Phylogenetic analysis of *Clostridium botulinum* type A by multi-locus sequence typing

Mark J. Jacobson,¹ Guangyun Lin,¹ Thomas S. Whittam²
and Eric A. Johnson¹

Correspondence
Eric A. Johnson
eajohnso@wisc.edu

¹Department of Bacteriology, Food Research Institute, University of Wisconsin-Madison, Madison, WI 53706, USA

²National Food Safety & Toxicology Center, Michigan State University, East Lansing, MI 48824-1314, USA

The genus *Clostridium* comprises a heterogeneous group of organisms for which the phylogeny and evolutionary relationships are poorly understood. The elucidation of these evolutionary relationships necessitates the use of experimental methods that can distinguish *Clostridium* lineages that are time and cost effective, and can be accurately and reproducibly employed in different laboratories. Multi-locus sequence typing (MLST) has been successfully used as a reproducible and discriminating system in the study of eukaryotic and prokaryotic evolutionary biology, and for strain typing of various bacteria. In this study, MLST was applied to evaluate the evolutionary lineages in the serotype A group of *Clostridium botulinum*. *C. botulinum* type A has recently been shown to produce multiple subtypes, suggesting that it is not monophyletic as previously reported, but comprises distinct lineages. For MLST analysis, we initially evaluated 14 housekeeping genes (*gapdh*, *tuf*, *sod*, *oppB*, *hsp60*, *dnaE*, *aroE*, *pta*, 23S *rDNA*, *aceK*, *rpoB*, 16S *rDNA*, *mdh* and *recA*) for amplification and sequence analysis. In the first phase of the analysis, 30 *C. botulinum* type A strains producing botulinum neurotoxin subtypes A1–A4 were examined. Results of this pilot study suggested that seven of the genes (*mdh*, *aceK*, *rpoB*, *aroE*, *hsp60*, *oppB* and *recA*) could be used for elucidation of evolutionary lineages and strain typing. These seven housekeeping genes were successfully applied for the elucidation of lineages for 73 *C. botulinum* type A strains, which resulted in 24 distinct sequence types. This strategy should be applicable to phylogenetic studies and typing of other *C. botulinum* serotypes and *Clostridium* species.

Received 17 January 2008

Revised 13 April 2008

Accepted 14 April 2008

INTRODUCTION

Clostridium botulinum produces a characteristic botulinum neurotoxin (BoNT), which is classified by the Centers for Disease Control and Prevention as one of the six highest-risk threat agents for bioterrorism ('Category A Agents') (Arnon *et al.*, 2001). BoNTs have traditionally been immunologically distinguished into seven serotypes

Abbreviations: BoNT, botulinum neurotoxin; HA, haemagglutinin; MLST, multi-locus sequence typing; NJ, neighbour-joining; ST, sequence type.

The GenBank accession numbers for genes *bont/a1*, *bont/a2*, *bont/a3*, *bont/a4*, *bont/b* and the silent *bont/b* are AF461539, AY953275, DQ185900, DQ185901, EU341304 and NCTC 2916, respectively. Accession numbers for sequences of the genes *rpoB*, *recA*, *oppB*, *mdh*, *hsp60*, *aceK* and *aroE* are EU372261–EU372269, EU372253–EU372260, EU372242–EU372252, EU372232–EU372241, EU372223–EU372231, EU372210–EU372222 and EU372197–EU372209, respectively.

A supplementary table of primers and a supplementary figure showing the genomic locations of the loci analysed are available with the online version of this paper.

(BoNT/A–G), among which BoNTs A, B, E and F are known to cause human botulism (Hatheway & Johnson, 1998; Smith & Sugiyama, 1988; Smith & Moryson, 1977). Type A is a representative of *C. botulinum* group I proteolytic strains as it demonstrates close similarity to proteolytic type F and B strains (Collins & Lawson, 1994). This is of particular importance and interest since it causes the most severe human botulism (Woodruff *et al.*, 1992), is considered to be the most significant bioterrorism threat and has been increasingly used as a pharmaceutical modality (Johnson *et al.*, 2005).

Recent findings have shown that BoNT/A has substantial sequence diversity and four subtypes have so far been identified (Arndt *et al.*, 2006; Kozaki *et al.*, 1995; Smith *et al.*, 2005). Such sequence variation is not limited to type A, as more than 46 subtypes exist for serotypes A–G (Gimenez & Gimenez, 1995; Smith *et al.*, 2005). With regard to the *bont/a* gene, it has been represented by *bont/a1* (NCBI accession number AF461539), *bont/a2* (AY953275), *bont/a3* (DQ185900) and *bont/a4* (DQ185901).

