AMERICAN SOCIETY FOR MICROBIOLOGY | mSphere®

RESEARCH ARTICLE

Comparative Genomic and Transcriptomic Analysis of *Naegleria fowleri* Clinical and Environmental Isolates

Sandeep J. Joseph,[a]* Subin Park,[b] Alyssa Kelley,[b] Shantanu Roy,[a] Jennifer R. Cope,[a] Ibne Karim M. Ali[a]

[a]Waterborne Disease Prevention Branch, 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]Eagle Medical Services, Atlanta, Georgia, USA

ABSTRACT Out of over 40 species of *Naegleria*, which are free-living thermophilic amebae found in freshwater and soil worldwide, only *Naegleria fowleri* infects humans, causing primary amebic meningoencephalitis (PAM), a typically fatal brain disease. To understand the population structure of *Naegleria* species and the genetic relationships between *N. fowleri* isolates and to detect pathogenic factors, we characterized 52 novel clinical and environmental *N. fowleri* genomes and a single *Naegleria lovaniensis* strain, along with transcriptomic data for a subset of 37 *N. fowleri* isolates. Whole-genome analysis of 56 isolates from three *Naegleria* species (*N. fowleri*, *N. lovaniensis*, and *Naegleria gruberi*) identified several genes unique to *N. fowleri* that have previously been linked to the pathogenicity of *N. fowleri*, while other unique genes could be associated with novel pathogenicity factors in this highly fatal pathogen. Population structure analysis estimated the presence of 10 populations within the three *Naegleria* species, of which 7 populations were within *N. fowleri*. The whole-nuclear-genome (WNG) phylogenetic analysis showed an overall geographical clustering of *N. fowleri* isolates, with few exceptions, and provided higher resolution in identifying potential clusters of isolates beyond that of the traditional locus typing. There were only 34 genes that showed significant differences in gene expression between the clinical and environmental isolates. Genomic data generated in this study can be used for developing rapid molecular assays and to conduct future population-based global genomic analysis and will also be a valuable addition to genomic reference databases, where shotgun metagenomics data from routine water samples could be searched for the presence of *N. fowleri* strains.

IMPORTANCE *N. fowleri*, the only known *Naegleria* species to infect humans, causes fatal brain disease. PAM cases from 1965 to 2016 showed <20 cases per year globally. Out of approximately 150 cases in North America since 1962, only four PAM survivors are known, yielding a >97% case fatality rate, which is critically high. Although the pathogenesis of *N. fowleri* has been studied for the last 50 years, pathogenetic factors that lead to human infection and breaching the blood-brain barrier remain unknown. In addition, little is known regarding the genomic diversity both within *N. fowleri* isolates and among *Naegleria* species. In this study, we generated novel genome sequences and performed comparative genomic and transcriptomic analysis of a set of 52 *N. fowleri* draft genome sequences from clinical and environmental isolates derived from all over the world in the last 53 years, which will help shape future genome-wide studies and develop sensitive assays for routine surveillance.

KEYWORDS *Naegleria fowleri*, comparative genomics, phylogenetic analysis, population structure, primary amebic meningoencephalitis, transcriptomics

Naegleria species are free-living thermophilic amebae found mainly in freshwater and soil worldwide (1, 2). Over 40 *Naegleria* species have been characterized based on genetic variations in the internal transcribed spacer (ITS) sequences on the

Citation Joseph SJ, Park S, Kelley A, Roy S, Cope JR, Ali IKM. 2021. Comparative genomic and transcriptomic analysis of *Naegleria fowleri* clinical and environmental isolates. mSphere 6: e00637-21. https://doi.org/10.1128/mSphere.00637-21.

Editor Mariana Castanheira, JMI Laboratories

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 Sandeep J. Joseph, lww9@cdc.gov, or Ibne Karim M. Ali, xzn5@cdc.gov.

* Present address: Sandeep J. Joseph, Laboratory Reference and Research Branch, Division of Sexually Transmitted Disease Prevention, National Center for HIV/AIDS, Viral Hepatitis, STD and TB Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia, USA.

Received 16 July 2021
Accepted 20 July 2021
Published 11 August 2021

ribosomal DNA and the mitochondrial small subunit (mtSSU) rRNA gene (3). Only one species, *Naegleria fowleri*, infects humans and causes the rare disease called primary amebic meningoencephalitis (PAM) (4, 5), a fast-progressing and typically fatal brain disease. *N. fowleri* has been isolated from rivers, lakes, hot springs, swimming pools, sewage, tap waters, soil, and dust in the air (6–11). Human infection occurs when water containing the ameba enters the nose and finds its way through the nasal mucosa and olfactory nerve, entering the brain (12, 13). The progress of the disease is rapid, with a median of 5 days from onset of symptoms to death (14). PAM patients suffer from severe inflammation of the central nervous system that leads to headache, fever, vomiting, seizures, and altered mental status (15). Mortality is very high, mainly due to delayed diagnosis of the causative organism, overwhelming inflammation of the central nervous system, and lack of effective treatment (10). Although the worldwide incidence of PAM is low, the presence of *N. fowleri* in human-associated environments is an important public health issue because of the rapid onset and high mortality of PAM. Recent trends suggest a spread of *N. fowleri* infections from traditionally warmer southern-tier states in the United States to colder northern-tier states such as Minnesota. More alarmingly, *N. fowleri* colonization of two public drinking water systems in Louisiana was documented for the first time in 2013. The contamination was linked to the death of a 4-year-old whose only known exposure was playing on a backyard water slide (9, 16).

In recent years, with very few published draft genomes of the three important *Naegleria* species, *N. fowleri* (17, 18), *Naegleria lovaniensis* (19), and *Naegleria. gruberi* (20), we sought to understand the genetic diversity among these three *Naegleria* species as well as to obtain insights on *N. fowleri*-specific protein clusters that might be responsible for the pathogenicity of the species. Although there are enormous genetic variations among *Naegleria* species, little is known about intraspecies genetic diversity, especially within the pathogenic *N. fowleri* species. To gain a better understanding of the population structure for *Naegleria* species, the genetic relationships between *N. fowleri* strains and with other *Naegleria* species, and to detect possible pathogenetic factors, we genome sequenced and characterized 52 clinical and environmental *N. fowleri* isolates (representing 49 unique *N. fowleri* isolates) and a single *N. lovaniensis* strain, isolated from all over the world, along with transcriptomic data for a subset of 37 *N. fowleri* strains.

RESULTS

Genome assembly, structure, and annotation of *N. fowleri* and *N. lovaniensis* genomes. As previously described (21), we made an attempt to generate a "close-to-complete" genome assembly from a 1969 *N. fowleri* TY isolate that caused PAM in a patient from Virginia using HiSeq Illumina and PacBio data. The genome assembly was verified using optical mapping data. This led to the final haploid assembly, which consisted of 37 chromosomes and had a total size of 27.9 Mb (27,994,426 bp) with chromosome size ranging from 1,206,962 bp to 537,351 bp. Similarly, a single isolate of *N. lovaniensis* (76-15-250) was also sequenced using the PacBio RS II sequencing platform using 4 single-molecule real-time (SMRT) cells, which resulted in 401,530 reads with an average read length of 6,065 bp. FALCON long read assembler reconstructed the best assembly, consisting of 199 contigs with an N_{50} of 455,122 bp and a total size of 30.8 Mb (30,830,598 bp). Benchmarking Universal Single-Copy Orthologs (BUSCO) v4.0.6 evaluation of gene completeness performed by searching 255 eukaryotic benchmarking universal single-copy orthologs showed that for *N. fowleri* TY and *N. lovaniensis* 76-15-250 draft genomes, 222/255 (87%) and 214/255 (84%) genes, respectively, were identified, either completely or in fragments (Table 1). Smaller numbers of complete BUSCOs were found for the previously published *N. fowleri* (ATCC 30894; 208/255 genes) and *N. fowleri* (ATCC 30863; 211/255 genes) genomes than for *N. fowleri* TY, while the number of complete BUSCOs identified in *N. lovaniensis* (76-15-250; 201/255 genes) was slightly smaller than that for the published *N. lovaniensis* (ATCC 30569; 205/255 genes) draft genome.

TABLE 1 BUSCO analysis of *Naegleria* species genomes[a]

BUSCO type	No. of BUSCOs in genomes of:					
	N. fowleri (TY)[b]	*N. fowleri* (ATCC 30894)[c]	*N. fowleri* (ATCC 30863)[c]	*N. lovaniensis* (ATCC 30569)[c]	*N. lovaniensis* (76-15-250)	*N. gruberi* (ATCC 30224)[c]
Complete	215	208	211	205	201	203
Duplicated	2	2	1	10	2	1
Fragmented	7	11	11	13	13	14
Missing	33	36	33	37	41	38

[a]Genome completeness was evaluated by analyzing 255 conserved BUSCOs (v4.0.6) of the Eukaryota odb10 data set.
[b]Out of 52 *N. fowleri* genomes sequenced, BUSCO analysis results are shown only for the "close-to-complete" genome of the *N. fowleri* TY isolate.
[c]Published *Naegleria* genomes.

The *N. fowleri* TY and *N. lovaniensis* 76-15-250 genomes were gene dense, with 71% and 75% of the genomes, respectively, being defined as coding sequences. Besides the nuclear genome, *Naegleria* spp. include a circular mitochondrial genome and an extrachromosomal plasmid encoding ribosomal RNAs. The mitochondrial genome of the *N. fowleri* TY isolate was 49,486 bp, while the size of the mitochondrial genome of *N. lovaniensis* 76-15-250 was 48,529 bp. Furthermore, the ribosomal DNA (rDNA) plasmids of *N. fowleri* TY and *N. lovaniensis* 76-15-250 were 15,976 bp and 13,153 bp, respectively.

De novo repeat prediction analysis using RepeatMasker v4.0.8 and RepeatModeler on *N. fowleri* TY genomes identified approximately 5.28% of the total genome was repetitive sequences (Table 2). Simple repeats represent 1.64% of the genome sequences, while 1.56% are Long interspersed nuclear elements (LINEs) and 1.65% are unclassified repeats. Approximately 10.78% of the total genome of *N. lovaniensis* 76-15-250 was identified as repetitive regions (Table 2). Repetitive sequences are thought to be important for the appropriate folding of the genome, and they confer unique identity on an organism by introducing genetic variation within a species (22).

