Gut Microbes 1:4, 269-276; July/August 2010; © 2010 Landes Bioscience

In-depth genetic analysis of *Clostridium difficile* PCR-ribotype 027 strains reveals high genome fluidity including point mutations and inversions

Richard A. Stabler,[1] Esmeralda Valiente,[1] Lisa F. Dawson,[1] Miao He,[2] Julian Parkhill[2] and Brendan W. Wren[1,*]

[1]London School of Hygiene and Tropical Medicine; London, UK; [2]Wellcome Trust Sanger Institute; Cambridge, UK

Previously, we demonstrated that the recently evolved PCR-ribotype 027 hypervirulent *Clostridium difficile* strain (R20291) has acquired five genetic regions compared to the historic 027 counterpart strain (CD196), that may in part explain phenotypic traits relating to survival, antimicrobial resistance and virulence. Closer scrutiny of the three genome sequences reveals that, in addition to gene gain/loss, point mutations and inversions appear to have accumulated. Inversions are located upstream of potential coding sequences and could affect expression of these. *C. difficile* has a highly fluid genome with multiple mechanisms to modify its genetic content and is continuing to evolve in our hospitals influenced by environmental changes and human activity.

Key words: *Clostridium difficile*, 027 ribotype, point mutations, inversions, hypervirulence

Submitted: 01/28/10

Revised: 03/09/10

Accepted: 03/16/10

Previously published online: www.landesbioscience.com/journals/gutmicrobes/article/11870

*Correspondence to:
Brendan W. Wren; Email: Brendan.Wren@lshtm.ac.uk

Addendum to: Stabler RA, He M, Dawson L, Martin M, Valiente E, Corton C et al. Comparative genome and phenotypic analysis of *Clostridium difficile* 027 strains provides insight into the evolution of a hypervirulent bacterium. Genome Biol 2009; 10:R102; PMID: 19781061; DOI: 10.1186/gb-2009-10-9-r102.

Introduction

Clostridium difficile is a Gram-positive, anaerobic, spore-forming bacillus that is the leading cause of nosocomial diarrhea worldwide.[1] *C. difficile* is a unique pathogen that often predominates in the bowel microflora as a result of the microbial compositional changes which follow antibiotic treatment. The hospital environment and patients undergoing antibiotic treatment provide a discrete ecosystem where *C. difficile* persists and virulent clones thrive. The continued rise of *C. difficile* infection (CDI) worldwide has been accompanied by the rapid emergence and transcontinental dissemination of a highly virulent clone, designated PCR-ribotype 027.[2] These strains have risen from obscurity to become the most frequently isolated *C. difficile* strain types. Additionally, patients infected with these strains often experience more severe diarrhea, more recurrent episodes and higher mortality.[3-7] The emergence of 027 strains partly explains the 35-fold increase in reported incidence of CDI in the United Kingdom in the last decade.

In a recent study, we compared the genomes of a historic 027 strain (CD196, isolated in France in 1984) with a modern hypervirulent strain (R20291, isolated in 2006 and the index case of epidemic 027 infection in the UK) and showed that this modern strain has five additional genetic regions compared to its historic counterpart. Furthermore both the 027 strains have an additional 234 genes compared to *C. difficile* 630 (a PCR-ribotype 012 strain, isolated from a patient in Zurich, Switzerland in 1982) and the only other reported full genome sequence of a *C. difficile* strain.[8] The implications of these studies are that the additional genes may account for the marked increase in disease capability (gain-of-trait-function). However, in bacteria there are other mechanisms of genetic variation, and perhaps counter intuitively gene re-arrangements and gene loss can be equally important in the evolution of virulence.[9] In this addendum we take a closer look at the 027 genome sequence data to reveal potential point mutations and inversions, which could contribute to 027 hypervirulence.

C. difficile Point Mutations

Prior to our sequence analysis it was known that some important genes in

Table 1. Potential disrupted genes.

