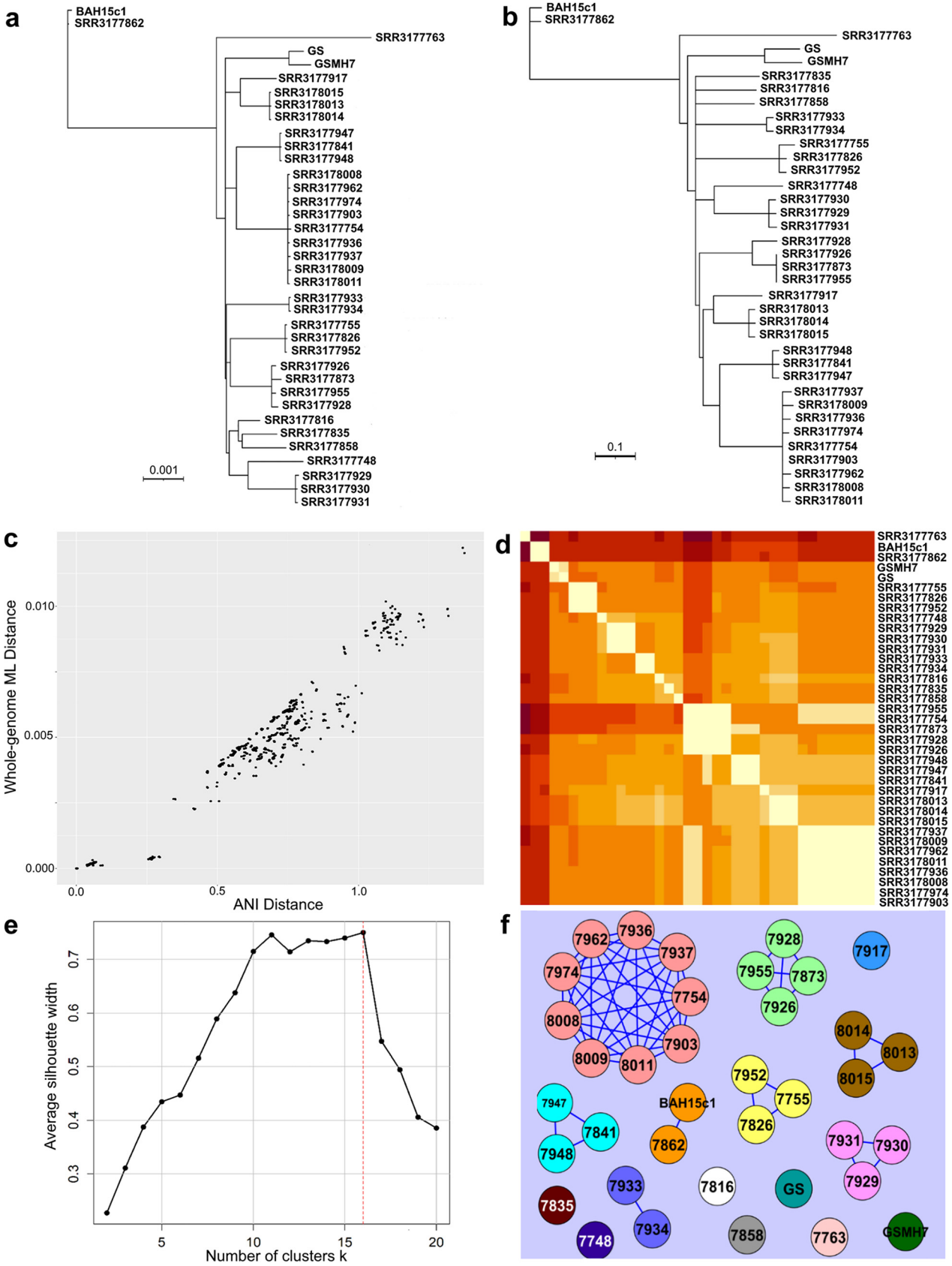

FIG 1 Comparisons of genomic relatedness between 37 *G. duodenalis* assemblage B genomes. (a) Maximum likelihood phylogeny (GTR+I+G model) of the core gene set shared among 37 whole-genome sequences utilized in this study, covering approximately 8.55 Mbp. (b) Whole-genome

(Continued on next page)

Downloaded from https://journals.asm.org/journal/aem on 17 July 2021 by 2a02:587:d449:5600:9805:7ecd:78fc:2edf.

99.9% identity threshold, we consistently observed only discrete cliques, which indicated that distinct boundaries are present between clonal complexes rather than a genetic continuum, despite overlapping hosts and/or sampling localities in some cases. We further explored this (see below) by typing and including twice as many genomes from different studies in the analysis.

Gene content analyses reveal further evidence in support of discrete clonal complexes versus a genetic continuum. Our results described above for the conserved core genome suggested that assemblage B is composed of distinct clusters of genomes at the nucleotide sequence level and occasionally occurring in the same hosts or locations. We additionally investigated the variable assemblage B pangenome to gain a better understanding of the gene content differences—if any—between clonal complexes. Only a small percentage of the reference genes (approximately 6.5%) were absent from any of the 37 available genomes. In order to obtain a broader insight into the pangenome gene pool and recover genes specific to or enriched in the clonal complexes, we trained a hidden Markov model (HMM) for *de novo* gene prediction with the Augustus software package using transcriptome sequencing (RNA-Seq) data from *G. duodenalis* assemblage A (a related cryptic species in the *G. duodenalis* complex), which shares approximately 77% ANI with *G. duodenalis* assemblage B (31), and protein data from NCBI reference genes for assemblage B. From the results of the Augustus model, we generated a pangenome consisting of 7,612 genes. A presence-absence matrix of these genes was subsequently computed, and the genes were further classified into core (present in all genomes), unique (only a single genome), and variable (at least two or more genomes but not all) categories (Fig. 2a).

We investigated gene content diversity among genomes and visualized this by constructing a neighbor-joining tree with 1,000 bootstraps of pairwise distances calculated using the presence-absence matrix and mapping associated clonal complex designation, sampling location, and host metadata to the leaves of the tree (Fig. 2b). The 34 genomes sampled from British Columbia form a single clade with very high bootstrap support, with all three reference genomes (GS and GSMH7 from Alaska and BAH15c1 from Australia) external to this clade. Within this clade, several subclades were recovered with strong bootstrap support, some of which are congruent with previously delineated clonal complexes based on sequence analysis.

We evaluated the strength of association using the Kendall tau rank correlation between clade assignment (i.e., p-distance) based on this gene content tree with (i) genomic sequence relatedness, measured by pairwise ANI, and (ii) geographic segregation, measured by computing Euclidean distance (in kilometers) between sampling sites provided for each of the genomes. Based on the topology of the tree, we analyzed (i) all pairs of 37 genomes, (ii) 36 genomes originating from North America (excluding BAH15c1 from Australia), and (iii) 34 genomes arising from the British Columbian polytomy, additionally excluding GS and GSMH7. In all comparisons, we observed moderate correlation, with tau values ranging between 0.25 and 0.49 (P values < 0.0001 for all comparisons) and the tau coefficients decreasing as the size of the data set (geographic breadth) compared decreased. Consistently, we found the correlation to be strongest between ANI and p-distance (tau = 0.47 to 0.49) and, likewise, weakest between geographic distance and p-distance (tau = 0.25 to 0.41; illustrated in File S1 in the supplemental material). This trend illustrated that diversity in gene con-

FIG 1 Legend (Continued)
ANI tree constructed using the neighbor-joining method based on the ANI distance matrix. In both trees, the 16 distinct lineages are readily discernible which correspond to the clonal complexes. No outgroup was included in this panel or in panel a since our focus was on differences at the intra-assemblage (subtype) level and the closest relative which has a genome available shares approximately 77% ANI with assemblage B (*G. duodenalis* assemblage E [27]), which is too divergent. Branch support for all branches in the ML phylogeny was 100% from 100 bootstrap replicates (values not shown for simplicity). (c) Scatterplot of pairwise maximum likelihood distances plotted against ANI distances. Each point represents a pair of genomes. (d) Heat map of ANI distances between pairs of genomes. The 16 clonal complexes can be distinguished along the diagonal of the heat map. (e) Comparison of results from PAM (partitioning around *k*-medoids) clustering to confirm the optimal division of genomes into clusters. The average silhouette width (*y* axis) for each iteration *k* of PAM (*x* axis) is plotted, with the maximum value ($k = 16$) representing the optimal clustering of genomes. (f) Graph network showing cliques (strongly connected, complete subgraphs) of genomes. Each genome is represented by a colored circle. Edges connecting two genomes indicate a pairwise ANI of $\geq$99.9%.