Ab initio gene prediction, incorporating transcriptome sequencing (RNAseq) data, of the *N. fowleri* TY nuclear genome using BRAKER2 identified 9,405 protein coding genes. Out of 9,405 genes predicted in the TY genome, 6,368 genes (68%) were functionally annotated using eggNOG-mapper, where ortholog assignments were made to the precomputed clusters and phylogenies present in the eggNOG 5 database, which also includes gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway information of each ortholog group. The number of *ab initio* protein coding gene predictions for the remaining *N. fowleri* draft genomes sequenced in this

TABLE 2 Summary of repetitive sequences identified in *N. fowleri* TY and *N. lovaniensis* (76-15-250)

Isolate or type of repeat element	No. of elements	Length occupied (bp)	% of the genome
N. fowleri TY			
Long interspersed nuclear elements (LINEs)	851	435,553	1.56
LTR elements	27	5,803	0.02
DNA elements	54	58,557	0.21
Unclassified	1,683	460,648	1.65
Small RNA	165	17,776	0.06
Simple repeats	10,470	457,811	1.64
Low complexity	827	39,272	0.14
N. lovaniensis 76-15-250			
Short interspersed nuclear elements (SINEs)	14	27,504	0.09
Long interspersed nuclear elements (LINEs)	71	105,310	0.34
LTR elements	81	70,415	0.23
DNA elements	157	317,286	1.03
Unclassified	2,501	1,550,165	5.03
Small RNA	226	921,552	2.99
Simple repeats	6,930	295,228	0.96
Low complexity	683	33,426	0.11

TABLE 3 BUSCO analysis of *Naegleria* species annotated protein sequences[a]

BUSCO type	No. of BUSCOs in:					
	N. fowleri (TY)[b]	*N. fowleri* (ATCC 30894)[c]	*N. fowleri* (ATCC 30863)[c]	*N. lovaniensis* (ATCC 30569)[c]	*N. lovaniensis* (76-15-250)	*N. gruberi* (ATCC 30224)[c]
Complete	220	219	210	220	220	203
Duplicated	3	2	7	16	4	1
Fragmented	7	7	8	10	9	14
Missing	28	29	37	25	26	38

[a]Genome completeness was evaluated by analyzing 255 conserved BUSCOs (v4.0.6) of the Eukaryota odb10 data set. Publicly available protein coding sequences were downloaded from figshare (https://doi.org/10.6084/m9.figshare.8313656) for *N. fowleri* (ATCC 30894), while for the rest of the published genomes, the reannotated protein sequences using Braker2 were used for the BUSCO analysis.
[b]Out of 52 *N. fowleri* genomes sequenced, BUSCO analysis is shown only for the "close-to-complete" genome of the *N. fowleri* TY isolate.
[c]Published *Naegleria* genomes.

study, either incorporating RNAseq data or not, ranged from 9,363 to 9,635 genes (see Table S1 in the supplemental material). *Ab initio* gene prediction of the *N. lovaniensis* 76-15-250 nuclear genome using Augustus identified 11,305 genes. eggNOG-mapper functionally annotated 7,323 genes (65%) in the *N. lovaniensis* 76-15-250 genome. The set of predicted proteins identified to be complete BUSCOs both in *N. fowleri* TY and *N. lovaniensis* 76-15-250 genomes were highly similar to their respective previously identified *Naegleria* species draft genome assemblies (Table 3). The number of genes assigned to the various clusters of orthologous groups (COGs) functions for both species are summarized in Fig. 1.

Identification of orthologous gene clusters among *Naegleria* species. To identify *N. fowleri* genes that are unique compared to those in *N. gruberi* and *N. lovaniensis*, proteins from each of the species (*N. fowleri* TY, *N. lovaniensis* 76-15-250, and *N. gruberi*) were initially clustered using Neptune, CD-HIT, and OrthoMCL, and the unique genus-specific genes and clusters were parsed out from each of the three tools. The genus formed 11,530 protein clusters (with at least 2 or more proteins present) and 7,511 orthologous clusters (with at least two species present). There were 5,749 protein clusters shared by all *Naegleria* species, out of which 4,019 clusters had exactly one protein from each of the three species. In total, 1,246 clusters are shared between *N. fowleri* and *N. lovaniensis*, while *N. fowleri* shares only 255 protein clusters with *N. gruberi* (Fig. 2). There were 638 and 3,018 proteins that did not cluster with any species (singletons) for

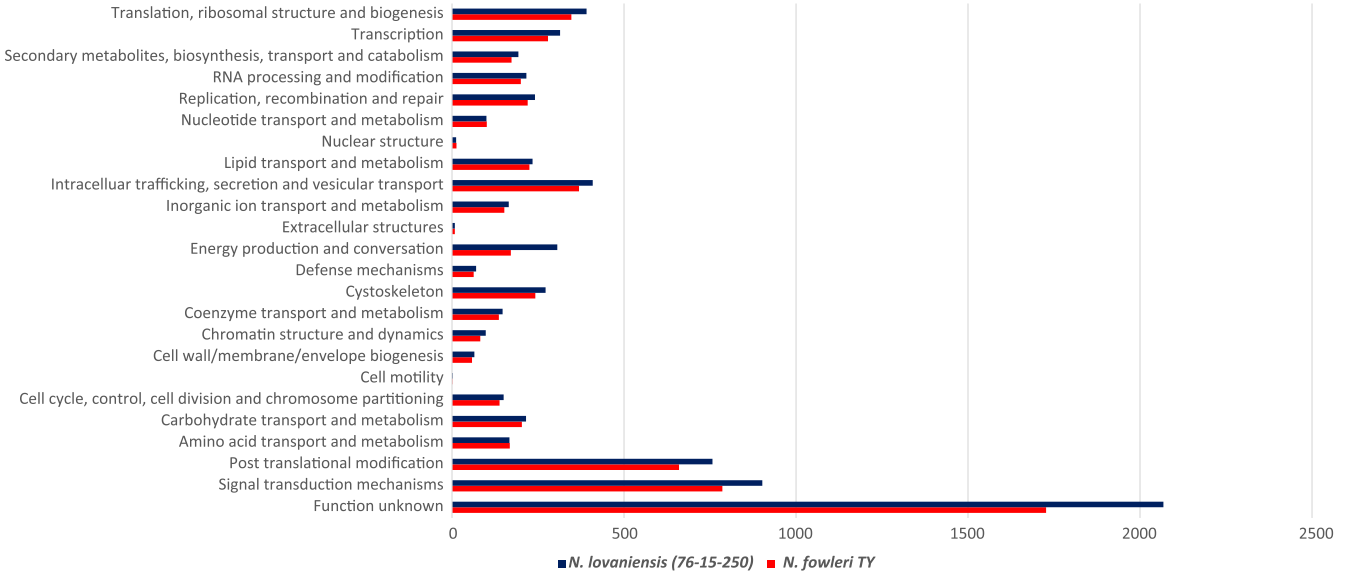

FIG 1 Summary of eggNOG-mapper results on the functional annotation of *N. fowleri* TY and *N. lovaniensis* 76-15-250 genomes. The number of genes assigned to the various clusters of orthologous groups (COGs) functions based on eggNOG-mapper for the *N. fowleri* TY and *N. lovaniensis* 76-15-250 genomes.

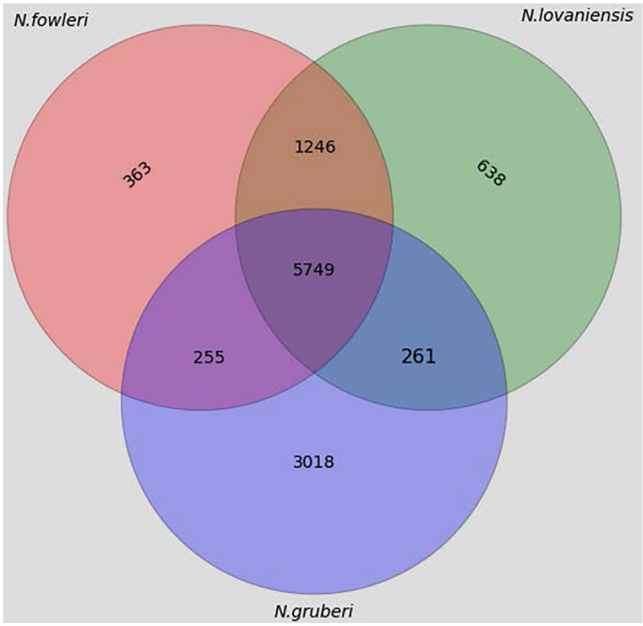

FIG 2 Results of cluster analysis showing the number of orthologous clusters shared between the three *Naegleria* species.

N. lovaniensis and N. gruberi, respectively (see Table S2 in the supplemental material). A total of 404 *N. fowleri* genes were identified that were unique compared to *N. gruberi* and *N. lovaniensis*, of which 80 genes have an annotation with known function and GO term assigned. GO term enrichment analysis on the unique *N. fowleri* genes showed that GO terms in each of the three categories, namely biological process (BP), molecular function (MF), and cellular component (CC), were significantly enriched after accounting for multiple testing. For the biological process category, GO terms describing the cytoskeleton, morphogenesis, and movements were enriched among *N. fowleri*-specific gene clusters, with *P* values of <0.05 (see Fig. S1a and Table S3 in the supplemental material). For the cellular component category, phagocytic cup, lysosome activity, and actin-related components were detected (Fig. S1b and Table S3). For the molecular function category, GO terms describing phosphatase and hydrolase activity, along with actin related binding and motor activity, were enriched among *N. fowleri*-specific gene clusters (Fig. S1c and Table S3).

To understand the protein diversity at the population level within the genus *Naegleria*, predicted proteins of all sequenced isolates (i.e., 56 *Naegleria* isolates representing 3 species) were also clustered into gene clusters/families using CD-HIT and Neptune. A total of 16,380 protein clusters were identified in *Naegleria* species. There was a total of 3,583 protein clusters shared among all 3 *Naegleria* species, where each cluster contained at least one protein sequence from the 53 *N. fowleri*, the two *N. lovaniensis*, and the single *N. gruberi* genomes. At the population level, 1,468 protein clusters were shared between *N. fowleri* and *N. lovaniensis*, while *N. fowleri* only shared 77 protein clusters with *N. gruberi*, and *N. lovaniensis* shared 283 protein clusters with *N. gruberi*. Also, there were 274 (containing 301 *N. fowleri* proteins), 3,619, and 7,648 protein clusters for *N. fowleri*, *N. lovaniensis*, and *N. gruberi*, respectively, that contained only single-species proteins (see Fig. S2 in the supplemental material). Out of 301 *N. fowleri*-specific proteins obtained from the population level analysis, only 51 genes were identified to be functionally annotated or contained a characterized domain. GO terms associated with cytoskeleton, lytic vesicle, actin-related components, transmembrane binding, and protease-related functions were mostly enriched among the *N. fowleri*-specific proteins obtained from the population-level comparison (uncorrected $P < 0.05$); however, after multiple correction only the GO terms associated with cellular response to toxic substances in biological processes was statistically significant

