Infection, Genetics and Evolution 30 (2015) 102–113

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Genomic sequences of six botulinum neurotoxin-producing strains representing three clostridial species illustrate the mobility and diversity of botulinum neurotoxin genes

Theresa J. Smith [a,*], Karen K. Hill [b], Gary Xie [b], Brian T. Foley [c], Charles H.D. Williamson [d], Jeffrey T. Foster [d], Shannon L. Johnson [b], Olga Chertkov [b], Hazuki Teshima [b], Henry S. Gibbons [e], Lauren A. Johnsky [e,1], Mark A. Karavis [e], Leonard A. Smith [f]

[a] Molecular and Translational Sciences Division, United States Army Medical Research Institute of Infectious Diseases, Fort Detrick, MD 21702, United States
[b] Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545, United States
[c] Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, United States
[d] Center for Microbial Genetics and Genomics, Northern Arizona University, Flagstaff, AZ 86011, United States
[e] Biosciences Division, Edgewood Chemical Biological Center, Aberdeen Proving Ground, MD 21010, United States
[f] Medical Countermeasures Technology, United States Army Medical Research and Material Command, United States Army Medical Research Institute of Infectious Diseases, Fort Detrick, MD 21702, United States

ARTICLE INFO

Article history:
Received 4 October 2014
Received in revised form 29 November 2014
Accepted 1 December 2014
Available online 6 December 2014

Keywords:
Clostridium botulinum
Clostridium argentinense
Clostridium baratii
Botulinum neurotoxin

ABSTRACT

The whole genomes for six botulinum neurotoxin-producing clostridial strains were sequenced to provide references for under-represented toxin types, bivalent strains or unusual toxin complexes associated with a *bont* gene. The strains include three *Clostridium botulinum* Group I strains (CDC 297, CDC 1436, and Prevot 594), a Group II *C. botulinum* strain (Eklund 202F), a Group IV *Clostridium argentinense* strain (CDC 2741), and a Group V *Clostridium baratii* strain (Sullivan). Comparisons of the Group I genomic sequences revealed close relationships and conservation of toxin gene locations with previously published Group I *C. botulinum* genomes. The *bont*/F6 gene of strain Eklund 202F was determined to be a chimeric toxin gene composed of *bont*/F1 and *bont*/F2. The serotype G strain CDC 2741 remained unfinished in 20 contigs with the *bont*/G located within a 1.15 Mb contig, indicating a possible chromosomal location for this toxin gene. Within the genome of *C. baratii* Sullivan strain, direct repeats of IS1182 insertion sequence (IS) elements were identified flanking the *bont*/F7 toxin complex that may be the mechanism of *bont* insertion into *C. baratii*. Highlights of the six strains are described and release of their genomic sequences will allow further study of unusual neurotoxin-producing clostridial strains.

Published by Elsevier B.V.

1. Introduction

Unlike the closely related *Clostridium tetani*, which is a single species that produces only one neurotoxin, at least four different botulinum neurotoxin-producing clostridial species have been described. *Clostridium botulinum* is a taxonomic designation for three Groups of anaerobic spore-forming bacteria that share the common characteristic of production of one or more botulinum neurotoxins (BoNTs) having conserved modes of action but widely varying genetic sequences (Collins and East, 1998). In addition, certain *Clostridium argentinense*, *Clostridium baratii* and *Clostridium butyricum* strains have been isolated that produce BoNTs (Aureli et al., 1986; Gimenez and Ciccarelli, 1970; Hall et al., 1985).

BoNTs are extremely potent toxins that are of concern in public health and food safety, and they are also monitored for national security reasons. The botulinum neurotoxin causes a flaccid paralysis known as botulism that can result from ingestion of the toxin (foodborne botulism) or inhalation and/or ingestion of neurotoxin-producing clostridial spores followed by colonization of the gut (infant botulism, adult toxicoinfections) or a contaminated wound (wound botulism) (Smith et al., 2012). Current therapies for botulism include human or equine antitoxins composed of combinations of serotype-specific antibodies (Arnon et al., 2006; CDC,

Abbreviations: BoNT, botulinum neurotoxin (protein); *bont*, botulinum neurotoxin (gene); orf, open reading frame; HA, hemagglutinin; NTNH, nontoxic non-hemagglutinin; kb, kilobase; bp, base pairs; SNP, single nucleotide polymorphism.