Downloaded from https://journals.asm.org/journal/aem on 17 July 2021 by 2a02:587:d449:5600:9805:7ecd:78fc:2edf.

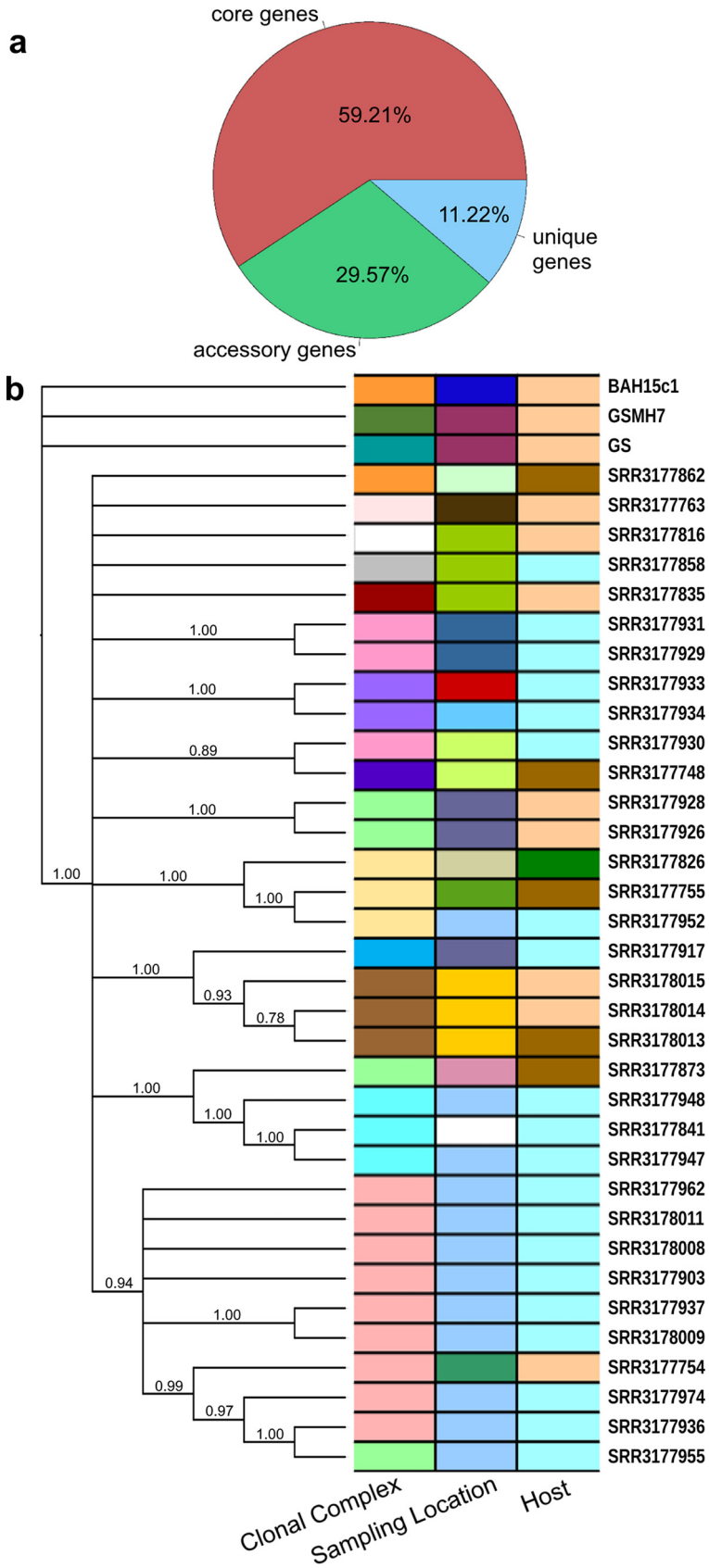

FIG 2 Evaluation of the pangenome of assemblage B. (a) Pie chart of genes classified into one of three categories: (i) core genes present in all genomes, (ii) accessory genes, missing from at least one

(Continued on next page)

Downloaded from https://journals.asm.org/journal/aem on 17 July 2021 by 2a02:587:d449:5600:9805:7ecd:78fc:2edf.

tent is more closely tied to genomic relatedness than to geography, which presumably indicates that genomes within each clonal complex share unique gene pools inherited through vertical descent as opposed—for instance—to horizontal gene transfer. Some phylogenetic signal is nonetheless exhibited in these data by geography, as evidenced by good correlation between ANI and geographic distance (tau = 0.35 to 0.44), suggesting that the genetic discontinuities between the clonal complexes probably result from a combination of geographic and genetic isolation.

We additionally examined whether genes unique to a clonal complex exhibit any functional trends by assigning them to KOG (euKaryotic Orthologous Groups, a eukaryote-specific version of Clusters of Orthologous Genes [COGs]) functional categories using protein homology searches against the eggNOG database (32). We identified a total of 328 genes which were unique to one of the 16 clonal complexes. Approximately 40% of these are functionally associated with the cell membrane, signaling, or the cytoskeleton (KOG classes S, T, and Z), with an additional 31% having no orthologous match in the eggNOG database and thus classified as hypothetical proteins (Fig. S1). We noted that in 11 of the 16 clonal complexes, at least one unique gene was annotated as a variable surface protein (VSP). *Giardia* VSPs have been shown to be diverse and comprise and estimated 4% of the *G. duodenalis* assemblage A genome (33). Further, it has been postulated that the diversity of VSPs in *Giardia* may be a factor in pathogenicity differences and ecological niches amenable to different subtypes of the parasite (33, 34).

Evaluating phylogenetic concordance of individual core genes relative to WGS. A total of 4,514 core genes identified in all 37 genomes were assigned a rank based on the similarity of the gene-based phylogenetic signal relative to that of the whole-genome-based patterns. We evaluated each individual gene in two ways: (i) by comparing gene-based distance matrices with the WGS-based ANI distance matrix using the Kendall tau rank correlation and (ii) tree-based comparisons using a Shimodaira-Hasegawa test (SH test), including branch lengths, in PAUP* (35). The data show that nearly all genes compared individually against WGS-based patterns exhibited a strong phylogenetic signal, with only 5 genes out of 4,387 showing nonsignificant ($P > 0.05$) tau rank correlations, which was also consistent with the lack of rampant recombination within assemblage B. Tau correlations could not be calculated for 12 genes due to undefined values in the gene-based distance matrix, which were caused by the absence of informative sites in the alignment. The distribution of tau correlations in assemblage B is approximately symmetrical, with a mean of 0.3534 and a standard deviation of 0.1113 (Fig. 3b). Weak but significant correlation ($r^2 = 0.117$; $P < 0.001$) existed between the length of the gene and the tau coefficient. There was also good correlation between Kendall tau coefficient with high consistency and retention indices (data not shown). However, a large percentage (86.78% [3,807 genes]) of gene ML trees significantly differed from the whole-genome ANI tree (i.e., $P < 0.05$), consistent with the fact that the SH test is sensitive to even small topological tree differences. Data for the top-ranking 30 genes with respect to their Kendall tau correlation plus selected genes which are currently in common use for molecular typing of *G. duodenalis* assemblages are shown in Table S2 (data for all genes are available from the authors on request).

We additionally examined whether genes with high Kendall tau correlations to ANI show any functional biases. For genes other than hypothetical (51% of the total), those associated with the cell membrane, cytoskeleton, signaling, and metabolism most of-

FIG 2 Legend (Continued)
genome, and (iii) unique genes, found in only one genome. (b) Neighbor-joining tree of pairwise distances calculated from the presence-absence matrix in panel a. The tree shows relationships between genomes based on shared gene content. Colored bars to the right of the tree highlight shared metadata (assignment to 1 of the 16 clonal complexes, sampling location, and host) between genomes.

Downloaded from https://journals.asm.org/journal/aem on 17 July 2021 by 2a02:587:d449:5600:9805:7ecd:78fc:2edf.