630	Interruption	R20291	Interruption	CD196	Interruption	Function/annotation
CD0541	F	CDR20291_0466	F	CD196_0481	FS	chemotaxis protein methyltransferase (*cheR*)
CD1181	F	CDR20291_1019	F	CD196_1041	FS	malonyl coa-acyl carrier protein transacylase (*fabD*)
CD2501	F	CDR20291_2393	F	CD196_2346	FS	putative hydrolase
CD2541	F	CDR20291_2428	F	CD196_2381	FS	sodium:dicarboxylate symporter family protein
CD0182	F	CDR20291_0183	FS	CD196_0195A	F	putative membrane protein precursor
CD1045	F	CDR20291_0901	FS	CD196_0922	F	putative membrane protein
CD1678A	F	CDR20291_1576	FS	CD196_1601	F	hypothetical protein
CD2475	F	CDR20291_2368	FS	CD196_2321	F	putative competence membrane protein
CD0682	F	CDR20291_0608	FS (I)	CD196_0626	FS (I)	putative sodium:solute symporter
CD2126	F	CDR20291_2033	FS	CD196_1990	FS	putative membrane protein precursor
CD0332	F [Q/caa]	CDR20291_0337	Ochre (I) [taa]	CD196_0351	Ochre (I) [taa]	putative exosporium glycoprotein
CD1761	F [Q/caa]	CDR20291_1656	Ochre [taa]	CD196_1681	Ochre [taa]	conserved hypothetical protein
CD0672	F [G/gga]	CDR20291_0595	Opal [tga]	CD196_0613	Opal [tga]	putative uncharacterized protein
CD0157	FS	CDR20291_0156	F	CD196_0169	F	putative membrane protein
CD0348	FS	CDR20291_0353	F	CD196_0367	F	conserved hypothetical protein
CD0525	FS	CDR20291_0451	F	CD196_0465	F	putative aminobenzoyl-glutamate transporter
CD1388	FS	CDR20291_1234	F	CD196_1257	F	putative transcriptional regulator
CD1426	FS	CDR20291_1273	F	CD196_1296	F	putative isochorismatase
CD1982	FS	CDR20291_1907	F	CD196_1864	F	conserved hypothetical protein
CD2267	FS	CDR20291_2166	F	CD196_2123	F	putative membrane-associated caaX amino terminal protease
CD3020	FS	CDR20291_2856	F	CD196_2809	F	conserved hypothetical protein
CD3156A	FS	CDR20291_3008	F	CD196_2961	F	conserved hypothetical protein
CD3185	FS	CDR20291_3041	F	CD196_2995	F	conserved hypothetical protein
CD3674	FS	CDR20291_3534	F	CD196_3488	F	methyltransferase (putative glucose inhibited division protein B)
CD0196	FS	CDR20291_0197	FS	CD196_0209	FS	conserved hypothetical protein
CD0440A	FS	Not annotated	(FS)	Not annotated	(FS)	regulatory protein (partial)
CD1718	I	CDR20291_1617	F	CD196_1642	FS	putative hydantoinase
CD1990A	Amber	Not annotated	(Amber)	Not annotated	(Amber)	putative regulatory protein
CD0857	Amber [tag]	CDR20291_0787	F [S/tcg]	CD196_0806	F [S/tcg]	oligopeptide ABC transporter, ATP-binding protein
CD3611	Ochre [taa]	CDR20291_3450	F [Q/caa]	CD196_3404	F [Q/caa]	putative multidrug resistance protein
CD0858	Ochre	CDR20291_0788	Ochre	CD196_0807	Ochre	putative transcription antiterminator
CD1741	Opal	CDR20291_1638	(Opal)	CD196_1663	(Opal)	sarcosine reductase complex component b beta subunit.
CD1809	Opal	CDR20291_1704	(Opal)	CD196_1729	(Opal)	putative multi-drug resistance efflux pump
CD2351	Opal	CDR20291_2239	(Opal)	CD196_2193	(Opal)	glycine reductase complex component B gamma subunit.
CD2352	Opal	CDR20291_2240	Opal	CD196_2194	Opal	glycine/sarcosine/betaine reductase complex component A.

Homologous CDSs between 630 (PCR-ribotype 012) and two PCR-ribotype 027s; CD196 (historic) and R20291 (epidemic). F, Uninterrupted CDS; FS, frame shift; Amber, point mutation (pm) resulting in a TAG stop codon; Ochre, TAA stop codon point mutation; Opal, TGA selenocysteine incorporation codon; I, interruption due to insertion of transposase-like protein B; (), different/missing CDS annotation but >98% amino acid identity (including interruption) was present in genome sequence, [X/xxx] indicates amino acid (X) and DNA sequence (xxx) present in uninterrupted CDS, [xxx] indicates sequence in interrupted CDS; (1), also truncated due to loss of repeats.