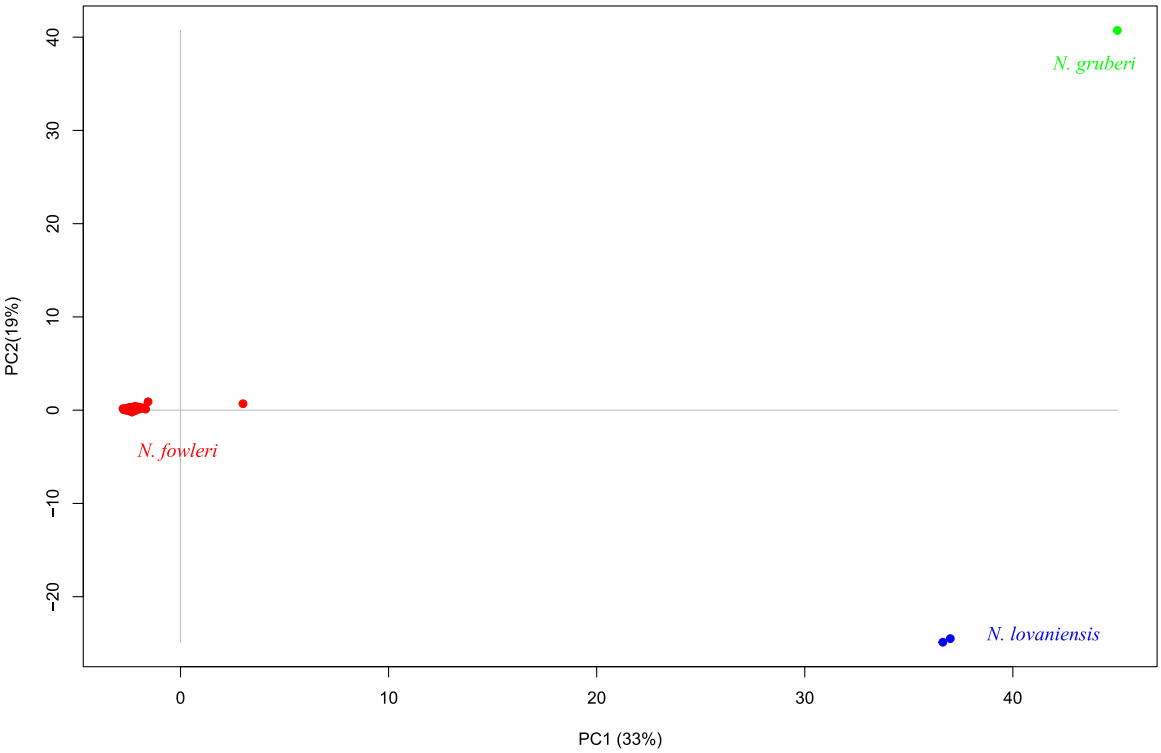

FIG 3 Principal-component analysis (PCA) for the three *Naegleria* species gene cluster/family distribution matrix. Each data point represents a *Naegleria* species genome in the two first principal components of the gene cluster distribution matrix. Percentages on the axis show how much of the total *Naegleria* species gene family matrix variation is seen along each principal component.

(Bonferroni $P = 0.03956$) (see Fig. S3 and Table S3 in the supplemental material). There were 118 *N. fowleri*-specific proteins that were common within the three *Naegleria* species-level genome comparison and the population-level comparison with 56 *Naegleria* genomes from 3 species, indicating that there is variability in the protein coding genes within *N. fowleri* isolates.

To investigate patterns of shared gene content, we ran a principal-component analysis (PCA) on the gene family distribution matrix of *Naegleria* species using the "panpca" function (*micropan* R package). Around 52% of the total variation among the *Naegleria* species genomes were seen along the two principal components (Fig. 3). Isolates from the same species formed nonoverlapping groups even after including all of the observed gene families in *Naegleria* species.

Population structure of *Naegleria* species. The clustered coancestry heatmap (Fig. 4), generated using ChromoPainter and fineSTRUCTURE analysis using genome-wide haplotype data from 56 *Naegleria* isolates, supported the existence of 3 species. This analysis estimated the presence of 10 populations within *Naegleria* species, of which 7 populations were within *N. fowleri*. Similarly to the phylogenetic analysis of *N. fowleri* isolates (described below), the population structure analysis also identified 4 major populations within *N. fowleri*, which correlated with the 4 genotypic phylogenetic clades. All *N. fowleri* isolates showed high shared coancestry with *N. gruberi*, while the two *N. lovaniensis* isolates showed the highest shared coancestry with *N. gruberi* and very limited shared coancestry with *N. fowleri*. Only one *N. fowleri* isolate, NF30863, showed some evidence of shared coancestry with *N. lovaniensis*. This analysis suggested that *N. gruberi* could be an ancestral species to both *N. lovaniensis* and *N. fowleri*, with more shared ancestry with *N. lovaniensis* than *N. fowleri* (Fig. 4).

Phylogenetic relationship among *N. fowleri* strains based on nuclear and mitochondrial genome data. Maximum-likelihood phylogenetic analysis of the whole nuclear genome (WNG) (Fig. 5) and mitochondrial genome of *N. fowleri* isolates (see Fig. S4 in the supplemental material) revealed that all isolates clustered primarily into

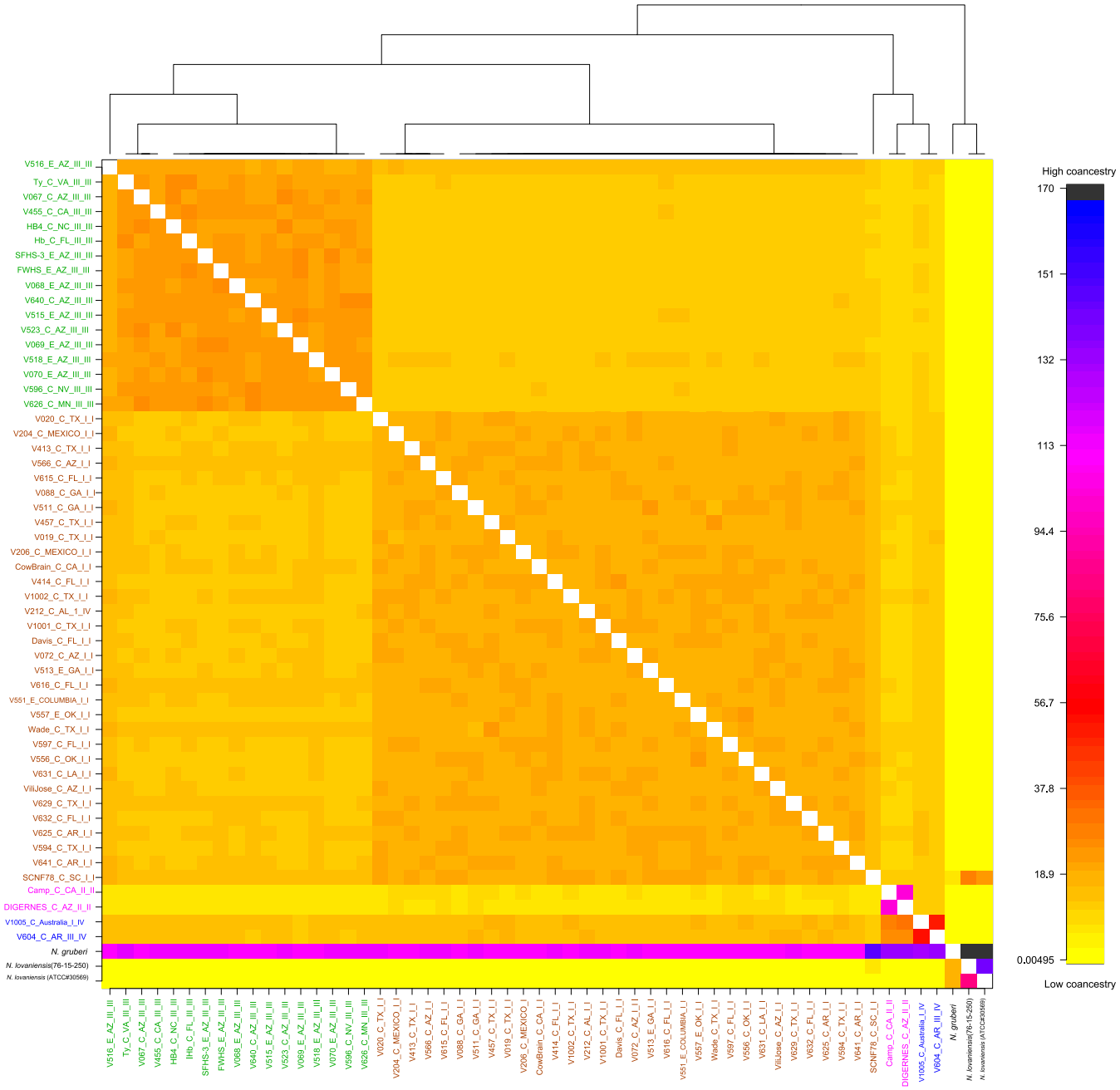

FIG 4 ChromoPainter coancestry matrix for *Naegleria* species with population structure assigned based on fineSTRUCTURE analysis. The heatmap shows the number of shared DNA chunks (coancestry) copied from a donor genome (*x* axis) to a recipient genome (*y* axis). This analysis estimated the presence of 10 populations within *Naegleria* species, of which 7 populations were within *N. fowleri*.

four and three phylogenetic clades, respectively. In fact, the WNG phylogenetic topological relationships among isolates were similar to the phylogenetic relationships previously inferred from both traditional mitochondrial small subunit (mtSSU) rRNA and ITS genotyping loci, where isolates belonging to a particular genotype formed a distinct clade. Essentially, the four main clades inferred in the WNG phylogeny consisted of isolates that corresponded to the four genotypes; clade 1 included 28 isolates with genotype I loci (shown in red), clade 2 consisted of 17 isolates with genotype III loci (shown in green), clade 3 included 2 isolates with genotype II loci (shown in purple), and clade 4 consisted of 2 isolates with genotype IV, based on the ITS locus only (shown in blue) (Fig. 5). The population structure analysis using ChromoPainter and

