Letter

Genome sequence of a proteolytic (Group I) *Clostridium botulinum* strain Hall A and comparative analysis of the clostridial genomes

Mohammed Sebaihia,[1] Michael W. Peck,[2] Nigel P. Minton,[3] Nicholas R. Thomson,[1] Matthew T.G. Holden,[1] Wilfrid J. Mitchell,[4] Andrew T. Carter,[2] Stephen D. Bentley,[1] David R. Mason,[2] Lisa Crossman,[1] Catherine J. Paul,[5] Alasdair Ivens,[1] Marjon H.J. Wells-Bennik,[2] Ian J. Davis,[3] Ana M. Cerdeño-Tárraga,[1] Carol Churcher,[1] Michael A. Quail,[1] Tracey Chillingworth,[1] Theresa Feltwell,[1] Audrey Fraser,[1] Ian Goodhead,[1] Zahra Hance,[1] Kay Jagels,[1] Natasha Larke,[1] Mark Maddison,[1] Sharon Moule,[1] Karen Mungall,[1] Halina Norbertczak,[1] Ester Rabbinowitsch,[1] Mandy Sanders,[1] Mark Simmonds,[1] Brian White,[1] Sally Whithead,[1] and Julian Parkhill[1,6]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom; [2]Institute of Food Research, Norwich Research Park, Colney, Norwich, NR4 7UA, United Kingdom; [3]Centre for Biomolecular Sciences, Institute of Infection, Immunity and Inflammation, School of Molecular Medical Sciences, University of Nottingham, Nottingham NG7 2RD, United Kingdom; [4]School of Life Sciences, Heriot-Watt University, Riccarton, Edinburgh EH14 4AS, United Kingdom; [5]Bureau of Microbial Hazards, Health Canada, Ottawa, Ontario, K1A 0L2, Canada

Clostridium botulinum is a heterogeneous Gram-positive species that comprises four genetically and physiologically distinct groups of bacteria that share the ability to produce botulinum neurotoxin, the most poisonous toxin known to man, and the causative agent of botulism, a severe disease of humans and animals. We report here the complete genome sequence of a representative of Group I (proteolytic) *C. botulinum* (strain Hall A, ATCC 3502). The genome consists of a chromosome (3,886,916 bp) and a plasmid (16,344 bp), which carry 3650 and 19 predicted genes, respectively. Consistent with the proteolytic phenotype of this strain, the genome harbors a large number of genes encoding secreted proteases and enzymes involved in uptake and metabolism of amino acids. The genome also reveals a hitherto unknown ability of *C. botulinum* to degrade chitin. There is a significant lack of recently acquired DNA, indicating a stable genomic content, in strong contrast to the fluid genome of *Clostridium difficile*, which can form longer-term relationships with its host. Overall, the genome indicates that *C. botulinum* is adapted to a saprophytic lifestyle both in soil and aquatic environments. This pathogen relies on its toxin to rapidly kill a wide range of prey species, and to gain access to nutrient sources, it releases a large number of extracellular enzymes to soften and destroy rotting or decayed tissues.

[Supplemental material is available online at www.genome.org. The sequence and annotation of the *Clostridium botulinum* chromosome and plasmid have been deposited in the EMBL database under accession nos. AM412317 and AM412318, respectively. Microarray data have been deposited in ArrayExpress under accession no. E-TABM-264.]

Clostridium botulinum is a Gram-positive organism, a member of the firmicutes, that produces one of several toxins collectively known as botulinum neurotoxin, which are the most potent toxins known to man and induce a potentially fatal paralytic condition in humans and various animal species known as "botulism." In humans, the most commonly reported types of botulism are food-borne botulism, infant botulism, and wound botulinum. Consumption of contaminated food in which neurotoxin has been produced can result in food-borne botulism, a severe disease with a high fatality rate. As little as 30 ng of neurotoxin can be fatal (Peck 2006). Infant botulism is an intestinal toxemia that affects children <12 mo of age; a similar disease also very rarely affects adults, and occurs when competing bacteria in the normal intestinal microbiota have been suppressed (e.g., by antibiotic treatment). Infant botulism has been reported in many countries, and in the United States, it is the commonest manifestation of the disease. Some reports suggest a link to sudden infant death syndrome (Arnon 2004; Fox et al. 2005). Wound botulism is an infection in which growth and neurotoxin formation occur in a wound in the body (Werner et al. 2000; Brett et al. 2004). Until recently, this disease was very rare; however, a significant number of cases have now been reported in many countries, primarily associated with intravenous drug abuse. For example, in the United Kingdom, wound botulism was not reported prior to 2000, but a total of 112 suspected cases were

[6]Corresponding author.
E-mail parkhill@sanger.ac.uk; fax 44-1223-494919.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6282807. Freely available online through the *Genome Research* Open Access option.

reported between 2000 and 2005, all involving heroin injection (Anonymous 2006).

The botulinum neurotoxins have been subdivided into seven distinct serotypes (types A to G), although variations within an individual serotype are evident. The ability to produce the botulinum neurotoxin is confined to the genus *Clostridium*. Although all botulinogenic clostridial strains have traditionally been classified as *C. botulinum*, it is recognized that *C. botulinum* contains four distinct genetic and physiological groupings: Group I (proteolytic *C. botulinum*) strains produce one or sometimes two toxins of type A, B or F; Group II (nonproteolytic *C. botulinum*) strains produce toxins of type B, E, or F; Group III strains produce toxins of type C or D; and Group IV strains produce toxin of type G (Lund and Peck 2000). In addition, strains of *Clostridium butyricum* and *Clostridium baratii* have also been isolated that produce type E and F neurotoxins, respectively. Food-borne botulism is most commonly caused by Group I and Group II *C. botulinum*, while infant and wound botulism are most frequently caused by Group I *C. botulinum* (Peck 2005). Each of the four *C. botulinum* groupings also has a nonneurotoxinogenic counterpart (e.g., *Clostridium sporogenes* for Group I, *Clostridium novyi* for Group III). In more recent years, the application of 16S rRNA sequencing technology has unequivocally demonstrated that the four groupings are composed of distinct species. In the case of those groups important to human botulism, Group I strains, regardless of toxin type, are highly related to one another (99.7%–100% 16S rRNA sequence identity) and together with *C. sporogenes* form a single phylogenetic unit (Hutson et al. 1993b). Neurotoxin-forming Group II strains (and their nontoxinogenic counterparts) form a distinct line that is quite separate from other saccharolytic clostridia and phylogenetically far removed from the Group I strains (Hutson et al. 1993a).

In this study, we have determined the genome sequence of a representative of Group I (proteolytic) *C. botulinum* (strain Hall A, ATCC 3502). Including *C. botulinum*, seven clostridial genomes are currently available; one nonpathogenic solventogenic species, *Clostridium acetobutylicum* (Nolling et al. 2001); and three toxigenic species, including three strains of *Clostridium perfringens* (Shimizu et al. 2002a; Myers et al. 2006), *Clostridium tetani* (Bruggemann et al. 2003), and *Clostridium difficile* (Sebaihia et al. 2006). These genomes provide an excellent opportunity for comparative analysis of the clostridia and will undoubtedly provide valuable insights into the lifestyle, metabolic diversity, pathogenicity, and evolution of these organisms.

Results and Discussion

General features of the genome and comparative genomics

The major features of the genome are listed in Table 1 and Figure 1. The genome of proteolytic *C. botulinum* strain Hall A (ATCC 3502) consists of a chromosome of 3,886,916 bp and a plasmid, pBOT3502, of 16,344 bp, which carry 3650 and 19 coding sequences (CDS), respectively. The %GC content (26.8%) of the plasmid is slightly lower than that of the chromosome (28.2%).

Plasmid pBOT3502, which is not similar to other sequenced clostridial plasmids, contains a large CDS (CBOP01), encoding a protein of 1194 amino acids that shows significant similarity with the alpha-subunit of DNA polymerase III (DnaE), the closest match being the chromosomally encoded DnaE of *C. perfringens*. This class of protein is not normally associated with plasmid replication. Homologs have been noted on Ti-plasmids, but in these cases, additional genes encoding other, more traditional, plasmid replication proteins are also present. The sole involvement of CBOP01 in replication of pBOT3502 has been demonstrated through the subcloning of the region encompassing this gene into replicon cloning vectors (M.H.J. Wells-Bennik, K. Medendorp, A.T. Carter, and M.W. Peck, unpubl.). Other CDSs are likely to be involved in plasmid stability (CBOP05) and mobilization (CBOP11 and CBOP12), but the most prominent CDSs are those encoding a biosynthetic and transport system apparently dedicated to the production of a bacteriocin (CBOP15-19) that shares 40% identity, at the amino acid level, with a boticin of proteolytic *C. botulinum* strain 213B (Dineen et al. 2000). Given the CDSs present in the plasmid, the major pressure for retention of pBOT3502 by the host is likely to be the production of boticin, which may offer a competitive advantage to the strain, outweighing the metabolic burden imposed by plasmid maintenance.

Unlike the highly mosaic genome of *C. difficile* (Sebaihia et al. 2006), there is no evidence of recent horizontal gene acquisition in the *C. botulinum* genome, apart from some selfish elements; two prophages, two prophage remnants, and 12 transposases (only one of which is intact), which are dispersed throughout the chromosome.

There is very little overall synteny between the genomes of the sequenced clostridia, confirming further the heterogeneity of the *Clostridium* genus. To identify sets of genes that are shared or unique to *C. botulinum*, reciprocal FASTA analysis of the *C. botulinum* CDSs was performed against four sequenced clostridial genomes, *C. acetobutylicum*, *C. perfringens* strain 13, *C. tetani*, and *C. difficile*. There are only 568 *C. botulinum* CDSs (16%) that are shared with all the other sequenced clostridia, while 1511 CDSs (41%) have orthologs in at least one, but not all, of the sequenced clostridia, and 1571 CDSs (43%) are unique to *C. botulinum*, compared to the other four sequenced clostridial genomes. The shared CDSs mainly encode core functions, whereas the CDSs that are unique to *C. botulinum* encode accessory functions (Supplemental Fig. S1). The distribution of *C. botulinum* unique genes is markedly nonhomogeneous around the genome (Fig. 1); there is no readily apparent reason for this, although a remarkably similar distribution was recently described in the genome of *C. perfringens* (Myers et al. 2006).