Previous studies have also identified extensive genotypic and phenotypic diversity in these strains (Hatheway & Johnson, 1998; Johnson & Bradshaw, 2001; Kozaki *et al.*, 1995; Smith *et al.*, 2005). To date, phylogenetic analysis and typing of *C. botulinum* by genetic methods has mainly focused on PFGE, ribotyping (rRNA analysis), amplified fragment length polymorphism (AFLP), randomly amplified polymorphic DNA analysis (RAPD) and repetitive element sequence-based PCR (Rep-PCR) (Lindstrom & Korkeala, 2006). Although these techniques have utility, they also have disadvantages. For instance, rRNA analysis focuses on a single locus only and may not represent the diversity of the genome in the species; and PFGE is based on determination of restriction digest polymorphisms which is labour intensive, may vary between different laboratories and is difficult to use for identifying phylogenetic lineages (Hill *et al.*, 2007; Johnson *et al.*, 2005; Noller *et al.*, 2003).

Therefore, a sequence-based system for assessing genetic relatedness among isolates would be useful in study of this pathogen. Multi-locus sequence typing (MLST) is a method currently being implemented in many laboratories as a means of determining the degree of evolutionary relatedness among various strains of bacterial and eukaryotic species (Gatei *et al.*, 2007; Jost *et al.*, 2006; Lacher *et al.*, 2007; Maiden, 2006; Vassileva *et al.*, 2006; Zadoks *et al.*, 2005). MLST was initially demonstrated in 1998 to be effective in studying the phylogeny of bacteria (Maiden *et al.*, 1998). Since this pioneering study, MLST has been shown to be a useful method for bacterial typing as it has broad applicability in both the range of organisms that can be studied and the breadth of practical and conceptual problems that can be addressed (Urwin & Maiden, 2003). MLST combines advances in high-throughput sequencing, population genetics and bioinformatics to provide a tool for the study of population and evolutionary biology of various organisms. In this study, we applied MLST analysis for assessing the genetic relatedness of *C. botulinum* type A strains.

METHODS

Bacterial strains and growth conditions. The *C. botulinum* strains included in this study (Table 1) are from E.A.J.'s laboratory strain collection. Cultures were grown in 10 ml TPGY medium (per litre: 50 g trypticase peptone, 5 g Bacto peptone, 4 g D-glucose, 20 g yeast extract, 1 g cysteine.HCl, pH 7.4) for 2 days at 37 °C.

Total genomic DNA isolation. Total genomic DNA was isolated from the *C. botulinum* strains by lysozyme and proteinase K treatment as described previously (Dineen *et al.*, 2003). DNA was then diluted to 50 ng µl⁻¹ and used for PCR amplification.

PCR amplification and sequencing. PCR amplifications were performed using the GeneAmp High Fidelity PCR System (Applied Biosystems). PCR cycles were as follows: 95 °C for 2 min, followed by 25 cycles of 95 °C for 1 min, an annealing step for 45 s, 72 °C for extension, followed by 1 cycle of 72 °C extension for 10 min. Annealing temperatures of the different primers are given in Table 2

and Supplementary Table S1, available with the online version of this paper. Extension time was dependent on the length of the amplification product (see Table S1); in general, a 1 min extension step was utilized to extend a 1 kb fragment. Following amplification, all PCR products were isolated using the PureLink PCR Purification kit (Invitrogen). Sequencing preparations were produced using conditions advised by the University of Wisconsin Biotechnology Center for the ABI PRISM BigDye Cycle Sequencing kit (Applied Biosystems). Sequencing analysis was performed at the University of Wisconsin Biotechnology Center and final sequencing results were analysed using the Vector NTI Suite Program (Invitrogen). Accession numbers for the resulting nucleotide sequences are as follows: *rpoB*, EU372261–EU372269; *recA*, EU372253–EU372260; *oppB*, EU372242–EU372252; *mdh*, EU372232–EU372241; *hsp60*, EU372223–EU372231; *aceK*, EU372210–EU372222; and *aroE*, EU372197–EU372209.

The genome sequence of *C. botulinum* type A ATCC 3502 from the Sanger Institute Genome Initiative (http://www.sanger.ac.uk/Projects/C_botulinum/) (Sebaihia *et al.*, 2007) was used to identify various loci for MLST analysis. Genes were chosen for MLST analyses mainly based on their utility in the studies of other bacterial species. Once the 14 genes were selected, the entire ORF of each was used to select primers for initial analysis.

In the initial analysis by MLST to assess its applicability to the *C. botulinum* group, 30 isolates (Table 1) were analysed by performing PCR on 14 genes (Table S1) which were sequenced using an overlapping gene sequencing approach. The primers used to amplify the complete coding frame and the internal primers used in subsequent sequencing reactions are listed in Table 2. After this analysis was complete, specific regions of all 14 genes were analysed for the 30 strains to determine which regions contained the optimum degree of divergence to allow for proper MLST analysis. While all 14 genes showed divergence, seven genes were chosen to facilitate further MLST analysis and enable the creation of subfamilies. Regions were evaluated and then entered into the Primer3 program (Rozen & Skaletsky, 2000) to amplify a 700–800 bp section (see Table 2 for primer sequence and gene-specific location). These primers were then used in the PCR amplification of designated gene products.

