Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

OPEN ACCESS   Check for updates

Analyses of publicly available *Hungatella hathewayi* genomes revealed genetic distances indicating they belong to more than one species

Laura E. Hernández-Juárez [a,b], Margarita Camorlinga[a], Alfonso Méndez-Tenorio[b], Judith Flores Calderón[c], B. Carol Huang[d], DJ Darwin R. Bandoy[d,e], Bart C. Weimer [d,*], and Javier Torres [a,*]

[a]Unidad de Investigación Médica en Enfermedades Infecciosas y Parasitarias, Hospital de Pediatría, Centro Médico Nacional Siglo XXI, Instituto Mexicano del Seguro Social, Ciudad de México, México; [b]Laboratorio de Bioinformática y Biotecnología Genómica, Departamento de Bioquímica. Escuela Nacional de Ciencias Biológicas, Unidad Profesional Lázaro Cárdenas, Instituto Politécnico Nacional, Ciudad de México, México; [c]Departamento de Gastroenterología, Hospital de Pediatría, Centro Médico Nacional Siglo XXI, Instituto Mexicano del Seguro Social, Ciudad de México, México; [d]Population Health and Reproduction Department, School of Veterinary Medicine, 100K Pathogen Genome Project, University of California Davis, Davis, CA, USA; [e]Department of Veterinary Paraclinical Sciences, College of Veterinary Medicine, University of the Philippines Los Baños, Laguna, Philippines

ABSTRACT

Hungatella hathewayi has been observed to be a member of the gut microbiome. Unfortunately, little is known about this organism in spite of being associated with human fatalities; it is important to understand virulence mechanisms and epidemiological prospective to cause disease. In this study, a patient with chronic neurologic symptoms presented to the clinic with subsequent isolation of a strain with phenotypic characteristics suggestive of *Clostridium difficile*. However, whole-genome sequence found the organism to be *H. hathewayi*. Analysis including publicly available *Hungatella* genomes found substantial genomic differences as compared to the type strain, indicating this isolate was not *C. difficile*. We examined the whole-genome of *Hungatella* species and related genera, using comparative genomics to fully examine species identification and toxin production. Orthogonal phylogenetic using the 16S rRNA gene and entire genome analyses that included genome distance analyses using Genome-to-Genome Distance (GGDC), Average Nucleotide Identity (ANI), and a pan-genome analysis with inclusion of available public genomes determined the speciation to be *Hungatella*. Two clearly differentiated groups were identified, one including a reference *H. hathewayi* genome (strain DSM-13,479) and a second group that was determined to be *H. effluvii*, which included our clinical isolate. Also, some genomes reported as *H. hathewayi* were found to belong to other genera, including *Clostridium* and *Faecalicatena*. We show that the *Hungatella* species have an open pan-genome reflecting high genomic diversity. This study highlights the importance of correctly assigning taxonomic identification, particularly in disease-associated strains, to better understand virulence and therapeutic options.

Introduction

In humans, anaerobic bacteria colonize many body sites, including the skin and multiple mucosal surfaces. Collectively, the microbiome is an important interacting, functionally active, and dynamic collection of microbes that impact human health [1]. The gut microbiome has variation in its composition between individuals but remains relatively stable over time [1]; although it is impacted by several factors that include, diet composition, age, and health status [2]. Several studies have established that the predominant bacteria in the human gut are members of the phyla *Firmicutes*

and *Bacteroidetes* [3,4]. Within the *Firmicutes* phylum, *Clostridium* is widely distributed in the intestinal microbiota of human and animals and has an important impact on host health status [5]. Steer et al. described *Clostridium hathewayi* that was originally isolated from a chemostat inoculated with human feces [6] and was classified within the rRNA gene cluster XIVa.

H. hathewayi is recognized as a nonpathogenic component of the human gut microbiome [7]. However, some reports indicated that *H. hathewayi* can cause septicemia in humans [5,8–11]. Importantly, a recent

---

gut microbiota study in patients with coronavirus SARS-CoV2 (COVID-19) revealed that high abundance of *H. hathewayi* is significantly associated with severe COVID-19 cases [12]. Recently, detailed phenotypic analyses and phylogenetic studies using the 16S rRNA gene found that the *C. hathewayi* type strain DSM-13479 was more closely related to *Hungatella effluvii* than to other clostridia, which led to the suggested reclassification of *C. hathewayi* to *Hungatella hathewayi* [13].

Classification of microbial genomes is becoming increasing common using whole-genome sequencing (WGS), which enables the analyses of the entire genome for relatedness and provides an approach that is scalable as new genomes are sequenced [14]. Use of WGS to examine the relatedness of organisms enables a new tool that expands the use of phenotypic differences with an accuracy that is superior to other typing tools [14]. In addition to taxonomic identification, genetic diversity, and virulence traits can be infer for disease potential. Additionally, allelic variants of core genes can be determined for disease presentation [15]. This is particularly useful in genomically diverse organisms with open pan-genomes and are common in the microbiome [15]. Previously, the genome sequence diversity of *Hungatella* was recognized to be larger than initially contemplated and led to the suggestion that the type strain is not representative of this species, as defined by the Human Microbiome Project [16]. WGS is emerging in clinical diagnostics to meet the accuracy demands for identification that do not exist for ecological studies [17]. Yet, WGS analyses provide very detailed and exhaustive analyses so as to place organisms into the correct taxonomy; this is especially important with *H. hathewayi* whose identity has been questioned relative to types strains that were previously defined using culturing tools and limited phenotyping assays [16].

In this study, a 12-year-old girl presented to the clinic with a 2 years-lasting episode of chronic abdominal pain. During the diagnostic process, we isolated an anaerobic sporulating bacilli that was initially identified as *C. difficile* using accepted clinical laboratory identification methods. Subsequently, a WGS comparison of this isolate and publicly available genomes identified the isolate as *H. hathewayi* was done. However, significant differences in the phenotype and genotype between the reported *H. hathewayi* genomes were observed, suggesting the new clinical isolate was substantially different to the historical reference *H. hathewayi* strain. With additional genomic comparisons, we provide evidence to support the need to reclassify members of this species and that the reference genome is not representative of the species. These studies demonstrated that the genomic diversity among reported *H. hathewayi* genomes is larger than previously appreciated, highlighting that WGS is a powerful tool for a proper taxonomic classification, which is crucial for clinically important bacterial identification that leads to therapy decisions.

Materials and methods

A 12-year-old girl with a history of partial epilepsy since the age of four presented to the hospital with a persistent two-year episode of chronic abdominal pain and loss of appetite, but no diarrhea. The patient was included in a protocol to study *C. difficile* infection in children, approved by the ethical committee of the hospital de Pediatria, IMSS, Mexico.

Isolation and characterization of a sporulated anaerobe from a 12-year-old girl

A stool sample was treated with 95% ethanol for 30 minutes and inoculated on cycloserine (250 μg/mL) cefoxitin (8 μg/mL) fructose agar (CCFA) (Becton Dickinson, USA), a selective medium for isolation of *C. difficile*. The plates were incubated for 48 h at 37°C in an anaerobic atmosphere (85% N_2, 5% H_2, 10% CO_2). Single colonies with distinctive morphology, odor, and Gram stain characteristic for *C. difficile* were selected for clinical diagnostics [18]. These colonies were subcultured on Casman Blood (Becton Dickinson, USA) in aerobic, microaerophilic, and anaerobic conditions to verify that the isolate was a strict anaerobe. Egg yolk agar (Becton Dickinson, USA) with Schaeffer–Fulton stain was used to observe spores [13].

Virulence factors and antibiotic resistance analysis

The recovered isolate (labeled IMSS-269) was tested for cytotoxicity in a mammalian cell culture, as described for *C. difficile* [19]. Briefly, human cervical carcinoma cells (HeLa) were cultured in DMEM (In vitro, Mexico), 1% inoculation into 96-well microtiter plates and allowed to reach 70% of confluency before testing, which was typically 24 h at 37°C with 5% CO_2 atmosphere. Bacterial inocula were grown in thioglycolate broth (Becton Dickinson, USA) for 24 h at 37°C anaerobically before using 50 μL to filter sterilize prior to inoculating the HeLa monolayers. The cytotoxic assay was examined at 24, 48, and 72 h by fixing the cells with 4% paraformaldehyde, Giemsa staining, and observing under a microscope at 100X magnification [20].

Table 1. GenBank accession number and statistics of each *Hungatella* genome downloaded and the genome of the clinical isolate IMSS-269.