CD2362	Opal	CDR20291_2249	Opal	CD196_2203	Opal	putative aliphatic sulfonates ABC transporter, permease protein
CD2496	Opal	CDR20291_2388	(Opal)	CD196_2341	(Opal)	selenide, water dikinase
CD3241	Opal	CDR20291_3101	Opal	CD196_3055	Opal	proline reductase
CD3317	Opal	CDR20291_3179	Opal	CD196_3133	Opal	formate dehydrogenase H (*fdhF*)

Homologous CDSs between 630 (PCR-ribotype 012) and two PCR-ribotype 027s; CD196 (historic) and R20291 (epidemic). F, Uninterrupted CDS; FS, frame shift; Amber, point mutation (pm) resulting in a TAG stop codon; Ochre, TAA stop codon point mutation; Opal, TGA selenocysteine incorporation codon; I, interruption due to insertion of transposase-like protein B; (), different/missing CDS annotation but >98% amino acid identity (including interruption) was present in genome sequence, [X/xxx] indicates amino acid (X) and DNA sequence (xxx) present in uninterrupted CDS, [xxx] indicates sequence in interrupted CDS; (1), also truncated due to loss of repeats.

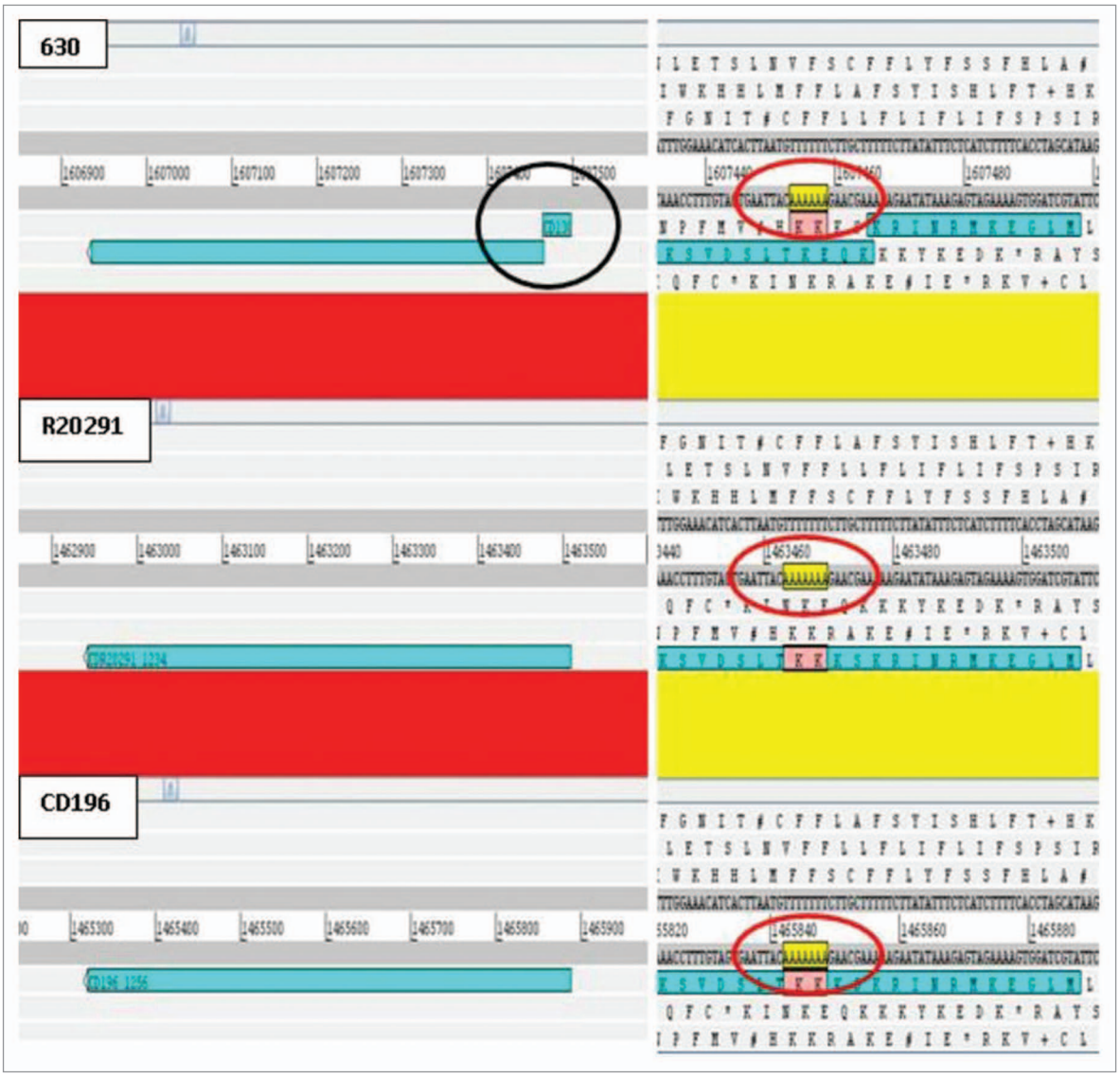

Figure 1. ACT comparison showing 630-specific point mutation in CD1388 resulting in frame shift. Homopolymeric adenosine tract in 630 contains 6 adenosine resulting in a frame shift, which is not present in both PCR ribotype 027 isolates R20291 and CD196 (7 adenosine residues). Red bars indicates ≥98% homology between DNA sequences.

C. difficile have undergone point mutations. These include multiple mutations in the actin-specific ADP-ribosylating toxin in several strains, point mutations in the negative toxin regulator particularly in 027 strains and point mutations in the *gyrA* gene of fluoroquinolone-resistant strains.[10] Re-analysis of the 630, R20291 and CD196 genomes

