frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 24 November 2025
DOI 10.3389/fpubh.2025.1651032

OPEN ACCESS

EDITED BY
Marc Jean Struelens,
Université libre de Bruxelles, Belgium

REVIEWED BY
Pedro Xavier-Elsas,
Federal University of Rio de Janeiro, Brazil
Enzo Guerrero-Araya,
University of Nottingham Ningbo China, China
Ana Rafaela Kruemmel,
Centers for Disease Control and Prevention (CDC), United States

*CORRESPONDENCE
Alexander J. Diaz
✉ alexander.diaz@health.ny.gov

RECEIVED 20 June 2025
ACCEPTED 31 October 2025
PUBLISHED 24 November 2025

CITATION
Diaz AJ, Centurioni DA, Lasek-Nesselquist E, Lapierre P, Egan CT and Perry MJ (2025) Whole genome sequencing of neurotoxin-producing *Clostridium* species in New York state to bolster epidemiological investigations and reveal patterns of diversity and distribution.
Front. Public Health 13:1651032.
doi: 10.3389/fpubh.2025.1651032

Whole genome sequencing of neurotoxin-producing *Clostridium* species in New York state to bolster epidemiological investigations and reveal patterns of diversity and distribution

Alexander J. Diaz[1]*, Dominick A. Centurioni[1], Erica Lasek-Nesselquist[1,2], Pascal Lapierre[1], Christina T. Egan[1] and Michael J. Perry[1]

[1]Wadsworth Center, New York State Department of Health, Albany, NY, United States, [2]Department Biomedical Sciences, University at Albany, Albany, NY, United States

Clostridia that produce neurotoxins are highly relevant organisms to public health. While cases of botulism [caused by *C. botulinum* and other organisms that produce botulinum neurotoxin (BoNT)] are rare, the severity of this disease necessitates robust epidemiologic surveillance to promptly identify and mitigate outbreaks. Next generation sequencing (NGS) can provide additional support to these investigations through single nucleotide polymorphism (SNP)-based analysis, phylogenetic reconstruction, toxin subtyping, and structural analysis. Until recently, testing for this disease was restricted to traditional culture or molecular methods such as polymerase chain reaction (PCR) to detect *bont* genes, while mouse bioassay and endopeptidase-mass spectrometry (Endopep-MS) methods confirmed the presence of enzymatically active toxin. The New York State Department of Health (NYSDOH) Wadsworth Center Biodefense Laboratory performed a retrospective whole genome sequence (WGS) analysis of approximately 240 *Clostridium* spp. isolates from the past 40 years to supplement traditional test results and further characterize these organisms. Genomic analyses identified seven BoNT serotypes/serotype combinations, including A4(B5), A5(B2'), and B5F2 that were uncharacteristic of samples typically received. Additionally, SNP-based analysis and *de novo* genome assemblies retrospectively validated several epidemiology links or differentiated samples previously tested with only traditional methods. Our work highlights the clinical utility of supplementing conventional data with NGS to further characterize BoNT-producing organisms and underscores the importance of incorporating WGS into laboratory workflows to support epidemiologic investigations. However, several obstacles still exist which may prevent implementation. These include the expertise needed to execute bioinformatic analyses and interpret the resulting data, a lack of standardized bioinformatic workflows, and difficulty in determining SNP-based thresholds to identify linked samples without incorporation of additional data and analyses. Supplementing or replacing short-read sequencing with long-read sequencing (LRS) and the use of metagenomic or capture-based enrichment for analysis of primary specimens could increase the leverage obtained from WGS in epidemiological investigations.

KEYWORDS

botulism, epidemiology, botulinum neurotoxin, whole genome sequencing, bioinformatics, *Clostridium*, public health

1 Introduction

Clostridium species are gram-positive bacteria with the ability to form endospores when subjected to sub-optimal conditions. This results in distinctive central, terminal, or subterminal swellings that allow the organism to persist after exposure to adverse conditions (1). These organisms can be classified into phylogenetically distinct groups/species through various methods such as Amplified Fragment Length Polymorphism (AFLP) analysis, DNA–DNA hybridization, and 16 s rDNA and multigene phylogenies (2–4). *Clostridium* species including *C. botulinum* (Group II), *C. parabotulinum* (Group I), *C. sporogenes* (Group I), *C. novyi sensu lato* (Group III), *C. argentinense* (Group IV), *C. baratii*, and *C. butyricum* produce a potent botulinum neurotoxin (BoNT), the causative agent of botulism (5). Botulism is a relatively rare illness - confirmed in only 243 individuals in the United States for 2021, six of whom were located in New York State (6). Because *Clostridium* spp. can exist under environmentally diverse conditions, such as those found in soils, marine sediments, and the intestinal tracts of animals, the possibility for food contamination is ever present, and often occurs due to the mishandling or under-processing of foods while home canning (7).

BoNTs are encoded by a *bont* toxin gene which, is part of either an *ha*+ or *orfx*+ neurotoxin gene cluster. These gene clusters are located in several genomic or extrachromosomal locations, based upon the strain and toxin serotype (8), and include genes that encode toxin-associated proteins and a nontoxic, non-hemagglutinin protein as well (3).

BoNTs can be classified into seven serotypes: BoNT/A-/G (9). Of these, BoNT/A, /B, /E, and /F most often affect humans while BoNT/C and /D often affect animals (10). BoNT/G-producing isolates have been recovered from clinical specimens but have not been definitively identified as the cause of illness (11). The different toxin serotypes can be further categorized by the soluble N-ethylmaleimide-sensitive factor attachment protein receptor (SNARE) they cleave. BoNT/B, /D, /F, and /G cleave the vesicle-associated membrane protein (VAMP) SNARE, BoNT/A and /E cleave synaptosomal-associated Protein 25 (SNAP25) and BoNT/C cleaves both SNAP25 and syntaxin (12).

Currently, serotypes are divided into more than 40 subtypes based on amino acid variation (13). Novel subtypes typically display at least 2.6% amino acid variation from presently known subtypes and should form monophyletic groups (3, 14). However, recent studies have shown that inter-subtype variation may not meet these criteria. For example, amino acid variation between serotype B subtypes ranges from 1.6%–7.1%. Some of these subtypes are chimeric or products of inter-subtype recombination and can be found in different genomic backgrounds (3, 15–18), demonstrating the role of horizontal gene transfer and reticulate evolution in shaping Clostridial genomes.

BoNTs and the *Clostridium* spp. organisms that produce them are considered Tier 1 select agents and toxins by the United States Centers for Disease Control and Prevention (CDC) and United States Department of Agriculture (USDA) Federal Select Agent Program, as they "have the potential to pose a severe threat to public, animal or plant health, or to animal or plant products" (19). Toxin activity leads to the acute, symmetric, descending, flaccid paralysis that is characteristic of botulism (20). While treatment with the antitoxin may halt disease progression and prevent death (21), it does nothing to reverse paralysis which often leads to a long recovery with supportive care (22).

Laboratory investigation of botulism cases includes isolation, identification, and characterization of botulinum neurotoxin-producing organisms from clinical specimens and environmental samples. The characterization provided through these investigations is essential considering the potential outcomes of BoNT intoxication, the effect of BoNT diversity on treatment options, and the potential for the deliberate misuse of BoNT. This information assists in determining the source of the exposure and assessing the risk to other individuals in epidemiologically confirmed cases of botulism analysis. Therefore, further characterization of the samples is required due to the impact of BoNT on public health.

Samples submitted to the New York State Department of Health (NYSDOH) for *Clostridium* spp. testing fall into several categories including clinical, clinical surveillance, animal, and environmental. Clinical specimens are those obtained from patients, often including stool and serum, and are used to confirm clinical cases (23). Clinical surveillance samples are those which are believed to have been the cause of illness and often include food or consumer product samples. These are used to determine the source of contamination, conduct outbreak tracing, and assess the risk to the general population. Animal specimens are often collected from carcasses when the cause of death is unknown, but botulism is suspected. These samples are used for sentinel surveillance to detect spillover of BoNT producing *Clostridium* spp. from natural reservoirs into areas which may affect people or livestock. Environmental samples are not directly clinically related but are often sent to the laboratory to assist in epidemiologic outbreak investigations and may include hay, leafy vegetation, soil, dust, and samples collected by law enforcement.

Whole genome sequencing (WGS) using next-generation sequencing (NGS) technologies can assist with investigations and support epidemiological inquiries by providing the resolution necessary to discriminate among closely related isolates. Genomic analyses of *Clostridium* spp. isolates have been performed to compare clinical and environmental isolates related to cases of infant botulism (24), foodborne outbreaks involving nacho cheese (25) or home-canned peas (26), and even wound botulism related to injection drug use (27). Frequently, genomic and phylogenetic analyses strengthen the epidemiological link between patient specimens and outbreak sources. Similar data have been generated for NYSDOH culture collection isolates to retrospectively identify or investigate outbreaks. In this study, the NYSDOH Biodefense laboratory has sequenced approximately 240 *Clostridium* spp. isolates to develop a workflow for the classification of botulinum toxin-producing isolates. Retrospective analysis of this culture collection has identified some rarely observed strains, revealed

potentially linked cases, and provided a framework for future analyses that assist in tracing the source of outbreaks in epidemiological investigations.

2 Materials and methods

No single method is adequate when testing for *Clostridium* spp. Complementary methods must be employed to obtain clinical results as quickly as possible and provide an accurate picture of the situation for diagnostic/epidemiologic concerns. Our laboratory has developed a workflow to optimize this process (Figure 1). By screening suspect botulism samples using an endopeptidase mass spectrometry-based assay (Endopep-MS) (28), real-time polymerase chain reaction (rtPCR) (29), and culture (30), positive samples are identified quickly with a high degree of certainty and subsequently sequenced.

Suspect botulism samples tested by our laboratory are processed to generate input material for rtPCR, culture, and Endopep-MS. Samples are further processed to ensure isolation if these conventional methods identify a positive sample. Once isolated, these samples can be batched into groups of 15 to 20 positive specimens to decrease sequencing costs.

2.1 Sample processing

For the current study, primary clinical specimens, animal specimens, foods, and other environmental samples including hay, grasses, and leafy vegetation were cut into sections, if necessary, then transferred to filtered stomacher bags. Gelatin diluent buffer (GDB) was added to samples at roughly a 1:1 ratio (g/mL).