Strain	Accession ID	Size (bp)	%GC	Predicted Genes	Predicted Proteins
2789STDY5608850	NZ_CYZE01000001.1	6,979,496	48.9	6117	5925
12489931	NZ_KB850950.1	6,873,024	49.3	5982	5761
AF19-13AC	NZ_QTJW01000001.1	7,421,149	49	6709	6421
AF33-11	NZ_QRQF01000001.1	7,522,004	48.6	6742	6440
AM35-8	NZ_QSHY01000001.1	7,857,595	48.7	7153	6084
ChathewayiLFYP18	NZ_CACRUH010000001.1	7,130,728	49.1	6348	6275
IMSS-269	SRR9298911	6,757,988	48.9	5920	5847
MGYG-HGUT-00032	NZ_CABIXC010000001.1	6,979,496	48.9	6130	5916
OM02-1	NZ_QSVU01000001.1	7,603,944	48.3	6917	6586
TF05-11AC	NZ_QSSQ01000001.1	7,545,577	48.6	6772	6441
TF09-11AC	NZ_QSRE01000001.1	7,708,434	48.8	6847	6600
TM09-12	NZ_QSON01000001.1	7,637,491	48.9	6747	6510
VE202-11	NZ_BAHY01000207.1	7,234,945	49	6757	4910
123Y-2 (*H. hathewayi*)	NZ_WNME01000001.1	6,646,572	48	5917	5708
2789STDY5834916	NZ_CZAZ01000001.1	6,759,576	48	6101	5898
AM39-16AC	NZ_QSGX01000001.1	7,048,369	48.1	6442	6144
AM58-2	NZ_QSDQ01000001.1	7,075,569	48.1	6606	5282
DSM-13479*	NZ_GG668320.1	7,163,884	48.1	6450	5770
MGYG-HGUT-00150	NZ_CABJBJ010000001.1	7,470,188	48	6908	6564
AF31-1	NZ_QVHZ01000001.1	6,510,599	49.2	6003	5818
MGYG-HGUT-01688	NZ_CABLBO010000001.1	5,654,180	50	5186	4942
WAL-18680*	NZ_CP040506.1	5,697,783	50	5185	4954
AF19-21	NZ_QVIA01000001.1	5,056,031	46	4541	4359

*Notates strains that are considered to be type strains.

Antibiotic susceptibility testing was done using the epsilometric method (E-test strip) with vancomycin, metronidazole, linezolid, ciprofloxacin, and levofloxacin (Diagnostic Liofilchem, Italy) [21–24]. Isolates were grown on reduced Casman agar (Becton Dickinson, USA) supplemented with 5% defibrinated sheep blood before adding the E-test strips. Susceptibility to clindamycin and chloramphenicol was done using the plate microdilution method [22,23]. The agar plates were incubated in an anaerobic atmosphere (85% N_2, 5% H_2, and 10% CO_2) at 37°C for 48 h before determining the susceptibility [25]. Minimal inhibitory concentration (MIC) interpretation was based on the European Committee on Antimicrobial Susceptibility (EUCAST) cutoff values for anaerobic bacteria.

Whole genome sequencing and genome assembly

Bacteria were grown in thioglycolate broth (Becton Dickinson, USA) for 48 h in anaerobic conditions (85% N_2, 5% H_2, and 10% CO_2) to obtain a cell pellet from 1 mL of culture. DNA was extracted using DNeasy Mini Kit (Qiagen, Hilden Germany) as previously described [15,26,27]. The quality and yield of genomic DNA was verified by agarose gel electrophoresis and a Nanodrop 2000 (ThermoFischer Scientific, USA). Whole-genome sequencing (WGS) was done using Illumina HiSeq 2000 with PE 150 plus index read (Illumina, San Diego, CA). Adapters and phiX Illumina standards were informatically removed before quality control of the reads was done using FastQC.

The coverage was calculated with the formula $C = LN/G$, where L is the read length, N is the number of reads and G is the genome length [28]. The genome sequences were assembled and annotated as previously described [15,16,26]. Briefly, the reads were assembled using Shovill (https://github.com/tseemann/shovill) with the default settings. Automated gene annotation was done with Prokka v.1.14.5 using default parameters [29].

Genomic isolate identification

Multiple approaches were used to confirm the identity of the isolated strain. The assembled genome was uploaded to OneCodex to perform a taxonomic analysis (https://www.onecodex.com/; version 2018), where every individual sequence is compared against the OneCodex database by alignment using 31 bp *k-mers*. Secondly, the species-level homology was determined based on the depth and coverage of sequencing against all publicly available reference genomes [15,30]. The genome of the clinical isolate was analyzed together with the 22 genome sequences of *H. hathewayi* and one *H. effluvii* available at NCBI – the only two *Hungatella* species available (Table 1). All genomes were analyzed via OneCodex.

Virtual fingerprint

The comparative WGS analyses were extended to include 130 public genomes of bacterial species of clinical importance from related orders (5), and an

additional set to enrich anaerobic species (32); these were combined with the 24 *Hungatella* (including our clinical isolate) to examine their genome relatedness. The 154 genomes were concatenated into a FASTA formatted file for analysis using the Virtual Genome Fingerprints (VGF) VAMPhyRE software (http://bio medbiotec.encb.ipn.mx/VAMPhyRE/). The analysis allowed one mismatch with examination of both strands. VGF uses a collection of 15,264 highly diverse 13-mer *probe* sequences for the virtual hybridization analyses and a genomic distance table is generated by comparing the VGF of each genome. The matrix of distances was built and used to generate a phylogenomic tree using MEGA 10.0 [31]. The tree was edited and annotated with iTOL v4 (Interactive Tree of Life) [32].

16S rRNA gene phylogenetic analysis
The 16S rRNA genes were extracted from 24 *Hungatella* genomes using the gene prediction with Prokka annotation before aligning the sequences with MEGA 10.0 to construct a 16S rRNA phylogenetic tree. The 16S rRNA gene of *Clostridium bolteae* and *Faecalicatena contorta* were also included based on results of the phylogenetic analyses and virtual fingerprint (see above). The tree was edited and annotated with iTOL.

Whole genome distances to delimit species
To assess whether the IMSS-269 clinical isolate can be classified as *H. hathewayi*, we used the *in-silico* DNA–DNA hybridization (DDH) in the Genome-to-Genome Distance Calculator (GGDC) (https://www.dsmz.de/services/online-tools/genome-to-genome-distance-calculator-ggdc) to estimate the similarity between the 24 *Hungatella* genomes. Model "formula 2" was chosen for analysis because it provides an estimation of DDH independent of the genome length and suggests boundaries for taxonomical differences of genus and species. The genome sequence of DSM-13479 was used for comparison of species location because it was the first genome reported as *H. hathewayi* [13]. It was also identified by the Human Microbiome Project as the reference strain for this organism. Genomes with a DDH similarity value >70% are considered to be the same species [33].

The Average Nucleotide Identity (ANI) was used to determine species differentiation. The algorithm first aligned genomes using the ANIm method with MUMmer to align multiple references and multiple query sequences. Matching regions were identified, and the percentage of nucleotide identity of the matching regions were calculated as a metric of similarity.

A ANI percentage threshold for species identity is 95% or greater [34].

Determination of the pan- and core-genome of the 23 H. hathewayi and the H. effluvii reported strains
The ORFs annotated using Prokka (v.1.14.5) for all of the public *Hungatella* genomes (24 total) were used as input for pan-genome analysis using Roary v3.13.0 (https://github.com/sanger-pathogens/Roary) (37). Core genes were considered to be those present in ≥ 99% of the genomes in the comparison [35]. The core genes were extracted and converted to protein sequences for an all-against-all comparison using BLASTP with by setting the sequence identity to 95%. Finally, the "roary_plots.py" script was used to visualize a matrix with the presence and absence of core and accessory genes. The gene diversity estimation was visualize using "create_pan genome_plots" with an R-script.

Informatic analyses of virulence and antimicrobial resistance genes
The virulome and resistome were determined using ABRicate (https://github.com/tseemann/abricate) with the 24 assembled genomes of *Hungatella*, using a percentage of identity of 60% and a coverage of 60% also. The Virulence Factors Database (VFDB) (http://www.mgc.ac.cn/VFs/) was used as the reference database for virulence genes. The Comprehensive Antibiotic Resistance Database (CARD) (https://card.mcmaster.ca/) was used as the reference database for antimicrobial resistance genes.

Results

Phenotypic characteristics of the clinical isolate
The patient-isolate (IMSS-269) was initially identified to be *C. difficile* based on clinically accepted phenotypic methods. While the isolate grew in selective CCFA and Casman blood agar with characteristic colony morphology observed with confirmed *C. difficile* isolates (Figure 1(a,b)), the Gram stain; however, revealed Gram-negative bacilli with a slimy capsule (Figure 1(c)) and a subterminal spore that deform the cell shape (Figure 1(d)). *In vitro* inoculation of sterile culture supernatant onto HeLa cells caused a cytopathic disruption characterized by cell detachment after 24 h (Figure 1(f)). The patient-isolate was susceptible to metronidazole, vancomycin, and clindamycin and resistant to ciprofloxacin (MIC >32 mg/μL) and chloramphenicol (MIC > 8 mg/μL). With these unusual
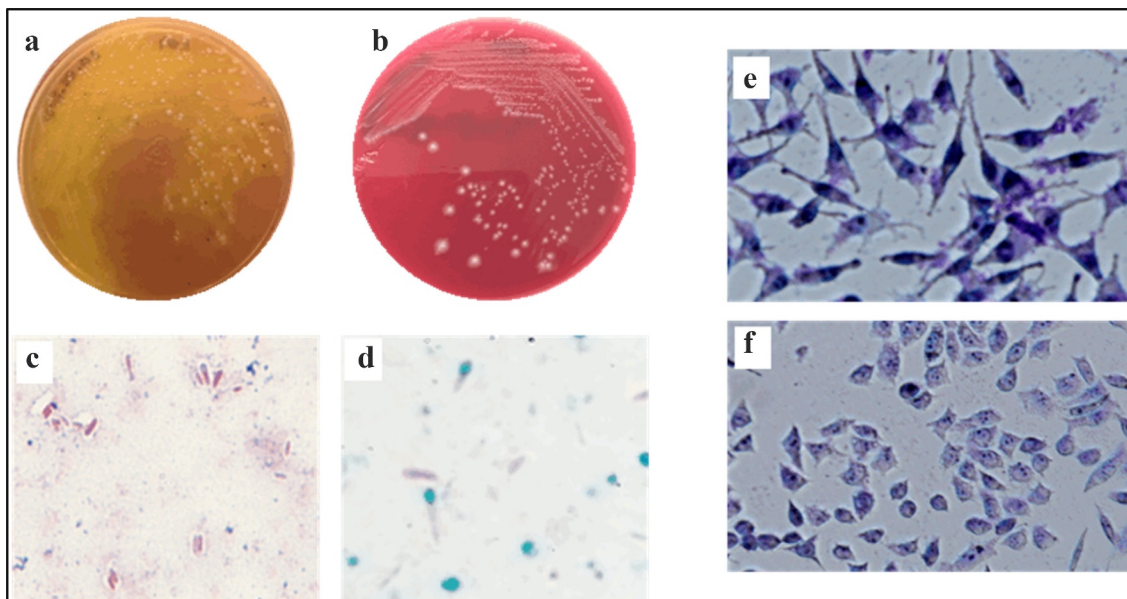
Figure 1. Phenotypic characteristics of the clinical isolate IMSS-269. (a) Colony morphology in CCFA. (b) Colony morphology in Casman blood agar. (c) Gram stain showed Gram-negative bacilli. (d) Subterminal spore. (e) HeLa cells negative control. (f) Cytotoxicity on HeLa cells.