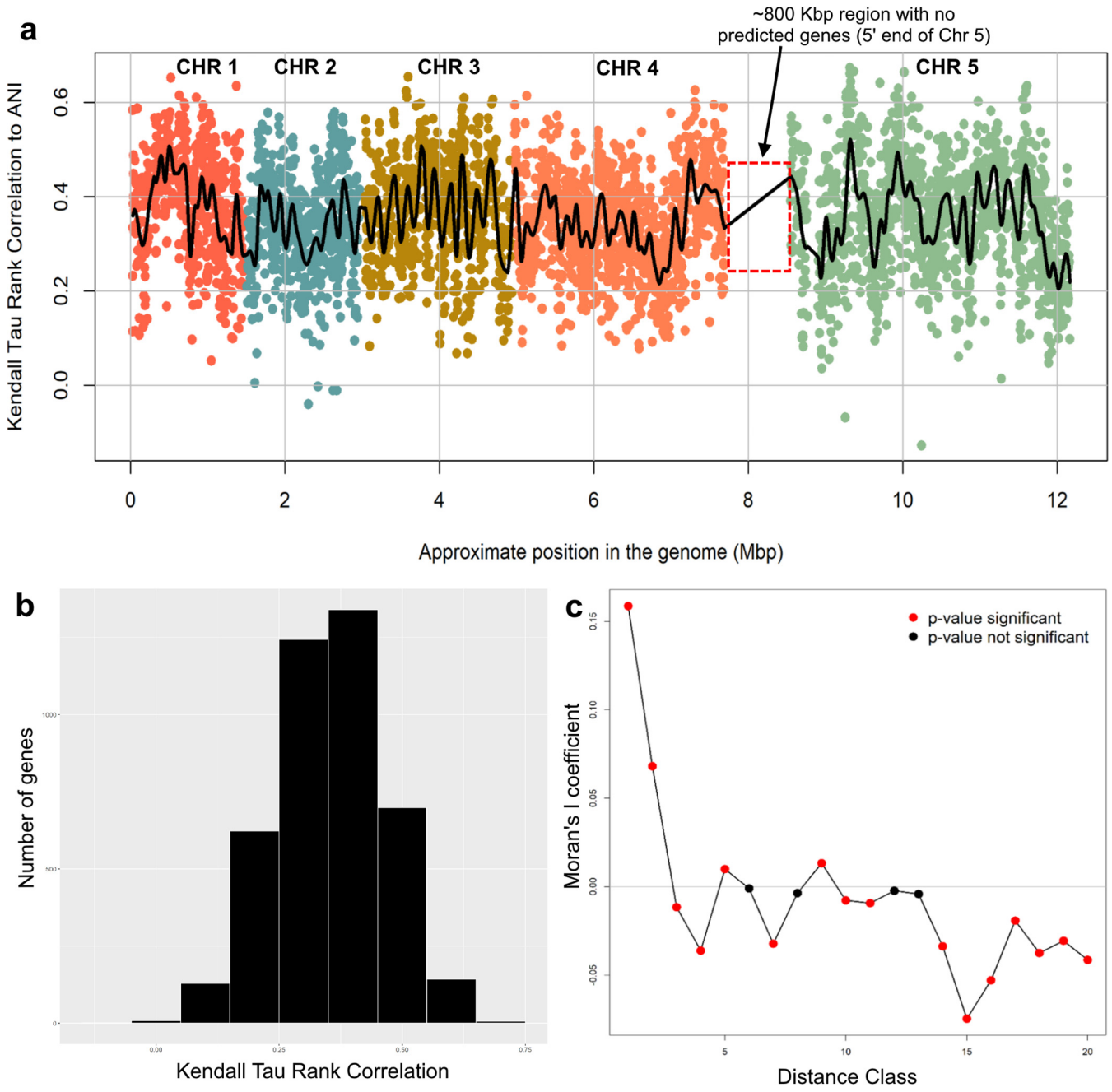

FIG 3 Spatial bias across the genome of individual genes that are robust phylogenetic markers. (a) Scatterplot matrix of Kendall tau values of the distance matrix for each gene against the corresponding ANI distances (represented by a point in the plot) plotted according to gene's approximate location in the most complete reference genome available (*G. duodenalis* assemblage A). Points (genes) are color-coded according to the chromosome (CHR) on which they are located. The black spline follows the local average Kendall tau value. (b) Distribution of Kendall tau values for all genes evaluated. (c) Correlogram of spatial autocorrelation between approximately equally sized distance classes of consecutive genes (approximately 50 genes each) based on their ordered location in the genome. Red points indicate a significant *P* value for a distance class.

ten show high Kendall tau correlations with species-level (i.e., WGS) evolutionary patterns. We investigated for biases in the spatial distribution of these genes across the genome using (i) a runs test for randomness, (ii) a Lomb-Scargle periodogram with 1,000 randomized replicates to test for periodicity in the data, and (iii) a spatial correlogram using Moran's I coefficient and equally sized distance classes of 50 genes. Plotting the tau rank correlations relative to their position in the genome revealed an interesting wave-like pattern throughout the genome, indicating that genes with high Kendall tau correlations to WGS data tend to be located near one another in clusters

Downloaded from https://journals.asm.org/journal/aem on 17 July 2021 by 2a02:587:d449:5600:9805:7ecd:78fc:2edf.

TABLE 1 Comparison of genes which have been used in subtyping MLST studies for *Giardia duodenalis* assemblage B and three of the best-performing genes identified in our study

Locus	Gene	Reference	Annotation	KOG	Kendall τ coefficient	SH test *P* value	CI	RI	No. of sites	Total no. of sites
KWX12433	*6-pgd*	This study	6-Phosphogluconate dehydrogenase		0.636	0.003	0.978	0.993	31	2,100
KWX13158		This study	Hypothetical protein		0.612	0.048	0.864	0.954	27	1,980
KWX15086	*phkg2*	This study	Phosphorylase B kinase gamma catalytic chain		0.610	<0.001	0.876	0.95	171	4,326
KWX12106	*tpi*	Sulaiman et al., 2003 (66)	Triose-phosphate isomerase	G	0.441	0.057	1.000	1.000	8	774
KWX15047	*gdh*	Read et al., 2004 (67)	Glutamate dehydrogenase	E	0.322	0.017	0.75	0.945	8	1,350
KWX12629	*bg*	Cacciò et al., 2008 (1)	Beta-giardin	Z	0.188	0.004	0.848	0.835	10	850

instead of being randomly distributed (Fig. 3a). The runs test for randomness around the mean Kendall tau value detected 1,335 runs and returned a *P* value of <0.0001, strongly corroborating the visually observed nonrandom dispersion pattern. Our Lomb-Scargle analysis of periodicity detected multiple significant periods within the data, with the most significant period occurring at 523 genes ($P < 0.0001$), approximately the same number of genes mapped to chromosome 1, indicating periodicity at the chromosomal level (see spline in Fig. 3a). Additional significant periods were detected at approximately 75, 200, 250, and 350 genes located on the same chromosome. Finally, we investigated the signal of autocorrelation between adjacent genes, which found significant autocorrelation coefficients as a series of peaks and valleys along increasing distance classes. The first distance class of 50 consecutive genes was the most significant; thus, the Kendall tau values (i.e., the genes) are not independent of one another within this distance class. The signal of significant autocorrelation (nonindependence) persisted through the first five distance classes, approximately 250 consecutive genes, which is concordant with the periodicity results (Fig. 3c).

A novel MLST scheme for genomotyping using the three best-performing genes. Current *Giardia* multilocus sequence typing (MLST) studies most commonly utilize approximately 500-bp-long amplicons from three genes, namely, the triose-phosphate isomerase (*tpi*), glutamate dehydrogenase (*gdh*), and beta-giardin (*bg*) genes. To test the robustness of these three genes against the whole-genome phylogeny reported here, the sequence of each of these genes in the 37 genome sequences analyzed were aligned and concatenated, providing a composite multilocus genotype (MLG) distance matrix. Phylogenetic analysis conducted using the concatenated alignment of these genes failed to robustly recover consistent subgroupings of assemblage B isolates, which was somewhat expected since these loci were originally chosen based on being informative for differentiating between assemblages and the availability of only a few sequences from which the original PCR assays were designed (36, 37). We next compared the results to those of the three best-performing genes based on our evaluation described above. Our selection of the gene markers was primarily informed by the Kendall tau correlations (above), with the presence of suitable sites for designing PCR primers and tree-based criteria (consistency and retention indices [CI and RI, respectively]) used as additional criteria. The three best performing from the overall top 30 genes in terms of Kendal tau coefficients were genes annotated as (i) 6-phosphogluconate dehydrogenase (NCBI:protein accession number KWX12433), (ii) a hypothetical protein (NCBI:protein accession number KWX13158), and (iii) phosphorylase B gamma catalytic chain kinase (NCBI:protein accession number KWX15086) (Table 1). Nested PCR primers were designed for amplicons of 1,028, 1,356, and 1,017 bp, respectively, for these genes. Neighbor-joining analysis using concatenated alignments of *in silico* PCR amplicons, a distance matrix of pairwise percent identities, and 1,000 bootstrap replicates resulted in a phylogeny which captures 13 well-supported groups, 12 of which are identifiable as clonal complexes on the whole-genome ML and ANI trees.