Table 1. General features of the clostridial genomes

	C. botulinum	*C. difficile*	*C. acetobutylicum*	*C. perfringens* strain 13	*C. tetani*
Size (bp)	3,886,916	4,290,252	3,940,880	3,031,430	2,799,250
G+C content (%)	28.24	29.06	30.93	28.60	28.75
Coding sequence	3650	3774	3740	2660	2368
Coding density	0.93	0.87	0.93	0.87	0.85
Average gene size (bp)	875	943	920	946	1011
rRNA operons	9	11	11	10	6
tRNA	80	87	73	96	54

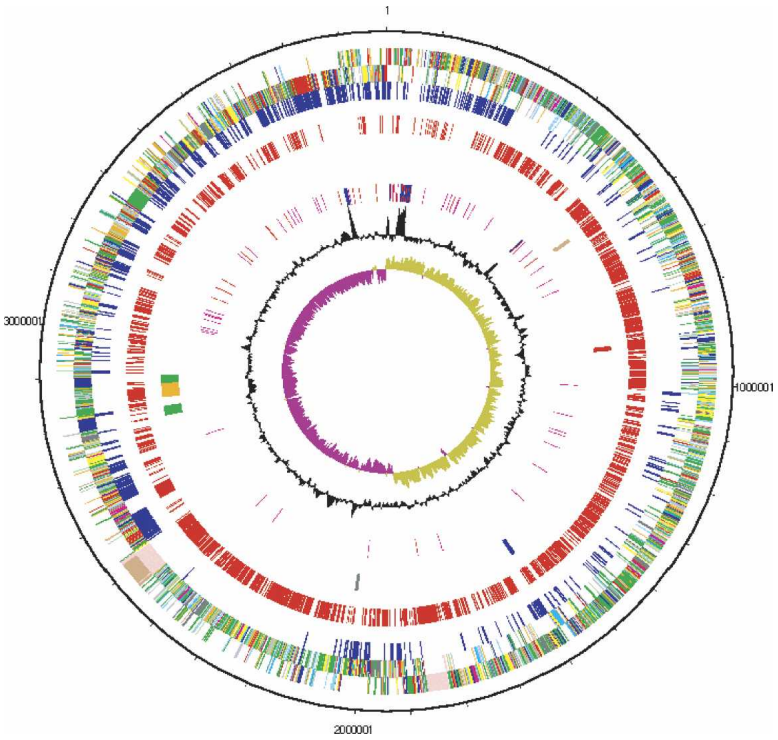

Figure 1. Circular representations of the genome of *C. botulinum*. The circles represent (from the outside in): (*1* and *2*) All CDS (transcribed clockwise and anticlockwise); (dark blue) pathogenicity/adaptation; (black) energy metabolism; (red) information transfer; (dark green) surface-associated; (cyan) degradation of large molecules; (magenta) degradation of small molecules; (yellow) central/intermediary metabolism; (pale green) unknown; (pale blue) regulators; (orange) conserved hypothetical; (brown) pseudogenes; (pink) phage and IS elements; (gray) miscellaneous. (*3*) (Blue) CDSs shared with sequenced clostridia. (*4*) (Red) *C. botulinum*-unique CDSs relative to the sequenced clostridia. (*5*) Virulence factors discussed in the text; (brown) streptolysin; (red) neurotoxins; (blue) six metalloproteases; (black) type IV pilus; (green) flagellar and chemotaxis operons; (orange) flagellar glycosylation island. (*6*) RNA genes; (blue) rRNAs; (red) tRNAs; (purple) stable RNAs. (*7*) G+C content (plotted using a 10-kb window). (*8*) GC deviation [(G − C)/(G + C) plotted using a 10-kb window]; (khaki) values >1; (purple) values <1.

Neurotoxin genes and virulence factors

The main virulence factor of *C. botulinum*, the neurotoxin, is produced as a noncovalently bound complex with two or more nontoxic protein components, hemagglutinin, and nontoxic nonhemagglutinin. To date, the phenotypic and genotypic designation of the proteins and genes involved in the production of clostridial neurotoxins (of both *C. botulinum* and *C. tetani*) has been both unconventional and has lacked consistency. In this annotation we have, therefore, adopted *cnt* (Clostridial NeuroToxin) as the prefix for genes involved in the production of botulinum, and tetanus, neurotoxins. All botulinum and tetanus neurotoxin genes are *cntA*. All nontoxic, nonhemagglutinin (NTNH) genes are *cntB*; the genes encoding hemagglutinin proteins are designated as *cntC*, *cntD*, and *cntE*, and replace *ha*34, *ha*17, and *ha*70, respectively; and the gene that encodes the regulatory protein that controls their expression is *cntR*. The genes of the various toxins and their accessory proteins are clustered either on the chromosome (Group I *C. botulinum* and Group II *C. botulinum*), a prophage (Group III *C. botulinum*), or a plasmid (Group IV *C. botulinum*). In Group I (proteolytic) *C. botulinum* strain Hall A (ATCC 3502), they are organized in two divergent transcriptional units on the chromosome, *cntAB* and *cntCDE*, which bracket the regulatory gene, *cntR*. The whole cluster is flanked on the right end by three transposases, and on the left end by a single transposase. The association of this locus with mobile elements (transposases) raises the possibility that the whole locus may be mobile, and it may have been laterally acquired, although there is no direct evidence that it has been transposed as a unit. It may also explain why strains can appear to lose the ability to form neurotoxin when repeatedly cultured in the laboratory (Lund and Peck 2000). The sequence of the neurotoxin gene and the organization of the neurotoxin gene cluster are typical of that for subtype A1 toxin, and closely resemble that of other strains (Dineen et al. 2003; Smith et al. 2005). There are no other full or partial neurotoxin genes in the genome.

Following synthesis of the neurotoxin, proteolytic cleavage of the toxin at one-third the distance from the N terminus, to produce the Heavy and Light Chains, is required for toxicity. The identity of the protease responsible is not known. One study has previously purified a 62-kDa protein from culture supernatant of *C. botulinum* that is believed to carry out this proteolytic nicking (Dekleva and Dasgupta 1990). Polyacrylamide gel electrophoresis revealed that the 62-kDa protease, which is specific for the arginyl peptidyl bond, is composed of 15.5- and 48-kDa polypeptides. Interestingly, the structure and substrate specificity of this enzyme are reminiscent of those of the secreted alpha-clostripain from *Clostridium histolyticum* (Dargatz et al. 1993), a homolog (74% amino acid identity) of which is present in *C. botulinum* (CBO1920). The *C. histolyticum* alpha-clostripain is a trypsin-like cysteine endopeptidase with strict specificity for the arginyl bond. It is synthesized as an inactive prepro-enzyme that undergoes an autocatalytic cleavage to generate 15.4- and 43-kDa polypeptides, which associate to form a heterodimeric active enzyme (Dargatz et al. 1993). Furthermore, both the *C. histolyticum* alpha-clostripain and the *C. botulinum* 62-kDa protease require a reducing agent and calcium for full activity and are susceptible to the same protease inhibitors. These data strongly suggest that the *C. botulinum* ortholog of alpha-clostripain (CBO1920) is the endogenous protease responsible for the proteolytic nicking of the neurotoxin of *C. botulinum*. At this stage it is not clear whether the sole purpose of this protease is to carry out the post-translational processing of the neurotoxin, or if it also contributes to the overall proteolytic activity of *C. botulinum* (see below). A gene encoding clostripain (CPE0846) is also present in *C. perfringens*, and has been found to be positively regulated by the two-component system VirR/VirS (Shimizu et al. 2002b).

A prominent feature of the *C. botulinum* genome that is absent from all the other sequenced clostridial genomes is the presence of a nine-gene cluster (CBO0486–0494) highly similar to the streptolysin S (SLS) biosynthetic operon (*sagA–I*) from *Streptococ-*

cus spp. The SLS is a cytolysin that is responsible for the hemolytic phenotype, and is an important virulence factor of *Streptococcus* spp. (Datta et al. 2005).

CBO2038 encodes a putative secreted protein containing a thrombospondin type 3 repeat (PF02412). This domain is associated with proteins that are involved in binding to components of extracellular matrices such as fibronectin and collagen. This CDS is part of a gene cluster comprising CBO2043, which encodes a protein similar to the myosin-cross-reactive antigen from *Streptococcus pyogenes* (Kil et al. 1994), orthologs of which are also present in *C. perfringens* (CPE0378) and *C. tetani* (CTC1855), but not in *C. acetobutylicum* and *C. difficile*.

Extracellular enzymes and general metabolism

Production of extracellular proteases by proteolytic clostridia has been associated with food spoilage and pathogenicity. The highly proteolytic nature of Group I (proteolytic) *C. botulinum* is reflected in the genome by the presence of several protease-encoding CDSs. In addition to the putative alpha-clostripain (CBO1920, mentioned above), the *C. botulinum* genome encodes several other proteases including six thermolysin-like metalloproteases (CBO1441–1446); a collagenase (CBO1620); two putative hemolysins (CBO1450 and CBO1589); a C-terminal peptidase (CBO3389); and a zinc-metallopeptidase (CBO3606). All these proteases appear to have an N-terminal signal sequence, which suggests that they are likely to be secreted.

Interestingly, the genes encoding the six thermolysin-like metalloproteases (CBO1441–1446) are tandemly arrayed, and the proteins they encode are highly similar (60%–80% amino acid identity) to each other. This six-gene locus is absent in all the other sequenced clostridia. Moreover, the block of DNA encompassing these protease genes displays an anomaly in the strand-specific GC bias (a bias toward G on the lagging strand, the reverse of that normally observed) (Fig. 1). Since there is no evidence that this DNA was recently horizontally acquired, the most likely explanation for this GC bias anomaly is that it is due to an inversion following a recent recombination event. It should be noted that the last CDS in this cluster, CBO1441, contains a frameshift mutation, and it is not clear if this protein is functional.