observations, we proceed to use WGS to determine the identification and genomic characteristics to augment the unexpected phenotypic results.

Taxonomic classification of the clinical isolate using whole-genome studies

WGS of the patient-isolate IMSS-269 contained a GC content of 48.9%, with a sequence coverage of 68.6X, 118 contigs, a chromosome size of 6,757,988 bp (6.8 Mbp), and 5,847 predicted coding sequences (CDS) (Table 1). To verify the identity of the patient-isolate we compared that genome with 22 public *H. hathewayi* and *H. effluvii* genomes using OneCodex (Figure 2). This analysis showed that almost 70% of the genome was homologous to *H. hathewayi* but other portions of the genome were homologous to other genera. For the other 21 strains most of the genome was homologous to *H. hathewayi*; although, some of the genomes contained fractions that were homologous to other genera. Two genomes previously named *H. hathewayi* (AF19-21 and AF31-1) had very low homology with the type strain DMS-13479, suggesting they belong to another genus. To further expand this observation, we summarized the genome characteristics of the 23 *H. hathewayi* genomes (Table 1) and found that they varied in key characteristics including size (varies from 5,056,031 to 7,857,595), GC content (ranged from 46 to 50%), and gene content (ranged from 4,541 to 7,153). This heterogeneity in genome features coupled to the genome

alignment diversity suggested diversity beyond that observed in the published genomes and indicated the need to review the classification of these isolates as well as the clinical isolate IMSS-269 before an accurate identification could be made.

These data led us to determine inconsistencies in classification between the phenotypic methods and the comparative genome analysis. Subsequently, we proceeded with multiple methods that used the entire genome to determine the genomic relatedness of *H. hathewayi* and other clinically important organisms. Phylogenomic analysis of the 24 *Hungatella* with 130 public genomes from different species was done using VAMPhyRE to generate a genome fingerprint (Figure 3). The *Hungatella* genomes distributed into multiple clades within the order *Clostridiales*. The majority of the *H. hathewayi* (19/23) genomes clustered in the groups designated as A or B (Figure 3). Two genomes (AF19-21 and AF31-1) clustered with another genus, *F. contorta* and *C. bolteae*, respectively. The type strains (WAL-18680 and MGYG-HGUT-01688) grouped in a cluster separated from the rest of *H. hathewayi* genomes, indicating that these strains were not genomically representative of this species, which agrees with a previous report questioning these to be the type strains [16].

Together these data support previous analyses that found AF19-21 and AF31-1 and the type strains were not representative of this species (Figure 2). To further investigate the taxonomical location, we calculated
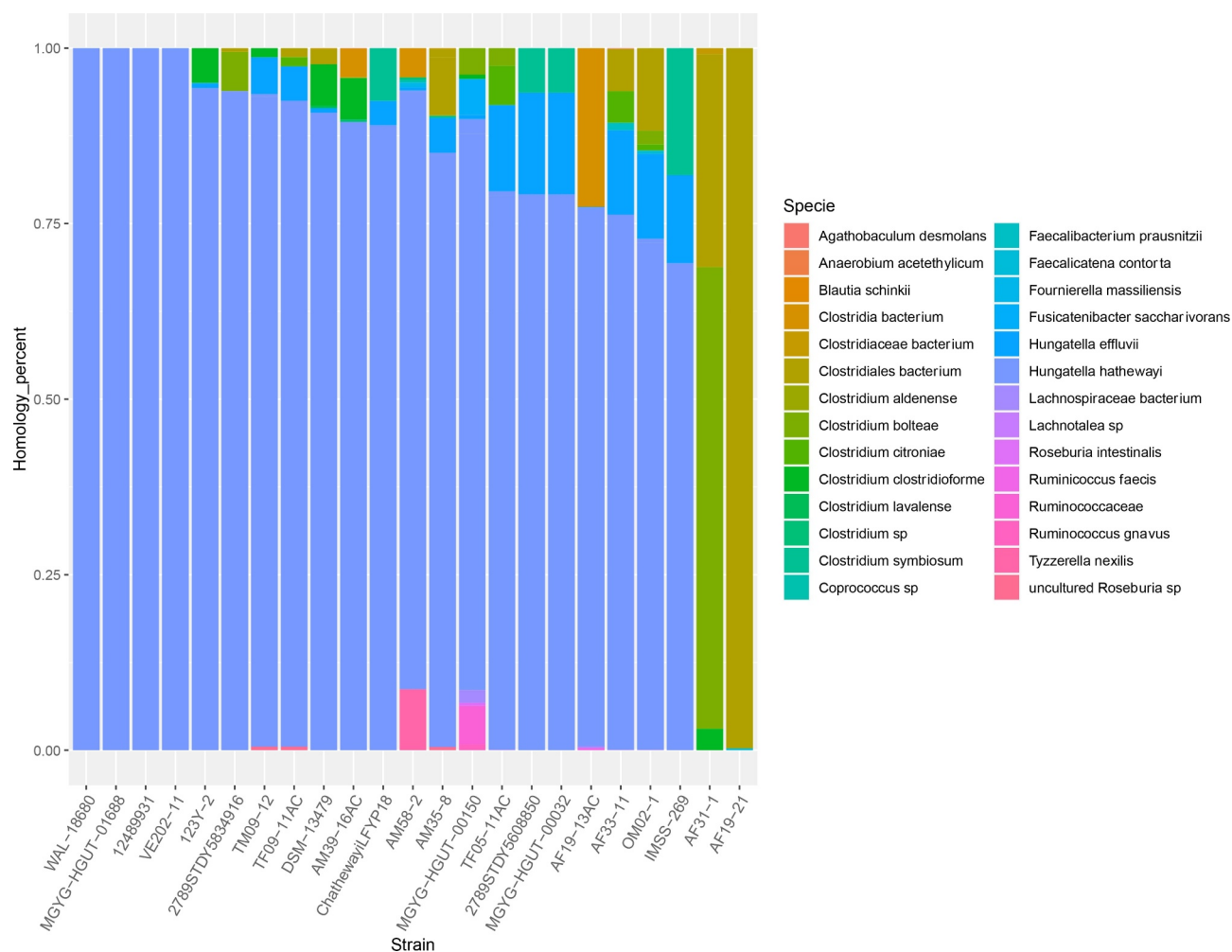
Figure 2. OneCodex taxonomic analyses of 22 *H. hathewayi* genomes and the genome of the patient-isolate IMSS-269. The analysis showed that the IMSS-269 genome was 70% homologous to the other reported *H. hathewayi* genomes.

a phylogenetic tree based on full-length gene sequences of the 16S rRNA gene, a well-accepted method for classification [36]. This analysis showed the formation of clusters concordant with groups B, C, D and E (Figure 3), but genomes from group A were distributed in at least three additional groups with one of them containing the *H. effluvii* genome (Figure 4). These results further suggested many of the isolates likely belong to species that were not aligned with the phenotypic identification. This analysis also confirmed that AF31-1 and AF19-21 belong to a group distant enough from A and B and were closely related to *C. bolteae* and *F. contorta*, respectively; probably belonging to a different genus, as did the strains WAL-1860 (ATCC type strain; Human Microbiome reference strain) and MGYG-HGUT-01688. Importantly, the patient-isolate fell within the group including *H. effluvii*.

To further substantiate these observations, we proceeded to examine the genomes using analytically orthogonal whole genome analysis methods that have shown utility to accurately delimit species assignation to determine the species boundaries (36).