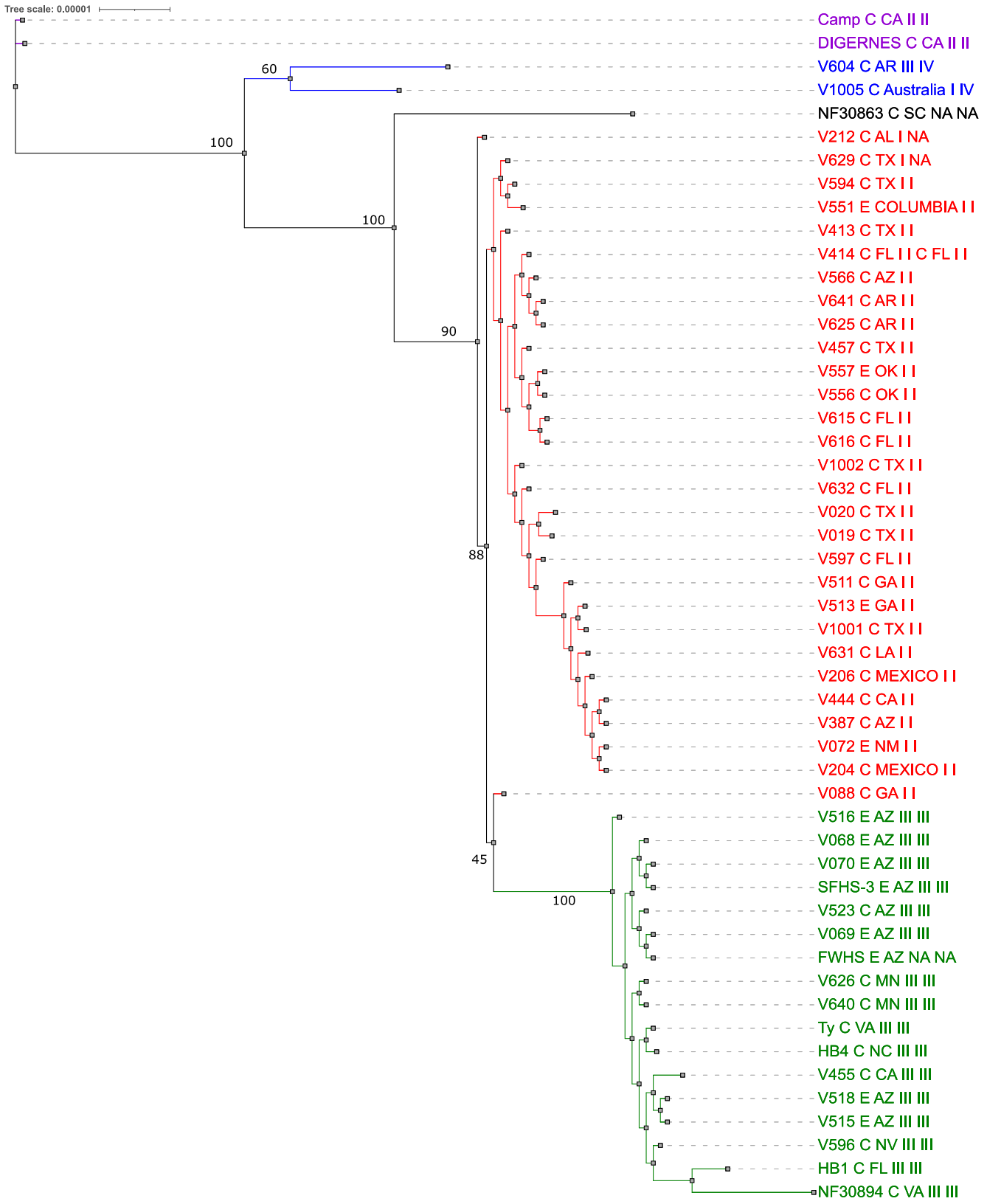

FIG 5 Whole-genome phylogeny of all the *N. fowleri* isolates. Isolate label contains isolate identifier (ID), clinical (C) or environmental (E) isolate, the U.S. state or country of origin, mitochondrial small subunit (mtSSU) rRNA gene, and internal transcribed space (ITS) genotype. "NA" represents missing data and, for genotyping data, cases in which the genotyping loci was not confidently identified from the draft genomes of those isolates. Only one isolate from each of the duplicated *N. fowleri* isolates was included in the phylogenetic analysis. A recently published *N. fowleri* isolate (ATCC 30894) by Liechti et al. (18) was included in this phylogenetic analysis. Bootstrap support estimates of major ancestral nodes are also shown.

fineSTRUCTURE also identified *N. fowleri* populations that correlated with the phylogenetic clades. Based on population structure analysis, clade 1 (genotype 1) isolates formed 2 populations. Similarly, clade 2 (genotype III) isolates formed three populations, with the V516 isolate forming a single population, which was also observed in the WNG phylogeny, where the V516 isolate was clustered as an outgroup for all clade 2 isolates. All of the clade 3 (genotype II) and clade 4 (genotype IV) isolates formed their own separate clades and populations in both the WNG phylogeny and population structure analysis and showed high levels of intrapopulation coancestry within their respective populations. The phylogenetic clades reconstructed using WNG partially agreed with previous reporting by Pelandakis et al. (23) that *N. fowleri* isolates tended to form geographical clusters, in which clade 2 isolates belonged to the American variant group and clade 4 included isolates from the South Pacific Chooz (SPC) variant (V1005 is isolated from Australia), while the rest of the clades consisted of isolates from the widespread Euro-American (WEA) variant. Clade 4 (genotype IV) not only included a South Pacific strain (V1005) from Australia but also an isolate (V604) from Arkansas, United States, an exception to the geographical specificity of identifying genotype IV isolates (Fig. 5). Similarly, the mitochondrial phylogeny formed three clades that correspond to their traditional genotyping loci (Fig. S4). Although clustered together with the WNG sequences, the Australian isolate V1005 and the Arkansas isolate V604 showed an identical ITS genotype (IV) but differed in the mtSSU genotypes (I and III, respectively). To our knowledge, these are rare instances, where ITS and mtSSU genotypes are showing disagreement with each other (24). This may suggest a possible horizontal gene transfer in the mitochondrial genome leading to these discrepancies. Similarly, the clade 2 isolates that clustered with other genotype III isolates in the WNG phylogenetic tree, V516 and V518, clustered with the clade 1 (genotype 1) isolates in the mitochondrial phylogeny, again suggesting possible horizontal gene transfer of the entire mitochondrial genome (Fig. S4).

Even though the topology of the WNG phylogenetic tree among the isolates was similar to previously inferred phylogenetic trees using traditional mtSSU rRNA and ITS typing loci, where isolates belonged to a particular genotype formed a distinct clade, the WNG phylogenetic tree provided sufficient resolution to distinguish strains even within a genotype. For example, within clade 2 isolates (genotype III), there is a subclade in which 6 clinical and environmental (V069, FWHS, V523, V070, SFHS-3, and V068), all from Arizona and isolated in 1987, clustered together, and the average core single-nucleotide polymorphism (SNP) difference within this subclade was 14 SNPs (4 to 29 SNPs), indicating all these 6 isolates might have originated from a common ancestor and suggesting that they could possibly represent a single *N. fowleri* isolate (Table S4). V523 being the only clinical isolate in this clade suggests that this clinical case could have been the result of acquiring any of the other three environmental isolates, but which isolate/strain was acquired by the infected individual could not be inferred from the genomic data alone without strong epidemiological data to support any possible exposure event. Also, two isolates from Oklahoma (both isolated in 2005) formed a two-isolate subclade within clade 1 (genotype I); one caused a clinical case (V556) and the other was an environmental isolate (V557), and they both showed high genetic similarity (11 SNP differences), suggesting these two isolates could be the same *N. fowleri* strain and that the most likely source of the clinical case could be the environmental isolate. Similarly, two clinical isolates from Florida (V615 and V616; clade 1 [genotype 1], both isolated in 2009) also showed high genetic similarity (5-SNP difference; Table S4). Similarly, we also noticed other evidence of geographical clustering within the U.S. isolates. For example, two isolates each from Arkansas (V625 [2010] and V641 [2013]; 8-SNP difference) and Minnesota (V626 [2010] and V640 [2012]; 3-SNP difference) could be the same isolate within each of their respective states. At the same time, not all geographically clustered isolates would necessarily be the same isolate. For example, two isolates each from California (DIGERNES [1980s] and Camp [1978]; 55-SNP difference) and Texas (V019 [1984] and V020 [1984]; 124-SNP difference)

cannot possibly be the same isolate due to the higher pairwise genetic differences. In contrast, we also noticed evidence of clustering of isolates originating from different states of the United States. For examples, the Arizona clinical isolate V387 (1996) was very similar to the clinical isolate V444 (2004), which was originally isolated from the brain of a cow from California (9-SNP difference), suggesting that the same *N. fowleri* isolate is capable of producing PAM in both humans and animals. Clinical isolate V204 (1990) from Mexico was very similar to environmental isolate V072 (1987) from New Mexico (20-SNP difference).

Only a limited number of genes showed significant transcriptome differences among clinical and environmental *N. fowleri* strains. RNAseq analysis carried out to study transcriptome changes among clinical ($N = 29$) and environmental ($N = 8$) *N. fowleri* strains identified 34 differentially expressed genes (false-discovery rate [FDR], $<1\%$; $q < 0.1$) (see Table S5 and Fig. S5 in the supplemental material). Among these, 14 genes had higher levels of gene expression in clinical strains compared to those in environmental isolates. These included genes with the following annotations: serine C-palmitoyl transferase, structural constituent of ribosome, DNA replication initiation, mitogen-activated protein (MAP) kinase activity, ion channel binding, and a gene with the Hint (Hog/INTein) auto-proteolytic protein-processing domain. At the same time, some of the genes that showed increased gene expression in the environmental isolates compared to the clinical isolates included genes responsible for ATP binding, regulation of the canonical Wnt signaling pathway, translation initiation factor activity, guanylate cyclase activity, cAMP biosynthetic process, methyltransferase small domain, cytoskeleton organization, and pyridoxal 5′-phosphate salvage. None of the GO terms associated with these differentially expressed genes were statistically significant in our GO term enrichment analysis, precluding us from assessing the functions and pathways of these genes in terms of the pathogenicity of *N. fowleri*.

DISCUSSION

Out of the 40 described *Naegleria* species, only *N. fowleri* is known to infect humans and cause death. Data on PAM cases from 1965 to 2016 show that fewer than 20 cases per year have been reported globally (25). Only four well-documented PAM survivors are known among approximately 150 cases that occurred in North America since 1962, yielding an astounding $>97\%$ case fatality rate, which is one of the highest among infectious diseases. Although the pathogenesis of *N. fowleri* has been studied for the last 50 years, the mechanisms or the pathogenetic factors that lead to human infection and breaching of the blood-brain barrier to invade the central nervous system remain unknown. In addition, little is known regarding the genomic diversity both within *N. fowleri* isolates and among *Naegleria* species. In this study, we performed comparative genomic and transcriptomic analysis of a set of 52 *N. fowleri* draft genomes from clinical and environmental isolates derived from all over the world in the last 53 years, which will help shape the future genome-wide studies of this pathogen. In addition, we genome sequenced one *N. lovaniensis* isolate and performed comparative genome analysis of 56 isolates from 3 *Naegleria* species, including the available published genomes.