Downloaded from https://journals.asm.org/journal/aem on 17 July 2021 by 2a02:587:d449:5600:9805:7ecd:78fc:2edf.

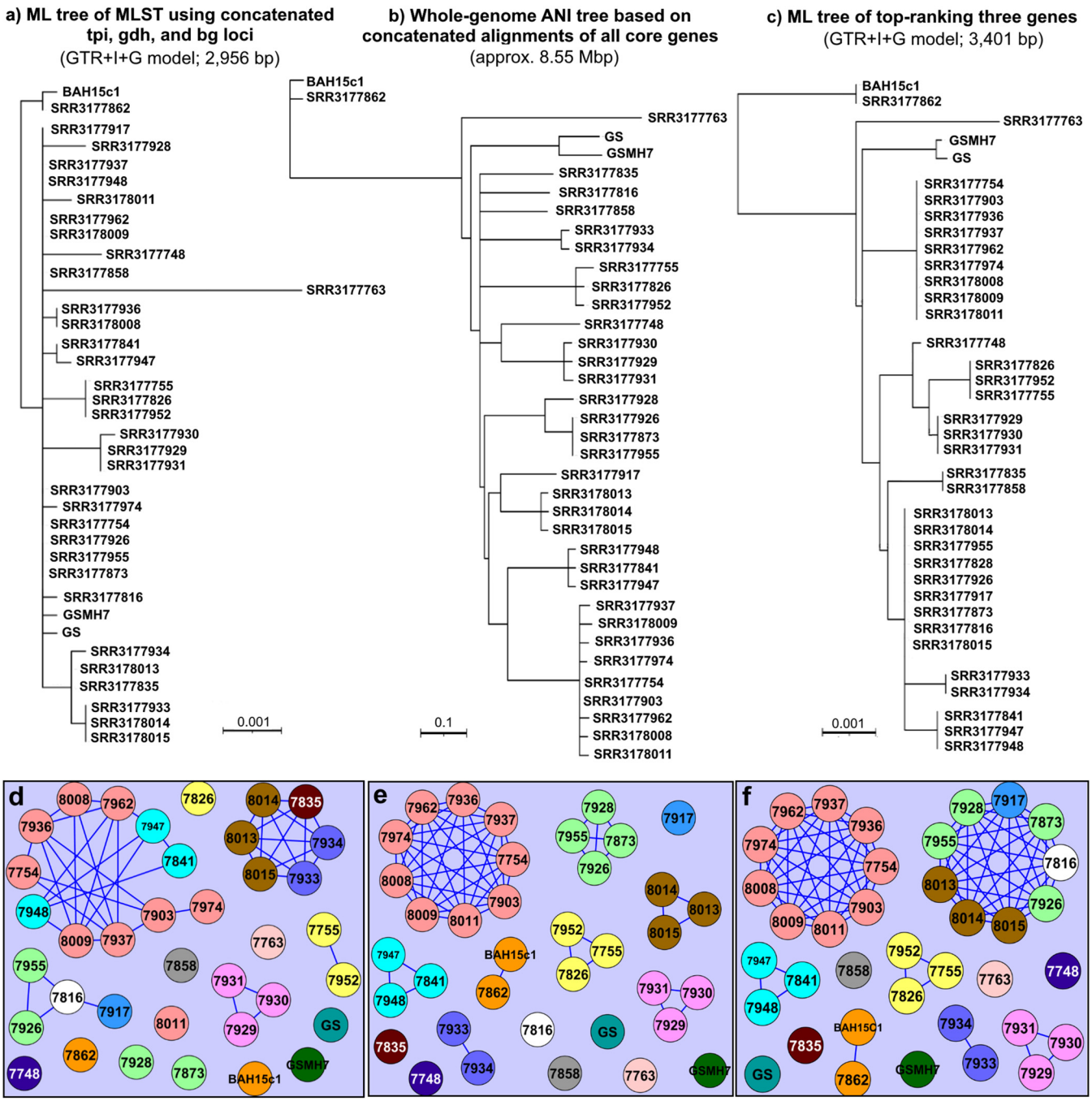

FIG 4 Comparison of traditional three-gene MLST scheme with the newly identified best-performing genes identified by this study. (a) The whole-genome ANI tree constructed from the concatenated superalignment of all 4,514 core genes (see Fig. 1a for the equivalent ML phylogeny). (b) ML phylogeny of the currently used MLST scheme for *Giardia*, constructed using full-length sequences of the *tpi*, *gdh*, and *bg* genes. (c) ML tree of novel MLST scheme described in our study, estimated from concatenated partial sequences of *6-pdg*, an unnamed gene annotated as a hypothetical protein, and *phkg2* (NCBI: protein accession numbers KWX12433, KWX13158, and KWX15086, respectively). (d to f) Graph networks constructed using pairwise nucleotide identities between sequences in the same alignments used to generate panels a, b, and c, respectively. Edges connecting two vertices indicate a shared nucleotide identity of 99.9% or greater. Vertices (genomes) are colored according to the whole-genome-based clonal complex assigned in our analysis.

The 13th group was comprised of all genomes from the remaining four (whole-genome) clonal complexes collapsed into one group (Fig. 4).

We next compared this phylogeny with the one constructed from the concatenated alignments of *tpi*, *gdh*, and *bg* sequences extracted from the genomes and to the whole-genome phylogeny. The results showed that the phylogeny built from our three selected markers is more congruent with the whole-genome phylogeny in terms of

Downloaded from https://journals.asm.org/journal/aem on 17 July 2021 by 2a02:587:d449:5600:9805:7ecd:78fc:2edf.

bootstrap support (only one node had <50% support), while the traditional MLST gene phylogeny showed little congruency with the whole-genome phylogeny (Fig. 4b). Comparison of MCE outputs from the same three data sets revealed that the traditional MLST only recovered five of the same clusters as the whole-genome data. Not counting singletons, only one clonal complex was recovered that is represented by a unique *tpi-gdh-bg* multilocus genotype (Fig. 4e, in pink). We observed no meaningful change in MLST resolution by increasing the minimum identity threshold or by computing maximal cliques from complete gene sequences (as opposed to *in silico* PCR amplicons). Further, the inclusion of additional high-ranking genes also did not yield any difference in clique computation until a total of 10 genes were included in the data set, and the improvement in resolution from adding a few more genes into the analysis was negligible. A concatenated alignment of all 30 top-ranking genes (the set from which we chose our markers) still was unable to perfectly match the resolution of the whole-genome data set (Fig. S2); therefore, based on this, we chose to limit our novel MLST to three loci, the same number as the existing MLST method, which has practical advantages.

Expanding the known genomic diversity based on the three best-performing genes. In order to expand the diversity of typed genomes and test the robustness of the 16 clonal complexes recovered among the 37 genomes, we employed the MLST genomotyping method based on the three best-performing targets mentioned above to type additional genomes available in CDC's collection. Using DNA from 68 assemblage B specimens gathered from CDC holdings and representing 6 continents, for a final total of 105 genomes represented, we sequenced amplicons from each of the three novel loci and constructed a neighbor-joining phylogeny as described above to identify a total of 55 distinct clusters (Fig. S3). We observed that specimens do segregate based on their geographic origins, with the exception of (i) Australian samples typed in this study ($n = 3$), which show a close relationship to European samples, and (ii) Algerian specimens, which split into two groups, one related to European samples and the other related to South American samples. Notably, we observed that the cluster assignment of the 37 whole genomes remained unchanged, indicating that the discrete threshold boundaries that we identified in the previous sections are robust to the inclusion of additional genomes, some of which were assigned to the existing clusters.