DNA-DNA hybridization (DDH) analyses. To better delimit taxonomy between the genomes, an *in silico* DDH analysis was done using the defined reference strains (Table 2). The analyses found that five genomes had a similarity >70% as compared to the reference strains (123Y-2, 2789STDY5834916, MGYG-HGUT -00150, AM58-2, and AM39-16AC), indicating that these six belonged to *H. hathewayi* species (group B in Figures 3 and 4). Whereas the other genomes (most of the 13 genomes in group A) had a similarity well below 70% with DSM-13479, suggesting they belong to a different species. Taking into account that in the previous analyses the genome of *H. effluvii*
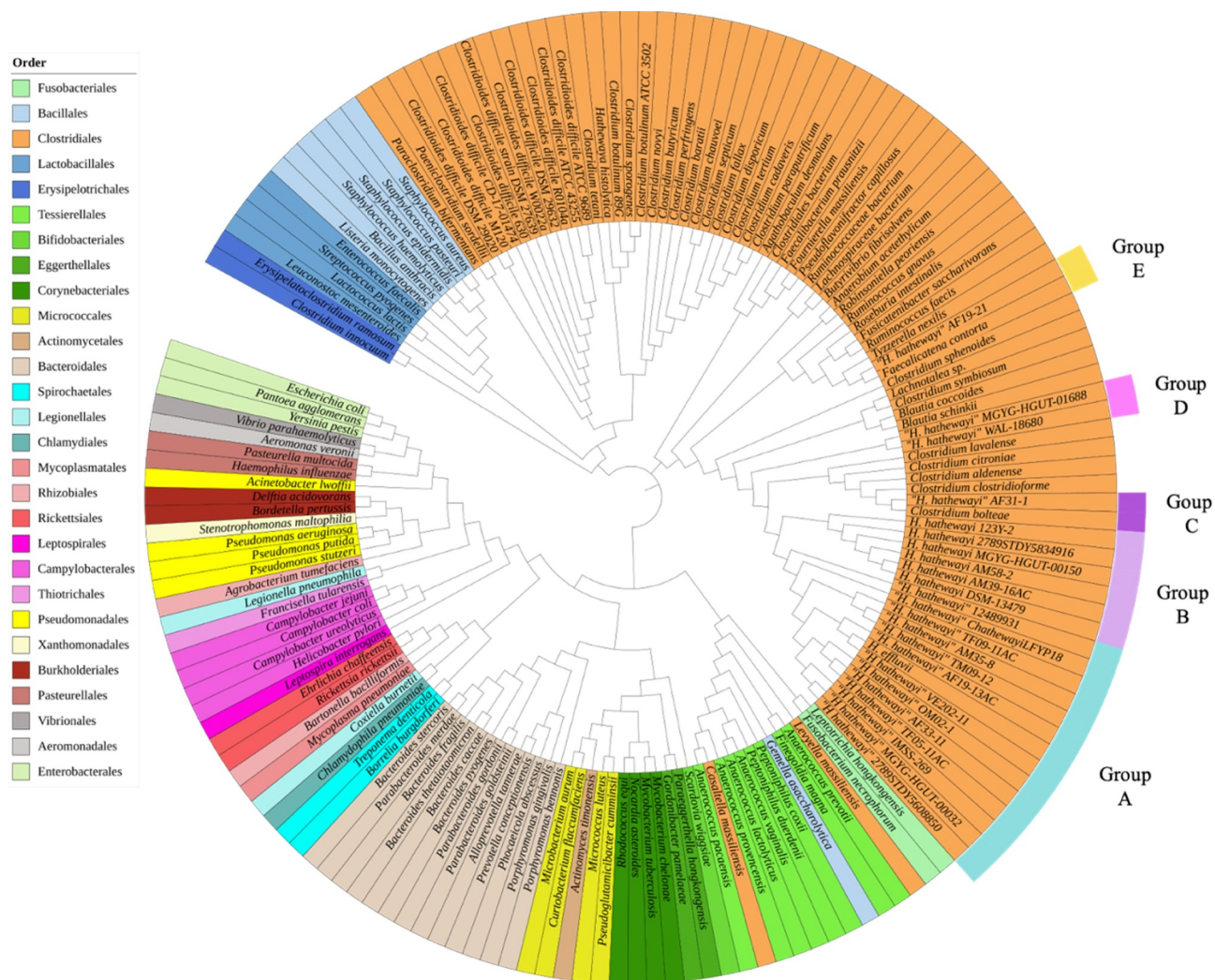
Figure 3. Phylogenomic relationships for *Hungatella* and other clinically relevant bacterial species. Phylogenomic relationships were calculated using virtual genome fingerprints using VAMPhyRE, a k-mer based approach. The tree was constructed used the entire genome of each species to determine the taxonomic localization relative to *H. hathewayi* and *H. effluvii* strains.

clustered with genomes of group A (Figures 3 and 4), we proceeded to conduct a DDH analyses using the genome of *H. effluvii* as the reference (Table 3). The analyses found that 7/13 genomes of group A had a similarity >70% with *H. effluvii* genome, whereas the other six were <70%, indicating several of the genomes from group A may belong to the *H. effluvii* species. The rest of the genomes had a lower similarity, suggesting that they are genomically a different species (Table 3).

Average Nucleotide Identity (ANI) analyses. To further clarify the genomic relatedness among the publicly available *H. hathewayi* genomes, an ANI analysis was done using the 23 *H. hathewayi* genomes along with genomes of other genera from the *Clostridiales* order (Figure 5), particularly those within the large

cluster were *Hungatella* were included in the phylogenomic analyses (Figure 3). The *H. hathewayi* reference strain (DSM-13479) clustered with five other *H. hathewayi* strains (Figure 5, group B) (also group B in Figures 3 and 4 and Tables 2 and 3), the group had an identity of >98% among its members. However, the identity of this group with the other *H. hathewayi* groups was <90%, low enough to be considered a different species. The other *Hungatella* strains (14 genomes) clustered in group A, including *H. effluvii* (also group A in Figures 3, 4 and 5), with an identity within the group of >96%, indicating they belong to the same species. These results suggest strains within this group do not belong to *H. hathewayi* but rather belong to the *H. effluvii* species, including the clinical isolate IMSS-269.
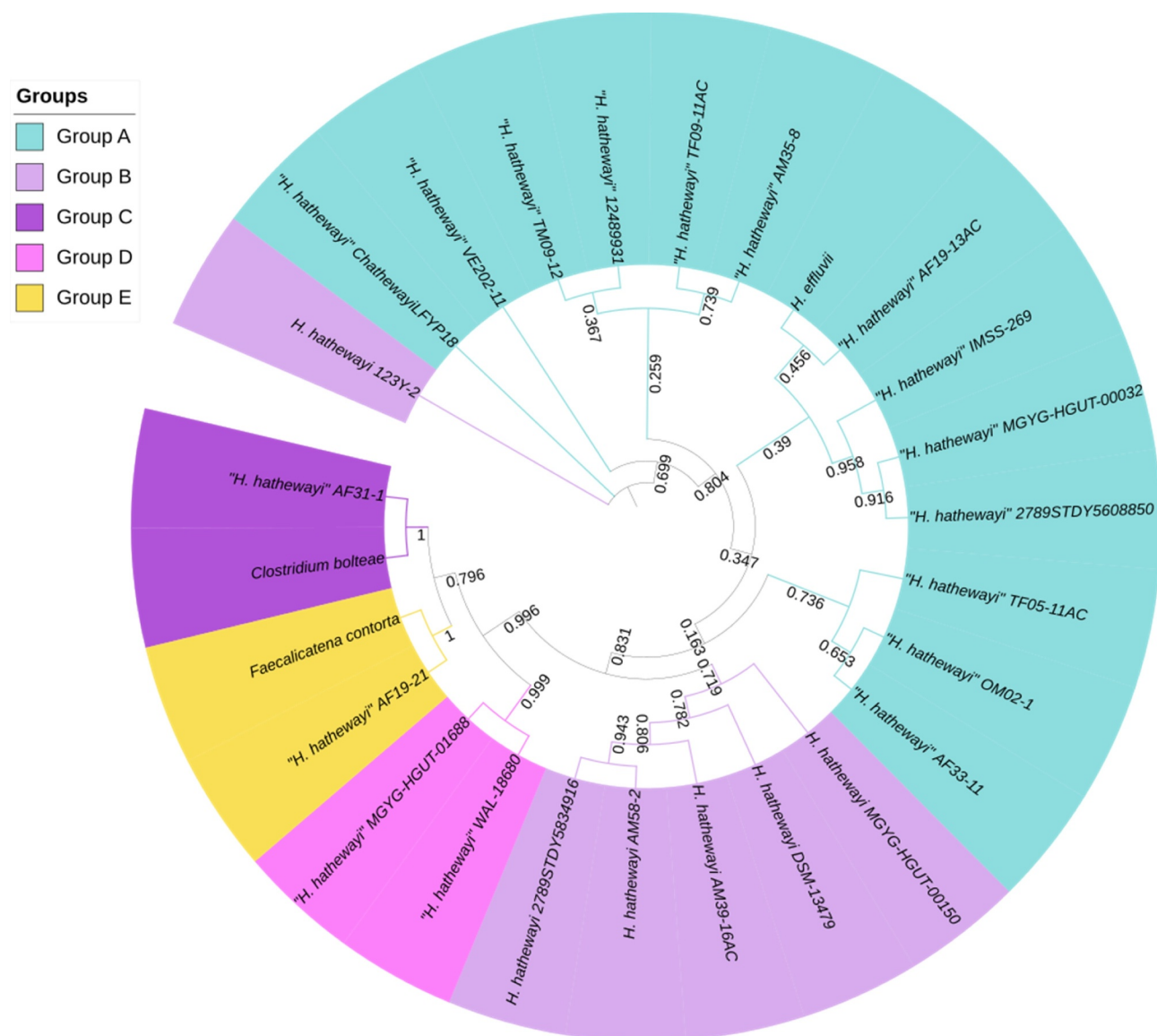
Figure 4. Neighbor joining phylogenetic tree using 16S rRNA of *H. hathewayi*, *H. effluvii*, *F. contorta* and *C. bolteae* strains. Distances between strains revealed several groups suggesting they belong to different species; numbers represent the bootstrap values.