The genome sequence of *N. fowleri* TY (21), generated based on scaffolding long-read PacBio *de novo*-assembled contigs through guided optical mapping data and polishing with high-quality Illumina reads, has generated a "near-to-complete" reference genome with the highest quality to date for the human-pathogenic ameba *N. fowleri*. We believe that the 37 final contigs generated can be considered equivalent to 37 chromosomes, even though karyotyping studies are needed for final confirmation. The recently published NF30894 genome (18), sequenced using Oxford Nanopore Technology (ONT) and assembled into 90 contigs has an estimated genome size of 29.54 Mb, which is longer than the TY genome, but NF30894 has a lower N_{50} value (717,491 bp versus 756,811 bp) than that of the TY genome. The TY genome showed the presence of 215 complete BUSCO genes, while the NF30894 genome only indicated the presence of 208 complete BUSCO genes along. The number of fragmented BUSCO genes was higher in the ONT-assembled NF30894 genome than that in the TY genome. Even though both ONT and

PacBio generate long contigs, the incorporation of optical mapping (26), which captures the labeling patterns of long DNA molecules using restriction enzymes, for scaffolding merged most of the TY contigs. The use of such optical mapping for scaffolding Illumina-polished ONT contigs is highly recommended (27) in determining the whole-genome sequences of complex eukaryotes like *N. fowleri*. Even though the *N. lovaniensis* 76-15-250 genome sequenced in this study has a very similar total genome length to that of a recently published *N. lovaniensis* (ATCC 30569) (19) genome, the number of contigs was higher (199 contigs) in the former genome isolate than in the latter isolate (111 contigs). The *N. lovaniensis* 76-15-250 genome also has a smaller N_{50} value (455,122 bp versus 657,933 bp). The incorporation of optical mapping for scaffolding these contigs would have improved the *N. lovaniensis* 76-15-250 draft genome.

Although fewer protein coding genes (9,405) were identified in the TY genome compared to those identified in all *Naegleria* species, especially in NF30894 (13,925) and NF30863 (17,252), we believe the predicted number of genes in the TY genome could possibly be a much more realistic estimate, mainly because of better genome scaffolding through optical mapping, as well as the incorporation of RNAseq data in the *ab initio* gene prediction. *Ab initio* gene prediction for NF30894 did not include transcriptomic data, while that for NF30863 included RNAseq data in gene calling but the genome lacked substantial enrichment of genomic DNA (17), which was reflected in the large number of contigs generated (1,729 contigs) and hence the greater number of predicted genes. However, the analysis of BUSCOs on all the protein coding genes showed a larger number of complete BUSCOs in the TY genome (220 genes) than in the NF30894 (219 genes) and NF30863 (210 genes) genomes, suggesting that the TY gene prediction could possibly be more realistic. *Ab initio* gene prediction of the *N. lovanensis* 76-15-250 gene predicted a smaller number than that predicted in the published *N. lovaniensis* (ATCC 30569) genome. The differences in the number of predicted genes might be due to lack of evidence from RNAseq data in our annotation, although both genomes have the same number of complete BUSCOs.

Recently, Herman et al. (28) sequenced two new *N. fowleri* isolates, V212 and 986, and by incorporating transcriptomic data for gene annotation they identified 12,677 and 11,599 genes, respectively. Even though the numbers of genes identified in these two genomes were a little higher than what we detected in the TY genome but lower than the numbers of genes detected in previously sequenced *N. fowleri* genomes, we believe that the scaffolding of the TY genome using optical mapping might have led to a realistic estimation of the number of genes present in *N. fowleri*. By comparing three *N. fowleri* (V212, 986, and NF30863) genomes against the *N. gruberi* ATCC 30224 genome, Herman et al. (28) detected 11,399 orthogroups (protein coding gene families), while our comparison of *N. fowleri* (52 isolates), *N. lovaniensis* (2 isolates), and *N. gruberi* (1 isolate) also estimated similar number of orthogroups (11,530). Similarly, Herman et al. (28) identified 7,656 genes shared among all genomes of the two *Naegleria* species, and 10,451 were shared by three *N. fowleri* isolates, whereas our analysis using multiple isolates from three *Naegleria* species estimated the presence of only 3,583 genes shared among the *Naegleria* species and 5,402 genes that were shared by all the 52 *N. fowleri* isolates. The disparity in the gene content could be mainly due to the greater genetic diversity of *N. fowleri* isolates included in this study, as well the fact that as our comparison included an additional *Naegleria* species, *N. lovaniensis*. Likewise, the numbers of unique *N. fowleri* genes identified in this study and by Herman et al. (28) were 274 and 458, respectively.

Considering that the high mortality rate of PAM comes from the combination of the intensive immune responses, tissue degradation during infection through overwhelming inflammation of the central nervous system, and parasite motility and phagocytosis, the results of GO enrichment analysis conducted on the unique *N. fowleri* genes in this study reflect the potential factors involved in the pathogenesis of PAM. For example, during the trophozoite free-living stage, *N. fowleri* uses a structure called a food cup to ingest bacteria and yeast, and in the human host, this same structure is used to

ingest red and white blood cells and tissue (29). We were able to identify *N. fowleri*-unique genes that could facilitate the functions of the food cup, including those encoding acid hydrolases (Ty_07029 and Ty_03291), phospholipases (Ty_08620), and phospholipolytic enzymes that might play an active role in nerve tissue destruction. GO term analysis of unique *N. fowleri* genes also showed the presence of statistically enriched GO terms associated with pathways potentially involved in food cup and phagocytic cellular components, such as early phagosome (GO:0032009), phagocytic cup (GO:001891), and phagolysosome (GO:0032010), which were enriched in seven unique *N. fowleri* genes (Ty_00511, Ty_01019, Ty_01138, Ty_02771, Ty_5276, Ty_06706, and Ty_07422). It is also known that *N. fowleri* secretes a pore-forming protein termed naegleriapore A (with the protein domain saposin B type; UniProt accession number Q9BKM2), as well as proteases, to traverse the extracellular matrix during infection and to kill and digest cells (18, 30). Four genes with saposin B-type domains (Ty_05913, Ty_06795, Ty_06806, and Ty_08201), which could potentially be naegleriapore A coding genes, were detected in the TY genomes, but none were unique to *N. fowleri*. The prominent protease cathepsin protease (Nf314 or cathepsin [Ty_06276]), which is highly associated with the pathogenic *N. fowleri* strains (31–34), was also identified as a unique *N. fowleri* gene in our analysis (see Table S3 in the supplemental material). Herman et al. (28) identified that gene expression of both the precursor protein for naegleriapore A and cathepsin A were upregulated in pathogenic *N. fowleri* strains during a transcriptomic study comparing the gene expression profiles of an *N. fowleri* strain passed through mice to the expression of the same strain grown in axenic medium. The endosomal vesicle trafficking gene Rab GTPase Rab14 (Ty_05736) (28), which regulates early steps during phagocytosis and endocytosis, was one of the *N. fowleri*-unique genes identified in this study. Genome-wide transcription profiling by Herman et al. (28) found that the gene expression of Rab32, a paralogue of Rab14, was upregulated in pathogenic *N. fowleri* strains.

Proteins associated with actin cytoskeletal rearrangements have been linked as a pathogenicity factor in *N. fowleri*, in particular NF-actin, which is involved in phagocytosis and its location in the food cup structure (17, 35). Furthermore, actin binding proteins and upstream regulators of actin polymerization were reported to correlate with virulence. In our GO term analysis of unique *N. fowleri* genes, several GO terms associated with actin were significantly enriched (GO:0070360: migration of symbiont within host by polymerization of host actin; GO:0070359: actin polymerization-dependent cell motility involved in migration of symbiont in host; and GO:0070358: actin polymerization-dependent cell motility; Table S3). Among the *N. fowleri*-unique genes, we detected four cytoskeleton protein coding genes, actin-binding protein profilin (Ty_09040), tubulin beta (Ty_01613), actin (Ty_06706), and tubulin-binding protein kinesin family member C1 (Ty_08296). Profilin is known to regulate the actin cytoskeletal rearrangements by inhibiting actin polymerization, and it controls the Rap1 signaling pathway, which in turn regulates plasma membrane receptors, cytoskeleton rearrangement, intracellular adhesion, and cell mobility. Considering that upstream regulation of actin polymerization was reported to be a virulence factor (36), the actin-binding protein profilin (Ty_09040) could be an important player in the pathogenicity of *N. fowleri*. As noted above, actin is also dependent on phagocytosis, and Dianokova et al. (37) showed that the actin binding myosins are concentrated around the phagocytic cups, and thus myosins may be involved in phagocytic process in *N. fowleri* (17). A few studies have detected ~11 myosin motor proteins, which were considered potential pathogenicity factors and include myosin II heavy chain and myosin Ie. We also detected several myosin motor protein coding genes in *N. fowleri*, including type I myosin protein coding genes ($n = 6$), myosin heavy chain ($n = 4$), myosin light chain kinase ($n = 1$), myosin XI tail binding ($n = 1$), and cofilin/tropomyosin-type actin binding protein ($n = 1$); however, none of the myosin protein coding genes were *N. fowleri*-specific genes. All of the myosin proteins are implicated in the organization and polarization of the actin cytoskeleton.

As previously identified (18), we also found several protein coding genes containing a DNAJ (HSP40) domain, which are known as regulators of HSP70 (38), a heat shock

protein (HSP) linked to the pathogenicity of *N. fowleri*. A few genes encoding HSP70, HSP90, and HSP20 were also identified in the *N. fowleri* genome, but none of the HSP coding genes were unique to *N. fowleri*.

Variations in virulence of different *N. fowleri* virulence isolates are poorly understood, due to the small number of reference *N. fowleri* isolates used in virulence studies. Ideally, in animal studies, randomly obtained *N. fowleri* isolates from freshwater environments would be used to investigate whether *N. fowleri* isolates have the same capacity to cause PAM. In the absence of these data, we studied population structures within 52 *N. fowleri* isolates obtained from human PAM patients and environmental samples. Forty-eight of the 52 *N. fowleri* isolates came from 14 U.S. states, 2 from Mexico, and 1 each from Australia and Colombia. However, one major limitation of this study was that the number of environmental samples available was much lower than for the clinical samples (12 and 40 for genome comparison and 8 and 29 for transcriptome analyses, respectively). We only had environmental isolates from 4 out of 14 states, namely, Arizona ($N = 8$), Georgia ($N = 1$), New Mexico ($N = 1$), and Oklahoma ($N = 1$). The only Colombian isolate was from an environmental sample, and no clinical isolates were available from Colombia for genome comparison. Eight environmental isolates were collected in connection with three clinical samples from Arizona, and the high-resolution WNG phylogenetic analysis indicated possible links of six environmental isolates to a single clinical case in Arizona. The high-resolution separation of isolates based on the WNG phylogeny within a given mtSSU or ITS genotype suggests that new loci could be identified for the development of more descriptive genotyping tools useful for infection source tracking investigations.