As a result of reduced peristalsis associated with botulism, it is often difficult to obtain large volumes of stool for testing. Minimal dilution of these samples is performed, when necessary, to avoid diluting BoNT or organism deoxyribonucleic acid (DNA) below the assay limit of detection but still provide enough volume for testing. Non-pipettable samples were transferred to stomacher bags,

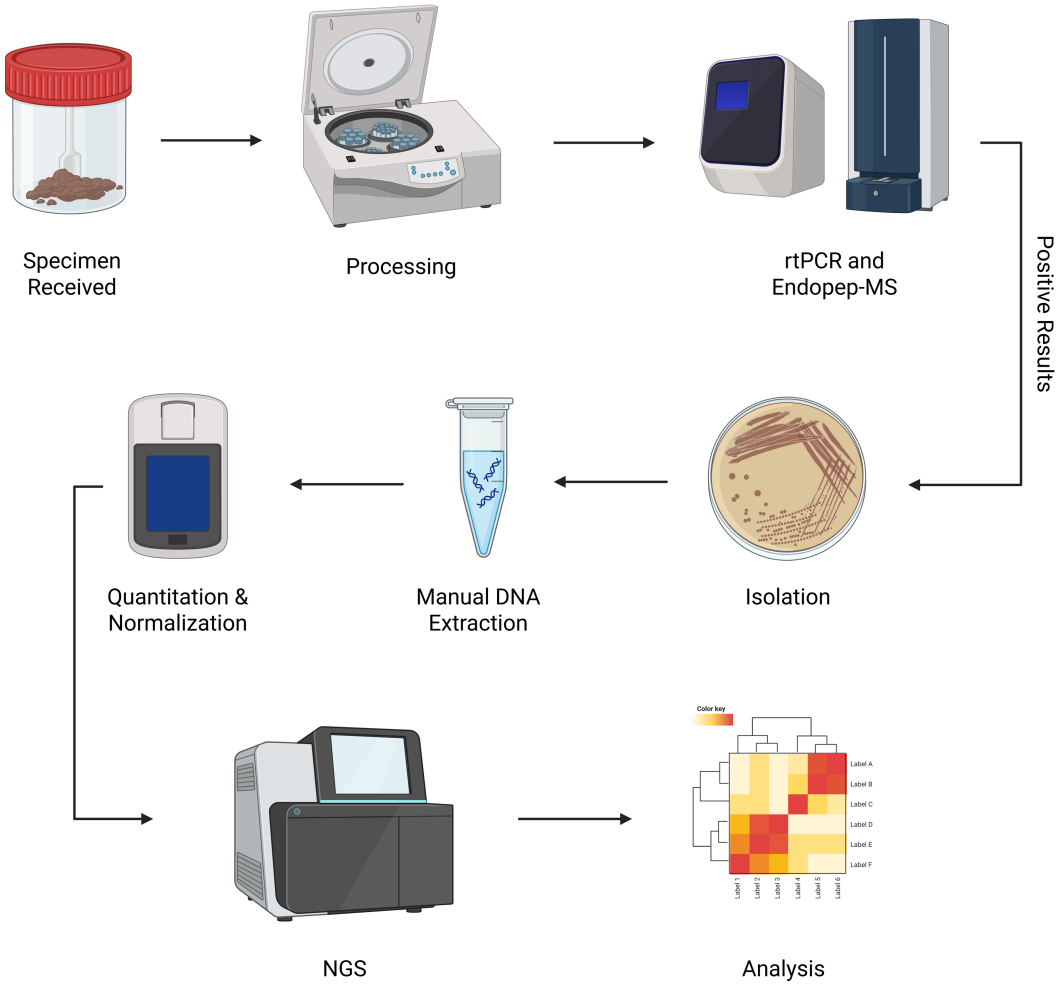

FIGURE 1
Clostridium spp. WGS workflow. Once laboratory testing for confirmation of botulism is approved by epidemiologic investigators, primary specimens received by the NYSDOH Wadsworth Center Biodefense Laboratory are processed and screened for toxin genes using rtPCR. Enzymatic activity is confirmed using Endopep-MS. Isolation of toxin-producing organisms and subsequent WGS is performed to further characterize isolates. Created in BioRender. Centurioni, D. (2025) https://BioRender.com/j4evmmp

homogenized in GDB (2 min, 200 rpm), and the resulting filtrate was used for downstream testing and culture.

Honey samples were processed by diluting 10 g aliquots in 25 mL of phosphate buffered saline. Next, samples were transferred to Nalgene™ Oak Ridge tubes and centrifuged (30 min, 12,000 rpm, 4 °C) using a sealed, fixed angle rotor. Aliquots of the supernatant were retained, and the pellets were resuspended in GDB for use in testing.

Soil samples were prepared by transferring 1 gram of soil to 10 mL of Tryptone Peptone Glucose Yeast Extract (TPGY) broth for enrichment. These samples were incubated anaerobically at 35 °C for 2 to 4 days before testing was performed.

2.2 Culture

Enrichment and subsequent isolation of BoNT-producing organisms from primary clinical specimens (excluding serum specimens) and environmental samples was performed before sequencing was attempted. All culture was performed under anaerobic conditions. Approximately 50 µL of processed primary specimen was used to inoculate solid media including Egg Yolk Agar (EYA), Trypticase Soy Agar with 5% Sheep's Blood Agar (SBA), and *Clostridium botulinum* Selective Medium (CBSM). Approximately 250 µL of processed specimen was used to inoculate TPGY Broth. All culture media was incubated at 35 °C; however, samples that screened positive for the *bont /B, /E,* or */F* toxin gene via rtPCR using the method developed by Davis et al. (29) were also incubated at 25 °C if no growth was observed at 35 °C. EYA and CBSM were checked at 24 to 48 h for lipase or lecithinase-reaction positive colonies, depending on preliminary rtPCR results. Some BoNT-producing organisms do not produce lipase reactions on EYA, which is taken into consideration when no lipase reaction is observed but a positive rtPCR result is generated. After 48 h, TPGY broth is screened for *bont* genes by rtPCR to confirm the presence of viable organism. Positive enrichment broths are then subcultured as described above.

In cases where it was difficult to isolate BoNT-producing organisms from primary clinical specimens or environmental samples, spore selection was performed in two ways. First, 500 µL cell suspensions of suspicious mixed growth or rtPCR positive enrichment broth samples were incubated on a rotary mixer (room temperature, 20 °C–25 °C, 400 rpm,) for 1 hour with 500 µL molecular grade ethanol. Separately, 500 µL cell suspensions of suspicious mixed growth or rtPCR positive enrichment broth samples were incubated at 80 °C for 10 min. Once completed, 25 to 50 µL of each mixture was used to inoculate media directly, while the remaining volume of sample was transferred to TPGY broth for 24 to 48 h, then plated, streaking for isolation.

Due to the Tier 1 select agent designation, use of BoNT-producing strains of *Clostridia* was logged and waste was segregated, destroyed, and disposed of appropriately. Work was performed in a select agent registered Biosafety Level 2 laboratory using appropriate personal protective equipment and enhanced precautions by trained personnel with Department of Justice select agent clearance.

2.3 Extraction

Once an axenic culture was obtained, a single, well-isolated colony was transferred into approximately 9 mL of TPGY broth and incubated anaerobically (35 °C) for 24 to 48 h. Cultures were pulse vortexed to obtain a homogenous cell suspension, then transferred to a 15 mL conical tube. Samples were centrifuged using sealed swing-bucket rotors for 10 min, or until a compact cell pellet formed, at 4,000 rpm, 4 °C. Next, all supernatant was removed and destroyed in a 20% bleach solution. Cell pellets were resuspended in 300 µL of a 20 mg/mL lysozyme solution by pipetting, then extracted using a modified Epicentre MasterPure Complete (Lucigen Biosearch Technologies, Hoddesdon, UK) DNA extraction as described by Halpin et al. (31).

2.4 Library preparation and sequencing

Library preparation was completed for Illumina sequencing platforms by the Wadsworth Center Advanced Genomic Technologies Cluster. Approximately 100X genome coverage was targeted for each isolate. Sequencing libraries were prepared using the Illumina DNA prep kit with modifications as outlined in Dickinson et al. (32). Samples were run on Illumina MiSeq (v2 500 cycle chemistry kits) or NextSeq instruments (NextSeq 500/550 Mid Output Kit v2.5 300 Cycles).

2.5 Genomic analysis

2.5.1 SNP-based analyses

To determine relatedness, each isolate library was compared to a database of 14 *Clostridium* spp. genomes. The reference genome database represents complete high-quality genomes for the following strains: Alaska E43 (CP001078.1), Ba4_657 (CP001083.1), BKT015925 (CP002410.1), BL5262 (GCA_000182605.1), CDC_67071 (CP013242.1), CDC_67190 (CP014148.1), Eklund 17B (CP001056.1), Hall (CP000727.1), Kyoto (CP001581.1), Langeland (CP000728.1), Loch Maree (CP000962.1), Okra (CP000939.1), Osaka05 (NZ_DF384213.1), and Sullivan (CP006905.1) (Supplementary Table S1). The closest matching reference genome for each isolate was identified by MinMash distances using Mash version 2.1.1 (33) and default parameters, except for increasing the number of sketches to 10,000. Species designations were assigned based on the closest matching reference genome (Supplementary Table S1). Isolates lacking a match to the reference database (Mash distance >0.1) were not included in the SNP-based analyses. Reads were filtered with BBDuk from the BBTools suite (version March 24, 2020) (34) and trimmed with Trimmomatic version 0.36 (35) under default parameters. Cleaned reads were mapped to the closest matching reference genome selected in the first step using BWA-MEM Version: 0.7.15-r1142-dirty (36). Alignment files were sorted, and duplicate reads were removed with Picard version 2.9.2-SNAPSHOT (37, 38). Variant positions were required to have a minimum mapping and base quality score of Q20, minimum 10X depth, a quality score (QUAL) > 100, and to be supported by ≥95% of the reads. Sites covered by ≥3X the mean genomic depth were masked to minimize duplicated, repetitive, or unreliable regions in the final consensus genome. Samples with less than 30X average read depth were not included in SNP-based analyses. High quality consensus genomes were generated with SAMTools/BCFtools version 1.4.1 (39, 40) and SNP alignments and

matrices were generated using snp-sites 2.5.1 (41) and snp-dists v.0.7.0 (42), respectively.

2.5.2 Toxin serotype/subtype detection by mapping and gene assembly

Toxin serotype and subtype were assigned by mapping reads to a separately curated *bont* subtype database using BWA v.0.7.12-r1039 (36). The database represented full length *bont* genes from each available serotype and subtype deposited in NCBI (Supplementary Table S2). Variants of each subtype were included to maximize the molecular diversity captured. Reads that aligned to any sequence in the database were extracted by Samtools v.1.2-242-g4d5647 (39, 40) and *de novo* assembled with MEGAHIT v.1.2.9 (43). Assembled contigs were then queried against the aforementioned *bont* subtype reference database with BLASTN v.2.13 (44). The top hit was identified and reported if the percent identity and coverage of the reference gene were greater than 70% and 50%, respectively. Cutoffs were set to ensure detection of variable *bont* genes. However, all *bont* genes assembled were highly conserved - with 99%–100% identity to genes from the reference database and 65%–100% coverage (data not shown). Toxin serotype was not reported if blast results could not confidently assign subtype. The presence or absence of *bont* genes was further supported by *de novo* genome assemblies (see below).

2.5.3 *De novo* genome assembly

De novo genome assemblies were generated by shovill v.1.1.0 (45) and annotated by Bakta v.1.8.1 (46) using the db-light v.5 database (47). Busco v.5.4.7 (48) determined assembly completeness using the clostridia_odb10.2020-03-06 database (49) and Quast v.5.3.0 (50) provided additional quality metrics. Kraken2 v.2.1.3 (51) assigned taxonomic classification to contigs to estimate assembly contamination levels. Assemblies with <60% Clostridial contigs or Busco scores <90% were not included in *de novo* assembly-based downstream analyses. MOB-suite v.3.0.3 (52) with the mob_recon option detected and classified plasmids, and Orthofinder v.3.0.1b1 (53) assigned proteins to orthogroups (homologous groups of proteins, which includes orthologs and paralogs). Multilocus sequence typing (ST) was performed with mlst v.2.23.0 (54) using the PubMLST database (55). Ribosomal Multilocus Sequence Typing (rST) was performed by the PubMLST webserver (55, 56).