The two other reference strains (WAL-1860 and MGYG-HGUT-01688) (Figure 5, group D) with an identity <90% with the other A and B groups indicating that they belong to a different species, as all previous analyses also indicated. Finally, two strains did not fit within the above groups and one, AF31-1 had an identity of 100% with *C. bolteae* and the other, AF19-21 an identity of 96% with *C. symbiosum*, confirming our observations with the previous analyses. Of note, identity of strains AF19-21 and AF31-1 with genomes from groups A, B and D is between 83 and 94%. A more detailed identity examination between the *Hungatella* group is illustrated in Figure 6 showing that within each group (A, B and D) the identity is ≥96%, whereas identity between groups is ≤ 90%.

Pangenome analyses. In summation of these WGS methods that focus on k-mers and SNPs along with the 16S rRNA gene analysis, we hypothesized that the pangenomes of the 20 confirmed *Hungatella* genomes (groups A and B) would be significantly different and provide additional insight into the genome conservation or diversity between types strains and the patient-isolate relative to all genomes. A pan-genome of 24,619 genes and a core of 246 genes was observed for the 20 *Hungatella* genomes (Figure 7), clearly indicating that different species were included in the analyses. The pan-genomes of this group were determined to be open, indicating that additional genomes are needed to define the entire genomic space of this group. Again, this analysis confirmed that the defined

Table 2. *In silico* DNA–DNA hybridization between *H. hathewayi* DSM-13479 strain as a reference and *H. hathewayi* and *H. effluvii* genomes.

Reference genome	Group	DDH	Model C.I.	Distance	Prob. DDH ≥ 70%	G + C difference	ANI value
2789STDY5608850	A	33.7	[31.3–36.2%]	0.1239	0.43	0.81	0.8836
12489931	A	33.5	[31.1–36%]	0.1247	0.4	1.13	0.8843
AF19-13AC	A	34.2	[31.8–36.7%]	0.1214	0.51	0.91	0.8867
AF33-11	A	34.4	[32–36.9%]	0.1208	0.54	0.45	0.8881
AM35-8	A	36	[33.5–38.5%]	0.1142	0.89	0.6	0.8936
ChathewayiLFYP18	A	33.5	[31–36%]	0.1249	0.39	1.01	0.8836
H. effluvii	A	33.1	[30.6–35.6%]	0.1268	0.34	0.84	0.8817
IMSS-269	A	33.7	[31.3–36.2%]	0.1238	0.43	0.74	0.8837
MGYG-HGUT-00032	A	33.7	[31.3–36.2%]	0.1239	0.43	0.81	0.8843
OM02-1	A	34.7	[32.3–37.2%]	0.1194	0.6	0.13	0.8890
TF05-11AC	A	34.4	[31.9–36.9%]	0.1208	0.54	0.43	0.8879
TF09-11AC	A	35.4	[33–38%]	0.1163	0.76	0.65	0.8914
TM09-12	A	35.4	[33–37.9%]	0.1164	0.75	0.81	0.8915
VE202-11	A	34	[31.6–36.5%]	0.1224	0.48	0.9	0.8861
123Y-2	B	88.7	[86.3–90.8%]	0.0135	95.34	0.17	0.9888
2789STDY5834916	B	89.1	[86.7–91.2%]	0.0131	95.5	0.11	0.9891
AM39-16AC	B	96.8	[95.6–97.7%]	0.0045	97.61	0.05	0.9967
AM58-2	B	92.5	[90.5–94.1%]	0.0094	96.57	0	0.9921
MGYG-HGUT-00150	B	90.8	[88.6–92.6%]	0.0112	96.06	0.14	0.9906
AF31-1	C	34.4	[31–35.9%]	0.1251	0.39	1.06	0.9261
MGYG-HGUT-01688	D	29.9	[27.5–32.4%]	0.1428	0.1	1.89	0.9087
WAL-18680	D	29.9	[27.5–32.4%]	0.1428	0.1	1.89	0.9087
AF19-21	E	30.9	[28.5–33.4%]	0.1373	0.15	2.13	0.9002

Table 3. *In silico* DNA–DNA hybridization between *H. effluvii* as a reference, and *H. hathewayi* genomes.

Reference genome	Group	DDH	Model C.I.	Distance	Prob. DDH ≥ 70%	G + C difference	ANI value
2789STDY5608850	A	64.9	[61.9–67.7%]	0.0436	66.94	0.03	0.9613
12489931	A	67.9	[64.9–70.7%]	0.0391	74.19	0.29	0.9571
AF19-13AC	A	67.2	[64.3–70.1%]	0.04	72.76	0.07	0.9610
AF33-11	A	65	[62–67.8%]	0.0435	67.2	0.39	0.9576
AM35-8	A	67.8	[64.8–70.7%]	0.0392	74.02	0.25	0.9614
ChathewayiLFYP18	A	67.5	[64.5–70.3%]	0.0397	73.33	0.16	0.9610
IMSS-269	A	64.1	[61.1–66.9%]	0.0449	64.77	0.11	0.9563
MGYG-HGUT-00032	A	64.9	[61.9–67.7%]	0.0436	66.94	0.03	0.9571
OM02-1	A	65.2	[62.3–68%]	0.0431	67.83	0.71	0.9574
TF05-11AC	A	65	[62–67.8%]	0.0435	67.22	0.41	0.9576
TF09-11AC	A	67.6	[64.7–70.5%]	0.0395	73.63	0.2	0.9614
TM09-12	A	67.8	[64.8–70.6%]	0.0392	73.96	0.04	0.9615
VE202-11	A	67.3	[64.3–70.1%]	0.04	72.82	0.06	0.9607
123Y-2	B	32.6	[30.2–35.1%]	0.1289	0.29	1.02	0.8799
2789STDY5834916	B	32.7	[30.3–35.2%]	0.1285	0.3	0.96	0.8802
AM39-16AC	B	32.6	[30.2–35.1%]	0.129	0.29	0.9	0.8800
AM58-2	B	32.5	[30.1–35%]	0.1293	0.28	0.84	0.8794
DSM-13479	B	33.1	[30.6–35.6%]	0.1268	0.34	0.84	0.8817
MGYG-HGUT-00150	B	32.6	[30.2–35.1%]	0.1289	0.29	0.99	0.8796
AF31-1	C	19.6	[17.4–22%]	0.2238	0	0.21	0.8384
MGYG-HGUT-01688	D	19.9	[17.7–22.3%]	0.2208	0	1.05	0.8306
WAL-18680	D	19.9	[17.7–22.3%]	0.2208	0	1.05	0.8306
AF19-21	E	20	[17.7–22.4%]	0.2203	0	2.98	0.8451

reference genomes were very different to other *Hungatella* species and that the clinical isolate was also different to these genomes. In total, all genomic analyses determined that the type strains were not genotypically representative of the groups and that the patient-isolate was not *C. difficile*.

Antibiotic resistance and virulence genes. The presence of genes associated with antibiotic resistance was determined using CARD. All strains of group A (*H. effluvii*) presented the aminoglycoside acetyl-transferase AAC (6′)-lad gene, the 23S rRNA methyl-transferase Llma, *vanRA*, and *vatB*. Whereas all strains

from group B (*H. hathewayi*) presented Llma and *poxtA* (Supplementary Table 1).

Virulence genes were examined using VFDB. Two genes, *clpP* and *glf*, were present only in group A strains. Whereas some strains of group A and all from group B contained *cps4I, ureB* and *ureG* (Supplementary Table 2).