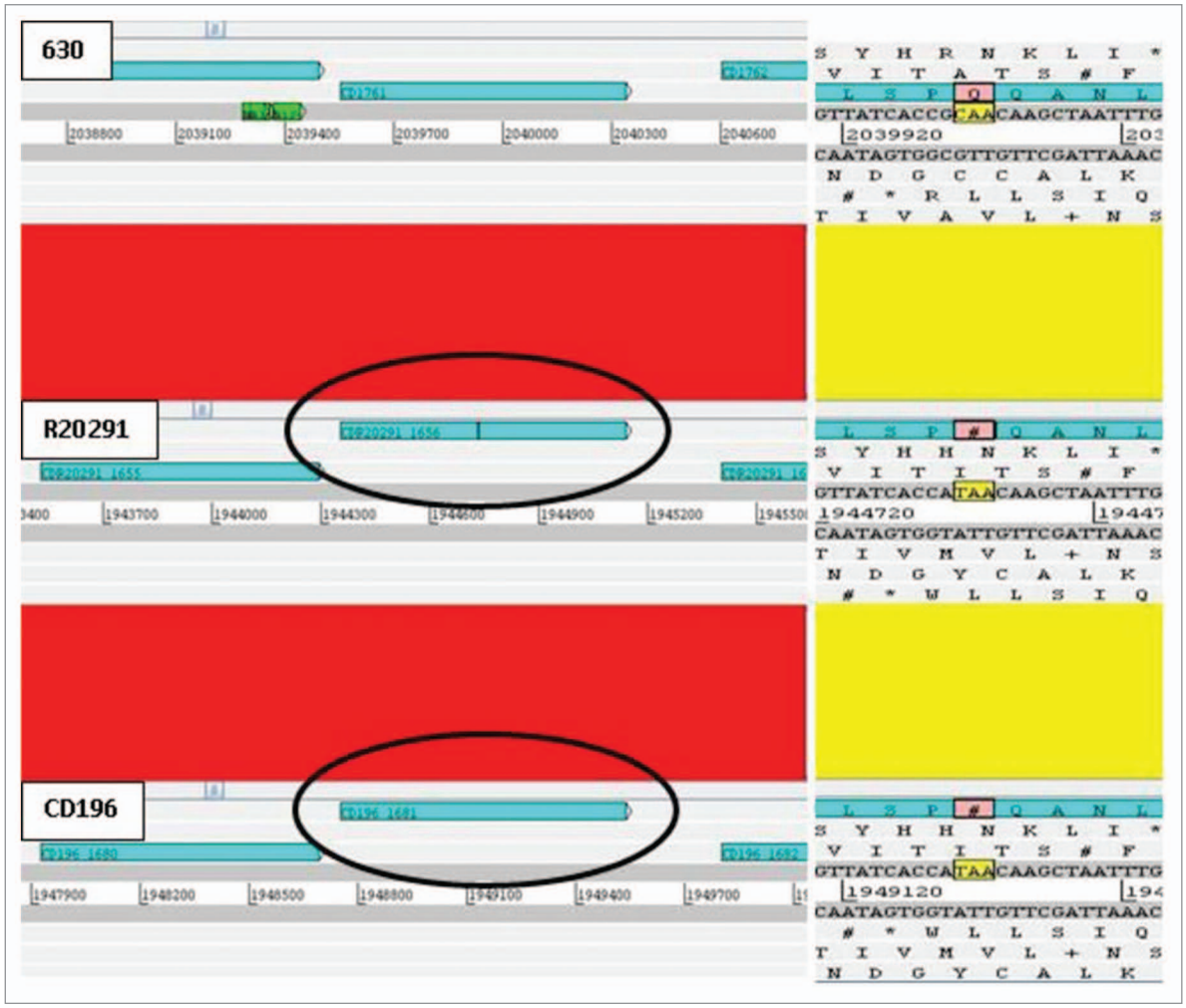

Figure 2. PCR ribotype 027 specific SNP results in premature translation stop. *C. difficile* 630 CDS CD1761 orthologues in R20291 and CD196 contain two adjacent single nucleotide polymorphisms (SNPs); a synonymous G→A and a non-synonymous C→T which results in the introduction of an ochre stop codon (TAA) at amino acid 126. Red bars indicates ≥99% homology between DNA sequences.

identified 39 coding sequences (CDSs) present in all three isolates, in which at least one orthologue contained an apparent inactivating mutation (**Table 1**). 13 CDSs contained interruptions in all three isolates e.g., CD1809 putative MDR efflux pump. 12 inactivating point mutations were specific to 630 [e.g., CD1388 putative transcriptional regulator (**Fig. 1** and **Suppl. Fig. 1**)], four specific to R20291 (e.g., CDR20291_2368 putative competence membrane protein) and four specific to CD196 (e.g., CD196_0481 chemotaxis protein methyltransferase *cheR*). Five interruptions were conserved in both 027 isolates (e.g., CDR20291_1656/CD196_1681 putative membrane protein precursor [**Fig. 2** and **Suppl. Fig. 2**)]. Interruptions were due to either frame shifts (23/39) or point mutations resulting in introduction of stop codons (four ochre, nine opal and two amber). A single CDS was functional