Genes annotated as "*bont*" by Bakta v1.8.1 (46, 47) were extracted and clustered at 100% identity by cdhit-est (57, 58) to retain single representatives of each unique sequence. Unique *bont* genes were aligned with those from the *bont* reference database by mafft v.7.453 (59, 60) and a *bont* genealogy was generated by IQ-TREE2 v2.4.0 (61) under a GTR + F + R3 substitution model with 1,000 ultrafast bootstrap replicates (62) to confirm subtyping assignments initially conferred by the mapping and assembly strategy.

2.5.4 SNP-based phylogenetic reconstruction

SNP-based phylogenetic trees for isolates in the Alaska, CDC_67071, and Hall reference groupings were calculated by IQ-TREE2 v2.4.0 (61) with a TVM + F + ASC + G4 substitution model selected by ModelFinder (63). Branch support was evaluated with 1,000 ultrafast bootstrap replicates (62). The number of positions included in the SNP alignments ranged from 90,750 to 161,391. Polymorphic positions due to recombination were identified and masked by Gubbins v2.3.4 (64) for whole genome alignments of Alaska, CDC_67071, and Hall groupings and all SNPs were extracted by snp-sites (41) to determine the effects of recombination on tree topology. Phylogenies were estimated in IQ-TREE2 (61) under the TVM + F + ASC + G4 substitution model. Topologies between initial trees (those based on all SNP positions) and those built by excluding recombinant positions were compared in IQ-TREE2 with the –rf (Robinson-Foulds) function. Because removing recombinant positions did not affect the overall topology of the Alaska, CDC_67071, or Hall phylogenies (with Robinson-Foulds distances ranging from 0–44) nor the relationships highlighted in the results and discussion sections, we do not report on these findings further and refer to the initial trees herein (SNP alignments and trees with and without recombination are available at doi:10.5061/dryad.2z34tmpzz).

2.5.5 Assembly-based phylogenetic reconstruction

A multigene phylogeny for 225 isolates (including reference genomes from SNP-based analyses) was generated from an alignment of 264 single-copy orthologues as identified by Orthofinder (53). Orthologous sequences were aligned by mafft v.7.453 (59, 60) and concatenated into a supermatrix with an in-house Python3 script. All alignments had complete isolate representation and yielded a supermatrix of 83,069 positions with no more than 10% missing data. IQ-TREE2 v.2.4.0 (61) estimated a maximum likelihood phylogeny under an LG + F + I + R4 model selected by ModelFinder (63) with 1,000 ultrafast bootstrap replicates (62). Additionally, a supertree from 900 protein trees for the same isolates was generated by Astral-Pro v.1.20.3.6 (65), which allows for the inclusion of orthologs and paralogs in species tree estimation by accounting for incomplete lineage sorting and gene duplication and loss. Multisequence alignments were generated by mafft v.7.453 (59, 60) and individual orthogroup trees were reconstructed in IQ-TREE2 v.2.4.0 (61) with the LG + F substitution model. Phylogenies were visualized with the ggtree package (66) for R v.4.4.1 (67) (alignments and trees are available at doi:10.5061/dryad.2z34tmpzz).

3 Results and discussion

3.1 Summary statistics

Over approximately 40 years, more than 240 specimens were archived at the NYSDOH Wadsworth Center Biodefense Laboratory and recently sequenced. While all samples submitted for testing were suspected to be linked to botulism cases in humans or animals, the laboratory did not play an active role in the selection or acquisition of these samples. Clinical presentation and epidemiologic investigation by submitting organizations were used to justify sample submission for laboratory confirmation. As a member of the CDC Laboratory Response Network, the NYSDOH Wadsworth Center may perform testing for other states in the northeastern portion of the country. After quality control, 220 samples remained for analysis, the majority of which originated from New York (162, 73.6%), but samples from the states of Pennsylvania (1, 0.5%), Connecticut (14, 6.4%), Massachusetts (6, 2.7%), Maine (1, 0.5%), New Hampshire (1, 0.5%), and samples of unknown origin (35, 15.9%) were included as well. Isolates obtained from clinical or clinical surveillance specimens

account for 62.7% of samples, while 33.6% were from animals, and 3.2% originated from environmental sources (Supplementary Table S3).

Within our sample set, seven serotype/serotype combinations were observed. Overall, analysis of *bont* genes extracted from WGS data allowed for the designation of 16 BoNT subtypes (Figure 2). The frequency of BoNT types in the collection ranged from as few as one sample producing BoNT/Bf or /C to as many as 70 specimens producing BoNT/B (Figure 2; Supplementary Table S3). All serotype B isolates were designated as Group I proteolytic (*C. sporogenes* or *C. parabotulinum*) except one Group II non-proteolytic (*C. botulinum*) subtype B4 isolate, which derived from an environmental hay sample (Figure 2; Supplementary Table S3). WGS revealed that 41 Group I *C. parabotulinum* genomes harbored *bont*/A1(B5) gene sequences (Figure 2; Supplementary Table S3). These BoNT A(B) isolates produce enzymatically active BoNT/A but not BoNT/B, which is consistent with other reports of silent *bont*/B5 genes in isolates which produce toxin (68). Infrequently observed combinations of *bont* genes included *C. sporogenes* A4(B5) isolates associated with a case of infant botulism in Monroe County, New York), two *C. parabotulinum* A5(B2') isolates, and one *C. parabotulinum* B5F2 isolate (isolated from another month-old infant from Queens, New York in 2019) (Figure 2; Supplementary Table S3). These rare combination cases appear to conform to previous observations of A4(B5) and B5F2 isolates in cases of very young infant botulism (69). The rare *bont*/B7 gene was detected in three *C. parabotulinum* isolates from clinical stool specimens that differed by 46–126 SNPs (Figure 2; Supplementary Table S3). One BoNT/B7-producing isolate was previously characterized and tied to infant botulism (70). One *Clostridium novyi sensu lato* isolate with a *bont*/C gene was associated with an animal specimen (Figure 2; Supplementary Table S3). Six *C. baratii* isolates collected over a period of at least 16 years harbored the *bont*/F7 gene, which is rarely identified in NYS (Figure 2; Supplementary Table S3). Five of the six F7 producing *C. baratii* isolates were obtained from individuals at least 59 years old (median age 68 years), which likely represent cases of adult toxicoinfections (17). No *C. argentinense* or *C. butyricum* isolates were detected.

3.2 *C. parabotulinum* investigations

Most clinical epidemiologically linked isolates were closely related genomically, differing by as little as 0 SNPs; however, as many as 259 SNPs differentiated *Clostridium* spp. isolates obtained from clinical primary specimens and associated clinical surveillance isolates. In one previously reported home-canning incident involving peas (26), *C. parabotulinum* A1(B5) (CDC_67190 reference grouping) was isolated from two clinical stool specimens from different patients, a salad bowl, and an empty jar that contained peas used in the salad. No SNPs were detected between these isolates, reinforcing the epidemiologically identified link (Supplementary Table S4a; Supplementary Figure S1).

Linked strains of *C. parabotulinum* A2, isolated from a primary stool specimen and a commercial honey sample (sourced from a combination of countries including Argentina, the US, and Canada), also showed minimal genomic variation. The clinical stool (IDR2100017857-01-01-2) and honey (IDR2100019300-01-02)

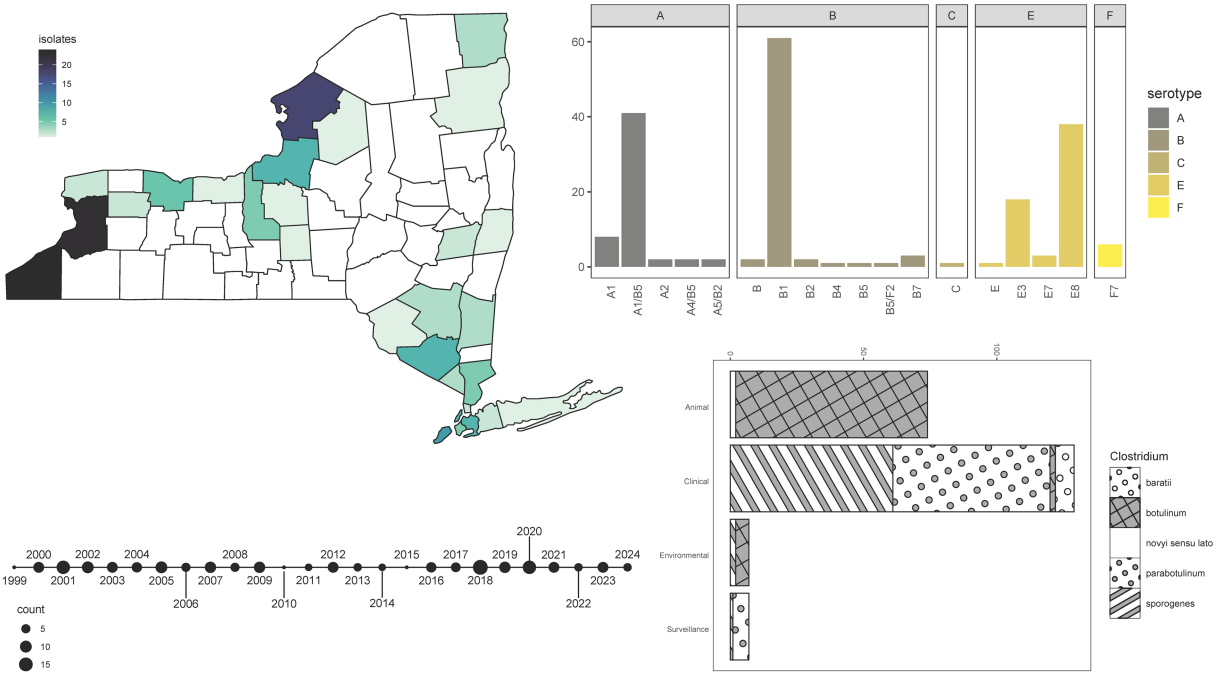

FIGURE 2
Isolate demographics and characterization. Summary information for 220 *Clostridium* isolates sequenced by the NYSDOH Wadsworth Center. Top left, heatmap, and geographic location for isolates (with available data) collected in New York State; bottom left, timeline reflecting the number of isolates collected for the years spanning 1999–2024 for specimens with collection years available; top right, barplot of the number of BoNT subtypes detected by WGS methods; bottom right, stacked barplot of isolate types and the species they represent with species assignments taken from MASH results. All plots were generated by the ggplot2 (92) package in R (67).

isolates differed by just two SNPs but were distinguished from the Kyoto reference genome by 1,545 and 1,647 SNPs, respectively (Supplementary Table S4b). Argentina was previously identified as a reservoir for BoNT/A2 strains (71–73), which underscores the importance of combining traditional epidemiological and WGS data to understand the global transmission and distribution of *Clostridium* spp.