Discussion

We report that a clinical isolate obtained from a patient with chronic health concerns was initially identified as
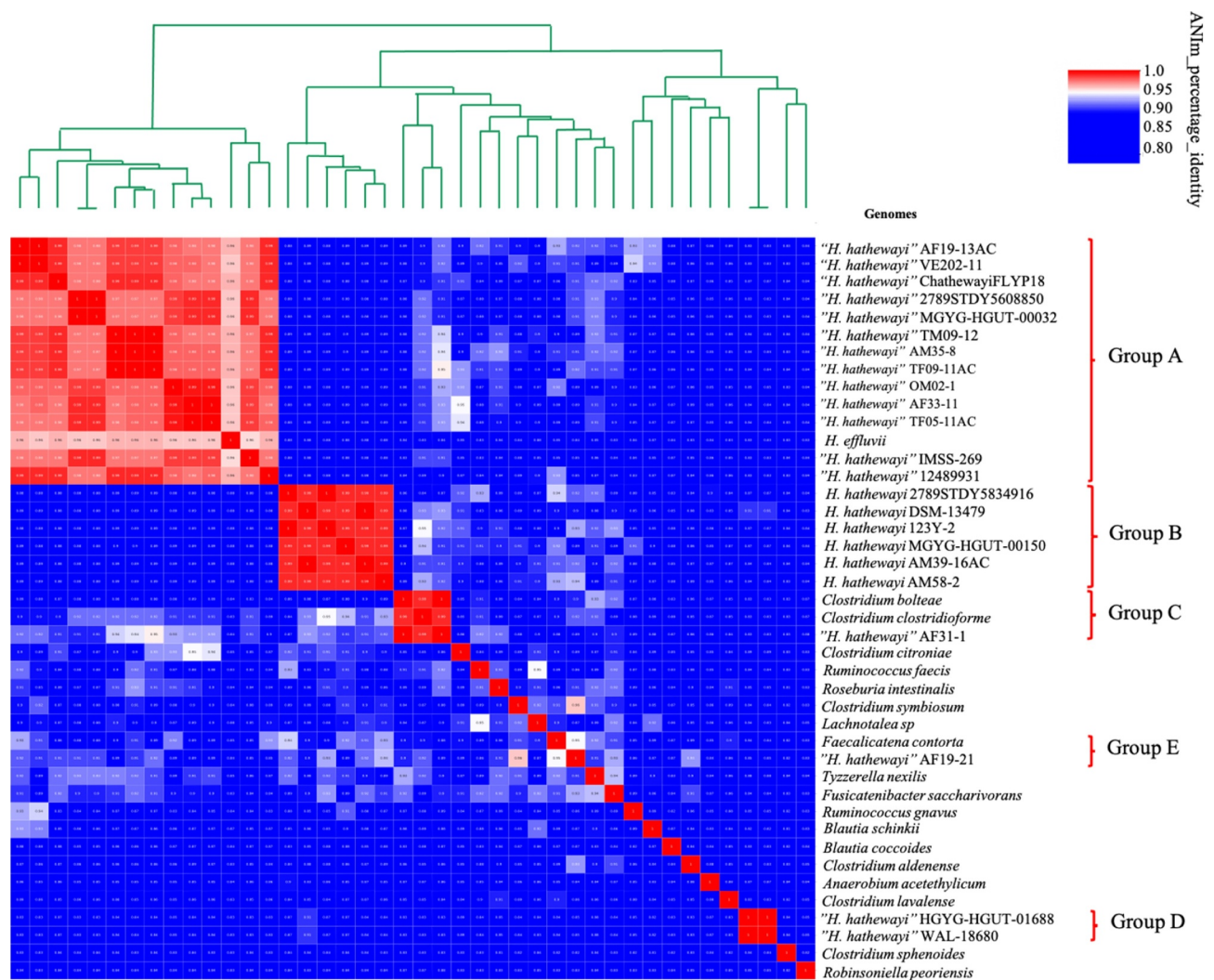
Figure 5. Average nucleotide identity among genomes of *Hungatella* and other members of the order Clostridiales. Two groups of *Hungatella* were formed, group A comprising the actual *H. hathewayi* species and another group with subclusters B and C which represent different species.

C. difficile using biochemical methods. Using WGS and comparative genomics it was later shown that the patient-isolate was not *C. difficile* and was more related to the *Hungatella* genus. The patient-strain was isolated on CCFA, a selective media for *C. difficile*, presented cytotoxic effects on HeLa cells, and was resistant to chloramphenicol and ciprofloxacin but was not ultimately identified to be this organism. The constellation of phenotype observations was in partial alignment with *C. difficile*. However, in total the phenotype characteristics, including spore-forming and biochemical characterization led to the conclusion this isolate did not agree with the taxonomic placement as *C. difficile*. This discord would impact therapeutic decisions in a clinical setting making it important to uncover the identity of the isolate in the 12-year-old patient.

Through the use of WGS and comparative genomics the clinical isolate, we found that this organism may be *Hungatella*. However, we observed distinct genomic differences and large genomic diversity between the published *H. hathewayi* genomes, making this identification more difficult, and suggesting the reference genomes were not representative of the genius or species. Identification of the clinical isolate was confounded due to the genomic diversity of the limited available WGS, which pointed out that type strains did not belong to a single species. The genomic diversity of this understudied organism was extremely large, as indicated by the open pan-genome, the substantial differences in GC content (ranging from 46 to 50%), whereas the genome size ranged from 5,056,031 to 7,857,595. These differences are too large to come from a single species [37].
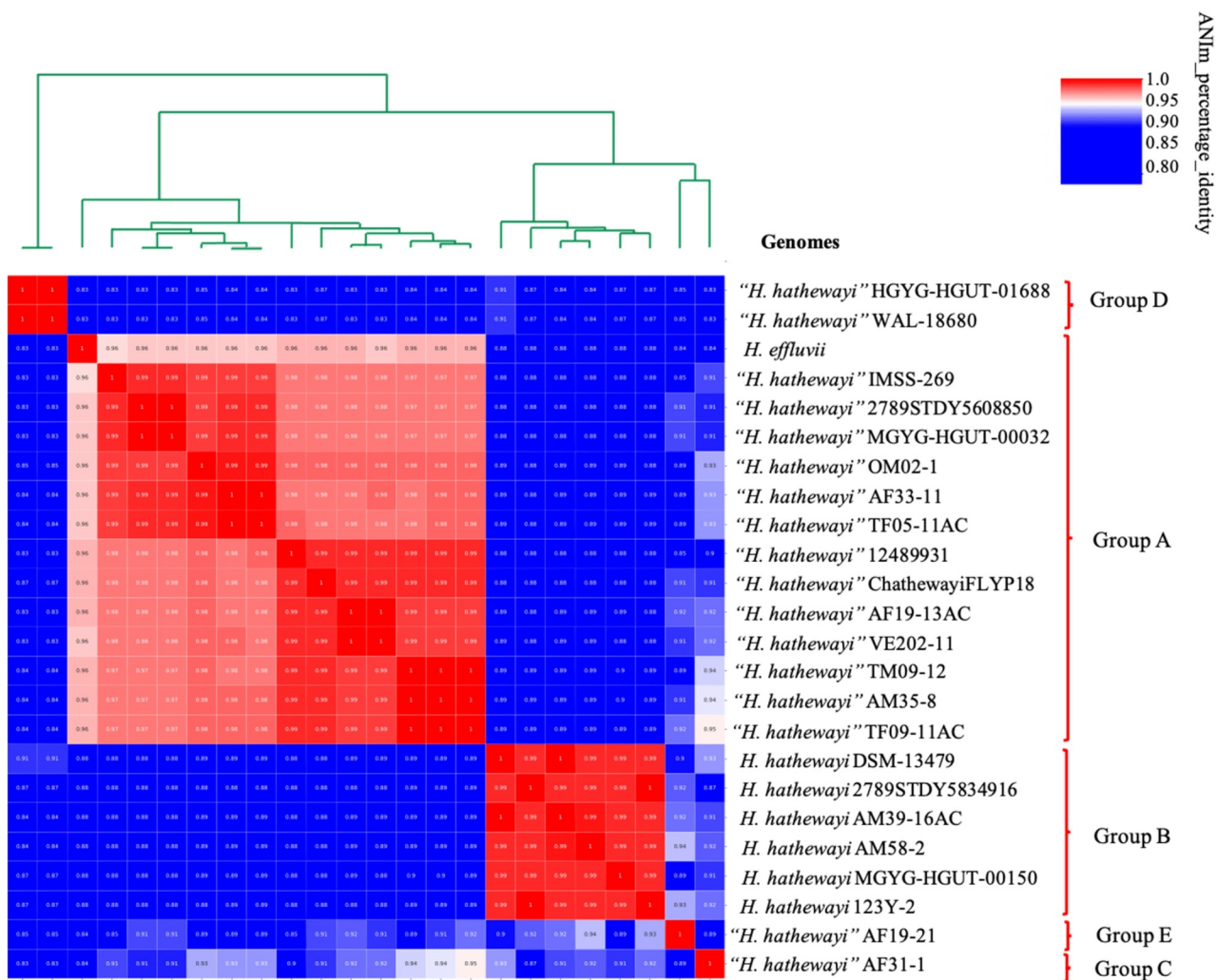
Figure 6. Average nucleotide identity in genomes of *H. hathewayi* and related species. Five groups were formed between the *H. hathewayi* genomes and, the IMSS-269 isolate.

Also, a large genomic difference observed between the organisms specifically led us to suspect that this organism was a chimera or at least not the same species as the references (Figure 2). To fully examine the taxonomic assignment we progressed beyond biochemical characterization to use multiple genetic and genomic methods that are based on different methods for taxonomic differentiation with the underlying hypothesis that genomic comparisons of this unusual isolate would provide a more accurate identification. The initial genomic identification suggested that this isolate may be *H. hathewayi*.

The importance of *H. hathewayi* recently increased because of its association with severe diseases, sporadic cases of bacteremia, fatal septicemia in adult patients, and severe cases of COVID-19 and it is critical that additional work is done to define the organism identification with the accuracy demanded in clinical diagnostics [5,8–12]. Additionally, it was reported that

a decrease in *H. hathewayi* was associated with fatal unruptured intracranial aneurysm, which is likely associated with the reported ability of *H. hathewayi* (strain DSM-13479, representing the actual *H. hathewayi* species according to the present study (Figures 3–4), to be a primary producer of circulating taurine in patients [38]. Interestingly, taurine supplementation reversed progression of unruptured intracranial aneurysm in a mouse model [38]. These examples of diverse disease association with this organism further points out the importance of defining the correct taxonomic assignment that can now be defined using the accuracy of WGS.