When final MLST primers had been designed, PCR reactions were performed on 73 isolates (Table 1). Most of these isolates were known type A strains from the A1, A2, A3, A4 and A(B) groups. *C. botulinum* type A neurotoxin was not used as a gene for MLST analysis in order to make the procedure as broad as possible and allow for potential analysis of *Clostridium* strains with other BoNT serotypes using the procedure.

Analysis of MLST data. Sequences were assembled from the resultant chromatograms using the ContigExpress program within Vector NTI (Invitrogen). For each of the seven loci, each sequence obtained was assigned a distinct allele number. Each isolate is defined by an allelic profile consisting of seven integers, which corresponds to the allele numbers at the seven loci of *recA*, *rpoB*, *oppB*, *hsp60*, *aceK*, *mdh* and *aroE*. The unique allelic profiles were assigned a sequence type (ST). The resulting STs were analysed using the program's sequence type analysis and recombination tests (START) (Jolley *et al.*, 2001) to organize the various data. Further analysis was conducted using MEGA3 (Kumar *et al.*, 2004) to identify relationships among the various strains. The final data were compiled and submitted for hosting at <http://pubmlst.org/cbotulinum/> (Jolley *et al.*, 2004).

Neighbour-joining (NJ) trees were constructed using the Kimura two-parameter model of nucleotide substitution with the MEGA3 software and the inferred phylogenies were each tested with 500 bootstrap replications. Phylogenetic network analysis was conducted with the SplitsTree 4 program (Huson & Bryant, 2006) using the neighbour-net algorithm (Bryant & Moulton, 2004) and untrans-

Table 1. List of strains used in MLST analysis

Strain no.*	Strain name†	<i>recA</i>	<i>rpoB</i>	<i>oppB</i>	<i>hsp60</i>	<i>aceK</i>	<i>mdh</i>	<i>aroE</i>	ST‡	Subtype§
1*	ATCC 3502	8	9	10	9	13	6	13	1	A1
2*	PHLS5	7	8	9	8	7	8	10	2	A2
3*	A33	7	8	9	8	6	10	9	4	A(B)
4*	A588	6	8	5	8	11	5	11	10	A(B)
5*	14842	1	1	1	1	1	1	1	5	A(B)
6*	14931	7	6	9	7	6	10	9	6	A(B)
7*	519	7	8	9	8	6	10	9	4	A(B)
8*	A4831	8	9	10	7	13	6	8	11	A1
9*	14843	1	1	1	1	1	1	1	5	A(B)
10*	14849	7	8	9	8	6	10	9	4	A(B)
11*	MDa10	7	8	9	8	6	10	9	4	A(B)
12*	14860	7	6	9	7	6	10	6	12	A(B)
13*	657Ba	5	2	4	5	2	2	4	7	A4
14*	657Ba	5	2	4	5	2	2	4	7	A4
15*	A2	7	8	9	8	6	10	9	4	A(B)
16*	CDC/A3	3	3	2	3	5	3	3	3	A3
17*	A4475	6	8	9	6	11	5	12	8	A1
18*	14862	7	8	9	8	6	10	9	4	A(B)
19*	14851	7	6	9	7	6	10	9	6	A(B)
20*	10758	4	7	6	9	8	8	7	13	A1
21*	14853	7	8	9	8	6	10	9	4	A(B)
22*	A3(S)	7	8	9	8	6	10	9	4	A(B)
23*	14844	1	1	1	1	1	1	1	5	A(B)
24*	14050	5	2	4	5	2	2	5	14	A(B)
25*	5311A	6	8	5	8	11	8	13	15	A(B)
26*	A7	8	9	10	9	13	6	13	1	A1
27*	A4567	7	8	9	8	6	10	9	4	A(B)
28*	A1	8	9	10	9	13	6	13	1	A1
29*	A661222	7	9	8	8	4	7	13	16	A1
30*	A207	2	4	3	2	3	4	2	17	A1
31	5328A	5	2	4	4	2	2	4	18	HA ⁻ /Orfx ⁺ A1
32	667A	7	8	9	8	6	10	9	4	A(B)
33	A4447	6	8	9	6	11	5	12	8	A1
34	A69	8	9	10	7	12	10	13	19	A1
35	A73	8	9	10	9	13	6	13	1	A1
36	A4	7	8	9	8	6	8	10	22	<i>sporo</i>
37	A3	8	9	7	9	13	6	13	21	A1
38	A4438	8	9	10	9	13	6	13	1	A1
39	14847	7	8	9	8	6	10	9	4	A(B)
40	14850	7	8	9	8	6	10	9	4	A(B)
41	726	8	9	10	9	13	6	13	1	A1
42	WHOA	8	9	10	9	9	6	13	9	A1
43	H Lysine	8	9	10	9	13	6	13	1	A1
44	603A	8	9	10	9	13	6	13	1	A(B)
45	547A	8	9	10	9	13	6	13	1	A1
46	10745	8	9	10	9	13	6	13	1	A(B)
47	A14	8	9	10	9	13	6	13	1	A1
48	14869	8	9	10	9	13	6	13	1	A1
49	Hall A1	8	9	10	9	13	6	13	1	A1
50	H formaldehyde	8	9	10	9	13	6	13	1	A1
51	A6	8	9	10	9	13	6	13	1	A1
52	AB Seattle	7	8	11	8	10	10	13	23	A1
53	599A	8	5	10	9	13	6	13	24	A1
54	14906	8	9	10	9	13	6	13	1	A1
55	14853	8	9	10	9	13	6	13	1	A1

Table 1. cont.