The proteinaceous products, peptides and amino acids, produced by the *C. botulinum* extracellular proteases can be taken up by a large number (40) of transporters. Several clostridia are able to ferment amino acids in a coupled oxidation–reduction reaction known as the Stickland reaction, in which the reduction of one amino acid (electron acceptor) is coupled to the oxidation of another amino acid (electron donor) (Stickland 1935). Analysis of the *C. botulinum* genome revealed the presence of fermentation pathways for several amino acids. Glycine is reduced by a glycine reductase complex (CBO1255–1264/Grd), and oxidized by a glycine cleavage system (CBO0696–699); thus it can serve both as an oxidizing and reducing agent in the Stickland reaction. Proline is reduced by a proline reductase complex (CBO2460–2490/Prd).

C. sporogenes, the non-neurotoxigenic counterpart of Group I *C. botulinum*, can ferment phenylalanine and leucine (Bader et al. 1982). Two gene clusters that encode key enzymes in the fermentation pathways of these two amino acids were identified in the *C. botulinum* genome. The first system, CBO2192–2199, is similar to 2-hydroxyisocaproyl-CoA dehydratase (HadAIBC) from *C. difficile*, in the pathway of leucine fermentation (Kim et al. 2005). The second, CBO3283–3292, which is missing in the other sequenced clostridial genomes, is highly similar to phenyllactate dehydratase (FldAIBC) from *C. sporogenes*, which is involved in the fermentation of phenylalanine (Dickert et al. 2002).

Although amino acids are a significant energy source for proteolytic clostridia, these organisms can also ferment sugars. Analysis of the genome revealed the presence of a large number of genes consistent with the degradation of complex polysaccharides and metabolism of a variety of sugars.

Chitin is the second most abundant polysaccharide after cellulose; it is an insoluble linear homopolymer of *N*-acetyl-D-glucosamine (GlcNAc), and is the major component of invertebrate exoskeletons and fungal cell walls. The *C. botulinum* genome encodes five putative secreted chitinases, CBO0728, CBO1057, CBO1081, CBO1212, and CBO2832, none of which have been found in other sequenced clostridial genomes. They all contain a catalytic domain classified in family 18 of glycosylhydrolases (PF00704), either alone or in association with additional domains such as a fibronectin type III (PF00041) and a carbohydrate-binding domain (PF02839). All but one (CBO1212) of these chitinases have an N-terminal signal sequence indicating that they are secreted into the extracellular environment (Fig. 2). One of these chitinase genes, CBO2832, is part of a gene cluster (CBO2832–2839) that is potentially involved in the transport and metabolism of GlcNAc, the product of chitin hydrolysis by chitinases. This gene is also convergently transcribed with another CDS, CBO2831, that encodes a predicted secreted protein that is weakly similar to chitinases but has no apparent catalytic or chitin-binding domains. We have also identified two additional CDSs, CBO1958 and CBO2105, that encode secreted proteins having a C-terminal chitin-binding domain, but lacking the chitinase catalytic domain. CBO1958 is also highly similar (79% amino acid identity) to another putatively secreted protein, CBO1966. However, the latter lacks both the chitin-binding and catalytic domains. Similar noncatalytic chitin-binding proteins have been found in chitinolytic organisms (Howard et al. 2003;

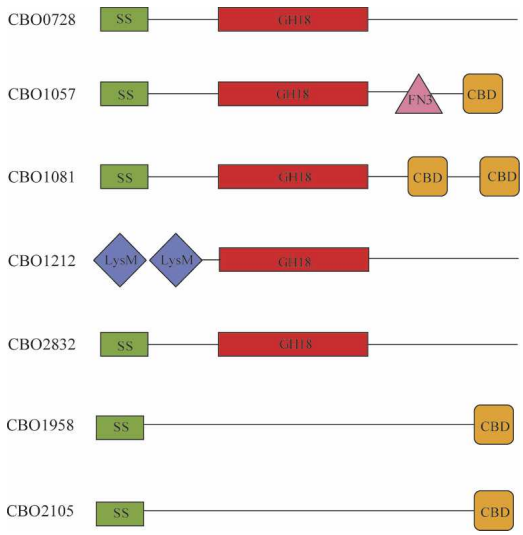

Figure 2. Schematic representation of the *C. botulinum* chitinolytic system. (CBD) Chitin-binding-domain; (SS) signal sequence; (FN3) fibronectin type III domain; (GH18) family 18 domain of glycosylhydrolases; (LysM) LysM domain.

Vaaje-Kolstad et al. 2005). Although the precise function of these proteins is unknown, it has been suggested that they may play a role in the initial interaction of the bacteria with chitin-containing surfaces (Montgomery and Kirchman 1994). In order to assess whether these proteins are involved in the degradation of chitin, we performed a chitinase assay. The presence of small zones of clearing on chitin-layered plates confirmed the ability of the sequenced Hall A strain and also proteolytic *C. botulinum* strain 213B to degrade chitin (Fig. 3). Chitin degradation was not, however, detected in nonproteolytic *C. botulinum* strain CDC 7854. Collectively, these data suggest that *C. botulinum* Hall A strain has an active chitinolytic system, enabling it to use chitin as a source of carbon and nitrogen. This is not surprising considering that *C. botulinum*, as a free-living organism, colonizes diverse soil and marine environments where chitin-containing organisms, such as insects, fungi, and crustaceans, are abundant.

Starch is another abundant polysaccharide that consists of a linear polymer, amylose, and a branched polymer, amylopectin. Complete degradation of starch requires the combined action of several enzymes, alpha-amylase, beta-amylase, pullulanase, and glucoamylase. Of these only one, a secreted beta-amylase (CBO1203), appears to be produced by *C. botulinum*. This enzyme catalyzes the removal of maltose molecules from the nonreducing ends of the starch polymers. The lack of other starch-hydrolyzing enzymes in *C. botulinum* suggests that this bacterium is unable to completely degrade whole starch but nevertheless can degrade starch-derived polysaccharides. This is consistent with earlier studies that showed that proteolytic *C. botulinum* strains were unable to hydrolyze starch in conventional tests (Dezfulian and Dowell 1980; Smith and Sugiyama 1988), yet starch and maltose could support growth (Whitmer and Johnson 1988).

Unlike some saccharolytic clostridia, proteolytic *C. botulinum* does not appear to have the capacity to degrade cellulose, the most abundant polysaccharide in nature.

The *C. botulinum* chromosome harbors two genes that code for two putatively secreted lipases, CBO0863 and CBO2061. Lipase activity leads to the production of a thin pearly layer on and around colonies formed on media containing egg yolk, and is exploited as a simple diagnostic test for the detection of *C. botulinum* (Mills et al. 1985; Lund and Peck 2000).

For uptake and phosphorylation of sugars and sugar derivatives, *C. botulinum* deploys 15 phosphoenolpyruvate (PEP)-dependent phosphotransferase systems (PTS), a greater number than have been found in any of the sequenced clostridial genomes with the exception of *C. difficile*. The *C. botulinum* PTS complement includes representatives of five of the seven known PTS families (Barabote and Saier 2005), and the individual proteins show conservation of domain structure and functional residues or motifs. The majority of phosphotransferase genes are located adjacent to genes encoding cognate enzymes for metabolism of the phosphorylated PTS products as well as transcriptional regulators, implying that expression is coordinately induced in the presence of the substrate. The *C. botulinum* genome contains a single gene encoding each of the general PTS energy-coupling proteins enzyme I (CBO3438) and HPr (CBO2398), which mediate the transfer of phosphate from the donor PEP to the substrate-specific enzyme II complexes. As in *C. acetobutylicum*, *C. perfringens*, and *C. tetani*, but in contrast to *C. difficile* and the majority of other bacteria, these genes in *C. botulinum* are not contiguous and appear to be expressed monocistronically.

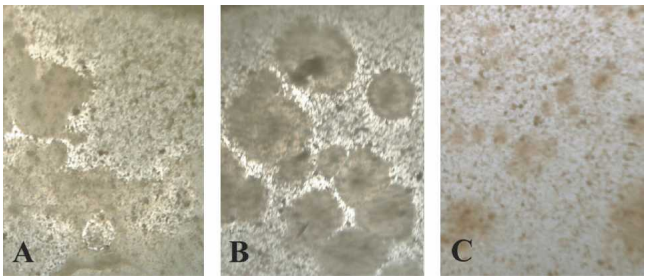

Figure 3. Chitinase assay. Chitin degradation was assessed on chitin-overlaid agar plates as described in Methods. Small zones of clearing are evident around colonies of proteolytic *C. botulinum* strains Hall A (*A*) and 213B (*B*), but not around colonies of nonproteolytic *C. botulinum* strain CDC 7854 (*C*).

Proteolytic *C. botulinum* strain Hall A has a complete glycolysis pathway, but an incomplete TCA cycle. *C. botulinum* appears to be capable of both acidogenic (production of acetate and butyrate) and solventogenic (production of butanol and ethanol) fermentations. Interestingly, there are two copies each of the genes encoding phosphotransbutyrylase (Ptb) (CBO3118 and CBO3427) and butyrate kinase (Buk) (CBO3426 and CBO3428), which constitute the butyrate formation pathway. Unlike *C. acetobutylicum*, *C. botulinum* is unable to produce acetone because it lacks the gene (*adc*) that encodes the terminal enzyme (acetoacetate decarboxylase) in the acetone formation pathway.

Components of the cell surface

Bacteria exhibit different cell surface structures according to the environmental niches they inhabit. These components play an important role in sensing and surviving environmental challenges and interaction with other organisms. Therefore, the characterization of the cell surface components of *C. botulinum* and comparison with other sequenced clostridia could provide an insight into their lifestyles and the environmental niches they colonize.