H. hathewayi is a species recently described but the number of WGS is relatively small. This restriction of WGS led to verification of the reference genomes and type strains as part of the identification process of this strain. This species was defined by the Human Microbiome Project to be a core organism in human
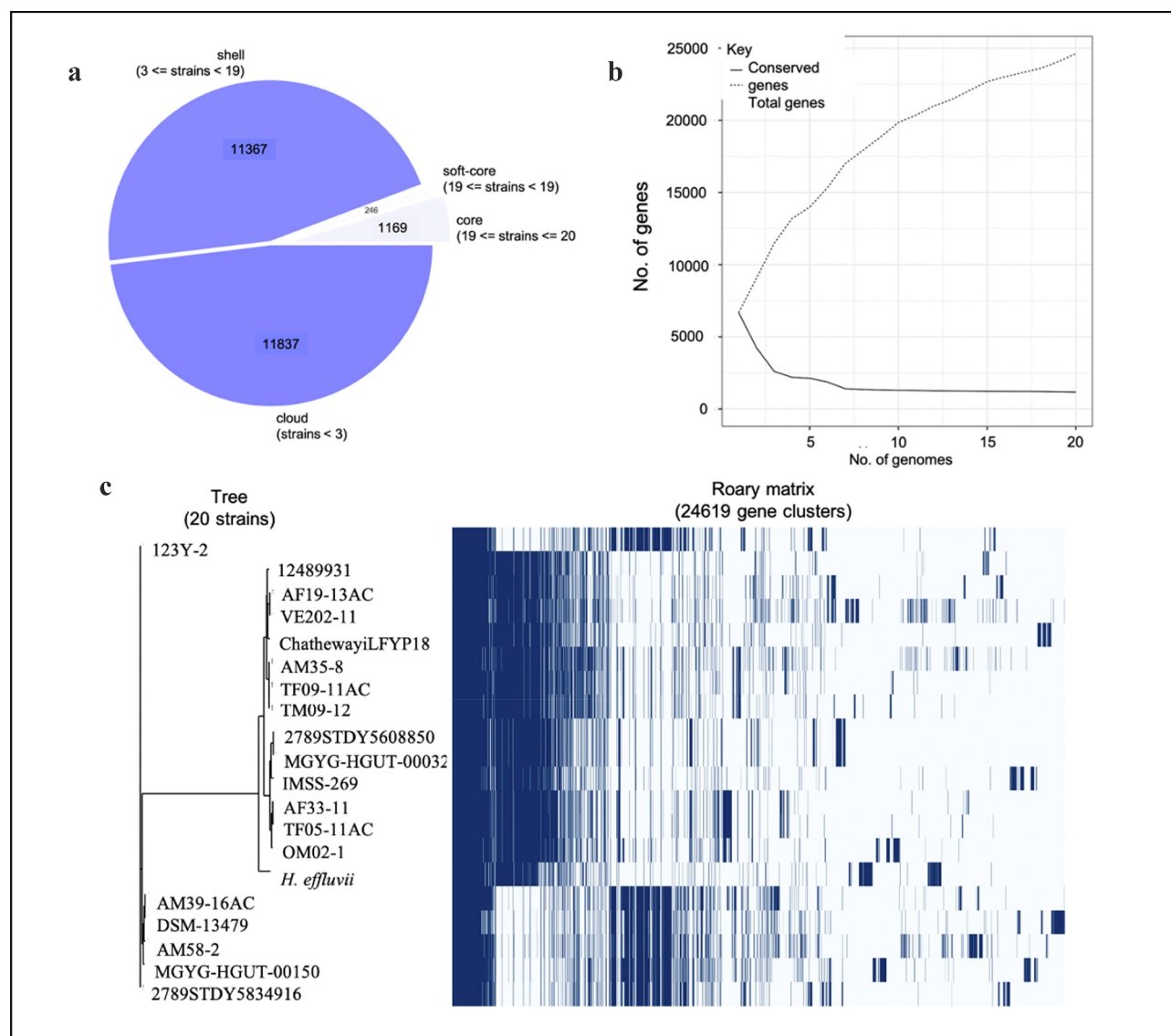
Figure 7. Pan-genome of 24 *Hungatella* genomes. (a) Summary of *Hungatella* pan-genome (39,518 genes). (b) Gene accumulation curve. The dashed line represents the pangenome, and the continuous line represents the conserved homologs genes. The pan-genome shows a core genome stabilization within 5 genomes. (c) This matrix represents the calculated pan-genome for 20 genomes of *Hungatella*. The presence (blue) and absence (white) of core and accessory genes of pan-genome are shown. The phylogenomic tree to the left of pan-genome plot was constructed using the core genome.

metagenomes and a common gut microbiome member [39–43] with strain WAL-18680 being defined as the reference of choice for metagenome studies that identify organisms in the community. Bandoy et al. [16] found that WAL-18680 to be misclassified and is certainly not representative of the species using a comparative genomic approach with all public WGS.

We studied the available *Hungatella* genomes with whole-genome tools aiming to better define the taxonomy of the strains. A whole-genome phylogenetic analyses of *Hungatella* genomes and genomes of other species of clinical interest confirmed *H. hathewayi*

belongs to the Firmicutes phyla and the *Clostridiales* order [13]. Analyses also showed that the reported *Hungatella* genomes grouped in five cluster, already suggesting four genera and two different *Hungatella* species; an observation that was also suggested by our results with the well-accepted phylogenetic analyses of the 16S rRNA gene [13,44]. The general characteristics of the genomes, size, number of genes and GC content also indicated we have more than a single species. In fact, these analyses indicated two of the genomes published as *H. hathewayi* belonged to another genera (*Faecalicatena* and *Clostridium*) and two others,

including the reference strain WAL-18680 [39] were not *H. hathewayi* species, as previously reported [16].

In order to get more accurate measurement of the variation between these genomes, we completed additional genomic analyses at various scales of resolution and bioinformatic approaches. The DDH similarity analyses suggest that genomes with a similarity >70% are considered to be a single species [33]. We successfully confirmed that strains from the group B, as defined by 16S rRNA and phylogenic analyses, were found to be *H. hathewayi* based on their similarity to the DMS-13,479 reference strain [13]. On the other hand, considering that the *H. effluvii* genome consistently clustered within the group A in the diverse analyses performed, we questioned if genomes in this group may actually belong to the *H. effluvii* species. The DDH analyses showed that several of the group A genomes had similarity values indicating that they belonged to the *H. effluvii* species; however, the other genomes of the group presented relatively moderate similarity values (64–68%) that fell short of the accepted threshold (Figure 5). Additional analyses were required to solve this difference in taxonomic location. Therefore, we turned to WGS and increased the scale of resolution of gene content, conservation of the core genome, and allelic determination.

Because of the more accessible platforms for NGS and because of its robustness, ANI has become the gold standard for definition of species and ANI values of 95–96% are considered to be the boundary between species [34,45]. Using this approach, we found that the 23 *Hungatella* genomes separated in four groups. Genomes within group B that included the reference strain DSM-13479 had an identity of >98% (Figure 6). Strains from this group may represent the *H. hathewayi* species, but additional availability of WGS will provide foundation to clearly define specific membership in this genomic cluster. The larger group A cluster had an identity <90% with group B, confirming they belong to different species as determined in this study. All genomes in this group presented an identity >96% with *H. effluvii* confirming their identity, taxonomic location, and indicating that they do not belong to *H. hathewayi* species. ANI identified two subgroups within group A with an identity between them >96%, suggesting they are the same species and may represent different subspecies.

The importance of getting the right taxonomic assignment is further illustrated by the clear differences we found in the content of antibiotic resistance and virulence genes between *H. hathewayi* and *H. effluvii* species, which impact clinical diagnostic capabilities that subsequently influence pathogenesis presentation, and treatment. Our results highlight for the first time that *H. effluvii* may be virulent to humans, an observation that should be considered by clinicians and microbiologist for future diagnostic applications.

Two genomes formed a group with an identity <94% with groups A and B, indicating these represent yet another species; also, these two genomes had the highest GC content, of 50%, and among the smallest genome size, 5.6 Mb. Of note, one of these genomes, WAL-1860, has been suggested as a reference strain for *H. hathewayi* [39], which is not supported by our results or other reports [16]. More genomes within this group are needed to better define species using genomic methods that indicate an open pan-genome.

In conclusion, we present genomic evidence that indicate reported *H. hathewayi* strains belong to different species and that several of the strains reported as *H. hathewayi* belong indeed to *H. effluvii*, including the clinical isolate described in this work. These results may be used as the bases to better identify new *Hungatella* strains, a genus that is recently recognized as part of the gut microbiota but also associated with lethal diseases, such as COVID-19, fatal septicemia and unruptured intracranial aneurysm. A correct taxonomic identification has important implications for appropriate treatment and better understanding of virulence and pathogenesis of the disease-associated pathogen.

Disclosure statement

Funding

Data availability

The data that support the findings of this study are openly available in GenBank and the reference number of each genome is described in Tables 1; 23 of the genomes were already available and the genome of our clinical isolate (strain IMSS-269) was recently uploaded.