Strain no.*	Strain name†	<i>recA</i>	<i>rpoB</i>	<i>oppB</i>	<i>hsp60</i>	<i>aceK</i>	<i>mdh</i>	<i>aroE</i>	ST‡	Subtype§
56	14846	8	9	10	9	9	6	13	9	A1
57	PHLS2	7	8	9	8	7	8	10	2	A2
58	9289	8	9	10	9	13	6	13	1	A1
59	A172	8	9	10	9	13	6	13	1	A1
60	A109	8	9	10	7	13	6	8	11	A1
61	A16	8	9	10	9	13	6	13	1	A1
62	13060	8	9	10	9	13	6	13	1	A1
63	11A	8	9	10	9	9	6	13	9	A1
64	77A	8	9	10	9	13	6	13	1	A1
65	1690	8	9	10	9	13	6	13	1	A1
66	A15	8	9	10	9	13	6	13	1	A1
67	14854	7	8	9	8	6	10	9	4	A(B)
68	56A	8	9	10	9	13	6	13	1	A1
69	A3955	8	9	10	9	13	6	13	1	A1
70	A608	8	9	10	9	13	9	13	20	A1
71	A5	8	9	10	9	13	6	13	1	A1
72	A9	8	9	10	9	13	6	13	1	A1
73	19A	8	9	10	9	13	6	13	1	A1

*The initial 30 strains used in the housekeeping genes screening study.

†Strain names refer to Eric A. Johnson's laboratory's stocks.

‡ST, Sequence type.

§A1 has the BoNT/A1 gene (NCBI accession number AF461539), A2 has the BoNT/A2 gene (NCBI accession number AY953275), A(B) has BoNT/A1 gene and a silent BoNT/B gene (NCTC 2916), A3 has the BoNT/A3 gene (NCBI accession number DQ185900), and A4 has the BoNT/A4 gene (NCBI accession number DQ185901) and BoNT/B gene (NCBI accession number EU341304). HA⁻/Orf⁺ A1 produces BoNT/A1 but with an ORFX cluster. *sporo*, *Clostridium sporogenes*.

formed distances (p distance). The number of synonymous substitutions per synonymous site (d_s) and the number of nonsynonymous substitutions per nonsynonymous site (d_N) were estimated by the modified Nei–Gojobori method using MEGA3 (Kumar *et al.*, 2004). Allelic sequences were fitted to a nucleotide substitution model using the Datamonkey website. Single likelihood ancestor counting (SLAC) was used to fit a codon model to detect selection on individual codons (Pond & Frost, 2005). The SLAC method was also used to calculate the ratio of d_N to d_s and estimate the 95% confidence interval. The ϕ_w recombination test (Bruen *et al.*, 2006), as implemented by SplitsTree 4, was used to distinguish recurrent mutation from recombination in generating genotypic diversity.

RESULTS

Initial MLST studies

Initial data indicated the presence of genetic linkages based on BoNT subtypes, since several housekeeping genes had sequence variations consistent with BoNT/A subtype identity [A1, A2, A3, A4 and A(B)]. Fourteen housekeeping candidate genes (Table S1) were initially amplified and sequenced. These were: *aroE*, *tuf*, *hsp60*, *mdh*, *recA*, *aceK*, *rpoB*, 23S *rDNA*, *oppB*, *pta*, 16S *rDNA*, *dnaE*, *sod* and *gapdh*. These genes were selected mainly because of their

use in previous bacterial MLST studies. The genes were then amplified by PCR, and sequenced for 30 *C. botulinum* type A strains (Table 1 and Supplementary Table S1). After analysing the sequence data, seven housekeeping genes (*rpoB*, *mdh*, *aroE*, *hsp60*, *aceK*, *oppB* and *recA*) were selected because of their inherent genetic variability and placement in the *C. botulinum* genome, so that the MLST scheme would have an approximately even distribution throughout the genome (Supplementary Fig. S1). Following this strategy, the seven housekeeping genes of 73 strains of *C. botulinum* and related clostridia were sequenced with appropriate primers (Tables 1 and 2).

Creation of MLST ST profiles

MLST analysis of 73 *C. botulinum* and related clostridia yielded locus frequencies that ranged from 8 to 13 alleles per locus (Table 1); 24 unique profile patterns or STs were identified. ST-1 encompassed 29 strains including subtype A1 and A(B) strains; ST-4 included 13 A(B) strains; ST-5 included 3 A(B) strains and ST-9 included 3 A1 strains. STs 2, 6, 7, 8 and 11 were represented by 2 strains, each covering a wide array of subtypes including A2, A(B), A4 and A1 (Table 1).