DISCUSSION

Giardia duodenalis assemblage B is a diplomonad parasite which causes giardiasis, a common gastrointestinal illness, in humans and animals worldwide. The high diversity and euryxenous range of this organism are already widely known, and these factors have confounded previous efforts to use molecular subtyping to aid in public health investigations and response. In order to advance our ability to link related isolates and the potential to further differentiate clinically relevant variants with different pathogenicities, we conducted phylogenomic and comparative gene content analyses to describe the relationships between 37 previously determined genomes as well as 68 newly typed genomes based on the best three gene markers identified by our analyses. We demonstrate here that assemblage B is composed of several genetically distinct, and stable to the addition of new genomes, lineages or clonal complexes. Further, several of these lineages are highly populated by genomes endemic to geographically narrow locations and are characterized by unique (nonhypothetical) gene content tuned to sensing the environment, signaling, and membrane proteins. These findings are also consistent with previously published hypotheses that differential gene content between *Giardia* subtypes may reflect important epidemiologic differences between them and/or adaptation to local environmental conditions (33, 34). Based on their distinct sequence and gene content diversity, it thus appears that these lineages may not represent functionally and ecologically neutral diversity and they are on their way to speciation. This hypothesis is generally consistent with the predominantly clonal evolution (PCE) model proposed by Tibayrenc and coauthors to describe patterns of relatedness among genomes of *Trypanosoma cruzi*, another parasitic

Downloaded from https://journals.asm.org/journal/aem on 17 July 2021 by 2a02:587:d449:5600:9805:7ecd:78fc:2edf.

protozoan (38, 39). Briefly, this model describes clonal evolution as limited or absent genetic recombination among populations of these parasitic protozoa, owing to the fact that it is uncommon for different *T. cruzi* strains to come into contact with one another during the reproductive stage of their life cycles, thus making genetic exchange between them rare (i.e., most sexual reproduction occurs among individuals of the same population or strain; any recombination is too infrequent to break the pattern of clonality).

A diverse variety of terms have been used to describe phylogenetic groupings such as those observed in our study, such as "stable, widespread [lineages], occurring in sympatry, including in the same host" or "stable phylogenetic clustering clouded by occasional recombination" (summarized in Table 1 in references 30 and 40). Further, these terms are often limited in scope/usage to specific organisms or taxa. We propose to use the term clonal complex to avoid confusion with other typing methods and the MLG subtypes reported in previous studies. In any case, however, our data collectively suggest that these clonal complexes represent important genomic, and thus phenotypic, diversity within assemblage B. This diversity most likely is important for fine-scale microdiversity and epidemiologic studies, and we have outlined a genome-based approach and associated three gene markers to robustly catalogue this diversity. It is also important to note that the genetic diversity observed within assemblage B based on our comparisons is extensive enough to qualify assemblage B as a distinct species of *Giardia*, even compared to the genomic standards used for species demarcation in prokaryotes (41, 42). This will also be consistent with earlier proposals to formally elevate the cryptic species in *G. duodenalis* to species rank (19, 20). Accordingly, in addition to a revision of the species-level taxonomy of *Giardia*, a suitable nomenclature should be implemented for the intraspecies clonal complexes to facilitate future ecological studies or public health investigations.

Using these genomic relationships as the reference data set, we further identified genes which best reflect clonal complex-level patterns of relatedness and establish their utility to inform future public health response and prevention strategies. We chose to limit our novel genomotyping scheme to three genes, the same number as the current typing method in use for *Giardia* infections in humans, due to the practical considerations of time/cost efficiency, and have further streamlined laboratory workflows by designing PCR assays that can be run in parallel using the same PCR master mix, substituting the appropriate primers, and same cycling conditions to amplify each target. We found that further meaningful improvement in typing resolution requires at a minimum 30 gene targets (or more when accounting for tau correlation to species-level patterns), which is inefficient to sequence using traditional PCR methods, essentially necessitating whole-genome sequencing.

We used the ANI distance matrix as the reference data set to evaluate individual genes for within-assemblage B resolution as performed previously for prokaryotic species (25). ANI has been used extensively in prokaryotic genome comparisons, as it offers robust resolution between strains of the same or closely related species (i.e., showing 80 to 100% ANI) and it is easy to compute (41). The ANI measure does not strictly represent core genome evolutionary relatedness, as orthologous genes can vary widely between pairs of genomes compared. Nevertheless, it closely reflects the traditional microbiological concept of DNA-DNA hybridization relatedness for defining prokaryotic species (42), as it takes into account the fluid nature of the bacterial gene pool and hence implicitly considers shared function. For these reasons, ANI has been recently used in place of the "gold standard" of prokaryotic taxonomy, the DNA-DNA hybridization for species-level resolution (42, 43). Recently, a new, accelerated kmer-based algorithm for computing ANI, FastANI, has been also described that provided essentially the same ANI values for the same genomes as the original, BLASTN-based implementation (44). Hence, ANI-based approaches for typing, coupled to maximal clique enumeration (MCE) to explore the landscape of genetic diversity as performed in this study for *Giardia* and previously for bacterial species and *T. cruzi* (45, 46), can scale

Downloaded from https://journals.asm.org/journal/aem on 17 July 2021 by 2a02:587:d449:5600:9805:7ecd:78fc:2edf.

up well with an increasing number of genomes, unlike core genome ML phylogenetic analysis (at least not without approximations for the ML algorithm that generally decrease accuracy of the resulting tree [47]). The minor discrepancies observed between core genome ML and ANI trees might be due to the fact that the relationship between ML and ANI is nonlinear, presumably due to multiple substitutions at the same sites, which ANI calculations are not sensitive in detecting (25). In addition, ambiguous base calls introduced during the pileup step of read mapping are assumed to have an effect on distance calculations similar to that of homoplasies. Nonetheless, these differences were rather minor overall and thus did not have any significant effects on the conclusions drawn here. Accordingly, our methodology can be implemented in any eukaryotic group with intragroup genetic diversity similar to (or lower than) that of assemblage B, as it has been successfully used for prokaryotic taxa previously (25), and our approach to design PCR primers and validate PCR assays can be used in any standard molecular lab equipped to conduct PCR and Sanger sequencing. Hence, our results should have implications for typing additional protozoan groups of interest that are difficult to genotype, like *Giardia*. The challenge of developing an affordable and easily scalable method for culture-free whole-genome sequencing of organisms like *Giardia* remains an open problem.

MATERIALS AND METHODS

Empirical data sets. Whole-genome sequencing reads from 34 isolates of *Giardia duodenalis* assemblage B were downloaded from the SRA database at NCBI. These genomes were originally derived from several hosts (human, beavers, and one dog) plus environmental samples and were collected between 1989 and 1995 in British Columbia, Canada, as part of either contemporary waterborne outbreak investigations or monitoring (48, 49). Additionally, contigs from three reference assemblies (isolates BAH15c1 from Australia and GS and GSMH7 from Alaska) and respective annotations were downloaded from the Assembly database at NCBI, for a total of 37 genomes in our final data set. This data set, to our knowledge, comprehensively represents the publicly available set of assemblage B genomes at the time of this study. Accompanying metadata for all genomes included in our comparisons are described in Table S1.

Conserved core genome identification. Data from the reference strain BAH15c1 were selected as representative nucleotide coding sequences and annotations. The following strategy was used for determining the conserved core genome of the group. Sequencing reads from each genome were mapped using BBMap (part of the BBTools package) (50) against the reference set of gene sequences to generate a core genome reference-based assembly for each isolate. Poor-quality bases less than Q20 (a Phred score of 20), Illumina adapter sequences, and reads shorter than 50 bp were removed using BBDuk (another component of BBTools) for quality control prior to mapping. Each reference gene from BAH15c1 was then searched against each of the remaining core genome assemblies using BLASTN (51). The best match was extracted from the subject assembly when it showed greater than 70% identity and covered at least 70% of the query length. Genes which were reciprocal best matches in all isolates were retained as the conserved gene core. Figure S4 details the workflow used in this section as well as the following section.

Whole-genome and individual gene phylogenetic and distance matrix comparisons. For all individual genes included in the conserved core of each group, an alignment was constructed and refined using MAFFT v7 and TrimAl with default settings (–auto and –automated1 parameters, respectively) (52, 53). The best-fitting evolutionary model for the alignment was estimated using jModelTest2 and the Akaike information criteria (AIC), followed by maximum likelihood (ML) analysis using the chosen model conducted in PAUP* 4.0a166 to compute the ML distance matrix (54, 55). Phylogenetic trees were constructed with PhyML v20130103 using the same evolutionary model plus the best starting tree from jModelTest2 (56). Alignments for all genes in the conserved gene cores for each group were concatenated to form a whole-genome alignment. The best-fitting evolutionary model, ML distance matrices, and ML phylogenies were computed for each whole-genome alignment as described above for individual genes. For ANI, the software FastANI was used to compute a similarity matrix from the draft genomes, which was converted to a distance/dissimilarity matrix (44). Clustering of genomes was accomplished by the PAM (partitioning around *k*-medoids) algorithm and the whole-genome ML distance matrix, followed by plotting the average silhouette widths against the number of clusters, where the maximum average silhouette width represented the optimal number of clusters (28, 29). The PAM algorithm and silhouette profiles were computed using the R library "cluster." ML distance matrices for individual genes were compared against the whole-genome ML and ANI distance matrices using the Kendall tau rank correlation to identify the genes with the highest correlation to the whole-genome distances. Custom Perl scripts automated the above stages of processing. Rank correlation calculations were conducted using the cor.test() function in R (57).