ORCID

Laura E. Hernández-Juárez http://orcid.org/0000-0001-6075-2528

Bart C. Weimer http://orcid.org/0000-0002-7471-1978

Javier Torres http://orcid.org/0000-0003-3945-4221

References

[1] Rinninella E, Raoul P, Cintoni M, et al. What is the healthy gut microbiota composition? A changing ecosystem across age, environment, diet, and diseases. Microorganisms. 2019;7(1):14.

[2] Riaz Rajoka MS, Shi J, Mehwish HM, et al. Interaction between diet composition and gut microbiota and its impact on gastrointestinal tract health. Elsevier B.V Food Sci Hum Wellness. 2017;6(3):121–130.

[3] Durbán A, Abellán JJ, Jiménez-Hernández N, et al. Assessing gut microbial diversity from feces and rectal mucosa. Microb Ecol. 2011;61(1):123–133.

[4] Eckburg PB, Bik EM, Bernstein CN, et al. Microbiology: diversity of the human intestinal microbial flora. Science. 2005;308(5728):1635–1638.

[5] Elsayed S, Zhang K. Human infection caused by clostridium hathewayi. Emerg Infect Dis. 2004;10 (11):1950–1952.

[6] Steer T, Collins MD, Gibson GR, et al. Clostridium hathewayi sp. nov., from human faeces. Syst Appl Microbiol. 2001;24(3):353–357.

[7] Liang Q, Chiu J, Chen Y, et al. Fecal bacteria act as novel biomarkers for noninvasive diagnosis of colorectal cancer. Clin Cancer Res. 2017;23(8):2061–2070.

[8] Dababneh AS, Nagpal A, Palraj BRV, et al. Clostridium hathewayi bacteraemia and surgical site infection after uterine myomectomy. BMJ Case Rep. 2014;2014. 2013–2015. DOI:10.1136/bcr-2013-009322

[9] Linscott AJ, Flamholtz RB, Shukla D, et al. Fatal septicemia due to Clostridium hathewayi and Campylobacter hominis. Anaerobe. 2005;11(1–2):97–98.

[10] Randazzo A, Kornreich A, Lissoir B. A Clostridium hathewayi isolate in blood culture of a patient with an acute appendicitis. Anaerobe. 2015;35:44–47.

[11] Woo PCY, Lau SKP, Woo GKS, et al. Bacteremia due to Clostridium hathewayi in a patient with acute appendicitis. J Clin Microbiol. 2004;42(12):5947–5949.

[12] Zuo T, Zhang F, Lui GCY, et al. Alterations in gut microbiota of patients with COVID-19 during time of hospitalization. Gastroenterology. 2020. DOI:10.1053/j.gastro.2020.05.048

[13] Kaur S, Yawar M, Kumar PA, et al. Hungatella effluvii gen. nov., sp. nov., an obligately anaerobic bacterium isolated from an effluent treatment plant, and reclassification of Clostridium hathewayi as Hungatella hathewayi gen. nov., comb. nov. Int J Syst Evol Microbiol. 2014;64(PART 3):710–718.

[14] Lawton SJ, Weis AM, Byrne BA, et al. Comparative analysis of Campylobacter isolates from wild birds and chickens using MALDI-TOF MS, biochemical testing, and DNA sequencing. J Vet Diagn Invest. 2018;30 (3):354–361.

[15] Bandoy DDR, Weimer BC. Biological machine learning combined with campylobacter population genomics reveals virulence gene allelic variants cause disease. Microorganisms. 2020;8(4):549.

[16] Bandoy DD, Huang BC, Weimer B (2019). Misclassification of a whole genome sequence reference defined by the human microbiome project: a detrimental carryover effect to microbiome studies. Misclassification of a Whole Genome Sequence Reference Defined by the Human Microbiome Project: A Detrimental Carryover Effect to Microbiome Studies, 19000489. 10.1101/19000489.

[17] Balloux F, Brønstad Brynildsrud O, Van Dorp L, et al. From theory to practice: translating whole-genome sequencing (WGS) into the clinic. Trends Microbiol. 2018;26(12):1035–1048. Elsevier Ltd.

[18] Camorlinga-Ponce M, Gamboa M, Barragan JJ, et al. Epidemiological aspects of Clostridium difficile in a pediatric hospital and its role in diarrheal disease. European J Clin Microbiol. 1987;6(5):542–546.

[19] Torres J, Camorlinga-Ponce M, Muñoz O. Sensitivity in culture of epithelial cells from rhesus monkey kidney and human colon carcinoma to toxins A and B from Clostridium difficile. Toxicon. 1992;30(4):419–426.

[20] Torres López J, Camorlinga Ponce M, Muñoz O. Neutralization of cytotoxic activity of Clostridium difficile with fecal flora. Archivos de Investigación Médica. 1987;18(4):315–317.

[21] Nagy E, Boyanova L, Justesen US. How to isolate, identify and determine antimicrobial susceptibility of anaerobic bacteria in routine laboratories. Elsevier B.V. Clin Microbiol Infect. 2018;24(11):1139–1148.

[22] Schaumann R, Ackermann G, Pless B, et al. In vitro activities of fourteen antimicrobial agents against obligately anaerobic bacteria. Int J Antimicrob Agents. 2000;16(3):225–232.

[23] Spigaglia P, Barbanti F, Morandi M, et al. Diagnostic testing for Clostridium difficile in Italian microbiological laboratories. Anaerobe. 2016;37:29–33.

[24] Tijerina-Rodríguez L, Villarreal-Treviño L, Baines SD, et al. High sporulation and overexpression of virulence factors in biofilms and reduced susceptibility to vancomycin and linezolid in recurrent Clostridium [Clostridioides] difficile infection isolates. PLoS ONE. 2019;14:7.

[25] Ramlachan N, Anderson RC, Andrews K, et al. Characterization of an antibiotic resistant Clostridium hathewayi strain from a continuous-flow exclusion chemostat culture derived from the cecal contents of a feral pig. Anaerobe. 2007;13(3–4):153–160.

[26] Arabyan N, Park D, Foutouhi S, et al. Salmonella degrades the host glycocalyx leading to altered infection and glycan remodeling. Sci Rep. 2016;6 (1):29525.

[27] Campbell E, Gerst M, Huang BC, et al. Draft genome sequence of bacillus velezensis CE2, which genetically encodes a novel multicomponent lantibiotic. Microbiol Resour ments. 2019;8:3.

[28] Draper JL, Hansen LM, Bernick DL, et al. Fallacy of the unique genome: sequence diversity within single Helicobacter pylori strains. MBio. 2017;8:1.

[29] Seemann T. Genome analysis Prokka: rapid prokaryotic genome annotation. BMC Bioinformatics. 2014;30 (14):2068–2069.

[30] Randle-Boggis RJ, Helgason T, Sapp M, et al. Evaluating techniques for metagenome annotation using simulated sequence data. FEMS Microbiol Ecol. 2016;92:95.

[31] Kumar S, Stecher G, Li M, et al. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. n.d.;35:1547–1549.

[32] Letunic I, Bork P. Interactive tree of life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47:W1.

[33] Meier-Kolthoff JP, Auch AF, Klenk H-P, et al. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics. 2013;14(1):60.

[34] Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 2009;106(45):19126–19131.

[35] Page AJ, Cummins CA, Hunt M, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31(22):3691–3693.

[36] Woese CR, Fox GE, Zablen L, et al. Conservation of primary structure in 16S ribosomal RNA. Nature. 1975;254(5495):83–86.

[37] Agashe D, Shankar N. The evolution of bacterial DNA base composition. J Exp Zool Part B. 2014;322(7):517–528.

[38] Li H, Xu H, Li Y, et al. Alterations of gut microbiota contribute to the progression of unruptured intracranial aneurysms. Nat Commun. 2020;11 (1):1–15.

[39] Almeida A, Mitchell AL, Boland M, et al. A new genomic blueprint of the human gut microbiota. Nature. 2019;568(7753):499–504.

[40] Aron-Wisnewsky J, Prifti E, Belda E, et al. Major microbiota dysbiosis in severe obesity: fate after bariatric surgery. Gut. 2019;68(1):70–82.

[41] Fontana A, Panebianco C, Picchianti-Diamanti A, et al. Gut microbiota profiles differ among individuals depending on their region of origin: an Italian pilot study. Int J Environ Res Public Health. 2019;16:21.

[42] Ohara T. Identification of the microbial diversity after fecal microbiota transplantation therapy for chronic intractable constipation using 16s rRNA amplicon sequencing. PLoS ONE. 2019;14(3):1–12.

[43] Schaubeck M, Clavel T, Calasan J, et al. Dysbiotic gut microbiota causes transmissible Crohn's disease-like ileitis independent of failure in antimicrobial defence. Gut. 2016;65(2):225–237.

[44] Bergey's Manual of Systematics of Archaea and Bacteria. Whitman WB, Rainey F, Kämpfer P, et al. editors. Bergey's manual of systematics of Archaea and Bacteria. Hungatella. John Wiley & Sons, NJ; 2015. pp 1–11. DOI:10.1002/9781118960608

[45] Konstantinidis KT, Ramette A, Tiedje JM. The bacterial species definition in the genomic era. Philos Trans R Soc B. 2006;361(1475):1929–1940.