Table 2. Details of primer sequences and the PCR conditions of the seven housekeeping genes used in MLST analysis

F, forward; R, reverse; *aroE*, shikimate dehydrogenase gene; *hsp60*, heat-shock protein gene; *mdh*, malate dehydrogenase gene; *recA*, RecA gene; *aceK*, isocitrate dehydrogenase gene; *rpoB*, RNA polymerase subunit B gene; *oppB*, oligopeptide/dipeptide ABC transporter gene.

Gene name and primer sequences	Annealing temp. (°C)	Product length (bp)	Product location
<i>rpoB</i> F 5'-ATG GTA CCC CTA AAG GAT TTA AGC A-3' R 5'-CGA GAT AGC GAT GGA GGA ATA GAT A-3'	50	790	3717117–3717906
<i>mdh</i> F 5'-TTC TCA GTA ACA ATA GCT GGT GGT G-3' R 5'-ATC TCC CTT ATC AAC CAC ATA TCC A-3'	48	693	329209–329901
<i>aroE</i> F 5'-TTG TTT CAA TTT ATC GTC CTC CTT T-3' R 5'-TTC AGA GTC ACC AGA AAT ACA CAA TAA-3'	48	749	2678553–2679301
<i>hsp60</i> F 5'-TTG GTG GTG TAT TTT CTT TTT CTG G-3' R 5'-AGC ACC AGG ATT TGG AGA TAG AAG-3'	48	767	3508084–3508850
<i>aceK</i> F 5'-TAG GAG CAG ATA TTA ATT GGC ATG T-3' R 5'-ACA AAT ACT TTT TCT ATA GCA TTT TCT-3'	44	835	500538–500562
<i>oppB</i> F 5'-GGT AAC TTA TGT ATG TTA GAA GTT AAA AAT-3' R 5'-TTA CTC TTT TAG AAG GAT GAT TTA TAG GAA-3'	44	811	1484690–1485500
<i>recA</i> F 5'-TTT GCA TTT TCT CTT CCT TGT CCT A-3' R 5'-AGC ATT AGG TAT AGG GGG AGT TCC T-3'	48	771	2532760–2533530

Evaluation of MLST ST profiles

To compare the level of sequence divergence as measured by MLST, we constructed a NJ dendrogram showing the genetic relatedness among the 24 STs (Fig. 1a). Bootstrap analysis classified the STs into four main groups with genetic distance greater than 0.01. The most divergent strains were in ST-5 and ST-17, which was surprising since ST-5 comprised a set of A(B) strains and ST-17 was a BoNT/A1-producing strain. These strains were expected to be more closely related to those possessing the same subtype of BoNT. Their relatively large differentiation from the other strains may indicate that they had split from the main family of *C. botulinum* early in the overall evolution of the species. The fact that they have the same neurotoxin sequence as the other strains supports the view that the evolution of BoNT is not linked to the evolution of the species *in toto*. Four STs (3, 7, 14 and 18) formed a separate group; ST-7, ST-14 and ST-18 were closely related, which was supported by 100 % bootstrap analysis. This cluster of STs is interesting since they are all composed of strains with different BoNT sequences; these are BoNT/A(B), BoNT/A4 and BoNT/A1 with an A2 cluster, respectively. The 18 remaining strains formed a closely related group with two subgroups; the largest group of nine STs (6, 12, 4, 2, 22, 23, 15, 8 and 10) had diverse BoNT profiles, while the A2 and A(B) groups had 90 % bootstrap support and seven STs (19, 9, 11, 1, 24, 20 and 21), largely composed of BoNT/A1-producing strains, had 88 % bootstrap support (Fig. 1a).

Based on these data, we sought to determine if sources of genetic variation other than classic evolution could have caused the observed variation among the genes. Specifically, we evaluated the role that recombination had in the genetic relationships among sequence types. To examine this, we constructed and examined a phylogenetic network (Fig. 1b) based on SplitsTree analysis (Bryant & Moulton, 2004). This analysis does not force the sequence data into a bifurcating tree and allows for numerous parallel paths indicative of the presence of phylogenetic incompatibilities in the divergence of STs. Such incompatibilities could arise from recombination or recurrent mutations in the MLST loci. To detect recombination, we used the ϕ_w test, which has been shown in SplitsTree analysis to discriminate between recurrent mutations and recombination in a variety of circumstances (Bryant *et al.*, 2006). In relation to the concatenated sequences of the 241 STs, there were 151 informative sites and the ϕ_w test was found to show statistically significant evidence of recombination ($P < 0.001$). Interestingly, although there is significant recombination among very clearly related STs, there is no evidence of recombination among the more distantly related strains matched by STs 3, 5, 7, 14, 17 and 18 (Fig. 1b). This result is intriguing considering the observation that there is a low degree of similarity between the strains in these STs at the BoNT level. In fact, all but two of these STs have their own unique BoNT profile, as only ST-5 and ST-14 have A(B) profiles. The remainder of the STs all represent unique strains with different neurotoxin profiles ranging from strains with a BoNT/A1 gene to more distinct