The chemotaxis process allows motile bacteria to sense environmental cues and to respond by moving toward attractants and away from repellents. Methyl-accepting chemotaxis proteins (MCP) play a key role in this process. MCPs have membrane-associated N-terminal sensor domains that bind to attractants and repellents (Scott et al. 1993) and intracellular methyl-accepting signaling domains that transduce the external signal to the intracellular chemotaxis machinery (Kim et al. 1999). *C. botulinum* encodes 24 putative MCPs, only one of which, CBO3558, designated SonO, has been characterized. SonO was recently identified as a sensor of nitric oxide (NO), which is extremely toxic to *C. botulinum,* and the protein crystal structure has been solved (Nioche et al. 2004). *C. botulinum* encodes more MCPs than *C. tetani* (15), *C. difficile* (1), and *C. perfringens* (0), but fewer than *C. acetobutylicum* (38), suggesting that *C. botulinum* and *C. acetobutylicum* have a relatively elaborate sensing capability, and can respond to a wider range of environmental cues (attractants and repellents) relative to *C. tetani*, *C. difficile*, and *C. perfringens*.

The *C. botulinum* genome carries 84 putative chemotaxis- and flagella-related proteins, 54 of which are predominantly organized into two operons (CBO2637–2666 and CBO2730–2753). Overall, the gene content and order in these two operons is conserved between *C. botulinum*, *C. acetobutylicum*, and *C. tetani*, but to a lesser extent in *C. difficile*. In contrast, *C. perfringens* lacks all the flagellar and chemotaxis genes, which is consistent with it

being defective in flagellar-mediated motility. The remaining 30 *C. botulinum* chemotaxis- and flagella-related CDSs are dispersed, singularly or in pairs, throughout the chromosome.

C. botulinum possesses five CDSs that putatively encode the structural subunits of the flagellar filament (CBO0242, CBO2666, CBO2695, CBO2730, and CBO2731), more than in *C. acetobutylicum* (4), *C. tetani* (4), *C. difficile* (1), and *C. perfringens* (0). *C. botulinum* has an additional CDS, CBO0798, that is similar to only the N-terminal domain of a flagellin structural subunit.

The two major flagellar and chemotaxis gene clusters of *C. botulinum* are separated by two gene clusters (CBO2678–2689 and CBO2696–2729) that potentially encode proteins involved in the biosynthesis, modification, polymerization, and export of polysaccharides. One of these clusters, CBO2696–2729, is flanked on the right by two almost identical (98.9% amino acid identity) flagellin structural genes, CBO2730 and CBO2731, that have been recently identified as the genes encoding the major structural components of the flagellar filament (Paul et al. 2007); and on the left by a third flagellin gene (CBO2695). The colocalization of this polysaccharide biosynthesis locus and the flagellin structural genes is similar to that of the flagellar glycosylation locus in *Campylobacter jejuni* (Szymanski et al. 2003).

Polysaccharide biosynthesis loci, located at an equivalent location to that of *C. botulinum*, were also identified in *C. acetobutylicum* and *C. tetani*, CAC2168–2202 and CTC1692–1714, respectively, but not in *C. perfringens* (defective in flagellar-mediated motility). However, although these loci share a few genes at their 5′-ends, the majority of their genes are either different or highly divergent, suggesting that they produce different glycan structures. In *C. difficile* only the ortholog (CD0240) of the first CDS of the *C. botulinum* polysaccharide biosynthesis cluster (CBO2729), which encodes a glycosyltransferase, is present. The *C. acetobutylicum* and *C. tetani* polysaccharide biosynthesis clusters and the glycosyltransferase (CD0240) of *C. difficile* also lie directly downstream from one flagellin gene, CAC2203/*flaC*, CTC1715, and CD0239/*fliC*, respectively.

Interestingly, motility accessory factors (MAF) of the Cj1318 family were also identified within the polysaccharide biosynthesis loci of *C. botulinum* (CBO2728), *C. acetobutylicum* (CAC2168, CAC2196, and CAC2202), and *C. tetani* (CTC1697 and CTC1714). In addition to *Campylobacter*, members of the Cj1318 family have only been identified in the genomes of *Helicobacter*, *Treponema*, and *Leptospira*; all of which were reported to modify their flagellins with glycan (Wyss 1998; Schirm et al. 2003).

The location of the *C. botulinum*, *C. acetobutylicum*, and *C. tetani* polysaccharide biosynthesis loci adjacent to the flagellar operon and the similarity in both gene content and organization to the known flagellar glycosylation locus of *C. jejuni* strongly suggest that these loci are most probably involved in the glycosylation of flagellins in these clostridial species. This view is supported by several lines of evidence: (1) The whole locus as well as the flagellar operon are missing in *C. perfringens*. (2) The modification of *C. acetobutylicum* flagellin protein, CAC2203/FlaC, is sensitive to neuraminidase treatment, suggesting it is glycosylated with a sialic-acid-like sugar (Lyristis et al. 2000). (3) The flagellin of *C. difficile* has been reported to be post-translationally modified (Tasteyre et al. 2000), and the modification is likely to be glycan, however, with significantly fewer putative glycosylation genes associated with the flagellin genes. The modification probably differs from that of *C. botulinum*, *C. tetani*, and *C. acetobutylicum*; (4) Flagellin glycosylation has also been reported in *Clostridium tyrobutyricum* (Arnold et al. 1998).

C. botulinum has a second gene cluster that is also potentially involved in the biosynthesis of a surface polysaccharide structure, CBO3092–3114. Both *C. acetobutylicum* (CAC2310–2337) and *C. tetani* (CTC2252–2270) have similar loci at an equivalent location to that of *C. botulinum*. However, putative polysaccharide biosynthesis loci in *C. perfringens* (CPE0613–0629) and *C. difficile* (CD2767–2801) are located elsewhere in their genomes. In addition, the genome of *C. perfringens* carries another gene cluster, CPE0461–0511, which is absent in all the other sequenced clostridia, and which potentially encodes the necessary functions for capsule biosynthesis. Among the sequenced clostridia, the presence of a capsule-like structure has been reported only in *C. perfringens* (Sheng and Cherniak 1997) and *C. difficile* (Davies and Borriello 1990; Baldassarri et al. 1991).

Type IV pili are important for the tight adhesion of bacteria to host cells and solid surfaces and mediate bacterial twitching motility. All the sequenced clostridia appear to produce type IV pili as indicated by the presence of gene clusters that encode the necessary functions for the assembly of this cell surface appendage; CBO1900–1909, CAC2096–2105, CPE1836–1844, CTC1595–1604, CD3290–3297, and CD3503–3513, in *C. botulinum*, *C. acetobutylicum*, *C. perfringens*, *C. tetani*, and *C. difficile*, respectively. The prepilins encoded in the sequenced clostridia exhibit similarities to type IV prepilins (Craig et al. 2004). Among the sequenced clostridia, the presence of pili has been previously reported in *C. difficile* (Borriello et al. 1988) and more recently in *C. perfringens* (Varga et al. 2006).

S-layer proteins form crystalline arrays on the surface of numerous bacteria and archaea (Sara and Sleytr 2000). A small region of the genome was identified that contained CDS that encode several putative S-layer proteins (CBO00374, CBO0378, CBO0379, and CBO0380). These proteins all contain three cell wall binding domains, the role of which is to anchor proteins to the cell surface, and are similar to the S-layer protein, SlpA, of *C. difficile* (Calabi et al. 2001). The amino acid compositions of CBO0378 and CBO380 are comparable to that determined for an antigenic cell wall protein of proteolytic *C. botulinum* type A strain 190L (Takumi et al. 1983), suggesting that these genes may encode the S-layer proteins for Hall A.

Sporulation and germination

Sporulation is a key event in the life cycle of *C. botulinum*. There are 111 CDS to which sporulation and germination functions have been assigned.

Initiation of the sporulation cascade in aerobic *Bacillus* species is dependent on the concentration of Spo0A and its phosphorylation state (Hoch 1993). As an orphan response regulator, the phosphorylation state of Spo0A is dependent on several histidine kinases (such as KinA-E in *Bacillus subtilis*) that phosphorylate Spo0A via a phosphorelay system (Spo0F and Spo0B) (Burbulys et al. 1991; Strauch et al. 1992). The phosphorelay system is also negatively influenced by the activity of various phosphatases including RapA, RapB, and RapE, which are specific to Spo0F~P (Perego et al. 1996). The initial signaling steps that lead to phosphorylation of Spo0A in *Clostridium* species appear to be different from those in *Bacillus* species. *C. botulinum* Hall A lacks homologs of the *kin* histidine kinases, the phosphorelay system (*spo0F* and *spo0B*), and the *rap* phosphatases. Virtually all of these genes are also absent from other sequenced clostridia (for review, see Paredes et al. 2005). The *C. botulinum* genome carries five orphan kinases (CBO0336, CBO0340, CBO0780, CBO1120, and

CBO2762) that could potentially phosphorylate Spo0A. A majority of the other key genes of the *B. subtilis* sporulation cascade are present within the *C. botulinum* genome and those of other sequenced clostridia.

Germination of spores is believed to be mediated by receptors that reside in the inner spore membrane, which are encoded by tricistronic operons. There are three tricistronic germinant receptor gene clusters in the genome of *C. botulinum* Hall A, CBO0123–0125 (*gerXA1-XB1-XC1*), CBO1975–1977 (*gerXA2-XB2-XC2*), and CBO2797–2795 (*gerXA3-XB3-XC3*). One of these gene clusters (CBO1975–1977/*gerXA2-XB2-XC2*) is flanked by two additional *gerXB* genes, CBO1974 and CBO1978. In addition, there is one orphan *gerXB* homolog elsewhere in the chromosome (CBO2300). Germinant receptor gene clusters have been described previously in proteolytic *C. botulinum* (strain NCTC 7273) and *C. sporogenes* (strain NCIMB 701792) (Broussolle et al. 2002). The operon described in proteolytic *C. botulinum* strain NCTC 7273 is most closely related to the CBO2797–2795 operon (89%–99% amino acid identity of encoded proteins), while interestingly, the *gerAA* and *gerAB* genes described in *C. sporogenes* strain NCIMB 701792 are most closely related to the CBO1975–1977 operon (82%–84% amino acid identity of encoded proteins). Multiple copies of the *ger* clusters are also present in *C. acetobutylicum* (three copies) and *C. tetani* (four copies). Interestingly, *C. perfringens* appears to have only one bicistronic *ger* cluster (CPE648–649, lacking a *gerXB* homolog). Surprisingly, there are no *ger* genes similar to those of *Clostridium* and *Bacillus* species in the *C. difficile* genome (Sebaihia et al. 2006), suggesting that the initiation of the germination process in *C. difficile* is different from that in other *Clostridium* and *Bacillus* species.