Identification of clonal complex boundaries using MCE. Delineation of clonal complex (=subtype) boundaries was accomplished through the use of a maximal clique enumeration (MCE) approach, using the Bron-Kerbosch algorithm to identify maximal cliques. For this, the ANI distance matrix was

Downloaded from https://journals.asm.org/journal/aem on 17 July 2021 by 2a02:587:d449:5600:9805:7ecd:78fc:2edf.

TABLE 2 Primer sequences for novel *Giardia duodenalis* assemblage B (genomotyping) MLST

Gene target	Primer name	Usage	Primer sequence (5′ → 3′)	Length	Estimated T_a (°C)[a]	Reference
KWX12433 (*6-pgd*)	KWX12433_118f	F1	GGR ATT RTT GCG CAR TCR CTT CC	23	58	This study
	KWX12433_157f	F2	GAC TAT AGY TCR CCA ATA GGC	21	56	This study
	KWX12433_1184r	R2	TTR TAT CTT GCA GKC AGC TGR CA	23	58	This study
	KWX12433_1641r	R1	CAG AGA TGT TCG YYT ACG AAA C	22	58	This study
KWX13158 (hypothetical)	KWX13158_339f	F1	GGT TAC YTT TCT AGG TGA YAT ATA	24	58	This study
	KWX13158_1820r	R1	CTR CAR AAC GGW AGR CTC ARG TC	23	56	This study
	KWX13158_1784r	R2	CCC GTG AAT ACR CAY AAG CTA T	22	58	This study
	KWX13158_426f	F2	CAG RGT GCC AAA TCT TTA CRC	21	56	This study
KWX15086 (*phkg2*)	KWX15086_1393f	F1	CTT GAC CTY AAT GCM TTY CTY ATG A	25	58	This study
	KWX15086_1548f	F2	AAT CTG TCC YCT YGA GAT TGC T	22	58	This study
	KWX15086_2586r	R1	GCT YTT GTT CTG YCC AAG GCT	21	58	This study
	KWX15086_2564r	R2	TGA AGA GCC TCC GAG AAR TC	20	58	This study

[a]T_a, annealing temperature.

transformed into a graph, where each genome is represented as a vertex and there exists a bidirectional edge connecting two vertices (genomes) if the similarity between them exceeds a minimum threshold. A maximal clique is defined as the largest possible subset of vertices in which each vertex displays complete connectivity to all other vertices in the subset (i.e., there exists an edge between a given vertex and all other vertices in the subset) and excludes all others in the complete set of vertices. To empirically determine minimum identity thresholds for subclade boundaries, maximal cliques were computed using a series of minimum thresholds, beginning with the minimum ANI (98.6% identity) between any pair of genomes in the matrix and incrementing by 0.1 ANI unit up to 100.0% identity, and visualized using Cytoscape v3.7.2. We assumed that monophyletic clades recovered by whole-genome phylogenies constituted biologically relevant (e.g., naturally occurring) groups; thus, the minimum value at which all distinct monophyletic clades observed on the phylogeny were subsequently recovered as maximal cliques was chosen as the minimum threshold for further analysis.

Gene content analyses. The *ab initio* gene finding software Augustus was used to predict genes in all 37 genomes. For this, an HMM model was trained following the procedure outlined by Hoff and Stanke using the NCBI reference genes and recently published RNA-Seq data sequenced from a related isolate of *G. duodenalis* assemblage A as sources of extrinsic evidence for estimating model parameters (58, 59). *De novo* gene predictions obtained using the new model from all 37 genomes were dereplicated using usearch with clustering parameters of 90% identity across 70% of the length of the protein sequence (60). This step resulted in 7,612 gene clusters represented by at least one sequence. When multiple sequences existed within a cluster, the sequence that best represented the cluster in terms of sequence identities (the cluster medoid) was exported and combined to form the predicted pangenome. A binary presence-absence matrix was generated by searching each gene's medoid against the genomes using BLASTN with a minimum threshold for a match of 70% identity across 70% of the query length parameters. A neighbor-joining tree with 1,000 bootstrap replicates was constructed from this binary matrix using the ape package in R. Functional annotation of genes was accomplished using both protein homology searches against the eggNOG database and existing annotations from NCBI (32). Annotations and gene identifiers derived from the NCBI reference files were given priority over eggNOG matches when a gene showed a reciprocal best match to a reference gene. Pairwise gene content comparisons across groups derived from genome metadata (i.e., host, sampling location, clonal complex assignment, etc.) were automated using custom Perl scripts (code available from the authors on request [kostas@ce.gatech.edu]). SRA accession numbers for RNA data used in this procedure are SRR8589747 to SRR8589752.

Statistical analyses. Statistical significance of the Kendall tau correlations of the individual gene ML distance matrices versus the whole-genome ML or ANI distance matrix was assessed using a delete-half jackknife strategy and 1,000 replicates for each group. In brief, for each replicate, 50% of the data from the gene-based distance matrix were randomly resampled without replacement and compared using Kendall tau to the corresponding data in the whole-genome ML or ANI matrix. The tau correlation for all 1,000 replicates was averaged, and this average tau value was used to compute a z-score and *P* value. To test for potential spatial biases across the genome of the best-performing individual genes, we sorted the genes by their order of appearance in the most complete reference genome available (*G. duodenalis* assemblage A WB, version 41 in GiardiaDB) (61) and filtered out genes which had bit scores of less than 75 and which did not map to the chromosomal scaffolds. A total of 4,191 genes remained after filters were applied. A runs test was used to estimate the randomness of the distribution of tau values along the mean gene position in the appropriate reference genome. The spatial autocorrelation and periodicity of genes were further tested using Moran's I correlogram and Lomb-Scargle periodogram. Lag classes for Moran's I correlogram were computed using an equal size of 50 consecutive genes (62). R packages spdep, lomb, and snpar were used to compute spatial correlograms, periodograms, and the runs test, respectively (63–65).

Downloaded from https://journals.asm.org/journal/aem on 17 July 2021 by 2a02:587:d449:5600:9805:7ecd:78fc:2edf.

Selection of gene targets for typing, primer development, and laboratory testing. In order to validate our MCE-based method of clonal complex assignment, we selected potential gene targets for laboratory testing from the top 30 core genes based on their tau rank correlation values. Potential genes were inspected manually for suitable sites for designing PCR primers with respect to an amplicon that produces an identical (or nearly identical) phylogeny compared to the phylogeny reconstructed from sequences of the entire length of the gene. Maximal cliques were computed from concatenated alignments of expected PCR amplicons to evaluate the level of resolution offered by a particular triplet of genes. The final selection of three gene targets was based on similarity of amplicon-based cliques when visually compared to cliques computed from WGS. Nested PCR primers capable of annealing to templates from both assemblages A and B at approximately the same temperatures were designed manually for three target genes. Primer sequences are provided in Table 2. Primary and secondary nested PCRs for all genes were carried out using the same cycling conditions: initial denaturation at 94°C for 5 min, followed by 35 cycles of 94°C for 45 s, annealing at 58°C for 45 s, extension at 72°C for 1 min 30 s, and a final extension at 72°C for 7 min. PCR master mixes were also identical for all three gene targets. Each primary reaction consisted of 22.35 μl of molecular-grade water, 2 μl of template DNA, 5 μl of 10$\times$ PCR buffer, 4 μl of 1.25 mM deoxynucleoside triphosphates (dNTPs), 1.25 μl each of 10 μM primers, 10 μl of 10-mg/ml bovine serum albumin (BSA), 4 μl of 25 mM $MgCl_2$ (final concentration of 2 mM), and, lastly, 0.15 μl of 5-U/μl *Taq* polymerase, for a final reaction volume of 50 μl. Secondary reactions are also carried out in 50-μl volumes with 29.85 μl of water, 2 μl of primary reaction template, 5 μl of PCR buffer, 2.5 μl of 10 μM primers, 4 μl of 25 mM $MgCl_2$ (final concentration of 2 mM again), and 0.15 μl of 5-U/μl *Taq* polymerase.