Fig. 1. Genetic relatedness among the 24 STs based on MLST sequence differences. (a) Genetic relatedness based on NJ dendrogram without recombination. (b) Tree-like neighbour-net for the 24 STs based on SplitsTree analysis that allows for recombination. In both trees, genetic distance is measured by *p*-distance from pairwise comparison of sequence differences.

strains, such as those that have a BoNT/A3, BoNT/A4 or a BoNT/A1 gene with an A2 cluster.

DISCUSSION

MLST was utilized to study the genetic diversity and phylogeny of *C. botulinum* serotype A, in which four distinct subtypes of BoNT have recently been elucidated (Arndt *et al.*, 2006). The genes chosen for MLST analysis in this study were selected because of their utility in previous MLST studies with other bacterial species and their distinct distribution in the *C. botulinum* chromosome. Initial analysis of 14 candidate genes was combined with the genomic location of each gene estimated from the genome sequence of *C. botulinum* ATCC 3502 to determine seven genes that provided a representation of the genetic relatedness of the strains on a genome scale. The loci used in the final analysis were *hsp60*, *rpoB*, *oppB*, *mdh*, *recA*, *aceK* and *aroE*.

Based on MLST studies of other pathogenic bacterial species, it is common to include a virulence gene as one of the loci used to determine genetic relationships within a bacterial group. In our analysis, this strategy was not followed since the BoNT gene does not have sufficiently distinct loci that could be used for MLST analysis. Further, it would limit the applicability of the MLST system to other *C. botulinum* serotypes and neurotoxicogenic clostridia of different species. At this time, five alleles of this gene have been described [A1, A2, A3, A4 and A1 in A(B) strains], which would not provide adequate genetic variation for MLST analysis.

There are several other genes associated with BoNT that could be used to study genetic variation, but were again not chosen for analysis; these include genes within the distinct toxin gene clusters in type A (Jacobson *et al.*, 2008). There are two primary types of neurotoxin clusters in *C. botulinum* type A (Jacobson *et al.*, 2008). The difference between the two basic clusters is substantial, as three of the four genes consist of either a set of haemagglutinin (HA) genes or a set of genes of unknown function called 'orf's'. The only gene that is common among the clusters is the nontoxic nonhaemagglutinin gene, *ntnh*. However, *ntnh* would pose problems if used as an MLST locus since two copies are present in strains with bivalent designations, such as A(B). Therefore, we selected seven housekeeping genes spread evenly across the genome for the MLST analysis and did not include the BoNT gene or genes within the BoNT clusters.

Several interesting results emerged from this study, specifically that there appears to be a significant amount of genetic association between the A1 and A(B) strains, since there was intersection between the two types in several ST groups, particularly ST-1 (Table 1). There are several possible hypotheses to explain this diversity. The most likely is that the evolution of the BoNT/A genes and their respective gene clusters differs significantly from the evolution of the species *in toto*. This hypothesis is supported by other observations related to evolution of BoNT/A and toxin gene clusters. Since BoNT acts solely on the nervous systems of higher eukaryotes, the selective pressure for its evolution is enigmatic. The BoNT gene and the structure of BoNT/A and BoNT/B have a highly mosaic composition (Arndt *et al.*, 2006). The composition of the toxin gene clusters may have resulted from acquisition of eukaryotic genes, e.g. by viral infection, and from gene transfer that occurred during evolution of *C. botulinum* (DasGupta, 2006; Johnson & Bradshaw, 2001). This would explain how members of this species, which differ widely in genetic and phenotypic properties, possess BoNT/A genes and an associated protein cluster that are highly conserved. It would also explain why certain *C. botulinum* isolates such as those possessing the BoNT/A3 and BoNT/A4 genes have such a low degree of relatedness compared to strains possessing BoNT/A1 gene, while the BoNT/A genes have relatively high homology of 80–90 %.

The MLST analysis also supports the theory that recombination was a prominent driving force contributing to the relatedness of the strains tested in this study. However, there were a few outliers (STs 3, 5, 7, 14, 17 and 18). The uniqueness of the BoNT sequences of most of these strains suggests that they may have been geographically or ecologically isolated, meaning that there was limited interaction with other strains within the *C. botulinum* species. This could in turn explain why the BoNT profiles of these strains are unique, as it is possible that they may have evolved to different degrees with respect to eukaryotic gene acquisition, gene transfer and recombination. This hypothesis is supported by the unique ST-3 pattern of BoNT/A3 strains, as only one outbreak of botulism has been attributed to BoNT/A3, whereas BoNT/A1 and BoNT/A2 have been involved in numerous botulism outbreaks.