Regulation

There have been few studies of the regulatory systems of *C. botulinum* to date. The genome contains 28 two-component systems, eight orphan histidine kinases, and eight orphan response regulators. The genome also contains 15 sigma factors, including six that are absent in other sequenced clostridia.

In low-GC Gram-positive bacteria, the PTS has been shown to play a crucial role in the regulation of catabolic genes and operons, as a result of the action of a metabolite-activated protein kinase (HPrK) that phosphorylates the PTS phosphocarrier protein HPr on a serine residue. HPr(ser)-P then interacts with the catabolite control protein CcpA to induce specific DNA binding and regulation of gene expression. In addition to HPr (CBO2398), genes encoding homologs of the essential elements of this mechanism of carbon catabolite repression, HPrK (CBO2608) and CcpA (CBO0100), are present in the *C. botulinum* genome, as is the case in other sequenced clostridia. The proteins of *C. botulinum*, *C. acetobutylicum*, and *C. perfringens* are closely related to each other (≥63% amino acid identity) but more distantly related to those in *C. difficile*. Nevertheless, it appears that all of these clostridia may exhibit a global mechanism of regulation of carbohydrate metabolism that is similar to that demonstrated in other Gram-positive organisms.

The one regulatory system that has been explored in any detail is that controlling neurotoxin synthesis. Expression of the *cntABCDE* genes of *C. botulinum* and *cntA* of *C. tetani* is regulated by CntR (previously designated BotR/TetR). These proteins function as specific alternative σ-factors that are seemingly related to a new subgroup of the sigma 70 family that includes TcdR of *C. difficile* and UviA of *C. perfringens*, which regulate production of toxins A and B and a bacteriocin, respectively (Raffestin et al. 2004; Dupuy and Matamouros 2006). The sequence to which CntR binds is very highly conserved (Dupuy and Matamouros 2006). A search of the genome reveals that there are no other examples of a sequence that closely conforms to this consensus, suggesting that the expression of those genes concerned with *C. botulinum* neurotoxin production are the only genes regulated by this specialized σ-factor.

Given the timing of neurotoxin production (late exponential and early stationary phase), it seems likely that neurotoxin gene expression may be subject to quorum sensing. Indeed, *C. botulinum* carries two pairs of genes, located in close proximity to one another, *agrB1/D1* (CBO0332 and CBO0332A) and *agrB2/D2* (CBO0338 and CBO0339), that encode homologs of components of the accessory gene regulator (*agr*) system of *Staphylococcus aureus* (Fig. 4). The staphylococcal locus comprises four genes—*agrC*, *agrA*, *agrB*, and *agrD*—and mediates the global regulation of a battery of virulence factors (for review, see Novick 2003). The secreted octapeptide signal molecule involved (AIP, autoinducer peptide) is derived from an internal fragment of AgrD through the action of the AgrB transmembrane protein. The *agrC* and *agrA* genes encode a sensor kinase and a response regulator of a two-component system, respectively. Although not identified during the annotation of the *C. perfringens* strain 13 genome, a gene encoding a protein equivalent to AgrD is present downstream from CPE1561/*agrB*, and for simplicity, this gene is termed CPE1560A (Fig. 4). An equivalent single pair of genes (*agrBD*) is also present in *C. acetobutylicum* (CAC0078 and CAC0079) and *C. difficile* (CD2750 and CD2749A) (Fig. 4) but not in *C. tetani*. All of

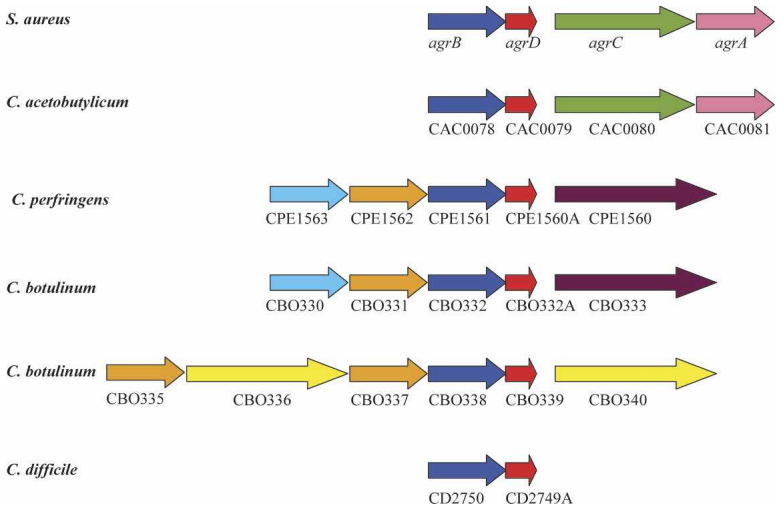

Figure 4. Organization of the accessory gene regulator (*agr*) loci in clostridia and *S. aureus*. Similar genes are shown in the same color. Significant functions are: (blue) autoinducer maturation protein (AgrB); (red) autoinducer peptide (AgrD); (green) two-component sensor kinase (AgrC); (pink) two-component response regulator (AgrA); (light blue and orange) membrane proteins; (purple) signal transduction proteins with GGDEF and HD domains; (yellow) two-component sensor kinases.

the clostridial AgrB homologs possess the Cys and His residues found in the conserved region within this class of protein (Qiu et al. 2005). Moreover, the putative AgrD peptides crucially retain the invariant Cys residue essential for post-translational production of AIPs, as well as the critical conserved C-terminal processing-signal motif "+EP+XP" (McDowell et al. 2001).

Consistent with a role in quorum sensing, CAC0078/*agrB* has recently been shown to be expressed at high levels at the late-exponential phase of growth in *C. acetobutylicum* (Alsaker and Papoutsakis 2005). This regulation is likely to be mediated via the structural equivalents of the staphylococcal *agrC* and *agrA* genes (CAC0080 and CAC0081) that are located immediately adjacent to CAC0079. In contrast, there are no two-component systems similar to the staphylococcal *agrC* and *agrA* genes within the vicinity of either of *C. botulinum argB1/D1* and *argB2/D2*, nor, indeed, of the *agrBD* homologs in *C. perfringens* or *C. difficile* (Fig. 4). In *C. botulinum*, the only two genes (CBO1186 and CBO1187) that encode a two-component system with significant similarity to AgrC and AgrA are present elsewhere in the genome.

However, it is noticeable that downstream from both the *C. botulinum argB1/D1* and its equivalent in *C. perfringens* lies one gene (CBO0333 and CPE1560, respectively; Fig. 4) that encodes a putative signal transduction protein containing the Pfam motifs GGDEF (PF00990) and HD (PF01966), both of which are found in signal transduction proteins. In addition, the second *agrBD* locus (*argB2/D2*) of *C. botulinum* is flanked by two genes encoding two orphan sensor kinases (CBO0336 and CBO0340) (Fig. 4) that are highly similar to each other (56% amino acid identity). It is not known at this stage if any of these proteins have any role in AIP sensing. If they do, this may occur via an unknown response regulator.

In the case of *C. perfringens*, the *virR*/*virS* two-component system plays a central role in virulence factor production (Shimizu et al. 2002b), through sensing of a small, diffusible molecule, substance "A," thought to be a peptide (Shimizu et al. 1997). The AgrD peptide (CPE1560A) may represent this signal molecule. In this respect, it is intriguing to note that the closest homologs of the *C. perfringens virR* gene in *C. difficile* and *C. botulinum* (CD3255/*rgaR* and CBO0575, respectively) are orphans. Interestingly, the orphan response regulator of *C. difficile* (CD3255/RgaR) has recently been shown to positively regulate the expression of the *agrB*/CD2750 gene (O'Connor et al. 2006).

Comparative genomic hybridizations

We have designed a DNA microarray that includes 94% of the predicted CDS of *C. botulinum* Hall A (ATCC 3502). Comparative genomic hybridization experiments were performed on nine proteolytic *C. botulinum* strains, producing different toxin types [A1, A3, A(B), B, and F], and two *C. sporogenes* strains, NCIMB 8053 and NCIMB 10696 (Fig. 5). A higher degree of relatedness was identified between strains of proteolytic *C. botulinum* than reported previously for *C. difficile* strains (Sebaihia et al. 2006). The two type A1 strains were most similar to Hall A, sharing 95%–96% of their CDS (Fig. 5). The other seven strains of proteolytic *C. botulinum* [toxin types, A3, A(B), B, and F] shared 87%–91% of the CDS, while the two nontoxigenic *C. sporogenes* strains shared 84%–87% of the CDS with Hall A. The strains generally clustered according to toxin type (Fig. 5), although it is interesting to note that *C. sporogenes* strain NCIMB 8053 associated more closely with several proteolytic *C. botulinum* strains than with *C. sporogenes* strain NCIMB 10696. The high divergence of the *Clostridium* genus is confirmed by the failure of DNA from *C. difficile* strain 630 and nonproteolytic *C. botulinum* strain Eklund 17B to hybridize to the array to any significant extent (data not shown). CDS for the two prophages (Φ-Cb1 and Φ-Cb2) represent ~3% of the Hall A genome, but are not present in any of the other strains, even the two type A1 neurotoxin producers. There is evidence, however, that isolated CDS within these prophages are shared, implying that other strains may harbor their own distinct