However, some putatively associated *C. parabotulinum* isolates (Hall reference grouping) displayed higher levels of variation. Epidemiologically linked *C. parabotulinum* A5(B2') stool and honey isolates (IDR2000276145-01-01 and IDR2000277026-01-05) were differentiated by 259 SNPs in another case of foodborne botulism (Figure 3 green highlight; Supplementary Table S4c). In contrast, the reference Hall genome deviated from a genome sequenced in this study (IDR2100035691-01-02) by just 187 SNPs (Figure 3, Supplementary Table S4c). The increased variation between honey and stool isolates could be due to the presence of polymorphic populations in the original samples (74), but also questions the legitimacy of the link. Incorporating additional A5(B2') genomes, such as those from closely related isolates predominantly associated with cases of wound botulism in the United Kingdom (69), would clarify the relationship among these isolates. Inspecting variation within a larger context reinforces the conclusion that SNP-based analyses cannot solely determine an epidemiologic link.

While WGS identified the toxin subtype for both samples as *bont*/A5(B2'), rtPCR and Endopep-MS results were positive for serotype A only. Further investigation revealed that these isolates contained a large deletion in the *bont*/B2 gene (Figure 4), which is targeted by the rtPCR assay, explaining the negative rtPCR and Endopep-MS results. Isolates harboring truncated, non-functional *bont*/B2 genes in this rare arrangement have been described previously (75), supporting the legitimacy of these partial deletions. Thus, NGS may enhance epidemiological investigations by clarifying discrepant results among methods.

3.3 *C. sporogenes* investigations

Retrospective WGS also supported investigations involving *C. sporogenes* (CDC_67071 reference grouping) - specifically, a case of botulism associated with the consumption of home-prepared fermented tofu made with ingredients obtained at the local supermarket. Both genomes from the clinical stool and fermented tofu isolates (IDR1200008265 and IDR1200008991) contained the *bont* B1 gene and formed a phylogenetically distinct cluster with identical consensus sequences (Figure 5, orange highlight; Supplementary Table S4d).

Neither genomic variation nor phylogenetic relationships among *C. sporogenes* isolates was correlated with collection date. Variation within one monophyletic group of *C. sporogenes* BoNT/B1 isolates (Figure 5, clade 1 in green highlight), with known collection dates between 2000 and 2024, ranged from 20 to 108 SNPs in comparison to the CDC_67071 reference genome (Supplementary Table S4d). The majority of these isolates could be traced back to New York City

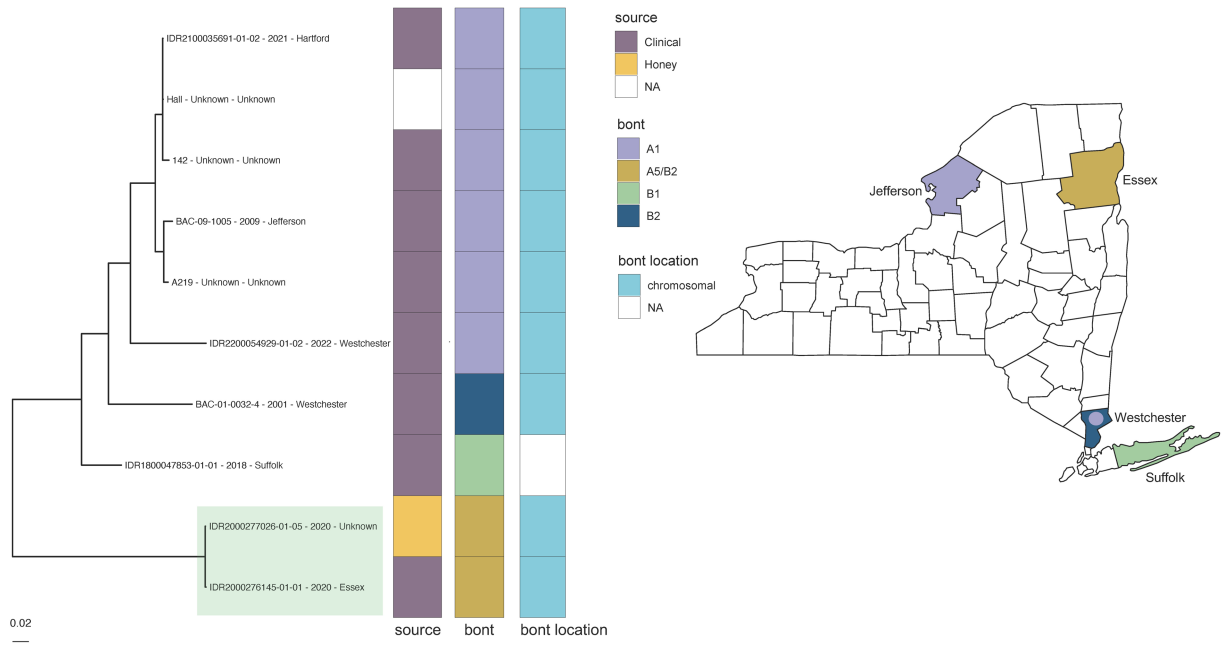

FIGURE 3
Clostridium parabotulinum phylogeny for Hall reference grouping and New York State (NYS) map. Maximum likelihood phylogeny generated by IQ-TREE2 (61) from an alignment of 90,750 SNPs under a TVM + F + ASC + G4 substitution model for *Clostridium parabotulinum* isolates and Hall reference genome. Support values were calculated using 1,000 ultrafast bootstrap replicates (62). Branch support was 100% for all bipartitions. Tips are labeled by isolate name - collection year – county location and are annotated by isolate source, *bont* type, and the predicted location of the *bont* gene. Phylogeny scale bar, substitutions per site. The collection location for NYS isolates with available demographic data are depicted by BoNT type in the NYS map. Epidemiologically linked A5 (B2') honey and clinical stool isolates (IDR2000277026-01-05 and IDR2000276145-01-01) associated with a case of foodborne botulism are highlighted in green. Trees and associated annotation were visualized with ggtree (66) and maps were created by ggplot2 (92) in R (67).

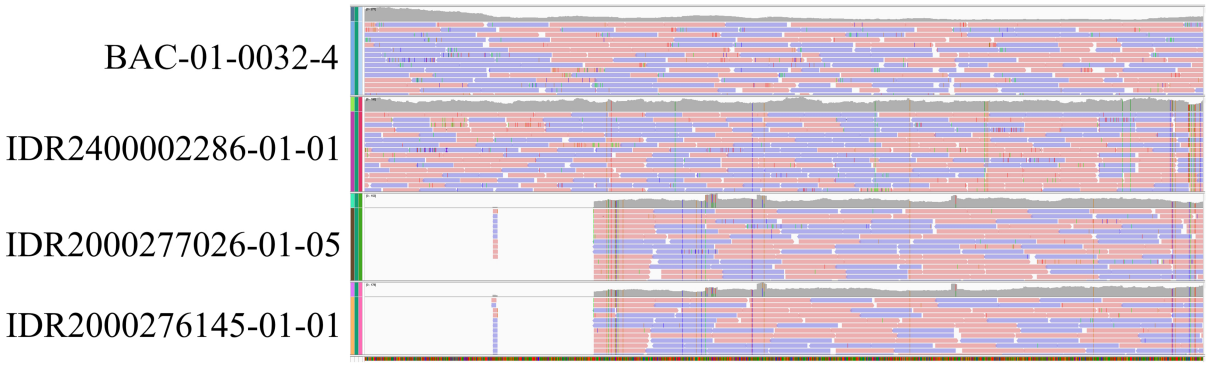

FIGURE 4
Bont B2 gene truncation confirmed by WGS. Illumina paired-end reads mapped to the *bont*/B2 gene from reference strain Su1036 (CP022397.1) for four *C. botulinum* isolates using Integrative Genomics Viewer (93). Isolates BAC-01-0032–4 and IDR2400002286-01-01 did not display a truncated B2 gene, while a large deletion was observed in the epidemiologically related isolates IDR2000277026-01-05 and IDR2000276145-01-01. *De novo* assembly and annotation were also performed, which supported the presence of a deletion.

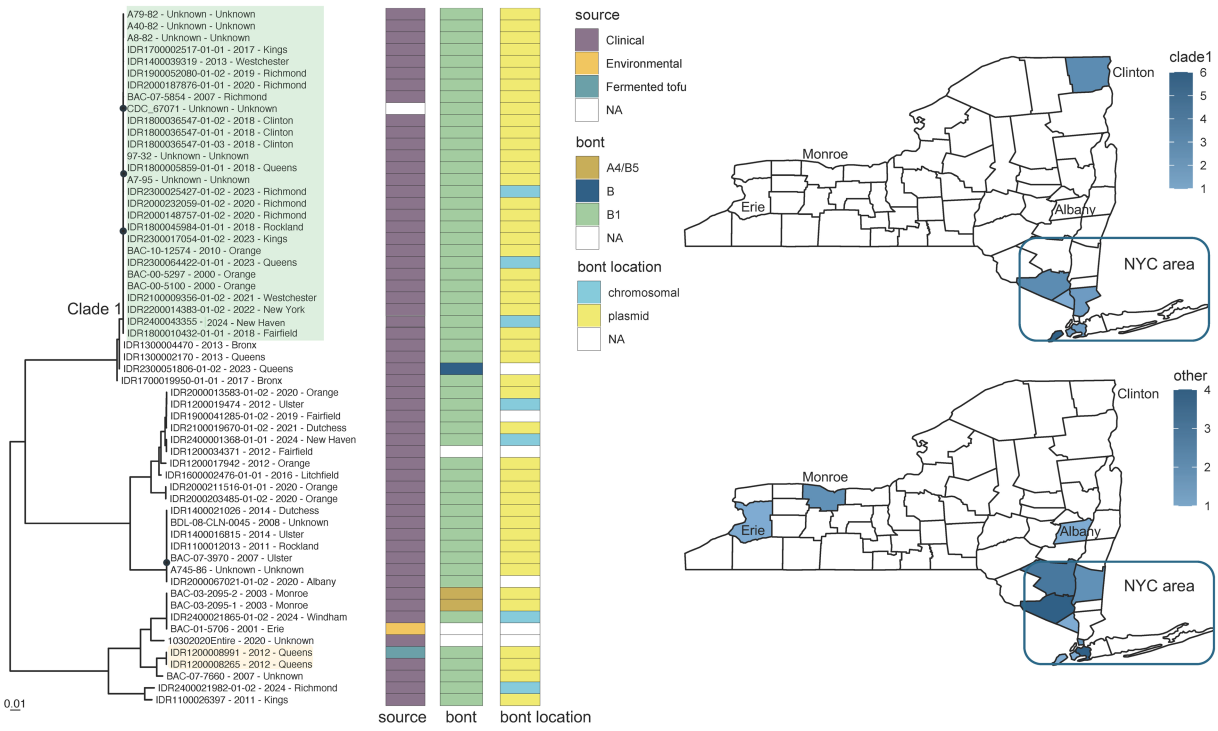

FIGURE 5
Clostridium sporogenes toxin subtype B1 phylogeny for CDC_67071 reference grouping and New York State (NYS) map. Maximum likelihood phylogeny generated by IQ-TREE2 (61) from an alignment of 161,391 SNP positions under a TVM + F + ASC + G4 substitution model for *Clostridium sporogenes* isolates harboring toxin subtype B1 genes and CDC_67071 reference genome. Support values were calculated using 1,000 ultrafast bootstrap replicates (62). Support values <80% are indicated by a black circle. Tips are labeled by isolate name - collection year – county location and are annotated by isolate source, *bont* type, and the predicted location of the *bont* gene. Phylogeny scale bar, substitutions per site. IDR1200008991 and IDR1200008265 associated with a case of foodborne botulism involving tofu are highlighted in light orange. Clade 1 (green highlight) represents a possible clonal expansion in New York City and surrounding metropolitan area. The two NYS maps show the collection locations for clade 1 isolates (top) and all other isolates (bottom) with the color intensity representing the number of isolates. Trees and associated annotation were visualized with ggtree (66) and maps were created by ggplot2 (92) in R (67).