RNAseq data showed only a limited number of genes ($N = 34$) that were differentially expressed among clinical and environmental *N. fowleri* isolates. Moreover, the expression fold change between the clinical and environmental isolates was low. Only 1.2- to 1.9-fold and 1.3- to 2.8-fold expression differences were noted among genes in clinical isolates that were more or less expressed than those in the environmental isolates (see Table S5 in the supplemental material). Perhaps this is not surprising, given the fact that (i) the number of environmental isolates was low ($N = 8$) compared to that from the clinical cases ($N = 29$), and (ii) all the isolates from both groups were maintained under identical laboratory conditions that are different from conditions inside the host (human brain) or warm freshwater environments. Moreover, it has already been shown (28) that the gene expression profile of a mouse-passaged *N. fowleri* strain compared to that of the same strain grown in axenic medium was different, suggesting that gene expression differed depending on whether the organism was in an infectious stage within the host, where it encounters host immune cells, or outside in a freshwater environment. Nevertheless, some of the differentially expressed genes between clinical and environmental isolates could be involved in pathogenesis and virulence, assuming that the clinical isolates represent only a minor population of *N. fowleri* in the environment and that they possess intrinsically different gene expression profile. When we looked at the chromosomal locations of some of the differentially expressed genes, we noticed something quite interesting (see Table S5 in the supplemental material). Nine out of 14 more highly expressed genes (64%) in clinical isolates were restricted to 3 chromosomes (chromosomes 3, 7, and 23). Likewise, 12 of 20 genes expressed at a lower level (60%) in clinical isolates were restricted to only 2 chromosomes (chromosomes 15 and 19). Since the chromosomal locations of a majority of the differentially genes in clinical strains (versus environmental strains) are mutually exclusive from each other, this seems to indicate an epigenetic regulation of gene expression that might regulate virulence in *N. fowleri*.

In summary, the genomic and transcriptomic data of this study will help shape future research in *N. fowleri* from basic biology to understanding of virulence and pathogenicity, global population structures, geographic origins, environmental exposure, and transmission dynamics. In addition, data generated in this study can be used for identifying novel genotyping loci for developing rapid molecular assays and will also be a valuable addition to genomic reference databases, where shotgun metagenomics data from routine water samples could be searched for the presence of *N. fowleri* strains.

MATERIALS AND METHODS

***Naegleria* species cultures and isolates.** The 52 *Naegleria fowleri* isolates and a nonpathogenic *Naegleria* species, *Naegleria lovaniensis* isolate (76-15-250) used in this study were cultured axenically in T-25 flasks in Nelson medium at 37°C (*N. fowleri* isolates) and 30°C (*N. lovaniensis*), respectively (39, 40). Forty-eight of the 52 *N. fowleri* isolates were originally from the United States (from 14 states) and the isolates included 37 from PAM patients and 11 from environmental water samples (from 4 states and Colombia). The remaining four *N. fowleri* isolates were from PAM patients outside the United States (2 from Mexico and 1 each from Australia and Colombia). Some of the metadata about the *N. fowleri* isolates are provided in Table S1 in the supplemental material.

Preparation of genomic DNA and RNA libraries. Genomic DNA and RNA was extracted after harvesting log-phase trophozoites ($\sim 2 \times 10^6$) by placing the culture flasks on ice for 10 min to dislodge the trophozoites from flask walls, followed by centrifugation at $1,500 \times g$ for 10 min. The resulting culture pellet was washed 3 times with 1× PBS and was used in DNA or RNA extraction according to the manufacturer's instructions (DNeasy blood and tissue kit and RNeasy minikit; Qiagen). Culture pellets were stored at −80°C with 1 ml of TRIzol after a 1× PBS wash. Extracted RNA was converted to cDNA and purified with a Qiagen PCR purification kit before library preparation for sequencing.

For Illumina library preparation, genomic DNA was sheared to a mean size of 600 bp using an LE220 focused ultrasonicator (Covaris, Inc., Woburn, MA). DNA fragments were AMPure (Beckman Coulter, Inc., Indianapolis, IN) cleaned and used to prepare dual-indexed sequencing libraries using NEBNext Ultra library prep reagents (New England Biolabs, Inc., Ipswich, MA) and barcoding indices synthesized in the CDC Biotechnology Core Facility. Libraries were analyzed for size and concentration, then pooled and denatured for loading onto flow cells for cluster generation. Sequencing was performed on an Illumina HiSeq2500 (or MiSeq) instrument using HiSeq Rapid SBS v2 250- × 250-cycle (or MiSeq 250- × 250-cycle) paired-end sequencing kits. On completion, sequence reads were filtered for read quality, base called, and demultiplexed using bcl2fastq (v2.19). A standard 20-kb PacBio library was size selected at 10 kb and then run on an RS II instrument for 360-min movies on 4 SMRTcells.

Processing of genomic data. We recently generated a "close-to-complete" genome of an *N. fowleri* TY isolate (ATCC 30107) using HiSeq (Illumina) and RS II (Pacific Biosciences, Menlo Park, CA) instruments and verified the assembled genome with optical mapping data (21). Briefly, genomic DNA was harvested from log-phase trophozoites ($\sim 2 \times 10^6$), and libraries were prepared for PacBio sequencing runs using the SMRTbell template prep kit 1.0 and polymerase binding kit P6, the filtered reads (minLength = 1000bps) were *de novo* assembled using Canu v1.6 (41), and the resulting consensus sequences were determined with Quiver (v1) (https://github.com/PacificBiosciences/GenomicConsensus). The assembly was confirmed by comparison to restriction digest optical maps obtained from log-phase trophozoites ($\sim 1 \times 10^6$) using the Argus system (OpGen) with MapSolver (v2.1.1; http://www.opgen.com). Final PacBio assembly of the TY strain was further polished by mapping Illumina reads using unicycler_polish (default parameters; Unicycler package v4.4) (42). This led to the identification of 37 contigs representing 37 chromosomes in *N. fowleri* (21). For all remaining *N. fowleri* isolates (see Table S1 in the supplemental material) described in this study, a combination of HiSeq/MiSeq (Illumina, San Diego, CA) libraries was prepared using the NEBNext Ultra library prep kit (New England Biolabs). The Illumina reads were trimmed with cutadapt (43) to remove adaptor sequences and reads below Q20 and 75 bp. Short-read assemblies were carried out using SPAdes (44) with the "-careful" option. The *de novo* assemblies were improved by scaffolding the best contigs with SSPACE (45) using the short-read Illumina sequences. The resultant scaffolds from each of the isolates were mapped against the "complete" *N. fowleri* TY PacBio genome using ABACAS (46), which uses the MUMmer alignment tool to identify syntenies of contigs against the reference genome and generates "pseudo chromosomes," taking overlapping contigs and gaps into account. Remaining sequence gaps were filled using GapFiller (47), which used short-read Illumina sequences to fill the gaps. Contigs from the previously published *N. fowleri* draft genome were also obtained from GenBank and scaffolded using ABACAS based on the *N. fowleri* TY genome. The isolate pairs Davis and V414, Wade and V457, and V067 and V523 were cryopreserved in different time frames and sequenced, so technically those isolate pairs are the same isolates with different names. A single *N. lovaniensis* isolate (76-15-250), sequenced using both PacBio and Illumina, was processed similarly as described above, but no restriction digest optical maps were generated for this isolate. For all *N. fowleri* isolates, the two traditional genotyping loci (24, 48, 49), mitochondrial small subunit rRNA (mtSSU rRNA), and the internal transcribed spacer (ITS) were identified by comparing their genome data to a local database of reference loci using custom Python scripts.

Genome completeness and annotation. To assess the completeness of the *N. fowleri* TY and *N. lovaniensis* 76-15-250 genomes, BUSCO v4.0.6 (50) was used to search the latest and updated set of 255 conserved eukaryotic Benchmarking Universal Single-Copy Orthologs (BUSCO) of the Eukaryota odb10 data set. To validate the output of BUSCO and for comparison between *Naegleria* species, the tool with the updated Eukaryota odb10 data set was reapplied to the previously published genomes of *N. fowleri* (17, 18), *N. gruberi* (20), and *N. lovaniensis* (ATCC 30569) (19).

Repetitive elements were predicted using a *de novo* approach by applying RepeatMasker v4.0.8 (51), which uses the Repeat Protein Database, which contains more than 7,400 entries and includes 16.1 million amino acids covering 133 subclasses of transposable elements, and by using RepeatModeler (52). The protein coding genes on the nuclear genome of all *N. fowleri* isolates were predicted using an *ab initio* approach taking into account RNAseq data by applying Braker2 (https://github.com/Gaius-Augustus/BRAKER) (53), which incorporates both GeneMark (54) and Augustus v3.3.2 (55) genome annotation tools. Briefly, Braker2 was initially run on the *N. fowleri* TY genome in *-esmode* without RNAseq evidence in order to train *N. fowleri*-specific Augustus parameters. Once the Augustus parameters were obtained,

a second Braker2 run was performed by reusing the pretrained parameters and providing the RNAseq evidence to Augustus only. For those isolates with RNAseq data, evidence was provided by mapping the RNAseq data (in .bam format) using TopHat2 (56) (-library-type fr-unstranded and the rest with default parameters) to the corresponding draft genome of the isolate. For isolates without RNAseq data, genome annotation was performed using Augustus using pretrained parameters without providing expression data evidence. The protein coding genes on the nuclear genomes of other *Naegleria* species (both newly sequenced and published) were reannotated using Augustus after obtaining respective species-specific training parameters as described above, except for *N. fowleri* (ATCC 30894), where gene annotations were publicly available. The completeness of the gene annotations of *N. fowleri* TY and *N. lovaniensis* 76-15-250 along with those of all previously published *Naegleria* species were assessed using BUSCO as described above. Functional annotation of all the *Naegleria* sp. predicted genes were obtained using eggNOG-mapper (57), which is based on eggNOG v4.5 orthology data.

Identifying orthologous genes among *Naegleria* species. To gain an overview of the gene repository of the genus *Naegleria* and to identify *N. fowleri* genes that are unique compared to *N. gruberi* and *N. lovaniensis*, proteins from each of the species were clustered using Neptune (58), CD-HIT (59), and OrthoMCL (60). The default parameter was used for Neptune to calculate the proper size of the *k*-mer, and both the filter length and filter percentage were set at 0.4. For CD-HIT (v4.6), the CD-HIT-2d comparison algorithm was used with a word size of 2 (-n 2) and 0.4 as the length difference cutoff (-s2). For OrthoMCL, the complete predicted protein coding sequences from all three *Naegleria* species genomes were searched against themselves using BLASTP with an E-value cutoff of 1E−10 and a minimum percent identity of 50% for significance. The best BLASTP scores were used for identifying orthologous groups using the OrthoMCL algorithm (60). A matrix with the information on the presence or absence of all orthologous genes present in all 56 isolates in the three *Naegleria* species was created and imported into the R package *micropan* (61), to perform principal-component analysis (PCA) and to generate visualizations to describe genome clustering. GO term enrichment analysis for the *N. fowleri* genes identified as unique compared to *N. gruberi* and *N. lovaniensis*, both at the genus and population level, was performed using GOATOOLS (62), and the statistically enriched GO terms were identified after Bonferroni multiple testing *P* value correction (p_bonferroni < 0.05). All of the statistically significant enriched GO terms within biological process (BP), molecular function (MF), and cell component (CC) categories were visualized using the Seaborn package in IPython (63).