Another set of interesting outliers in this study were ST-5 and ST-17. At this time, there is little known about the strains possessing these STs. The three strains in ST-5 are A(B) strains, while the one ST-17 is an A1 strain: each is distinct from other analysed strains. Initial experiments performed in our laboratory have indicated that the ST-5 bacteria possess a unique neurotoxin cluster arrangement compared to the standard clusters observed in the literature (Jacobson *et al.*, 2008).

Additionally, strain 5328A (ST-18) has a BoNT/A1 gene associated with an *orfx* cluster arrangement that is similar to that seen in *C. botulinum* strains that have the BoNT/A2 gene. This type of cluster arrangement is present in about half of BoNT/A1 strains, but only when a BoNT/B silent gene is also present and associated with a HA cluster. Strain 5328A lacks the BoNT/B silent gene cluster and has only a BoNT/A1 cluster. This appears to be unusual but becomes more revealing when compared with an A(B) strain and the BoNT/A4 strain, which also has a BoNT/B gene cluster. The implications of both the lack of a cryptic BoNT/B and its relatedness to these other strains have yet to be completely explained and will require further analysis.

In summary, MLST is a nucleotide sequence-based approach with many advantages for subtyping and phylogenetic analysis of various organisms. We show in this study that MLST is an efficient and discriminatory method for strain differentiation and phylogenetic analysis of *C. botulinum*. Twenty-four unique ST lineages were identified from analysis of 73 *C. botulinum* type A strains. In future studies, we will expand this MLST procedure to other BoNT-producing bacteria including serotypes B–G. This will be of value in further elucidating and understanding the genetic relatedness in this diverse species. Lastly, this strategy may also be applicable to phylogenetic studies of other *Clostridium* species.

ACKNOWLEDGEMENTS

This work was sponsored by the NIH/NIAID Regional Center of Excellence for Biodefense and Emerging Infectious Diseases Research

(RCE) Programs. The authors wish to acknowledge membership within, and support from, the Pacific South-West Regional Center of Excellence grant U54 AI065359. Additional funding for this project was provided by the NIAID cooperative agreement U01 AI056493 and the University of Wisconsin-Madison. We also thank Pat Schloss for discussion of MLST.

REFERENCES

- Arndt, J. W., Jacobson, M. J., Abola, E. E., Forsyth, C. M., Tepp, W. H., Marks, J. D., Johnson, E. A. & Stevens, R. C. (2006). A structural perspective of the sequence variability within botulinum neurotoxin subtypes A1–A4. *J Mol Biol* **362**, 733–742.
- Arnon, S. S., Schechter, R., Inglesby, T. V., Henderson, D. A., Bartlett, J. G., Ascher, M. S., Eitzen, E., Fine, A. D., Hauer, J. & other authors (2001). Botulinum toxin as a biological weapon: medical and public health management. *JAMA* **285**, 1059–1070.
- Bruen, T. C., Philippe, H. & Bryant, D. (2006). A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172**, 2665–2681.
- Bryant, D. & Moulton, V. (2004). Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol* **21**, 255–265.
- Collins, M. D. & Lawson, P. A. (1994). The phylogeny of the genus *Clostridium*: proposal of five new genera and eleven new species combinations. *Int J Syst Bacteriol* **44**, 812–826.
- DasGupta, B. R. (2006). Botulinum neurotoxins: perspective on their existence and as polyproteins harboring viral proteases. *J Gen Appl Microbiol* **52**, 1–8.
- Dineen, S. S., Bradshaw, M. & Johnson, E. A. (2003). Neurotoxin gene clusters in *Clostridium botulinum* type A strains: sequence comparison and evolutionary implications. *Curr Microbiol* **46**, 345–352.
- Gatei, W., Das, P., Dutta, P., Sen, A., Cama, V., Lal, A. A. & Xiao, L. (2007). Multilocus sequence typing and genetic structure of *Cryptosporidium hominis* from children in Kolkata, India. *Infect Genet Evol* **7**, 197–205.
- Gimenez, D. F. & Gimenez, J. A. (1995). The typing of botulinum neurotoxins. *Int J Food Microbiol* **27**, 1–9.
- Hatheway, C. L. & Johnson, E. A. (1998). *Clostridium*: the spore-bearing anaerobes. In *Topley & Wilson's Microbiology and Infections*, 9th edn, vol. 2, pp. 731–782, Systematic Bacteriology. Edited by L. Collier, A. Balows & M. Sussman. London: Arnold.
- Hill, K. K., Smith, T. J., Helma, C. H., Ticknor, L. O., Foley, B. T., Svensson, R. T., Brown, J. L., Johnson, E. A., Smith, L. A. & other authors (2007). Genetic diversity among botulinum neurotoxin-producing clostridial strains. *J Bacteriol* **189**, 818–832.
- Huson, D. H. & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* **23**, 254–267.
- Jacobson, M. J., Lin, G., Raphael, B., Andreadis, J. & Johnson, E. A. (2008). Analysis of neurotoxin cluster genes in *Clostridium botulinum* strains producing botulinum neurotoxin serotype A subtypes. *Appl Environ Microbiol* **74**, 2778–2786.
- Johnson, E. A. & Bradshaw, M. (2001). *Clostridium botulinum* and its neurotoxins: a metabolic and cellular perspective. *Toxicon* **39**, 1703–1722.
- Johnson, E. A., Tepp, W. H., Bradshaw, M., Gilbert, R. J., Cook, P. E. & McIntosh, E. D. (2005). Characterization of *Clostridium botulinum* strains associated with an infant botulism case in the United Kingdom. *J Clin Microbiol* **43**, 2602–2607.
- Jolley, K. A., Feil, E. J., Chan, M. S. & Maiden, M. C. (2001). Sequence type analysis and recombinational tests (START). *Bioinformatics* **17**, 1230–1231.