Data availability. Sequences generated as part of this study have been submitted to GenBank and are available as accession numbers MW289320 to MW289523.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 1.1 MB.

ACKNOWLEDGMENTS

We thank Anson Koehler, Una Ryan, and Alireza Zahedi for their kind contributions of DNA extracts for PCR testing.

This work was partly funded by the U.S. National Science Foundation, award number 1759831 (to K.T.K.).

The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

REFERENCES

1. Cacciò SM, Beck R, Lalle M, Marinculic A, Pozio E. 2008. Multilocus genotyping of *Giardia duodenalis* reveals striking differences between assemblages A and B. Int J Parasitol 38:1523–1531. https://doi.org/10.1016/j.ijpara.2008.04.008.
2. Sprong H, Cacciò SM, van der Giessen JW, ZOOPNET network and partners. 2009. Identification of zoonotic genotypes of *Giardia duodenalis*. PLoS Negl Trop Dis 3:e558. https://doi.org/10.1371/journal.pntd.0000558.
3. Roellig DM, Salzer JS, Carroll DS, Ritter JM, Drew C, Gallardo-Romero N, Keckler MS, Langham G, Hutson CL, Karem KL, Gillespie TR, Visvesvara GS, Metcalfe MG, Damon IK, Xiao L. 2015. Identification of *Giardia duodenalis* and *Enterocytozoon bieneusi* in an epizoological investigation of a laboratory colony of prairie dogs, *Cynomys ludovicianus*. Vet Parasitol 210:91–97. https://doi.org/10.1016/j.vetpar.2015.03.022.
4. Caccio SM, de Waele V, Widmer G. 2015. Geographical segregation of *Cryptosporidium parvum* multilocus genotypes in Europe. Infect Genet Evol 31:245–249. https://doi.org/10.1016/j.meegid.2015.02.008.
5. Chalmers RM, Caccio S. 2016. Towards a consensus on genotyping schemes for surveillance and outbreak investigations of *Cryptosporidium*, Berlin, June 2016. Euro Surveill 21:30338. https://doi.org/10.2807/1560-7917.ES.2016.21.37.30338.
6. Squire SA, Yang R, Robertson I, Ayi I, Ryan U. 2017. Molecular characterization of *Cryptosporidium* and *Giardia* in farmers and their ruminant livestock from the Coastal Savannah zone of Ghana. Infect Genet Evol 55:236–243. https://doi.org/10.1016/j.meegid.2017.09.025.
7. Geurden T, Levecke B, Caccio SM, Visser A, De Groote G, Casaert S, Vercruysse J, Claerebout E. 2009. Multilocus genotyping of *Cryptosporidium* and *Giardia* in non-outbreak related cases of diarrhoea in human patients in Belgium. Parasitology 136:1161–1168. https://doi.org/10.1017/S0031182009990436.
8. Wielinga C, Ryan U, Thompson RA, 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 multilocus genotyping with conserved and variable genes. Int J Parasitol 41:495–503. https://doi.org/10.1016/j.ijpara.2010.11.007.
9. Ankarklev J, Franzén O, Peirasmaki D, Jerlström-Hultqvist J, Lebbad M, Andersson J, Andersson B, Svärd SG. 2015. Comparative genomic analyses of freshly isolated *Giardia intestinalis* assemblage A isolates. BMC Genomics 16:697. https://doi.org/10.1186/s12864-015-1893-6.
10. Wielinga C, Thompson RA, Monis P, Ryan U. 2015. Identification of polymorphic genes for use in assemblage B genotyping assays through comparative genomics of multiple assemblage B *Giardia duodenalis* isolates. Mol Biochem Parasitol 201:1–4. https://doi.org/10.1016/j.molbiopara.2015.05.002.
11. Kaper JB, Nataro JP, Mobley HL. 2004. Pathogenic *Escherichia coli*. Nat Rev Microbiol 2:123–140. https://doi.org/10.1038/nrmicro818.
12. Van Ert MN, Easterday WR, Huynh LY, Okinaka RT, Hugh-Jones ME, Ravel J, Zanecki SR, Pearson T, Simonson TS, U'Ren JM, Kachur SM, Leadem-Dougherty RR, Rhoton SD, Zinser G, Farlow J, Coker PR, Smith KL, Wang B, Kenefic LJ, Fraser-Liggett CM, Wagner DM, Keim P. 2007. Global genetic population structure of *Bacillus anthracis*. PLoS One 2:e461. https://doi.org/10.1371/journal.pone.0000461.
13. Pena-Gonzalez A, Rodriguez-R LM, Marston CK, Gee JE, Gulvik CA, Kolton CB, Saile E, Frace M, Hoffmaster AR, Konstantinidis KT. 2018. Genomic characterization and copy number variation of *Bacillus anthracis* plasmids pXO1 and pXO2 in a historical collection of 412 strains. mSystems 3:e00065-18. https://doi.org/10.1128/mSystems.00065-18.
14. Hanevik K, Bakken R, Brattbakk HR, Saghaug CS, Langeland N. 2015. Whole genome sequencing of clinical isolates of *Giardia lamblia*. Clin Microbiol Infect 21:192.e1–192.e3. https://doi.org/10.1016/j.cmi.2014.08.014.

Downloaded from https://journals.asm.org/journal/aem on 17 July 2021 by 2a02:587:d449:5600:9805:7ecd:78fc:2edf.