(Bronx, Kings, New York, Richmond, and Queens counties), and the surrounding metropolitan area [Westchester, Orange, and Rockland counties in NY and Fairfield County in Connecticut (CT)], when demographic data were available. Comparatively, isolates falling outside of this cluster varied by approximately 1,620 to 77,184 SNPs when mapped to the reference (Figure 5; Supplementary Table S4d). Although no known epidemiological links exist, highly similar monophyletic isolates collected over 20 years in a circumscribed geographic area resembles a pattern of clonal expansion. Clade 1 could be endemic to the region with limited export to other locations (such

10.3389/fpubh.2025.1651032

as Clinton County, NY) or it could represent an importation event of a successful, more widely distributed clone. The distribution and phylogeographic origins of this variant would require a more global representation of genome sequences from *C. sporogenes*, which is beyond the scope of this paper.

3.4 *Clostridium novyi sensu lato* investigation

One *Clostridium novyi sensu lato* BoNT/C isolate was obtained from a cow liver specimen (IDR1800052927-02-03) on a dairy farm about 30 miles east of the Lake Ontario shoreline. While BoNT/C is often associated with cases of botulism in waterfowl (76), mammals are also susceptible (77). An animal from the farm or Lake Ontario may have contaminated livestock feed, similar to findings in other recent investigations (78, 79).

3.5 *Clostridium botulinum* investigations

Animal isolates offered insight into the evolutionary dynamics of *Clostridium botulinum* serotype E organisms in New York State. *C. botulinum* was isolated from only 1.5% of clinical specimens (2/131) in contrast to 97% of all animal specimens (72/74) sampled (Figure 2), despite clinical specimens being collected from a wider geographic distribution and longer timeframe (Supplementary Table S3). *C. botulinum* isolates derived predominantly from various bird species in New York counties that border Lake Erie and Lake Ontario but also from three fish species and sediment samples. Isolates from different animal species were broadly distributed throughout the phylogeny (Figure 6), supporting a lack of host specificity (15). Two isolates collected from double-crested cormorants on the Lake Ontario shore (IDR1900041093-01-02 and IDR2000247142-02-02) contained *bont*/E8 genes and differed by only 6 SNPs while a third specimen, collected from an eagle only about two miles inland (IDR1800013236-01-02), contained the *bont*/E3 gene and differed from the shore samples by ~27,600 SNPs (Figure 6; Supplementary Table S4e). *C. botulinum* subtype E8 and E3 isolates were also obtained from a single long-tailed duck (Figure 6). Collectively, these results highlight the presence of multiple BoNT/E subtypes in a phylogenetically diverse community of *C. botulinum*, which could contribute to avian botulism in a single geographic region (80). *C. botulinum* outbreaks due to contaminated fish from the Great Lakes (such as Lake Ontario and Lake Erie) have been recorded since the 1960s. Our study suggests that birds may be highly susceptible to infection from fish as well (81).

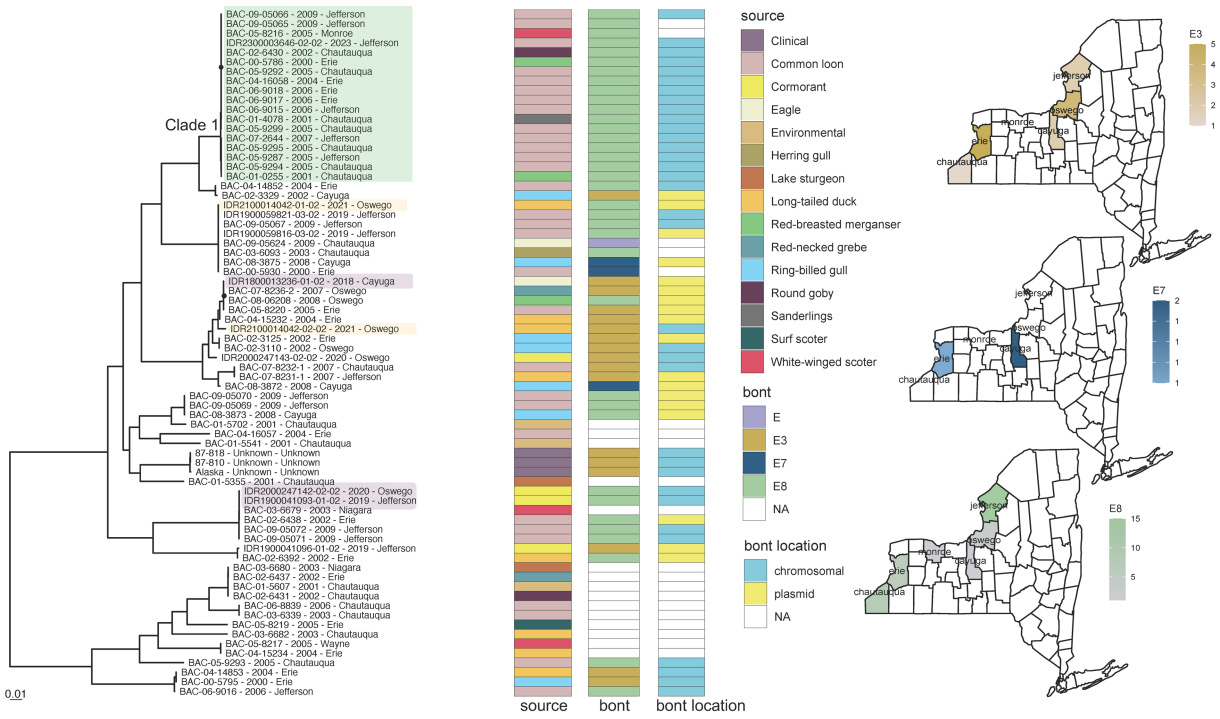

FIGURE 6
Clostridium botulinum toxin serotype E phylogeny Alaska reference grouping and New York State (NYS) map. Mid-point rooted maximum likelihood phylogeny generated by IQ-TREE2 (61) from an alignment of 121,606 SNP positions under a TVM + F + ASC + G4 substitution model for *Clostridium botulinum* serotype E isolates and Alaska reference genome. Support values were calculated using 1,000 ultrafast bootstrap replicates (62). Support values <80% are indicated by a black circle. Tips are labeled by isolate name - collection year – county location and are annotated by isolate source, *bont* type, and the predicted location of the *bont* gene. Isolates from similar geographic areas separated by 6 to 27,600 SNPs (IDR1800013236-01-02, IDR1900041093-01-02 and IDR2000247142-02-02) are highlighted in light purple. Isolates from the same duck with different BoNT subtypes are highlighted in light orange (IDR2100014042-01-02 and IDR2100014042-02-02). Phylogeny scale bar, substitutions per site. The three maps depict the distribution and number of *bont*/e subtypes collected in NYS in this study. Trees and associated annotation were visualized with ggtree (66) and maps were created by ggplot2 (92) in R (67).

Some *C. botulinum* serotype E isolates collected decades apart from environmental and animal sources showed limited genomic variation, such as members of a large BoNT/E8 clade (Clade 1, Figure 6) that differed by 0 to 123 SNPs but by approximately 5,500 to 40,000 SNPs compared to other isolates within the Alaska reference grouping (Supplementary Table S4e). These E8 isolates derived from four New York counties spanning approximately 300 miles, all bordering naturally linked Lake Erie and Lake Ontario. The low levels of variation among these geographically constrained, temporally diverse E8 isolates in addition to the chromosomal location of their *bont* genes might reflect stable environmental reservoirs that have facilitated expansion and long-term dormancy, which limits the introduction of mutations via replication (82).

However, several clades across the *C. botulinum* tree were comprised of members with different BoNT subtypes. The presence of different BoNT subtypes in closely related isolates as well as the presence of the same subtype in diverse genomic backgrounds in both chromosomal and plasmid locations support the influence of horizontal gene transfer and recombination in the wide-scale transmission of these genes. Indeed, *bont* E8 itself was determined to be the product of recombination as it shares regions of similarity with subtypes E2, E6, and E7 (15). Group II spores (including *C. botulinum*) are usually less resilient, particularly against temperature (83) and heat (84), with lower optimal growth temperatures than Group I proteolytic organisms. Such factors might contribute to the prevalence of serotype E organisms in the northwestern parts of New York, but lack of environmental and animal sampling from other regions of the state prevents any strong conclusions.

3.6 *De novo* assembly results

De novo assemblies were 84%–100% complete according to Busco scores with contamination levels ranging from 0%–94%. Mixed or contaminated libraries were also evidenced by duplication level and assembly size (Supplementary Table S3). Most assemblies fell within the standard 2.5–6 MB size range for members of the *Clostridium* genus but some exhibited sizes exceeding 8–9 MB, indicative of genetic material from other sources (Supplementary Table S3). While contaminated assemblies were excluded from downstream analyses (i.e., orthology assignment and phylogenetic reconstruction), mapping-based assemblies distinguished between *Clostridium* and non-*Clostridium* reads, and in some cases, enabled SNP-based analyses to proceed if genome coverage was acceptable. However, the isolates discarded from both analyses were largely consistent with each other (data not shown). As these were clonal isolates, we did not anticipate the contamination levels detected in some of the assemblies. In the future, we would employ a filtering step before *de novo* assembly to avoid data loss.

Analysis of annotated *bont* genes from *de novo* assemblies confirmed the results of the mapping and gene assembly strategy. *De novo* assembly and annotation identified 187 isolates with *bont* genes, which were clustered at 100% identity into 33 groups, reflecting a high degree of sequence conservation within the dataset (Supplementary Table S5). Representative sequences were selected from each group to generate a bont gene tree and evaluate subtyping assignments. With the exception of two sequences (from IDR2100017857-01-01-2 and IDR2100035691-01-02), annotated genes clustered with those from the *bont* reference database according to the subtypes initially assigned by the mapping and gene assembly method (Supplementary Figure S2). Only partial toxin genes were recovered from the *de novo* assemblies of IDR2100017857-01-01-2 and IDR2100035691-01-02. Although blast analyses supported toxin A2 and A1 subtype designations for these genes, both fell on long branches sister to larger E/F or C/D clades due to their truncated nature (Supplementary Figure S2).

For several isolates, no toxin gene was detected by *de novo* genome assembly or by mapping reads to a *bont* gene reference database. This is likely due to the presence of non-toxin producing isolates in our collection, as was also noted in Hannett et al. (80), or loss of *bont*-containing plasmids through subculture. It is also possible that genome assemblies were incomplete, although Busco completeness scores for 93% of *de novo* assemblies lacking *bont* genes ranged from 93%–100% (Supplementary Table S3). In fact, assembling reads extracted from reference-based *bont* alignments appeared more sensitive in detecting *bont* genes and assigning subtypes than *de novo* assembly. The former detected *bont* genes in 15 isolates and assigned subtype designations for 14 [A1(B5), A1, B1, E7, E8] while the annotation from the latter method showed no *bont* genes present (Supplementary Table S3). Moreover, the mapping and gene assembly-based method was able to detect dual toxin isolates that were not recovered by *de novo* assembly.