- Jolley, K. A., Chan, M. S. & Maiden, M. C. (2004). mlstdbNet – distributed multi-locus sequence typing (MLST) databases. *BMC Bioinformatics* 5, 86.
- Jost, B. H., Trinh, H. T. & Songer, J. G. (2006). Clonal relationships among *Clostridium perfringens* of porcine origin as determined by multilocus sequence typing. *Vet Microbiol* 116, 158–165.
- Kozaki, S., Nakaue, S. & Kamata, Y. (1995). Immunological characterization of the neurotoxin produced by *Clostridium botulinum* type A associated with infant botulism in Japan. *Microbiol Immunol* 39, 767–774.
- Kumar, S., Tamura, K. & Nei, M. (2004). MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* 5, 150–163.
- Lacher, D. W., Steinsland, H., Blank, T. E., Donnenberg, M. S. & Whittam, T. S. (2007). Molecular evolution of typical enteropathogenic *Escherichia coli*: clonal analysis by multilocus sequence typing and virulence gene allelic profiling. *J Bacteriol* 189, 342–350.
- Lindstrom, M. & Korkeala, H. (2006). Laboratory diagnostics of botulinum. *Clin Microbiol Rev* 19, 298–314.
- Maiden, M. C. (2006). Multilocus sequence typing of bacteria. *Annu Rev Microbiol* 60, 561–588.
- Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K. & other authors (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* 95, 3140–3145.
- Noller, A. C., McEllistrem, M. C., Pacheco, A. G., Boxrud, D. J. & Harrison, L. H. (2003). Multilocus variable-number tandem repeat analysis distinguishes outbreak and sporadic *Escherichia coli* O157:H7 isolates. *J Clin Microbiol* 41, 5389–5397.
- Pond, S. L. & Frost, S. D. (2005). Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics* 21, 2531–2533.
- Rozen, S. & Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132, 365–386.
- Sebaihia, M., Peck, M. W., Minton, N. P., Thomson, N. R., Holden, M. T., Mitchell, W. J., Carter, A. T., Bentley, S. D., Mason, D. R. & other authors (2007). Genome sequence of a proteolytic (group I) *Clostridium botulinum* strain Hall A and comparative analysis of the clostridial genomes. *Genome Res* 17, 1082–1092.
- Smith, G. R. & Moryson, C. J. (1977). A comparison of the distribution of *Clostridium botulinum* in soil and in lake mud. *J Hyg (Lond)* 78, 39–41.
- Smith, L. D. S. & Sugiyama, H. (1988). *Botulism. The Organism, its Toxin, the Disease*. Springfield, IL: Charles C. Thomas.
- Smith, T. J., Lou, J., Geren, I. N., Forsyth, C. M., Tsai, R., Laporte, S. L., Tepp, W. H., Bradshaw, M., Johnson, E. A. & other authors (2005). Sequence variation within botulinum neurotoxin serotypes impacts antibody binding and neutralization. *Infect Immun* 73, 5450–5457.
- Urwin, R. & Maiden, M. C. (2003). Multi-locus sequence typing: a tool for global epidemiology. *Trends Microbiol* 11, 479–487.
- Vassileva, M., Torii, K., Oshimoto, M., Okamoto, A., Agata, N., Yamada, K., Hasegawa, T. & Ohta, M. (2006). Phylogenetic analysis of *Bacillus cereus* isolates from severe systemic infections using multi-locus sequence typing scheme. *Microbiol Immunol* 50, 743–749.
- Woodruff, B. A., Griffin, P. M., McCroskey, L. M., Smart, J. F., Wainwright, R. B., Bryant, R. G., Hutwagner, L. C. & Hatheway, C. L. (1992). Clinical and laboratory comparison of botulism from toxin types A, B, and E in the United States, 1975–1988. *J Infect Dis* 166, 1281–1286.
- Zadoks, R. N., Schukken, Y. H. & Wiedmann, M. (2005). Multilocus sequence typing of *Streptococcus uberis* provides sensitive and epidemiologically relevant subtype information and reveals positive selection in the virulence gene *pauA*. *J Clin Microbiol* 43, 2407–2417.

Edited by: S. D. Bentley