* Corresponding author at: Molecular and Translational Sciences Division, United States Army Medical Research Institute of Infectious Diseases, 1425 Porter Street, Fort Detrick, MD 21702, United States. Tel.: +1 301 619 4229.
E-mail address: theresa.j.smith.civ@mail.mil (T.J. Smith).
[1] Current address: PEO C3T, PD COMSEC Cryptographic Systems SFAE-CCC-CMC, Aberdeen Proving Ground, MD 21005, United States.

http://dx.doi.org/10.1016/j.meegid.2014.12.002
1567-1348/Published by Elsevier B.V.

2010). While in most cases these products provide effective treatment for the patient, they are less effective when given later in the course of the intoxication (Leclair et al., 2013), and protection may vary depending on differences between the toxin causing the intoxication and the toxin used to produce the antibodies (Smith et al., 2005).

The neurotoxin-producing clostridia are genetically diverse by 16S *rrn* comparisons (Collins and East, 1998; Hill et al., 2009). They are organized into six clades (Groups I–VI) that also contain non-neurotoxigenic species. Each clade represents a different clostridial species, but Groups I–III continue to be defined as *C. botulinum* based on their production of botulinum neurotoxins. Group I includes BoNT/A-, /B-, and /F-producing *C. botulinum* and non-neurotoxic *Clostridium sporogenes*; Group II includes BoNT/B-, /E-, and /F-producing *C. botulinum*; Group III includes BoNT/C- or /D-producing *C. botulinum* and non-neurotoxic *Clostridium novyi*; Group IV includes BoNT/G-producing *C. argentinense* and *Clostridium subterminale*; Group V includes neurotoxigenic BoNT/F-producing and non-neurotoxigenic *C. baratii*; and Group VI includes neurotoxigenic BoNT/E-producing and non-neurotoxigenic *C. butyricum* (Collins and East, 1998). The presence of the same toxin type in different Groups (for example BoNT/B in Groups I and II, BoNT/E in Groups II and VI, BoNT/F in Groups I, II and V) illustrates the horizontal transfer of *bont* genes encoding the different toxin serotypes within and between the *C. botulinum* Groups (Hill et al., 2009).

The Group designations describe differences at the species level, and this level of diversity is also evident in the neurotoxins. Variation in the nucleotide and corresponding protein sequences within the toxin has been reported in multiple publications (Hill and Smith, 2013; Hill et al., 2007; Raphael et al., 2010; Smith et al., 2005). Genetic variants or subtypes within each serotype are generally given numerical designations following the toxin type, for example BoNT/A1 or BoNT/A2, where the toxin proteins differ by 10% (Hill and Smith, 2013). There are bivalent strains that produce two toxins, such as Ab, Af, Ba, Bf, with the capital letter designating the toxin produced in greater amounts (Hill and Smith, 2013). Recently, strain Af84 was shown to contain three *bont* genes (Dover et al., 2013). There are also reports of partial toxin cluster genes within strains and a "silent" B gene (designated (B)) that does not produce a toxin due to a mutation introducing a premature stop codon within the gene (Carter et al., 2010, 2013; Dover et al., 2009; Kirma et al., 2004).

Horizontal gene transfer of the toxin genes is evident from the presence of the same toxin types in the different bacterial species (Groups) and the presence of toxin operons in plasmids and phages as well as within the chromosome. Interestingly, the toxin operon does not appear to integrate into the bacterial plasmids or chromosomes at random; rather, it can be located at specific sites (Carter et al., 2013; Dover et al., 2013; Hill and Smith, 2013). The toxin operon is very often associated with transposases, helicases, insertion sequence (IS) elements and other hallmarks of mobile genetic elements (Hill and Smith, 2013; Hill et al., 2007).