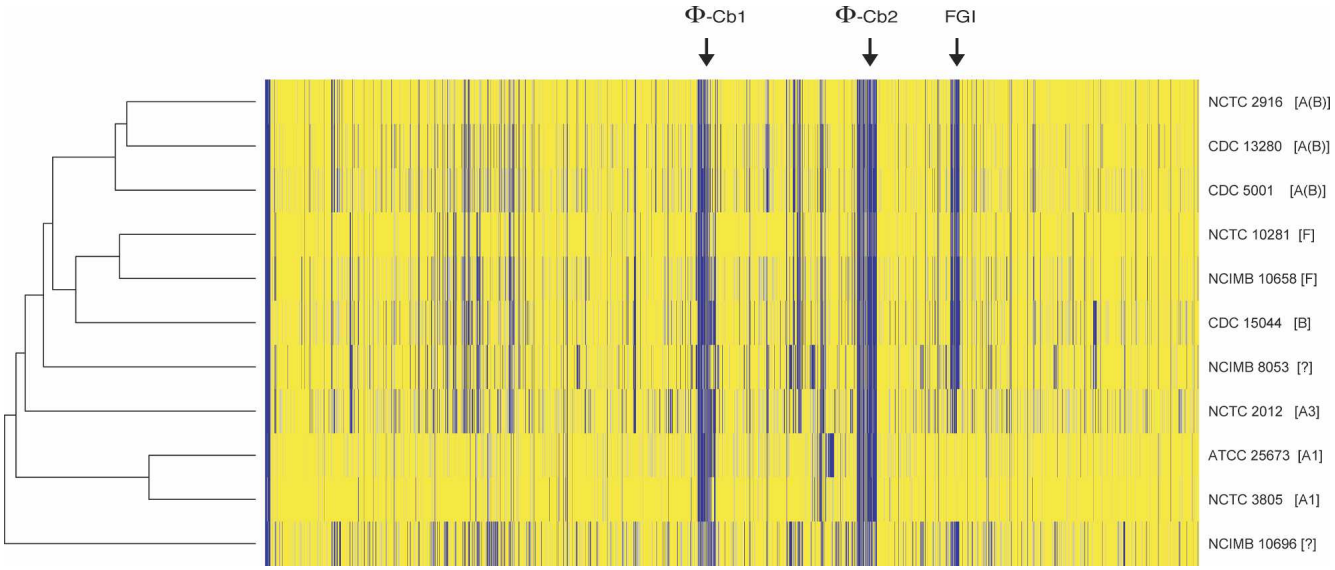

Figure 5. Comparative genomic analysis of strains of proteolytic *C. botulinum* and *C. sporogenes* strains using microarrays. Horizontal colored bars indicate array competitive hybridization. Vertical lines represent CDS present (yellow lines), absent or highly diverged (blue lines), and uncertain (gray lines). The plasmid CDS are on the *left*, followed immediately by the first gene in the chromosome (CBO001), and the last chromosomal CDS (CBO3650) is on the *right*. For strains of proteolytic *C. botulinum*, the toxin types are included in square brackets; two strains of *C. sporogenes* (NCIMB 8053 and NCIMB 10696) are also included. The position of the two prophages (Φ-Cb1 and Φ-Cb2) and the potential flagellin glycosylation island (FGI) are indicated by arrows.

prophages. Interestingly, the potential flagellar glycosylation island of Hall A strain (CBO2696–CBO2729, discussed above, and marked FGI in Fig. 5) is also present in the two other strains forming type A1 neurotoxin, but is missing in the other strains of *C. botulinum* and *C. sporogenes* examined. These data suggest that either the flagellins of the non-A1 neurotoxin strains are not post-translationally modified by glycosylation, or are glycosylated with different glycan structures, the synthesis of which is encoded by flagellar glycosylation genes that are substantially different from those of type A1 neurotoxin-producing strains. Polymorphism in the flagellin glycosylation islands and structural diversity of the glycan structures have been observed in different strains of *Pseudomonas aeruginosa* (Arora et al. 2004; Schirm et al. 2004).

Conclusion

Analysis of the *C. botulinum* genome revealed novel and interesting aspects of its lifestyle. The genome supports the view of *C. botulinum* as an essentially saprophytic organism that uses its toxin to rapidly kill a host for subsequent saprophytic utilization. The fact that it does not spend long periods associated with a living host may explain its relatively stable genome, contrasting with that of *C. difficile*, which can spend long periods coexisting with its host, and has a highly variable genome correlating with this highly dynamic niche. *C. botulinum* is both a proteolytic and chitinolytic bacterium, producing several extracellular proteases and chitinases. The combined action of the adhesins, extracellular matrix-binding proteins, proteases, and cytolysin may contribute to the softening and extensive destruction of tissues of rotting carcasses of dead animals; whereas the chitinolytic system may be deployed to degrade chitin-containing invertebrate species such as insects, fungi, and crustaceans. *C. botulinum* may well deploy the plasmid-encoded boticin to defend this nutrient-rich source against different microbial competitors. In addition to metabolism of peptides and amino acids, the bacterium has a significant capacity for uptake and metabolism of sugars and related molecules.

Like many of the clostridial genome projects before it, the data generated have provided a fascinating insight into the physiology of this pathogen. Exploitation of this information in hypothesis-driven research has, however, until now been severely impeded by a general absence of clostridial mutational tools for functional genomic studies. This deficiency has recently been solved through the generation of a gene knockout system for *C. difficile* (O'Connor et al. 2006) and a new system that has been shown to be applicable to a range of clostridial species (N.P. Minton, unpubl.). The availability of such systems will considerably aid future analysis of clostridial genomes.

Methods

ATCC 3502 is one of the most widely studied *C. botulinum* strains. It was deposited at the ATCC in 1940 by I.C. Hall, and is listed as Hall Strain 174. This strain belongs to *C. botulinum* Group I (proteolytic *C. botulinum*) and forms type A1 neurotoxin. The Health Protection Agency Center for Emergency Responsiveness and Preparedness at Porton Down (formerly the Center for Applied Microbiology, CAMR) obtained this strain from the ATCC in 1987 as freeze-dried ampoules from Batch MED-38 (prepared in 1981). The strain has been deposited with the NCTC as NCTC 13319.

Cloning and sequencing

The initial genome assembly was obtained from 69,632 paired end sequences (giving 9.15-fold coverage) derived from four genomic shotgun libraries (all in pUC18 with insert sizes of 1.5–2.0 kb and 2.0–2.2 kb, 2.2–2.5 kb, and 2.5–4.0 kb) using dye terminator chemistry on ABI3700 automated sequencers; 1604 paired-end sequences from one pBACe3.6 library with insert sizes of 15–23 kb (a clone coverage of 3.9-fold) were used as a scaffold. A further 9343 directed sequencing reads were generated during finishing.

Sequence analysis and annotation

The sequence was assembled, finished, and annotated as described previously (Sebaihia et al. 2006), using Artemis (Rutherford et al. 2000) to collate data and facilitate annotation. The DNA and predicted protein sequences of *C. botulinum* were compared to the sequenced clostridial genomes using the Artemis Comparison Tool (ACT) (Carver et al. 2005). Orthologous gene sets were calculated by reciprocal best-match FASTA comparisons with subsequent manual curation. Pseudogenes had one or more mutations that would prevent translation; each of the inactivating mutations was checked against the original sequencing data. Horizontally acquired DNA was identified by anomalies in G+C content and GC deviation, combined with the presence of mobility elements such as transposases and integrases. All such regions were subject to manual checking and curation.

Microarray design

A 100–500 bp probe for each CDS was produced by PCR using genomic DNA of *C. botulinum* Hall A (ATCC 3502) as template. PCR products were checked by agarose gel electrophoresis and purified by precipitation with isopropanol. Array printing buffer was 0.3× saline sodium citrate, 50% dimethyl sulfoxide (DMSO). The probes were spotted onto gamma amino propyl silane-coated GAPS II slides (Corning) using a Stanford arrayer (Thompson et al. 2001). Owing partly to the very low G+C content and to multiple-copy CDS, several loci failed primer and/or predicted PCR product specificity analysis (assessed using BLAST). It was only possible to include probes for 3433 CDS (out of 3650 predicted, ~94%). The array also included a probe for each CDS identified for the plasmid pBOT3502.

Preparation of genomic DNA

Genomic DNA was purified from exponentially growing cells by lysis in lysozyme/mutanolysin, incubation in proteinase K/SDS, followed by a standard phenol/chloroform extraction procedure. Final pellets were dissolved in TE (10 mM Tris-HCl at pH 8.0, 1 mM EDTA). Genomic DNA was digested to completion with Sau3A1 (Promega) and purified by phenol/chloroform extraction. Genomic DNA was labeled overnight (~14 h) with Cy5-dCTP or Cy3-dCTP fluorescent nucleotides (Amersham) using a BioPrime Array CGH kit (Invitrogen) and purified for hybridization using QiaQuick PCR cleanup columns (QIAGEN).

Microarray hybridization and data analysis

For each experiment, 2 µg of Cy5-labeled Hall A (reference) DNA and 2 µg of Cy3-labeled test DNA were hybridized onto the array under a glass coverslip in a humidified dual microarray hybridization chamber (Monterey Industries) overnight at 54°C. DNA microarrays were then scanned using an Axon GenePix 4000A microarray laser scanner (Axon Instruments) using GenePixPro 5.1 software. All data were subsequently analyzed using the R statistical language and environment (http://www.r-project.org),

specifically with the tools available from the Bioconductor Project (http://www.bioconductor.org), predominantly the LIMMA Bioconductor package (Smyth 2004; http://www.bioconductor.org/packages/bioc/1.9/html/limma.html). Fluorescence data were background-subtracted using a Bayesian model-based algorithm (Kooperberg et al. 2002), then normalized within each array, using "printtiploess" to correct for spatial and other artifacts. Scatterplots and/or box plots of the raw and normalized fluorescence values were manually inspected prior to proceeding with linear model fitting, according to the appropriate design matrix, with empirical Bayesian smoothing/moderation of the standard errors. The linear modeling was output as toptables, comprising $\log_2$-based coefficients (test vs. reference fold ratios), with appropriate significance statistics for each locus on the array in each genome tested. Coefficients were imported into GACK (Kim et al. 2002), and a trinary analysis was performed to assign present/uncertain/absent or diverged status to all loci in all samples. EPP thresholds of 0% and 100% were applied. GACK output was plotted as a heatmap, with loci ordered by physical location in the *C. botulinum* (Hall A) genome.