15. Ryan U, Cacciò SM. 2013. Zoonotic potential of Giardia. Int J Parasitol 43:943–956. https://doi.org/10.1016/j.ijpara.2013.06.001.
16. Einarsson E, Ma'ayeh S, Svärd SG. 2016. An up-date on *Giardia* and giardiasis. Curr Opin Microbiol 34:47–52. https://doi.org/10.1016/j.mib.2016.07.019.
17. Savioli L, Smith H, Thompson A. 2006. *Giardia* and *Cryptosporidium* join the 'neglected diseases initiative.' Trends Parasitol 22:203–208. https://doi.org/10.1016/j.pt.2006.02.015.
18. Olson ME, Goh J, Phillips M, Guselle N, McAllister TA. 1999. *Giardia* cyst and *Cryptosporidium* oocyst survival in water, soil, and cattle feces. J Environ Qual 28:1991–1996. https://doi.org/10.2134/jeq1999.00472425002800060040x.
19. Monis PT, Caccio SM, Thompson RA. 2009. Variation in *Giardia*: towards a taxonomic revision of the genus. Trends Parasitol 25:93–100. https://doi.org/10.1016/j.pt.2008.11.006.
20. Thompson RCA, Monis PT. 2004. Variation in *Giardia*: implications for taxonomy and epidemiology. Adv Parasitol 58:69–137. https://doi.org/10.1016/S0065-308X(04)58002-8.
21. Feng Y, Xiao L. 2011. Zoonotic potential and molecular epidemiology of *Giardia* species and giardiasis. Clin Microbiol Rev 24:110–140. https://doi.org/10.1128/CMR.00033-10.
22. Thompson RCA, Ash A. 2016. Molecular epidemiology of *Giardia* and *Cryptosporidium* infections. Infect Genet Evol 40:315–323. https://doi.org/10.1016/j.meegid.2015.09.028.
23. Xiao L, Feng Y. 2017. Molecular epidemiologic tools for waterborne pathogens *Cryptosporidium* spp. and *Giardia duodenalis*. Food Waterborne Parasitol 8–9:14–32. https://doi.org/10.1016/j.fawpar.2017.09.002.
24. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A 95:3140–3145. https://doi.org/10.1073/pnas.95.6.3140.
25. Tibayrenc M, Ayala FJ. 2014. *Cryptosporidium*, *Giardia*, *Cryptococcus*, *Pneumocystis* genetic variability: cryptic biological species or clonal near-clades? PLoS Pathog 10:e1003908. https://doi.org/10.1371/journal.ppat.1003908.
26. Konstantinidis KT, Ramette A, Tiedje JM. 2006. Toward a more robust assessment of intraspecies diversity, using fewer genetic markers. Appl Environ Microbiol 72:7286–7293. https://doi.org/10.1128/AEM.01398-06.
27. Jerlström-Hultqvist J, Franzén O, Ankarklev J, Xu F, Nohýnková E, Andersson JO, Svärd SG, Andersson B. 2010. Genome analysis and comparative genomics of a *Giardia intestinalis* assemblage E isolate. BMC Genomics 11:1–15. https://doi.org/10.1186/1471-2164-11-543.
28. Konstantinidis KT, Rosselló-Móra R, Amann R. 2017. Uncultivated microbes in need of their own taxonomy. ISME J 11:2399–2406. https://doi.org/10.1038/ismej.2017.113.
29. Kaufman L, Rousseeuw PJ. 1987. Clustering by means of medoids, p 405–416. *In* Dodge Y (ed), Statistical data analysis based on the L1 norm and related methods. North Holland, Amsterdam, the Netherlands.
30. Kaufman L, Rousseeuw PJ. 1990. Finding groups in data: an introduction to cluster analysis, p 68–125. Wiley-Interscience, Hoboken, NJ.
31. Franzen O, Jerlström-Hultqvist J, Castro E, Sherwood E, Ankarklev J, Reiner DS, Palm D, Andersson JO, Andersson B, Svärd SG. 2009. Draft genome sequencing of Giardia intestinalis assemblage B isolate GS: is human giardiasis caused by two different species? PLoS Pathog 5:e1000560. https://doi.org/10.1371/journal.ppat.1000560.
32. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, von Mering C, Bork P. 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 44:D286–D293. https://doi.org/10.1093/nar/gkv1248.
33. Adam RD, Nigam A, Seshadri V, Martens CA, Farneth GA, Morrison HG, Nash TE, Porcella SF, Patel R. 2010. The *Giardia lamblia* vsp gene repertoire: characteristics, genomic organization, and evolution. BMC Genomics 11:424–414. https://doi.org/10.1186/1471-2164-11-424.
34. Ankarklev J, Jerlström-Hultqvist J, Ringqvist E, Troell K, Svärd SG. 2010. Behind the smile: cell biology and disease mechanisms of *Giardia* species. Nat Rev Microbiol 8:413–422. https://doi.org/10.1038/nrmicro2317.
35. Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol 16:1114. https://doi.org/10.1093/oxfordjournals.molbev.a026201.
36. Monis PT, Andrews RH, Mayrhofer G, Ey PL. 1999. Molecular systematics of the parasitic protozoan *Giardia intestinalis*. Mol Biol Evol 16:1135–1144. https://doi.org/10.1093/oxfordjournals.molbev.a026204.
37. Cacciò SM, De Giacomo M, Pozio E. 2002. Sequence analysis of the β-giardin gene and development of a polymerase chain reaction-restriction fragment length polymorphism assay to genotype *Giardia duodenalis* cysts from human faecal samples. Int J Parasitol 32:1023–1030. https://doi.org/10.1016/S0020-7519(02)00068-1.
38. Tibayrenc M, Kjellberg F, Ayala FJ. 1990. A clonal theory of parasitic protozoa: the population structures *of Entamoeba*, *Giardia*, *Leishmania*, *Naegleria*, *Plasmodium*, *Trichomonas*, and *Trypanosoma* and their medical and taxonomical consequences. Proc Natl Acad Sci U S A 87:2414–2418. https://doi.org/10.1073/pnas.87.7.2414.
39. Tibayrenc M, Ayala FJ. 2020. Genomics and high-resolution typing confirm predominant clonal evolution down to a microevolutionary scale in *Trypanosoma cruzi*. Pathogens 9:356. https://doi.org/10.3390/pathogens9050356.
40. Tibayrenc M, Ayala FJ. 2012. Reproductive clonality of pathogens: a perspective on pathogenic viruses, bacteria, fungi, and parasitic protozoa. Proc Natl Acad Sci U S A 109:e3305–e3313. https://doi.org/10.1073/pnas.1212452109.
41. Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A 102:2567–2572. https://doi.org/10.1073/pnas.0409727102.
42. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 57:81–91. https://doi.org/10.1099/ijs.0.64483-0.
43. Richter M, Rossello-Mora R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A 106:19126–19131. https://doi.org/10.1073/pnas.0906412106.
44. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9:5114. https://doi.org/10.1038/s41467-018-07641-9.
45. Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, Pati A. 2015. Microbial species delineation using whole genome sequences. Nucleic Acids Res 43:6761–6771. https://doi.org/10.1093/nar/gkv657.
46. Arnaud-Haond S, Moalic Y, Barnabe C, Ayala FJ, Tibayrenc M. 2014. Discriminating micropathogen lineages and their reticulate evolution through graph theory-based network analysis: the case of *Trypanosoma cruzi*, the agent of Chagas disease. PLoS One 9:e103213. https://doi.org/10.1371/journal.pone.0103213.
47. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490. https://doi.org/10.1371/journal.pone.0009490.
48. Prystajecky N, Tsui CKM, Hsiao WW, Uyaguari-Diaz MI, Ho J, Tang P, Isaac-Renton J. 2015. *Giardia* spp. are commonly found in mixed assemblages in surface water, as revealed by molecular and whole-genome characterization. Appl Environ Microbiol 81:4827–4834. https://doi.org/10.1128/AEM.00524-15.
49. Tsui CKM, Miller R, Uyaguari-Diaz M, Tang P, Chauve C, Hsiao W, Isaac-Renton J, Prystajecky N. 2018. Beaver fever: whole-genome characterization of waterborne outbreak and sporadic isolates to study the zoonotic transmission of giardiasis. mSphere 3:e00090-18. https://doi.org/10.1128/mSphere.00090-18.
50. Bushnell B. 2014. BBMap: a fast, accurate, splice-aware aligner (no. LBNL-7065E). Lawrence Berkeley National Lab, Berkeley, CA.
51. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. https://doi.org/10.1093/nar/25.17.3389.
52. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010.
53. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973. https://doi.org/10.1093/bioinformatics/btp348.
54. Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9:772. https://doi.org/10.1038/nmeth.2109.
55. Swofford DL. 2002. PAUP*: Phylogenetic Analysis Using Parsimony (and other methods) 4.0. Sinauer Associates, Sunderland, MA.
56. Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704. https://doi.org/10.1080/10635150390235520.

Downloaded from https://journals.asm.org/journal/aem on 17 July 2021 by 2a02:587:d449:5600:9805:7ecd:78fc:2edf.

57. R Core Team. 2013. R: a language and environment for statistical computing. http://www.R-project.org/.
58. Hoff KJ, Stanke M. 2019. Predicting genes in single genomes with Augustus. Curr Protoc Bioinformatics 65:e57. https://doi.org/10.1002/cpbi.57.
59. Kim J, Shin MY, Park SJ. 2019. RNA-sequencing profiles of cell cycle–related genes upregulated during the G2-phase in *Giardia lamblia*. Korean J Parasitol 57:185–189. https://doi.org/10.3347/kjp.2019.57.2.185.
60. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461. https://doi.org/10.1093/bioinformatics/btq461.
61. Aurrecoechea C, Brestelli J, Brunk BP, Carlton JM, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, Innamorato F, Iodice J, Kissinger JC, Kraemer E, Li W, Miller JA, Morrison HG, Nayak V, Pennington C, Pinney DF, Roos DS, Ross C, Stoeckert CJ, Jr, Sullivan S, Treatman C, Wang H. 2009. GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens *Giardia lamblia* and *Trichomonas vaginalis*. Nucleic Acids Res 37:D526–D530. https://doi.org/10.1093/nar/gkn631.
62. Borcard D, Gillet F, Legendre P. 2018. Numerical ecology with R. Springer, New York, NY.
63. Bivand R, Altman M, Anselin L, Assunção R, Berke O, Bernat A, Blanchet G. 2015. R package 'spdep.' https://cran.r-project.org/package=spdep.
64. Ruf T. 1999. The Lomb-Scargle periodogram in biological rhythm research: analysis of incomplete and unequally spaced time-series. Biol Rhythm Res 30:178–201. https://doi.org/10.1076/brhm.30.2.178.1422.
65. Qiu D. 2014. snpar: supplementary non-parametric statistics methods. R package version 1. https://CRAN.R-project.org/package=snpar.
66. Sulaiman IM, Fayer R, Bern C, Gilman RH, Trout JM, Schantz PM, Das P, Lal AA, Xiao L. 2003. Triosephosphate isomerase gene characterization and potential zoonotic transmission of *Giardia duodenalis*. Emerg Infect Dis 39:1444–1452. https://doi.org/10.3201/eid0911.030084.
67. Read CM, Monis PT, Thompson RA. 2004. Discrimination of all genotypes of *Giardia duodenalis* at the glutamate dehydrogenase locus using PCR-RFLP. Infect Genet Evol 4:125–130. https://doi.org/10.1016/j.meegid.2004.02.001.

Downloaded from https://journals.asm.org/journal/aem on 17 July 2021 by 2a02:587:d449:5600:9805:7ecd:78fc:2edf.