Population structure analysis of *Naegleria* species. To elucidate the possible population structure within the 56 *Naegleria* isolates representing three *Naegleria* species, the ChromoPainter algorithm (64) was applied to the genome-wide haplotype data, generated from a core genes nucleotide alignment generated by concatenating all the genes shared among the 3 *Naegleria* species, using the linkage model. A recombination map file was created by specifying a uniform recombination rate per site per generation using a Perl script called makeuniformrecfile.pl (http://www.paintmychromosomes.com). Each isolate is reconstructed using the haplotypes of each of the other isolates in the sample as possible donors. The donor in each region is interpreted as the isolate with which the haplotype shares the most recent common ancestor for that stretch of DNA. The output from ChromoPainter is a coancestry matrix that summarizes the similarity between isolates in terms of the number of haplotype "DNA chunks" used to reconstruct the recipient isolate from each donor isolate. The fineSTRUCTURE algorithm (64) used the coancestry matrix generated by ChromoPainter to perform model-based clustering using a Bayesian Markov chain Monte Carlo (MCMC) approach to explore the population structure. FineSTRUCTURE was run for 400,000 iterations; the first 200,000 iterations were discarded as MCMC burn-in. The thinning interval was specified at 100.

Phylogenetic analysis of *N. fowleri* isolates/strains. *N. fowleri* Illumina reads were mapped to the *N. fowleri* TY "complete" reference genome using BWA-MEM v0.7.12 (65) for short-read mapping. SNPs were called using Freebayes v1.0.2 (https://github.com/ekg/freebayes), requiring a minimum read coverage of 10× with a Phred quality score of at least 20. For each SNP that passes these criteria in any one isolate, consensus base calls for the SNP loci were extracted from all genomes mapped, and if the Phred quality scores were under 20, that allele was treated as unknown and represented with a gap (N) character. Chromosomal SNPs with confident homozygous calls in >90% of the genomes mapped were concatenated to form an SNP alignment. A maximum-likelihood (ML) phylogenetic tree was inferred from the chromosomal SNP alignment using RAxML (v8.2.9) (66) with a generalized time-reversible model and gamma distribution model with site-specific rate variations (GTRGAMMA nucleotide model in RAxML with ascertainment correction [-asc-corr=stamatakis] and 100 bootstrap pseudoreplicates used to assess branch support for the ML phylogeny). Only one isolate from each of the pairs of isolates (Davis and V414, Wade and V457, and V067 and V523) that were cryopreserved at different times were used in the phylogenetic analysis. The aforementioned chromosomal SNP alignment was also used to calculate the SNP differences among closely related *N. fowleri* isolates or phylogenetic clades. Phylogenetic analysis of all of the *N. fowleri* mitochondrial genomes were also performed as described above.

Differential gene and transcript expression analysis of clinical and environmental *N. fowleri* isolates. Thirty-seven of the 52 *N. fowleri* isolates were available for differential gene expression analysis, including 29 from clinical PAM patients and 8 from environmental samples (see Table S1 in the supplemental material). RNAseq reads in FASTQ format for each isolate were first aligned to the *N. fowleri* TY reference genome using TopHat2 (56), and then transcript assembly was performed using Cufflinks (67). All transcript assemblies were merged using Cuffmerge to create a single merged transcriptome annotation. Counts for each gene were quantified and normalized across all samples using Cuffnorm v2.2.1.35. Differentially expressed genes in either clinical or environmental groups were identified by Cuffdiff using the following cutoffs: $\log_2$ fold change > 1 or < -1 and false-discovery rate [FDR] $< 1\%$.

Data availability. All sequencing and transcriptomic data associated with this study were submitted to the National Center for Biotechnology Information's Sequence Read Archive (SRA) under BioProject accession identifier PRJNA642022. All of the gene sequences of *N. fowleri TY* and functional annotations of all genes and of the *N. fowleri* unique genes are available in the figshare repository (https://figshare.com/projects/Comparative_genomic_and_transcriptomic_analysis_of_Naegleria_fowleri_clinical_and_environmental_isolates/118200).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, PDF file, 1 MB.

FIG S2, EPS file, 0.4 MB.

FIG S3, PDF file, 2.6 MB.

FIG S4, PDF file, 0.02 MB.

FIG S5, EPS file, 0.4 MB.

TABLE S1, XLSX file, 0.02 MB.

TABLE S2, XLSX file, 0.1 MB.

TABLE S3, XLSX file, 0.9 MB.

TABLE S4, XLSX file, 0.02 MB.

TABLE S5, XLSX file, 0.01 MB.

ACKNOWLEDGMENTS

We thank Govinda Visvesvara, founder of the CDC Free-Living Ameba (FLA) laboratory, who relentlessly isolated and cryopreserved the majority of the *N. fowleri* strains used in this study over a period of 4 decades. We thank the members of the CDC core facility who performed the genome sequencing and optical mapping procedures.

I.K.M.A. conceived the study. S.J.J. and I.K.M.A. designed the study. S.J.J., A.K., and S.P. performed all of the bioinformatics and data analysis. S.R. performed the laboratory experiments. J.R.C. provided the epidemiological data. S.J.J., with contributions from I.K.M.A. and J.R.C., wrote and prepared the first draft of the manuscript, which was reviewed by all authors for revisions.

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

We declare that we do not have any competing interests.

REFERENCES

1. Martinez AJ. 1993. Free-living amebas: infection of the central nervous system. Mt Sinai J Med 60:271–278.
2. Martinez AJ, Visvesvara GS. 1997. Free-living, amphizoic and opportunistic amebas. Brain Pathol 7:583–598. https://doi.org/10.1111/j.1750-3639.1997.tb01076.x.
3. De Jonckheere JF. 2014. What do we know by now about the genus *Naegleria*? Exp Parasitol 145:S2–S9. https://doi.org/10.1016/j.exppara.2014.07.011.
4. Carter RF. 1970. Description of a *Naegleria* sp. isolated from two cases of primary amoebic meningo-encephalitis, and of the experimental pathological changes induced by it. J Pathol 100:217–244. https://doi.org/10.1002/path.1711000402.
5. Visvesvara GS, Stehr-Green JK. 1990. Epidemiology of free-living ameba infections. J Protozool 37:25S–33S. https://doi.org/10.1111/j.1550-7408.1990.tb01142.x.
6. Yoder JS, Eddy BA, Visvesvara GS, Capewell L, Beach MJ. 2010. The epidemiology of primary amoebic meningoencephalitis in the USA, 1962–2008. Epidemiol Infect 138:968–975. https://doi.org/10.1017/S0950268809991014.
7. Izumiyama S, Yagita K, Furushima-Shimogawara R, Asakura T, Karasudani T, Endo T. 2003. Occurrence and distribution of *Naegleria* species in thermal waters in Japan. J Eukaryot Microbiol 50:514–515. https://doi.org/10.1111/j.1550-7408.2003.tb00614.x.
8. Sbeeban KB, Ferris MJ, Benson JM. 2003. Detection of *Naegleria* sp. in a thermal, acidic stream in Yellowstone National Park. J Eukaryotic Microbiology 50:263–265. https://doi.org/10.1111/j.1550-7408.2003.tb00132.x.
9. Kemble SK, Lynfield R, DeVries AS, Drehner DM, Pomputius WF, 3rd, Beach MJ, Visvesvara GS, da Silva AJ, Hill VR, Yoder JS, Xiao L, Smith KE, Danila R. 2012. Fatal *Naegleria fowleri* infection acquired in Minnesota: possible expanded range of a deadly thermophilic organism. Clin Infect Dis 54:805–809. https://doi.org/10.1093/cid/cir961.
10. Yoder JS, Straif-Bourgeois S, Roy SL, Moore TA, Visvesvara GS, Ratard RC, Hill VR, Wilson JD, Linscott AJ, Crager R, Kozak NA, Sriram R, Narayanan J, Mull B, Kahler AM, Schneeberger C, da Silva AJ, Poudel M, Baumgarten KL, Xiao L, Beach MJ. 2012. Primary amebic meningoencephalitis deaths associated with sinus irrigation using contaminated tap water. Clin Infect Dis 55:e79–e85. https://doi.org/10.1093/cid/cis626.
11. Cursons RT, Brown TJ, Keys EA, Gordon EH, Leng RH, Havill JH, Hyne BE. 1979. Primary amoebic meningo-encephalitis in an indoor heat-exchange swimming pool. N Z Med J 90:330–331.
12. Marciano-Cabral F, Cabral GA. 2007. The immune response to *Naegleria fowleri* amebae and pathogenesis of infection. FEMS Immunol Med Microbiol 51:243–259. https://doi.org/10.1111/j.1574-695X.2007.00332.x.
13. Lozano-Alarcon F, Bradley GA, Houser BS, Visvesvara GS. 1997. Primary amebic meningoencephalitis due to *Naegleria fowleri* in a South American tapir. Vet Pathol 34:239–243. https://doi.org/10.1177/030098589703400312.
14. Capewell LG, Harris AM, Yoder JS, Cope JR, Eddy BA, Roy SL, Visvesvara GS, Fox LM, Beach MJ. 2015. Diagnosis, clinical course, and treatment of primary amoebic meningoencephalitis in the United States, 1937–2013. J Pediatric Infect Dis Soc 4:e68–e75. https://doi.org/10.1093/jpids/piu103.
15. Visvesvara GS, Moura H, Schuster FL. 2007. Pathogenic and opportunistic free-living amoebae: *Acanthamoeba* spp., *Balamuthia mandrillaris*, *Naegleria fowleri*, and *Sappinia diploidea*. FEMS Immunol Med Microbiol 50:1–26. https://doi.org/10.1111/j.1574-695X.2007.00232.x.