Chitin degradation assay

The ability to degrade chitin was determined by spotting *C. botulinum* cultures onto layered peptone-yeast extract-glucose-starch (PYGS) agar plates (Stringer et al. 2005), in which the upper layer additionally contained 2% finely ground chitin. Clearing was assessed after incubating for 8 d at 30°C under a headspace of H_2/CO_2 (90/10). Cell extract from the chitinolytic bacterium *Streptomyces griseus* was included for comparison.

Acknowledgments

We acknowledge the support of the Wellcome Trust Sanger Institute core sequencing and informatics groups. This work was funded by the Wellcome Trust, a competitive strategic grant from the BBSRC and a CRTI-IRTC operating grant.

References

Alsaker, K.V. and Papoutsakis, E.T. 2005. Transcriptional program of early sporulation and stationary-phase events in *Clostridium acetobutylicum*. *J. Bacteriol.* **187:** 7103–7118.

Anonymous. 2006. Wound botulism in injecting drug users in the United Kingdom. *CDR Weekly* http://www.camr.org.uk/cdr/archives/archive06/News/news1306.htm#bot.

Arnold, F., Bedouet, L., Batina, P., Robreau, G., Talbot, F., Lecher, P., and Malcoste, R. 1998. Biochemical and immunological analyses of the flagellin of *Clostridium tyrobutyricum* ATCC 25755. *Microbiol. Immunol.* **42:** 23–31.

Arnon, S.S. 2004. Infant botulism. In *Textbook of pediatric infectious disease*, 5th ed. (eds. R.D. Feigen and J.D. Cherry), pp. 1758–1766. Saunders, Philadelphia.

Arora, S.K., Wolfgang, M.C., Lory, S., and Ramphal, R. 2004. Sequence polymorphism in the glycosylation island and flagellins of *Pseudomonas aeruginosa*. *J. Bacteriol.* **186:** 2115–2122.

Bader, J., Rauschenbach, P., and Simon, H. 1982. On a hitherto unknown fermentation path of several amino acids by proteolytic clostridia. *FEBS Lett.* **140:** 67–72.

Baldassarri, L., Donelli, G., Cerquetti, M., and Mastrantonio, P. 1991. Capsule-like structures in *Clostridium difficile* strains. *Microbiologica* **14:** 295–300.

Barabote, R.D. and Saier Jr., M.H. 2005. Comparative genomic analyses of the bacterial phosphotransferase system. *Microbiol. Mol. Biol. Rev.* **69:** 608–634.

Borriello, S.P., Welch, A.R., Barclay, F.E., and Davies, H.A. 1988. Mucosal association by *Clostridium difficile* in the hamster gastrointestinal tract. *J. Med. Microbiol.* **25:** 191–196.

Brett, M.M., Hallas, G., and Mpamugo, O. 2004. Wound botulism in the UK and Ireland. *J. Med. Microbiol.* **53:** 555–561.

Broussolle, V., Alberto, F., Shearman, C.A., Mason, D.R., Botella, L., Nguyen-The, C., Peck, M.W., and Carlin, F. 2002. Molecular and physiological characterisation of spore germination in *Clostridium botulinum* and *C. sporogenes*. *Anaerobe* **8:** 89–100.

Bruggemann, H., Baumer, S., Fricke, W.F., Wiezer, A., Liesegang, H., Decker, I., Herzberg, C., Martinez-Arias, R., Merkl, R., Henne, A., et al. 2003. The genome sequence of *Clostridium tetani*, the causative agent of tetanus disease. *Proc. Natl. Acad. Sci.* **100:** 1316–1321.

Burbulys, D., Trach, K.A., and Hoch, J.A. 1991. Initiation of sporulation in *B. subtilis* is controlled by a multicomponent phosphorelay. *Cell* **64:** 545–552.

Calabi, E., Ward, S., Wren, B., Paxton, T., Panico, M., Morris, H., Dell, A., Dougan, G., and Fairweather, N. 2001. Molecular characterization of the surface layer proteins from *Clostridium difficile*. *Mol. Microbiol.* **40:** 1187–1199.

Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.A., Barrell, B.G., and Parkhill, J. 2005. ACT: The Artemis comparison tool. *Bioinformatics* **21:** 3422–3423.

Craig, L., Pique, M.E., and Tainer, J.A. 2004. Type IV pilus structure and bacterial pathogenicity. *Nat. Rev. Microbiol.* **2:** 363–378.

Dargatz, H., Diefenthal, T., Witte, V., Reipen, G., and von Wettstein, D. 1993. The heterodimeric protease clostripain from *Clostridium histolyticum* is encoded by a single gene. *Mol. Gen. Genet.* **240:** 140–145.

Datta, V., Myskowski, S.M., Kwinn, L.A., Chiem, D.N., Varki, N., Kansal, R.G., Kotb, M., and Nizet, V. 2005. Mutational analysis of the group A streptococcal operon encoding streptolysin S and its virulence role in invasive infection. *Mol. Microbiol.* **56:** 681–695.

Davies, H.A. and Borriello, S.P. 1990. Detection of capsule in strains of *Clostridium difficile* of varying virulence and toxigenicity. *Microb. Pathog.* **9:** 141–146.

Dekleva, M.L. and Dasgupta, B.R. 1990. Purification and characterization of a protease from *Clostridium botulinum* type A that nicks single-chain type A botulinum neurotoxin into the di-chain form. *J. Bacteriol.* **172:** 2498–2503.

Dezfulian, M. and Dowell Jr., V.R. 1980. Cultural and physiological characteristics and antimicrobial susceptibility of *Clostridium botulinum* isolates from foodborne and infant botulism cases. *J. Clin. Microbiol.* **11:** 604–609.

Dickert, S., Pierik, A.J., and Buckel, W. 2002. Molecular characterization of phenyllactate dehydratase and its initiator from *Clostridium sporogenes*. *Mol. Microbiol.* **44:** 49–60.

Dineen, S.S., Bradshaw, M., and Johnson, E.A. 2000. Cloning, nucleotide sequence, and expression of the gene encoding the bacteriocin boticin B from *Clostridium botulinum* strain 213B. *Appl. Environ. Microbiol.* **66:** 5480–5483.

Dineen, S.S., Bradshaw, M., and Johnson, E.A. 2003. Neurotoxin gene clusters in *Clostridium botulinum* type A strains: Sequence comparison and evolutionary implications. *Curr. Microbiol.* **46:** 345–352.

Dupuy, B. and Matamouros, S. 2006. Regulation of toxin and bacteriocin synthesis in *Clostridium* species by a new subgroup of RNA polymerase sigma-factors. *Res. Microbiol.* **157:** 201–205.

Fox, C.K., Keet, C.A., and Strober, J.B. 2005. Recent advances in infant botulism. *Pediatr. Neurol.* **32:** 149–154.

Hoch, J.A. 1993. The phosphorelay signal transduction pathway in the initiation of *Bacillus subtilis* sporulation. *J. Cell. Biochem.* **51:** 55–61.

Howard, M.B., Ekborg, N.A., Taylor, L.E., Weiner, R.M., and Hutcheson, S.W. 2003. Genomic analysis and initial characterization of the chitinolytic system of *Microbulbifer degradans* strain 2-40. *J. Bacteriol.* **185:** 3352–3360.

Hutson, R.A., Thompson, D.E., and Collins, M.D. 1993a. Genetic interrelationships of saccharolytic *Clostridium botulinum* types B, E and F and related clostridia as revealed by small-subunit rRNA gene sequences. *FEMS Microbiol. Lett.* **108:** 103–110.

Hutson, R.A., Thompson, D.E., Lawson, P.A., Schocken-Itturino, R.P., Bottger, E.C., and Collins, M.D. 1993b. Genetic interrelationships of proteolytic *Clostridium botulinum* types A, B, and F and other members of the *Clostridium botulinum* complex as revealed by small-subunit rRNA gene sequences. *Antonie Van Leeuwenhoek* **64:** 273–283.

Kil, K.S., Cunningham, M.W., and Barnett, L.A. 1994. Cloning and sequence analysis of a gene encoding a 67-kilodalton myosin-cross-reactive antigen of *Streptococcus pyogenes* reveals its similarity with class II major histocompatibility antigens. *Infect. Immun.* **62:** 2440–2449.

Kim, K.K., Yokota, H., and Kim, S.H. 1999. Four-helical-bundle structure of the cytoplasmic domain of a serine chemotaxis receptor. *Nature* **400:** 787–792.

Kim, C.C., Joyce, E.A., Chan, K., and Falkow, S. 2002. Improved analytical methods for microarray-based genome-composition analysis. *Genome Biol.* doi: 10.1186/gb-2002-3-11-research0065.

Kim, J., Darley, D., and Buckel, W. 2005. 2-Hydroxyisocaproyl-CoA dehydratase and its activator from *Clostridium difficile*. *FEBS J.* **272:** 550–561.

Kooperberg, C., Fazzio, T.G., Delrow, J.J., and Tsukiyama, T. 2002. Improved background correction for spotted DNA microarrays. *J. Comput. Biol.* **9:** 55–66.

Lund, B.M. and Peck, M.W. 2000. *Clostridium botulinum*. In *The microbiological safety and quality of food* (eds. B.M. Lund et al.), pp. 1057–1109. Aspen, Gaithersburg, MD.

Lyristis, M., Boynton, Z.L.B., Petersen, D., Kan, K., Bennett, G.N., and Rudolph, F.B. 2000. Cloning, sequencing, and characterization of the gene encoding flagellin, *flaC*, and the post-translational modification of flagellin, FlaC, from *Clostridium acetobutylicum* ATCC824. *Anaerobe* **6:** 69–79.

McDowell, P., Affas, Z., Reynolds, C., Holden, M.T., Wood, S.J., Saint, S., Cockayne, A., Hill, P.J., Dodd, C.E., Bycroft, B.W., et al. 2001. Structure, activity and evolution of the group I thiolactone peptide quorum-sensing system of *Staphylococcus aureus*. *Mol. Microbiol.* **41:** 503–512.