3.7 Supermatrix and supertree phylogenies

The topologies recovered by the supermatrix and supertree strategies were roughly equivalent (Robinson-Foulds distance calculated by IQ-TREE2: 292) and both confidently differentiated species into phylogenetically distinct clades (Supplementary Figure S3). Isolates clustered with their respective reference genomes as identified by MASH distances in the SNP-based workflow, except for two isolates assigned to the CDC_67190 reference grouping that clustered with Hall isolates and two Hall isolates that formed a clade with a CDC_67190 isolate (Supplementary Figure S3). Members of the CDC_67190 group formed several distinct clades that were paraphyletic with the Kyoto clade; however, the monophyly of *C. parabotulinum* (which includes CDC_67190, Hall, and Kyoto) was maintained (Supplementary Figure S3). The patterns of orthogroup presence or absence also differentiated isolates by species (Figure 7) and further validated the initial groupings assigned by the SNP-based analyses (Supplementary Figure S4).

Within our dataset, *C. parabotulinum* showed the greatest diversity of serotypes and the greatest mobility of *bont* genes, with some subtypes distributed across multiple *C. parabotulinum* clades (Figure 7). In general, most isolates with the same toxin subtype clustered together [such as the *C. parabotulinum* A1(B5) or *C. sporogenes* B1 clusters] but examples of closely related isolates with different BoNT types appeared throughout the tree (Figure 7). For example, *C. botulinum* subtype E3, E7, and E8 isolates formed a closely related monophyletic group and putatively harbored 6 plasmid types, with E3, E7, and E8 subtypes observed on the same plasmid background (Figure 7). Five *C. botulinum* genomes contained a plasmid (CP017254) first described in *C. taeniosporum* (85). The majority of *C. sporogenes* isolates had *bont* B1 genes with predicted locations on the same plasmid type (Figure 7). However, toxin

10.3389/fpubh.2025.1651032

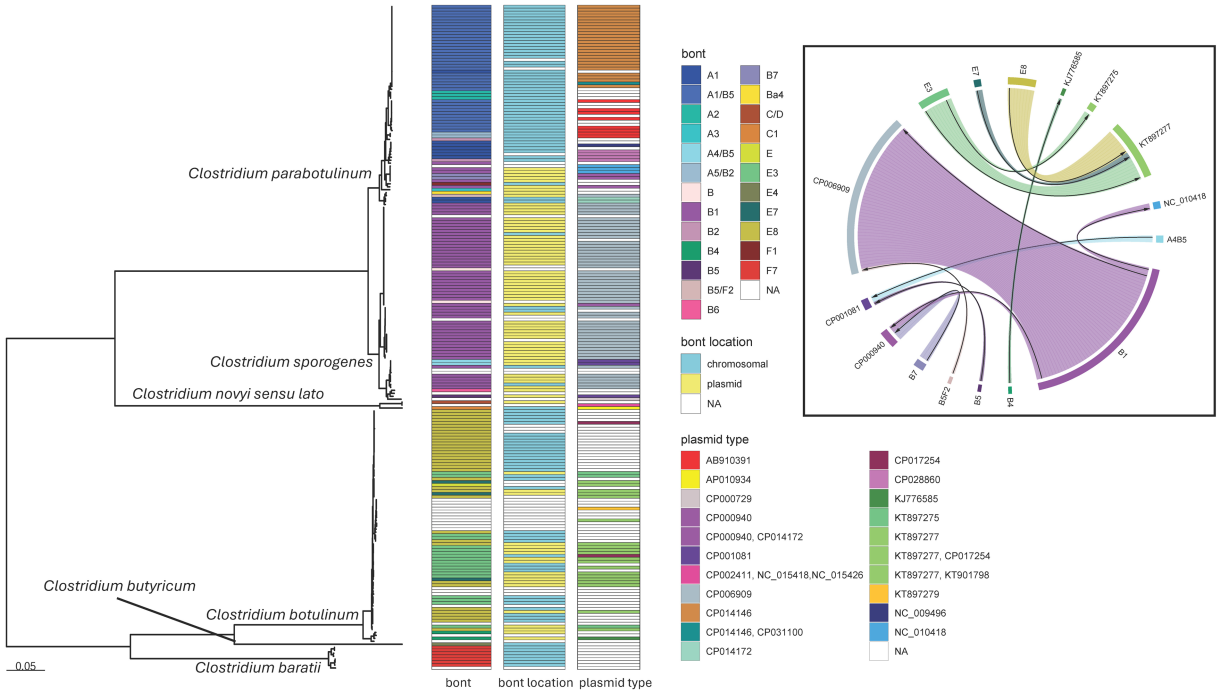

FIGURE 7
Clostridium species maximum likelihood phylogeny and *bont* plasmid distribution. Mid-point rooted maximum likelihood phylogeny of 225 *Clostridium* isolates representing six species. The tree was generated in IQ-TREE2 (61) using an LG + F + R4 substitution model and an 83,069 character supermatrix of 264 single-copy orthologues. Support values were calculated using 1,000 ultrafast bootstrap support replicates (62). Phylogeny scale bar, substitutions per site. All species were recovered with 100% support (not shown). *bont*, *bont* gene serotype or subtype; *bont* location, putative location of the *bont* gene as evaluated by MOB-suite; plasmid type, the GenBank accessions of the closest matching plasmids to those identified in the *de novo* assemblies. Boxed inset, chordogram diagram connecting *bont* gene subtypes to their putative plasmid locations. The direction of the arrow is from toxin subtype to plasmid. Trees and associated annotation were visualized with the ggtree (66) package and the chordogram was generated with the circlize (94) package for R (67).

serotype B genes were predicted on multiple plasmid types with the B1 subtype putatively located on three different plasmids (Figure 7). The plasmid and chromosomal locations for the same *bont* genes, their distribution throughout the tree, and the diversity of plasmids carrying these genes suggests a high degree of genetic mobility for both Group I (*C. parabotulinum* and *C. sporogenes*) and Group II (*C. botulinum*) organisms.

4 Conclusion

Our study increases the WGS data available for *Clostridium* spp., particularly for clinical and animal isolates. The combination of WGS for 220 additional *Clostridium* spp. isolates and their associated metadata will provide contextual and reference genomes for future epidemiological investigations and more broadly assist with understanding the evolution, diversity, and global distribution of these organisms.

SNP-based analyses assigned *Clostridium* species to multiple groups determined by MASH distances to a reference genome database. Separating isolates of the same species into different reference groupings ensures that a highly similar reference genome is employed for accurate mapping and SNP detection, which is crucial for the high resolution required in epidemiological investigations. No definitive SNP-based cutoff has been established for linking epidemiological isolates and our results indicate that no single threshold will be appropriate across *Clostridium* species. Although most isolates in related cases differed by very few SNPs, our results uniquely showed that the number of SNPs differentiating unrelated isolates varied by 0 to 4 orders of magnitude. This variation contributes to the challenges faced in determining evolutionary and taxonomic diversity of *Clostridium* spp. (86). Therefore, each reference grouping should be examined independently and phylogenetic relationships along with epidemiological data, should be considered when determining the connection between an isolate and a potential outbreak. Additionally, this approach fails to capture the broader patterns of genomic evolution and phylogeographic distribution within and between species and limits conclusions when potentially linked isolates demonstrate higher levels of variation.

Limited sample collection related to confirmed clinical botulism cases hinders our ability to establish links between samples that may have an unknown, related, local origin. While our dataset substantially expands the genomic representation of *Clostridium* species, we lack sampling from much of the middle and southern regions of New York State. Thus, conclusions regarding transmission, distribution, and endemicity among regions are limited. Increased efforts for active surveillance through in-home and surrounding environmental sample collection would be beneficial to retrospectively identify sources or provide warning for potential outbreaks. Even with an extensive

local environmental database, epidemiologic efforts that rely solely on WGS to elucidate sample origins may be confounded by the importation of organisms in various products sourced from around the world. Therefore, a local environmental database has the potential to mislead investigators without contextualizing genomes within a more global framework.

Incorporation of long read sequencing (LRS) may improve *de novo* assembly of genomes and resolve regions, such as repetitive elements, which challenge short read sequencing (SRS) technology (87). Consequently, LRS platforms have been used in combination with SRS platforms to close genomes with high confidence (88). Some *Clostridium* spp. genomes sequenced with Illumina technology contained areas of poor coverage that LRS technology may improve, including the assembly of plasmids, which may harbor *bont* genes. Additionally, LRS platforms may be run directly within biosafety cabinets in secure areas, such as select agent registered spaces. This reduces turnaround time by eliminating the need for inactivation and verification of sterility prior to transfer of extracts to an external sequencing laboratory.

Complex matrices associated with botulism testing can result in poor coverage of target genomes if metagenomic sequencing is performed on primary samples. While multiplexed PCR-based methods have been developed that can detect *bont* genes, toxin gene clusters, or speciate isolates (4, 29), we are unaware of targeted enrichment or amplicon-based NGS assays for detection or characterization of BoNT-producing *Clostridia*. Such assays could provide important information if primary specimens containing no viable organism are received. It is possible that increased output through non-targeted deep sequencing may allow for characterization of low-level pathogens and eventually approach levels of sensitivity similar to rtPCR.

One of the more difficult aspects of sequencing is results interpretation. Bioinformatic training is necessary to carry out and properly interpret the nuances of data analysis, as current pipelines are not standardized across public health institutions and can be complex. Although botulism is a rare disease, we believe it would be beneficial to establish a national or international database and standardized, freely available, version-controlled analysis for BoNT-producing *Clostridium* species. With the availability of numerous tools and associated parameters for each step of the bioinformatic analysis, workflows must be thoroughly evaluated and compared before deciding which would be used in the standardized approach. While this is time consuming, without a standardized method for comparison, changes to pipeline components will alter results and make comparisons between data sets difficult. Previously described WGS analyses include whole genome multilocus sequence typing, toxin gene cluster analysis, and genome-wide average nucleotide identity (27). We envision that a standardized analysis may also take demographic or product information and sequence analysis results into consideration for comparisons and provide outbreak alerts or epidemiological connections. Automatic recognition of novel serotypes or subtypes could also be incorporated, based on established guidelines (13).

Incorporation of WGS into laboratory testing algorithms has allowed us to further characterize isolates in the NYSDOH Wadsworth Center culture collection. Retrospective analysis led to identification of rare subtypes and elucidated challenges in linking potentially related isolates. We believe that incorporation of NGS into public health laboratories will benefit epidemiological investigations to improve public health outcomes.

Data availability statement

Original datasets are available in a publicly accessible repository. Sample sequence data has been submitted to the NCBI Sequence Read Archive (SRA) under BioProject ID (PRJNA1308155). Orthogroup alignments and associated phylogenies are available at: https://doi.org/10.5061/dryad.2z34tmpzz.