In their natural state, toxins are not found as single proteins, but rather as protein complexes composed of the toxin and several non-toxic proteins. The genes that encode these associated proteins are encoded within a bi-directional operon (Dineen et al., 2004; Dupuy et al., 2006). In various strains the toxin gene operons range from approximately 9 to 13 kb in length and contain multiple genes or predicted open reading frames (orf) associated with the ~3.8 kb *bont* gene. The toxin gene (*bont*) is always preceded by a gene encoding a nontoxic non-hemagglutinin protein (*ntnh*). The *ntnh*/*bont* are accompanied by either *ha* or *orfX* genes (*ha*+ or *orfX*+ gene clusters), which are always encoded on the opposite strand, preceding the *ntnh* gene. A *botR* gene encoding a p21 sigma factor and/or a *p47* gene are located in the center of the operon. *Bont*/A, /B, /C, /D or /G genes may be found within *ha*+ operons, while the *orfX*+ operons encode *bont*/A, /E, or /F (Popoff and Marvaud, 1999). Toxin subtype *bont*/A1 is unique in that it can be associated with either the *ha*+ or *orfX*+ gene cluster. The NTNH protein is known to be tightly bound to the toxin until the toxin complex leaves the acidic stomach compartment and enters the alkaline small intestine environment. The HA proteins form a complex that interacts with an epitope on the NTNH part of the NTNH-BoNT protein heterodimer (Gu and Jin, 2013). While it is known that the *orfX* genes encode proteins that interact loosely with the NTNH-BoNT complex, the details of these interactions have not been determined. The *botR* and *bont* genes are clearly homologous with the *tetR* and *tetx* gene from *C. tetani* (Dupuy et al., 2006; Popoff and Marvaud, 1999).

Because of the diversity that is observed among the BoNT-producing clostridia, the toxin complex proteins, and the toxin itself, multiple genomes have been sequenced in an effort to increase our understanding of the underlying mechanisms of this diversity. Here, three Group I strains, one Group II strain, one Group IV BoNT/G-producing *C. argentinense* strain, and one Group V BoNT/F7-producing *C. baratii* strain were sequenced in order to provide complete genomes representing uncommon serotypes and strains with unusual toxin types. The sequences provide references for comparison with draft short-read genomic sequences, and analysis of these new genomes provide additional insights into *C. botulinum* Groups I, II, IV and V strains.

2. Materials and methods

2.1. Genomic sequencing and analysis

DNA preparations from the clostridial strains were prepared as previously described (Hill et al., 2007). Draft genomes were produced using both Roche 454 (Margulies et al., 2005) and Illumina GAII data (Bennett, 2004). Table 1 lists the genome coverage for each data type used in the final assemblies. The raw data from each draft were assembled using the native assembler for that type-Newbler (Roche) for 454, Velvet for Illumina (Zerbino and Birney, 2008), and the assemblies were merged into single circular chromosome or plasmid contigs (Parallel Phrap), except for serotype G strain CDC 2741, which remained unfinished in 20 contigs.

Final assemblies for all strains except CDC 2741 were brought to closed and finished status through both manual and computational finishing efforts. Annotations were completed using an automated system utilizing the Ergatis workflow manager (Hemmerich et al., 2010). Each annotation was manually reviewed and submitted to NCBI for deposit in the GenBank database. Sequencing information and accession numbers for each strain are listed in Table 1. Final annotations were uploaded to RAST for comparative review (Aziz et al., 2008). Note that the assembled sequences provided a single contig for either the chromosome or plasmid (if present); however the Prevot 594 chromosome contains one region that is estimated to be 50 nucleotides (represented by a series of N's) where the sequence is unknown. The length of the gap is based upon Roche 454 long insert data.

2.2. 16S rrn gene phylogeny

Newly sequenced and published genome assemblies were annotated with Prokka (Seemann, 2014), and 16S *rrn* gene sequences for each genome assembly were parsed from Prokka output files. As multiple 16S *rrn* gene sequences were present in some assemblies, a representative sequence was chosen for each genome after clustering sequences (99% identity) from each assembly with USearch (Edgar, 2010). Representative sequences were aligned and masked with ssu-align (Nawrocki and University, 2009). A maximum-like-