Mills, D.C., Midura, T.F., and Arnon, S.S. 1985. Improved selective medium for the isolation of lipase-positive *Clostridium botulinum* from feces of human infants. *J. Clin. Microbiol.* **21:** 947–950.

Montgomery, M.T. and Kirchman, D.L. 1994. Induction of chitin-binding proteins during the specific attachment of the marine bacterium *Vibrio harveyi* to chitin. *Appl. Environ. Microbiol.* **60:** 4284–4288.

Myers, G.S., Rasko, D.A., Cheung, J.K., Ravel, J., Seshadri, R., DeBoy, R.T., Ren, Q., Varga, J., Awad, M.M., Brinkac, L.M., et al. 2006. Skewed genomic variability in strains of the toxigenic bacterial pathogen, *Clostridium perfringens*. *Genome Res.* **16:** 1031–1040.

Nioche, P., Berka, V., Vipond, J., Minton, N., Tsai, A.L., and Raman, C.S. 2004. Femtomolar sensitivity of a NO sensor from *Clostridium botulinum*. *Science* **306:** 1550–1553.

Nolling, J., Breton, G., Omelchenko, M.V., Makarova, K.S., Zeng, Q., Gibson, R., Lee, H.M., Dubois, J., Qiu, D., Hitti, J., et al. 2001. Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*. *J. Bacteriol.* **183:** 4823–4838.

Novick, R.P. 2003. Autoinduction and signal transduction in the regulation of staphylococcal virulence. *Mol. Microbiol.* **48:** 1429–1449.

O'Connor, J.R., Lyras, D., Farrow, K.A., Adams, V., Powell, D.R., Hinds, J., Cheung, J.K., and Rood, J.I. 2006. Construction and analysis of chromosomal *Clostridium difficile* mutants. *Mol. Microbiol.* **61:** 1335–1351.

Paredes, C.J., Alsaker, K.V., and Papoutsakis, E.T. 2005. A comparative genomic view of clostridial sporulation and physiology. *Nat. Rev. Microbiol.* **3:** 969–978.

Paul, C.J., Twine, S.M., Tam, K.J., Mullen, J.A., Kelly, J.F., Austin, J.W., and Logan, S.M. 2007. Flagellin diversity in *Clostridium botulinum* Groups I and II: A new strategy for strain identification. *Appl. Environ. Microbiol.* **73:** 2963–2975.

Peck, M.W. 2005. *Clostridium botulinum*. In *Understanding pathogen behaviour* (ed. M. Griffiths), pp. 531–548. Woodhead Press, Cambridge, UK.

Peck, M.W. 2006. *Clostridium botulinum* and the safety of minimally heated, chilled foods: An emerging issue? *J. Appl. Microbiol.* **101:** 556–570.

Perego, M., Glaser, P., and Hoch, J.A. 1996. Aspartyl-phosphate phosphatases deactivate the response regulator components of the sporulation signal transduction system in *Bacillus subtilis*. *Mol. Microbiol.* **19:** 1151–1157.

Qiu, R., Pei, W., Zhang, L., Lin, J., and Ji, G. 2005. Identification of the putative staphylococcal AgrB catalytic residues involving the proteolytic cleavage of AgrD to generate autoinducing peptide. *J. Biol. Chem.* **280:** 16695–16704.

Raffestin, S., Marvaud, J.C., Cerrato, R., Dupuy, B., and Popoff, M.R. 2004. Organization and regulation of the neurotoxin genes in *Clostridium botulinum* and *Clostridium tetani*. *Anaerobe* **10:** 93–100.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B. 2000. Artemis: Sequence visualization and annotation. *Bioinformatics* **16:** 944–945.

Sara, M. and Sleytr, U.B. 2000. S-Layer proteins. *J. Bacteriol.* **182:** 859–868.

Schirm, M., Soo, E.C., Aubry, A.J., Austin, J., Thibault, P., and Logan, S.M. 2003. Structural, genetic and functional characterization of the flagellin glycosylation process in *Helicobacter pylori*. *Mol. Microbiol.* **48:** 1579–1592.

Schirm, M., Arora, S.K., Verma, A., Vinogradov, E., Thibault, P., Ramphal, R., and Logan, S.M. 2004. Structural and genetic characterization of glycosylation of type a flagellin in *Pseudomonas aeruginosa*. *J. Bacteriol.* **186:** 2523–2531.

Scott, W.G., Milligan, D.L., Milburn, M.V., Prive, G.G., Yeh, J., Koshland Jr., D.E., and Kim, S.H. 1993. Refined structures of the ligand-binding domain of the aspartate receptor from *Salmonella typhimurium*. *J. Mol. Biol.* **232:** 555–573.

Sebaihia, M., Wren, B.W., Mullany, P., Fairweather, N.F., Minton, N., Stabler, R., Thomson, N.R., Roberts, A.P., Cerdeno-Tarraga, A.M., Wang, H., et al. 2006. The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. *Nat. Genet.* **38:** 779–786.

Sheng, S. and Cherniak, R. 1997. Structure of the capsular polysaccharide of *Clostridium perfringens* Hobbs 10 determined by NMR spectroscopy. *Carbohydr. Res.* **305:** 65–72.

Shimizu, T., Okabe, A., and Rood, J.I. 1997. Regulation of toxin production in *Clostridium perfringens*. In *The Clostridia: Molecular biology & pathogenesis* (eds. J.I. Rood et al.), pp. 451–470. Academic Press, London.

Shimizu, T., Ohtani, K., Hirakawa, H., Ohshima, K., Yamashita, A., Shiba, T., Ogasawara, N., Hattori, M., Kuhara, S., and Hayashi, H. 2002a. Complete genome sequence of *Clostridium perfringens*, an anaerobic flesh-eater. *Proc. Natl. Acad. Sci.* **99:** 996–1001.

Shimizu, T., Shima, K., Yoshino, K., Yonezawa, K., Shimizu, T., and Hayashi, H. 2002b. Proteome and transcriptome analysis of the virulence genes regulated by the VirR/VirS system in *Clostridium perfringens*. *J. Bacteriol.* **184:** 2587–2594.

Smith, L.D.S. and Sugiyama, H. 1988. *Botulism: The organism, its toxins, the disease*. Charles C. Thomas, Springfield, IL.

Smith, T.J., Lou, J., Geren, I.N., Forsyth, C.M., Tsai, R., Laporte, S.L., Tepp, W.H., Bradshaw, M., Johnson, E.A., Smith, L.A., et al. 2005. Sequence variation within botulinum neurotoxin serotypes impacts antibody binding and neutralization. *Infect. Immun.* **73:** 5450–5457.

Smyth, G.K. 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Applic. Genet. Mol. Biol.* http://www.bepress.com/sagmb/vol3/iss1/art3.

Stickland, L.H. 1935. Studies in the metabolism of the strict anaerobes (genus *Clostridium*): The oxidation of alanine by *Cl. sporogenes*. IV. The reduction of glycine by *Cl. sporogenes*. *Biochem. J.* **29:** 889–898.

Strauch, M.A., de Mendoza, D., and Hoch, J.A. 1992. *Cis*-unsaturated fatty acids specifically inhibit a signal-transducing protein kinase required for initiation of sporulation in *Bacillus subtilis*. *Mol. Microbiol.* **6:** 2909–2917.

Stringer, S.C., Webb, M.D., George, S.M., Pin, C., and Peck, M.W. 2005. Heterogeneity of times required for germination and outgrowth from single spores of nonproteolytic *Clostridium botulinum*. *Appl. Environ. Microbiol.* **71:** 4998–5003.

Szymanski, C.M., Logan, S.M., Linton, D., and Wren, B.W. 2003. *Campylobacter*—A tale of two protein glycosylation systems. *Trends Microbiol.* **11:** 233–238.

Takumi, K., Takeoka, A., and Kawata, T. 1983. Purification and characterization of a wall protein antigen from *Clostridium botulinum* type A. *Infect. Immun.* **39:** 1346–1353.

Tasteyre, A., Barc, M.C., Karjalainen, T., Dodson, P., Hyde, S., Bourlioux, P., and Borriello, P. 2000. A *Clostridium difficile* gene encoding flagellin. *Microbiology* **146:** 957–966.

Thompson, A., Lucchini, S., and Hinton, J.C. 2001. It's easy to build your own microarrayer! *Trends Microbiol.* **9:** 154–156.

Vaaje-Kolstad, G., Horn, S.J., van Aalten, D.M., Synstad, B., and Eijsink, V.G. 2005. The non-catalytic chitin-binding protein CBP21 from *Serratia marcescens* is essential for chitin degradation. *J. Biol. Chem.* **280:** 28492–28497.

Varga, J.J., Nguyen, V., O'Brien, K., Rodgers, D.K., Walker, R.A., and Melville, S.B. 2006. Type IV pili-dependent gliding motility in the Gram-positive pathogen *Clostridium perfringens* and other Clostridia. *Mol. Microbiol.* **62:** 680–694.

Werner, S.B., Passaro, D., McGee, J., Schechter, R., and Vugia, D.J. 2000. Wound botulism in California, 1951–1998: Recent epidemic in heroin injectors. *Clin. Infect. Dis.* **31:** 1018–1024.

Whitmer, M.E. and Johnson, E.A. 1988. Development of improved defined media for *Clostridium botulinum* serotypes A, B, and E. *Appl. Environ. Microbiol.* **54:** 753–759.

Wyss, C. 1998. Flagellins, but not endoflagellar sheath proteins, of *Treponema pallidum* and of pathogen-related oral spirochetes are glycosylated. *Infect. Immun.* **66:** 5751–5754.

Received January 14, 2007; accepted in revised form April 10, 2007.

Genome sequence of a proteolytic (Group I) *Clostridium botulinum* strain Hall A and comparative analysis of the clostridial genomes

Mohammed Sebaihia, Michael W. Peck, Nigel P. Minton, et al.

Genome Res. published online May 22, 2007
Access the most recent version at doi:10.1101/gr.6282807