Ethics statement

The studies involving human specimens were approved by New York State Department of Health Institutional Review Board-approved protocol (03-037). The studies were conducted in accordance with the local legislation and institutional requirements. The ethics committee/institutional review board waived the requirement of written informed consent for participation from the participants or the participants' legal guardians/next of kin because specimens were submitted as part of clinical patient testing.

Author contributions

AD: Writing – review & editing, Writing – original draft, Methodology, Data curation, Investigation, Visualization. DC: Supervision, Methodology, Visualization, Investigation, Data curation, Writing – review & editing, Writing – original draft. EL-N: Methodology, Data curation, Software, Writing – original draft, Writing – review & editing, Formal analysis, Visualization. PL: Methodology, Writing – review & editing, Software, Visualization, Formal analysis, Data curation. CE: Conceptualization, Writing – review & editing. MP: Conceptualization, Methodology, Writing – review & editing, Supervision, Visualization, Project administration.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Acknowledgments

We would like to thank all past and present members of the NYSDOH Wadsworth Center Biodefense Laboratory for assistance with testing throughout the years which made this project possible. Additionally, thank you to the members of the NYSDOH Wadsworth Center Advanced Genomic Technologies Cluster (AGTC) Sequencing Core Laboratory for their work preparing and sequencing libraries used for data analysis. Thank you to Kimberly McClive-Reed for review of the manuscript. Finally, we would like to thank the Centers for Disease Control and Prevention for the reference samples provided.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2025.1651032/full#supplementary-material

SUPPLEMENTARY FIGURE S1
Phylogeny of *C. parabotulinum* and CDC_67190 reference genome. Mid-point rooted maximum likelihood phylogeny generated by IQ-TREE2 (61) from an alignment of 97,907 SNP positions under a TVM+F+ASC substitution model. Support values were estimated using 1000 ultrafast bootstrap replicates (62). Isolates related to the cases of botulism involving home-canned peas are highlighted in green. The tree was visualized with FigTree (89).

SUPPLEMENTARY FIGURE S2
Bont gene tree. A maximum likelihood tree of representative *bont* genes from *de novo* assemblies and *bont* genes from the reference database was generated by IQ-TREE2 (61) under a GTR+F+R3 substitution model with 1000 ultrafast bootstrap replicates (62). Genes from *de novo* assemblies represent unique sequences within the Wadsworth dataset and are labeled by isolateName_geneNumber (colored text). Toxin genes from the reference database are labeled by subtype_AccessionInformation (black text). Subtype clades containing genes from Wadsworth isolates are highlighted. Nodes are labeled with bootstrap support values. Scale bar, substitutions per site. All Wadsworth *bont* genes cluster by the subtypes initially assigned with the mapping and gene assembly method except two truncated sequences (gray text).

SUPPLEMENTARY FIGURE S3
Clostridium species supermatrix and supertree phylogenies. A multigene phylogeny for 225 isolates was generated from an alignment of 264 single-copy orthologues by IQ-TREE2 (61) under an LG+F+I+R4 substitution model. A supertree from 900 protein trees for the same isolates was generated by Astral-Pro. Trees were visualized in R (67) with the ggtree (66) package. Tips are colored by the species assigned by MASH distances from the SNP-based analysis. The tip points for isolates in the CDC_67190, Hall, and Kyoto reference groupings assigned by the SNP-based analysis are colored to highlight their paraphyletic nature.

SUPPLEMENTARY FIGURE S4
Heatmap of orthogroup presence / absence data for *Clostridium* isolates. Isolates (columns) clustered by the presence (black) and absence (white) of orthogroups (rows). The heatmap was generated by the ComplexHeatmap (90, 91) package in R (67). The distinct patterns of orthogroups reveals good agreement with species designations assigned by SNP-based analyses.

References

1. Blaschek HP. CLOSTRIDIUM | introduction In: CA Batt and ML Tortorello, editors. Encyclopedia of food microbiology (second edition). Oxford: Academic Press (2014). 444–8.

2. Collins MD, East AK. Phylogeny and taxonomy of the food-borne pathogen Clostridium botulinum and its neurotoxins. *J Appl Microbiol*. (1998) 84:5–17. doi: 10.1046/j.1365-2672.1997.00313.x

3. Hill KK, Smith TJ, Helma CH, Ticknor LO, Foley BT, Svensson RT, et al. Genetic diversity among botulinum neurotoxin-producing clostridial strains. *J Bacteriol*. (2007) 189:818–32. doi: 10.1128/JB.01180-06

4. Williamson CHD, Vazquez AJ, Hill K, Smith TJ, Nottingham R, Stone NE, et al. Differentiating botulinum neurotoxin-producing Clostridia with a simple, multiplex PCR assay. *Appl Environ Microbiol*. (2017) 83:e00806–17. doi: 10.1128/AEM.00806-17

5. Smith T, CHD W, Hill K, Sahl J, Keim P. Botulinum neurotoxin-producing bacteria. Isn't it time that we called a species a species? *mBio*. (2018) 9:e01469–18. doi: 10.1128/mBio.01469-18

6. Centers for Disease Control and Prevention. National botulism surveillance. Available online at: https://www.cdc.gov/botulism/php/national-botulism-surveillance/2021.html

7. Bintsis T. Foodborne pathogens. *AIMS Microbiology*. (2017) 3:529–63. doi: 10.3934/microbiol.2017.3.529

8. Smith TJ, Tian R, Imanian B, Williamson CHD, Johnson SL, Daligault HE, et al. Integration of complete plasmids containing *Bont* genes into chromosomes of *Clostridium parabotulinum, Clostridium sporogenes*, and *Clostridium argentinense*. *Toxins*. (2021) 13:473. doi: 10.3390/toxins13070473

9. Dong M, Masuyer G, Stenmark P. Botulinum and tetanus neurotoxins. *Annu Rev Biochem*. (2019) 88:811–37. doi: 10.1146/annurev-biochem-013118-111654

10. Benoit RM. Botulinum neurotoxin diversity from a gene-centered view. *Toxins*. (2018) 10:310. doi: 10.3390/toxins10080310

11. Sonnabend O, Sonnabend W, Heinzle R, Sigrist T, Dirnhofer R, Krech U. Isolation of Clostridium botulinum type G and identification of type G botulinal toxin in humans: report of five sudden unexpected deaths. *J Infect Dis*. (1981) 143:22–7. doi: 10.1093/infdis/143.1.22

12. Zanetti G, Azarnia Tehran D, Pirazzini M, Binz T, Shone CC, Fillo S, et al. Inhibition of botulinum neurotoxins interchain disulfide bond reduction prevents the peripheral neuroparalysis of botulism. *Biochem Pharmacol*. (2015) 98:522–30. doi: 10.1016/j.bcp.2015.09.023

13. Peck MW, Smith TJ, Anniballi F, Austin JW, Bano L, Bradshaw M, et al. Historical perspectives and guidelines for botulinum neurotoxin subtype nomenclature. *Toxins*. (2017) 9:38. doi: 10.3390/toxins9010038

14. Hill K.K., Smith T.J., Genetic diversity within Clostridium botulinum serotypes, botulinum neurotoxin gene clusters and toxin subtypes, in Botulinum neurotoxins, Rummel A., Binz T., Editors. (2013), Springer: Berlin Heidelberg. p. 1–20.

15. Macdonald Thomas E, Helma CH, Shou Y, Valdez YE, Ticknor LO, Foley BT, et al. Analysis of *Clostridium botulinum* serotype E strains by using multilocus sequence typing, amplified fragment length polymorphism, variable-number tandem-repeat analysis, and botulinum neurotoxin gene sequencing. *Appl Environ Microbiol*. (2011) 77:8625–34. doi: 10.1128/AEM.05155-11

16. Skarin H, Håfström T, Westerberg J, Segerman B. *Clostridium botulinum* group III: a group with dual identity shaped by plasmids, phages and mobile elements. *BMC Genomics*. (2011) 12:185. doi: 10.1186/1471-2164-12-185

17. Smith TJ, Hill KK, Xie G, Foley BT, Williamson CHD, Foster JT, et al. Genomic sequences of six botulinum neurotoxin-producing strains representing three clostridial species illustrate the mobility and diversity of botulinum neurotoxin genes. *Infect Genet Evol*. (2015) 30:102–13. doi: 10.1016/j.meegid.2014.12.002

18. Smith TJ, Williamson CHD, Hill KK, Johnson SL, Xie G, Anniballi F, et al. The distinctive evolution of *orfX Clostridium parabotulinum* strains and their botulinum neurotoxin type a and F gene clusters is influenced by environmental factors and gene interactions via Mobile genetic elements. *Front Microbiol*. (2021) 12:566908. doi: 10.3389/fmicb.2021.566908

19. Centers for Disease Control and Prevention. Federal select agent program. Atlanta, GA: Centers for Disease Control and Prevention (2024).

20. Lonati D, Schicchi A, Crevani M, Buscaglia E, Scaravaggi G, Maida F, et al. Foodborne botulism: clinical diagnosis and medical treatment. *Toxins*. (2020) 12:509. doi: 10.3390/toxins12080509

21. Machamer JB, Vazquez-Cintron EJ, O'Brien SW, Kelly KE, Altvater AC, Pagarigan KT, et al. Antidotal treatment of botulism in rats by continuous infusion with 3,4-diaminopyridine. *Mol Med*. (2022) 28:61. doi: 10.1186/s10020-022-00487-4

22. Torgeman A, Diamant E, Dor E, Schwartz A, Baruchi T, Ben David A, et al. A rabbit model for the evaluation of drugs for treating the chronic phase of botulism. *Toxins*. (2021) 13:679. doi: 10.3390/toxins13100679

23. Centers for Disease Control and Prevention. Botulism (Clostridium botulinum) 2011 case definition. Atlanta, GA: Centers for Disease Control and Prevention (2021).

24. Gladney L, Halpin JL, Lúquez C. Genomic characterization of strains from a cluster of infant botulism type a in a small town in Colorado, United States. *Front Microbiol*. (2021) 12:688240. doi: 10.3389/fmicb.2021.688240

25. Rosen HE, Kimura AC, Crandall J, Poe A, Nash J, Boetzer J, et al. Foodborne botulism outbreak associated with commercial nacho cheese sauce from a Gas Station market. *Clin Infect Dis*. (2019) 70:1695–700. doi: 10.1093/cid/ciz479

26. Bergeron G, Latash J, da Costa-Carter CA, Egan C, Stavinsky F, Kileci JA, et al. Notes from the field: botulism outbreak associated with home-canned peas - new York City, 2018. *MMWR Morb Mortal Wkly Rep*. (2019) 68:251–2. doi: 10.15585/mmwr.mm6810a5

27. Halpin JL, Foltz V, Dykes JK, Chatham-Stephens K, Lúquez C. Clostridium botulinum type B isolated from a wound botulism case due to injection drug use resembles other local strains originating from Hawaii. *Front Microbiol*. (2021) 12:678473. doi: 10.3389/fmicb.2021.678473

28. Hoyt KM, Barr JR, Hopkins AO, Dykes JK, Lúquez C, Kalb SR. Validation of a clinical assay for botulinum neurotoxins through mass spectrometric detection. *J Clin Microbiol*. (2024) 62:e01629–3. doi: 10.1128/jcm.01629-23

29. Davis S, Kelly-Cirino C, Cirino N, Hannett G, Musser K, Egan C. A 10 year analysis of the use of multiplex real-time PCR screening for botulinum neurotoxin-producing Clostridium species. *J Bacteriol Mycol*. (2016) 3:1030.