16. Cope JR, Ratard RC, Hill VR, Sokol T, Causey JJ, Yoder JS, Mirani G, Mull B, Mukerjee KA, Narayanan J, Doucet M, Qvarnstrom Y, Poole CN, Akingbola OA, Ritter JM, Xiong Z, da Silva AJ, Roellig D, Van Dyke RB, Stern H, Xiao L, Beach MJ. 2015. The first association of a primary amebic meningoencephalitis death with culturable *Naegleria fowleri* in tap water from a US treated public drinking water system. Clin Infect Dis 60:e36–e42. https://doi.org/10.1093/cid/civ017.
17. Zysset-Burri DC, Müller N, Beuret C, Heller M, Schürch N, Gottstein B, Wittwer M. 2014. Genome-wide identification of pathogenicity factors of the free-living amoeba *Naegleria fowleri*. BMC Genomics 15:496. https://doi.org/10.1186/1471-2164-15-496.
18. Liechti N, Schürch N, Bruggmann R, Wittwer M. 2019. Nanopore sequencing improves the draft genome of the human pathogenic amoeba *Naegleria fowleri*. Sci Rep 9:16040. https://doi.org/10.1038/s41598-019-52572-0.
19. Liechti N, Schürch N, Bruggmann R, Wittwer M. 2018. The genome of *Naegleria lovaniensis*, the basis for a comparative approach to unravel pathogenicity factors of the human pathogenic amoeba *N. fowleri*. BMC Genomics 19:654. https://doi.org/10.1186/s12864-018-4994-1.
20. Fritz-Laylin LK, Prochnik SE, Ginger ML, Dacks JB, Carpenter ML, Field MC, Kuo A, Paredez A, Chapman J, Pham J, Shu S, Neupane R, Cipriano M, Mancuso J, Tu H, Salamov A, Lindquist E, Shapiro H, Lucas S, Grigoriev IV, Cande WZ, Fulton C, Rokhsar DS, Dawson SC. 2010. The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. Cell 140:631–642. https://doi.org/10.1016/j.cell.2010.01.032.
21. Ali IKM, Kelley A, Joseph SJ, Park S, Roy S, Jackson J, Cope JR, Rowe LA, Burroughs M, Sheth M, Batra D, Loparev V. 2021. Draft chromosome sequences of a clinical isolate of the free-living ameba *Naegleria fowleri*. Microbiol Resour Announc 10:e01034-20. https://doi.org/10.1128/MRA.01034-20.
22. Shapiro JA, von Sternberg R. 2005. Why repetitive DNA is essential to genome function. Biol Rev Camb Philos Soc 80:227–250. https://doi.org/10.1017/s1464793104006657.
23. Pélandakis M, Serre S, Pernin P. 2000. Analysis of the 5.8S rRNA gene and the internal transcribed spacers in Naegleria spp. and in N. fowleri. J Eukaryot Microbiol 47:116–121. https://doi.org/10.1111/j.1550-7408.2000.tb00020.x.
24. Zhou L, Sriram R, Visvesvara GS, Xiao L. 2003. Genetic variations in the internal transcribed spacer and mitochondrial small subunit rRNA gene of *Naegleria* spp. J Eukaryot Microbiol 50:522–526. https://doi.org/10.1111/j.1550-7408.2003.tb00617.x.
25. Gharpure R, Bliton J, Goodman A, Ali IKM, Yoder J, Cope JR. 2020. Epidemiology and clinical characteristics of primary amebic meningoencephalitis caused by *Naegleria fowleri*: a global review. Clin Infect Dis 73:e19–e27. https://doi.org/10.1093/cid/ciaa520.
26. Jiao W-B, Accinelli GG, Hartwig B, Kiefer C, Baker D, Severing E, Willing E-M, Piednoel M, Woetzel S, Madrid-Herrero E, Huettel B, Hümann U, Reinhard R, Koch MA, Swan D, Clavijo B, Coupland G, Schneeberger K. 2017. Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. Genome Res 27:778–786. https://doi.org/10.1101/gr.213652.116.
27. Deschamps S, Zhang Y, Llaca V, Ye L, Sanyal A, King M, May G, Lin H. 2018. A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. Nat Commun 9:4844. https://doi.org/10.1038/s41467-018-07271-1.
28. Herman EK, Greninger AL, van der Geizen M, Ginger ML, Ramirez-Macias I, Miller HC, Morgan MJ, Tsaousis AD, Velle K, Vargova R, Najle SR, MacIntyre G, Muller B, Wittwer M, Zysset-Burri DC, Elias M, Slamovits C, Weirauch M, Fritz-Laylin L, Marciano-Cabral F, Puzon GJ, Walsh Y, Chiu C, Dacks JB. 2020. A comparative 'omics approach to candidate pathogenicity factor discovery in the brain-eating amoeba *Naegleria fowleri*. bioRxiv https://doi.org/10.1101/2020.01.16.908186.
29. Sohn H-J, Kim J-H, Shin M-H, Song K-J, Shin H-J. 2010. The *Nf-actin* gene is an important factor for food-cup formation and cytotoxicity of pathogenic *Naegleria fowleri*. Parasitol Res 106:917–924. https://doi.org/10.1007/s00436-010-1760-y.
30. Herbst R, Ott C, Jacobs T, Marti T, Marciano-Cabral F, Leippe M. 2002. Pore-forming polypeptides of the pathogenic protozoon *Naegleria fowleri*. J Biol Chem 277:22353–22360. https://doi.org/10.1074/jbc.M201475200.
31. Aldape K, Huizinga H, Bouvier J, McKerrow J. 1994. *Naegleria fowleri*: characterization of a secreted histolytic cysteine protease. Exp Parasitol 78:230–241. https://doi.org/10.1006/expr.1994.1023.
32. Hu WN, Band RN, Kopachik WJ. 1991. Virulence-related protein synthesis in *Naegleria fowleri*. Infect Immun 59:4278–4282. https://doi.org/10.1128/iai.59.11.4278-4282.1991.
33. Serrano-Luna J, Cervantes-Sandoval I, Tsutsumi V, Shibayama M. 2007. A biochemical comparison of proteases from pathogenic *Naegleria fowleri* and non-pathogenic *Naegleria gruberi*. J Eukaryot Microbiol 54:411–417. https://doi.org/10.1111/j.1550-7408.2007.00280.x.
34. Hu WN, Kopachik W, Band RN. 1992. Cloning and characterization of transcripts showing virulence-related gene expression in *Naegleria fowleri*. Infect Immun 60:2418–2424. https://doi.org/10.1128/iai.60.6.2418-2424.1992.
35. Lee Y-J, Kim J-H, Jeong S-R, Song K-J, Kim K, Park S, Park M-S, Shin H-J. 2007. Production of Nfa1-specific monoclonal antibodies that influences the *in vitro* cytotoxicity of *Naegleria fowleri* trophozoites on microglial cells. Parasitol Res 101:1191–1196. https://doi.org/10.1007/s00436-007-0600-1.
36. Velle KB, Fritz-Laylin LK. 2020. Arp2/3 complex-mediated actin assembly drives microtubule-independent motility and phagocytosis in the evolutionarily divergent amoeba *Naegleria*. bioRxiv https://doi.org/10.1101/2020.05.12.091538.
37. Diakonova M, Bokoch G, Swanson JA. 2002. Dynamics of cytoskeletal proteins during Fcγ receptor-mediated phagocytosis in macrophages. Mol Biol Cell 13:402–411. https://doi.org/10.1091/mbc.01-05-0273.
38. Fan C-Y, Lee S, Cyr DM. 2003. Mechanisms for regulation of Hsp70 function by Hsp40. Cell Stress Chaper 8:309–316. https://doi.org/10.1379/1466-1268(2003)008<0309:MFROHF>2.0.CO;2.
39. Marciano-Cabral F. 1988. Biology of *Naegleria* spp. Microbiol Rev 52:114–133. https://doi.org/10.1128/mr.52.1.114-133.1988.
40. Schuster FL. 2002. Cultivation of pathogenic and opportunistic free-living amebas. Clin Microbiol Rev 15:342–354. https://doi.org/10.1128/CMR.15.3.342-354.2002.
41. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Res 27:722–736. https://doi.org/10.1101/gr.215087.116.
42. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595. https://doi.org/10.1371/journal.pcbi.1005595.
43. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17:10–12. https://doi.org/10.14806/ej.17.1.200.
44. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.
45. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27:578–579. https://doi.org/10.1093/bioinformatics/btq683.
46. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. 2009. ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics 25:1968–1969. https://doi.org/10.1093/bioinformatics/btp347.
47. Nadalin F, Vezzi F, Policriti A. 2012. GapFiller: a *de novo* assembly approach to fill the gap within paired reads. BMC Bioinformatics 13(Suppl 14):S8. https://doi.org/10.1186/1471-2105-13-S14-S8.
48. Johan F. 2002. A century of research on the amoeboflagellate genus *Naegleria*. Acta Protozool 41:309–342.
49. De Jonckheere JF. 1987. Characterization of *Naegleria* species by restriction endonuclease digestion of whole-cell DNA. Mol Biochem Parasitol 24:55–66. https://doi.org/10.1016/0166-6851(87)90115-0.
50. Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation completeness. Methods Mol Biol 1962:227–245. https://doi.org/10.1007/978-1-4939-9173-0_14.
51. Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics Chapter 4:Unit 4.10.
52. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A 117:9451–9457. https://doi.org/10.1073/pnas.1921046117.
53. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32:767–769. https://doi.org/10.1093/bioinformatics/btv661.

54. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res 33:6494–6506. https://doi.org/10.1093/nar/gki937.
55. Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res 33:W465–7. https://doi.org/10.1093/nar/gki458.
56. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14:R36. https://doi.org/10.1186/gb-2013-14-4-r36.
57. Huerta-Cepas J, Forslund K, Pedro Coelho L, Szklarczyk D, Juhl Jensen L, von Mering C, Bork P. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol Biol Evol 34:2115–2122. https://doi.org/10.1093/molbev/msx148.
58. Marinier E, Zaheer R, Berry C, Weedmark KA, Domaratzki M, Mabon P, Knox NC, Reimer AR, Graham MR, Chui L, Patterson-Fortin L, Zhang J, Pagotto F, Farber J, Mahony J, Seyer K, Bekal S, Tremblay C, Isaac-Renton J, Prystajecky N, Chen J, Slade P, Van Domselaar G. 2017. Neptune: a bioinformatics tool for rapid discovery of genomic variation in bacterial populations. Nucleic Acids Res 45:e159. https://doi.org/10.1093/nar/gkx702.
59. Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152. https://doi.org/10.1093/bioinformatics/bts565.
60. Li L, Stoeckert CJ, Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189. https://doi.org/10.1101/gr.1224503.
61. Snipen L, Liland KH. 2015. micropanmicropan: an R-package for microbial pan-genomics. BMC Bioinformatics 16:79. https://doi.org/10.1186/s12859-015-0517-0.
62. Klopfenstein DV, Zhang L, Pedersen BS, Ramírez F, Warwick Vesztrocy A, Naldi A, Mungall CJ, Yunes JM, Botvinnik O, Weigel M, Dampier W, Dessimoz C, Flick P, Tang H. 2018. GOATOOLS: a Python library for gene ontology analyses. Sci Rep 8:10872. https://doi.org/10.1038/s41598-018-28948-z.
63. Hunter JD. 2007. Matplotlib: a 2D graphics environment. Comput Sci Eng 9:90–95. https://doi.org/10.1109/MCSE.2007.55.
64. Lawson DJ, Hellenthal G, Myers S, Falush D. 2012. Inference of population structure using dense haplotype data. PLoS Genet 8:e1002453. https://doi.org/10.1371/journal.pgen.1002453.
65. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 1303.3997 [q-bio.GN]. https://arxiv.org/abs/1303.3997.
66. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033.
67. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7:562–578. https://doi.org/10.1038/nprot.2012.016.