30. Centurioni DA, Egan CT, Perry MJ. Current developments in diagnostic assays for laboratory confirmation and investigation of botulism. *J Clin Microbiol*. (2022) 60:e0013920. doi: 10.1128/jcm.00139-20

31. Halpin JL, Dykes JK, Katz L, Centurioni DA, Perry MJ, Egan CT, et al. Molecular characterization of Clostridium botulinum Harboring the bont/B7 gene. *Foodborne Pathog Dis*. (2019) 16:428–33. doi: 10.1089/fpd.2018.2600

32. Dickinson MC, Wirth SE, Baker D, Kidney AM, Mitchell KK, Nazarian EJ, et al. Implementation of a high-throughput whole genome sequencing approach with the goal of maximizing efficiency and cost effectiveness to improve public health. *Microbiol Spectr*. (2024) 12:e0388523. doi: 10.1128/spectrum.03885-23

33. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol*. (2016) 17:132. doi: 10.1186/s13059-016-0997-x

34. Ghignone S. BBTools. (2020). Available online at: https://github.com/sghignone/BBTools.

35. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. (2014) 30:2114–20. doi: 10.1093/bioinformatics/btu170

36. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. (2013). doi: 10.48550/arXiv.1303.3997

37. Broad Institute. Picard toolkit GitHub repository (2019).

38. Xiang G, Giardine B, An L, Sun C, Keller CA, Heuston EF, et al. Snapshot: a package for clustering and visualizing epigenetic history during cell differentiation. *BMC Bioinformatics*. (2023) 24:102. doi: 10.1186/s12859-023-05223-1

39. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. (2009) 25:2078–9. doi: 10.1093/bioinformatics/btp352

40. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. *GigaScience*. (2021) 10:giab008. doi: 10.1093/gigascience/giab008

41. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom*. (2016) 2:e000056. doi: 10.1099/mgen.0.000056

42. Seemann T. Source code for snp-dists software (0.6.2) Zenodo (2018).

43. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*. (2015) 31:1674–6. doi: 10.1093/bioinformatics/btv033

44. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. (2009) 10:421. doi: 10.1186/1471-2105-10-421

45. Seemann T. Shovill: faster SPAdes assembly of Illumina reads (2017). Available at: https://github.com/tseemann/shovill

46. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb Genom*. (2021) 7:000685. doi: 10.1099/mgen.0.000685

47. Schwengers O. Bakta database Zenodo (2025).

48. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. (2015) 31:3210–2. doi: 10.1093/bioinformatics/btv351

49. Team, B. BUSCO lineage dataset: clostridia_odb10. Geneva: EZLab, University of Geneva (2020).

50. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. (2013) 29:1072–5. doi: 10.1093/bioinformatics/btt086

51. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with kraken 2. *Genome Biol*. (2019) 20:257. doi: 10.1186/s13059-019-1891-0

52. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom*. (2018) 4:e000206. doi: 10.1099/mgen.0.000206

53. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. (2019) 20:238. doi: 10.1186/s13059-019-1832-y

54. Seemann T. mlst. (2025); Available online at: https://github.com/tseemann/mlst (Accessed October 2025).

55. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, et al. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology*. (2012) 158:1005–15. doi: 10.1099/mic.0.055459-0

56. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res*. (2018) 3:124. eng. No competing interests were disclosed. doi: 10.12688/wellcomeopenres.14826.1

57. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. (2012) 28:3150–2. doi: 10.1093/bioinformatics/bts565

58. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. (2006) 22:1658–9. doi: 10.1093/bioinformatics/btl158

59. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. (2002) 30:3059–66. doi: 10.1093/nar/gkf436

60. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. (2013) 30:772–80. doi: 10.1093/molbev/mst010

61. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol*. (2020) 37:1530–4. doi: 10.1093/molbev/msaa015

62. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol*. (2018) 35:518–22. doi: 10.1093/molbev/msx281

63. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. (2017) 14:587–9. doi: 10.1038/nmeth.4285

64. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using gubbins. *Nucleic Acids Res*. (2015) 43:e15–5. doi: 10.1093/nar/gku1196

65. Zhang C, Mirarab S. ASTRAL-pro 2: ultrafast species tree reconstruction from multi-copy gene family trees. *Bioinformatics*. (2022) 38:4949–50. doi: 10.1093/bioinformatics/btac620

66. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. Ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol*. (2017) 8:28–36. doi: 10.1111/2041-210X.12628

67. Team RC. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing (2023).

68. Halpin Jessica L, Hill K, Johnson SL, Bruce DC, Shirey TB, Dykes JK, et al. Finished whole-genome sequences of two Clostridium botulinum type a(B) isolates. *Genome Announc*. (2017) 5:e00381–17. doi: 10.1128/genomeA.00381-17

69. Brunt J, van Vliet AHM, Carter AT, Stringer SC, Amar C, Grant KA, et al. *Diversity of the genomes and neurotoxins of strains of Clostridium botulinum* group I and *Clostridium sporogenes* associated with foodborne, infant and wound botulism. *Toxins*. (2020) 12:586. doi: 10.3390/toxins12090586

70. Kalb SR, Baudys J, Rees JC, Smith TJ, Smith LA, Helma CH, et al. De novo subtype and strain identification of botulinum neurotoxin type B through toxin proteomics. *Anal Bioanal Chem*. (2012) 403:215–26. doi: 10.1007/s00216-012-5767-3

71. Lúquez C, Bianco MI, de Jong LI, Sagua MD, Arenas GN, Ciccarelli AS, et al. Distribution of botulinum toxin-producing clostridia in soils of Argentina. *Appl Environ Microbiol*. (2005) 71:4137–9. doi: 10.1128/AEM.71.7.4137-4139.2005

72. Sagua MD, Lúquez C, Barzola CP, Bianco MI, Fernández RA. Phenotypic characterization of Clostridium botulinum strains isolated from infant botulism cases in Argentina. *Rev Argent Microbiol*. (2009) 41:141–7.

73. Williamson CHD, Sahl JW, Smith TJ, Xie G, Foley BT, Smith LA, et al. Comparative genomic analyses reveal broad diversity in botulinum-toxin-producing Clostridia. *BMC Genomics*. (2016) 17:180. doi: 10.1186/s12864-016-2502-z

74. Douillard François P, Derman Y, Woudstra C, Selby K, Mäklin T, Dorner MB, et al. Genomic and phenotypic characterization of Clostridium botulinum isolates from an infant botulism case suggests adaptation signatures to the gut. *mBio*. (2022) 13:e02384–21. doi: 10.1128/mbio.02384-21

75. Mazuet C, Legeay C, Sautereau J, Ma L, Bouchier C, Bouvet P, et al. Diversity of group I and II Clostridium botulinum strains from France including recently identified subtypes. *Genome Biol Evol*. (2016) 8:1643–60. doi: 10.1093/gbe/evw101

76. Badagliacca P, Pomilio F, Auricchio B, Sperandii AF, di Provvido A, di Ventura M, et al. Type C/D botulism in the waterfowl in an urban park in Italy. *Anaerobe*. (2018) 54:72–4. doi: 10.1016/j.anaerobe.2018.07.010

77. Rasetti-Escargueil C, Lemichez E, Popoff MR. Public health risk associated with botulism as foodborne zoonoses. *Toxins*. (2020) 12:17. doi: 10.3390/toxins12010017

78. Guizelini CC, Lemos RAA, de Paula JLP, Pupin RC, Gomes DC, Barros CSL, et al. Type C botulism outbreak in feedlot cattle fed contaminated corn silage. *Anaerobe*. (2019) 55:103–6. doi: 10.1016/j.anaerobe.2018.11.003

79. Le Maréchal C, Hulin O, Macé S, Chuzeville C, Rouxel S, Poëzevara T, et al. A case report of a botulism outbreak in beef cattle due to the contamination of wheat by a roaming cat carcass: from the suspicion to the management of the outbreak. *Animals*. (2019) 9:1025. doi: 10.3390/ani9121025

80. Hannett George E, Stone WB, Davis SW, Wroblewski D. Biodiversity of Clostridium botulinum type E associated with a large outbreak of botulism in wildlife from Lake Erie and Lake Ontario. *Appl Environ Microbiol*. (2011) 77:1061–8. doi: 10.1128/AEM.01578-10

81. Lindström M, Kiviniemi K, Korkeala H. Hazard and control of group II (non-proteolytic) Clostridium botulinum in modern food processing. *Int J Food Microbiol*. (2006) 108:92–104. doi: 10.1016/j.ijfoodmicro.2005.11.003

82. Espelund M, Klaveness D. Botulism outbreaks in natural environments - an update. *Front Microbiol*. (2014) 5:287. doi: 10.3389/fmicb.2014.00287

83. Portinha IM, Douillard FP, Korkeala H, Lindström M. Sporulation strategies and potential role of the exosporium in survival and persistence of Clostridium botulinum. *Int J Mol Sci*. (2022) 23:754. doi: 10.3390/ijms23020754

84. Lynt RK, Kautter DA, Solomon HM. Differences and similarities among proteolytic and nonproteolytic strains of Clostridium botulinum types a, B, E and F: a review. *J Food Prot*. (1982) 45:466–75. doi: 10.4315/0362-028X-45.5.466

85. Cambridge JM, Blinkova AL, Salvador Rocha EI, Bode Hernández A, Moreno M, Ginés-Candelaria E, et al. Genomics of Clostridium taeniosporum, an organism which forms endospores with ribbon-like appendages. *PLoS One*. (2018) 13:e0189673. doi: 10.1371/journal.pone.0189673

86. Grenda T, Jarosz A, Sapała M, Stasiak K, Grenda A, Domaradzki P, et al. Molecular diversity of BoNT-producing Clostridia—a still-emerging and challenging problem. *Diversity*. (2023) 15:392. doi: 10.3390/d15030392

87. Tedersoo L, Albertsen M, Anslan S, Callahan B. Perspectives and benefits of high-throughput long-read sequencing in microbial ecology. *Appl Environ Microbiol*. (2021) 87:e00626–1. doi: 10.1128/AEM.00626-21

88. Gonzalez-Escalona N, Sharma SK. Closing Clostridium botulinum group I genomes using a combination of short- and long-reads. *Front Microbiol*. (2020) 11:239. doi: 10.3389/fmicb.2020.00239

89. Rambaut A. FigTree v1.4.4. Edinburgh: Institute of Evolutionary Biology, University of Edinburgh (2018).

90. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*. (2016) 32:2847–9. doi: 10.1093/bioinformatics/btw313

91. Gu Z. Complex heatmap visualization. *iMeta*. (2022) 1:e43. doi: 10.1002/imt2.43

92. Wickham H. ggplot2: Elegant graphics for data analysis. New York: Springer-Verlag (2016).

93. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. *Nat Biotechnol*. (2011) 29:24–6. doi: 10.1038/nbt.1754

94. Gu Z, Gu L, Eils R, Schlesner M, Brors B. Circlize implements and enhances circular visualization in R. *Bioinformatics*. (2014) 30:2811–2. doi: 10.1093/bioinformatics/btu393