in R20291, contained a frame shift due to an additional adenosine in CD196 (AAT GCA to AAT **A**GC A) and was disrupted by a copy of transposase-like protein B in 630 (**Fig. 3**). Interestingly *bclA1,* which is fully functional in 630, contained both a point mutation and loss of repeats in both 027 isolates. In order to confirm this was not an error of sequence assembly, PCR and sequencing analysis was performed (unpublished).

C. difficile Putative Phase Variation

Recently, confirmation of the first example of phase variation has been demonstrated in *C. difficile* (strain 630).[11] Expression of *cwpV* (CD0514) that encodes a surface protein (CwpV) is switched on or off via DNA inversion by a site-specific recombinase.[11] Comparative analysis of the three genomes revealed putative inversions. Three intergenic inversions, including the *cwpV* inversion (**Fig. 4**), were detected in all three strains, present in both orientations (**Table 2**). Interestingly *C. difficile* inversion (Cdi) 1 was annotated in the 'off' position in 630 and R20291 but 'on' in CD196 (**Fig. 4**). The other two additional inversions were located upstream of putative signaling proteins. Cdi2 was located 872 bp upstream of CD0757/CDR20291_0685/CD196_0704 (**Fig. 5**), in the same orientation in 630 and R20291, but inverted in CD196. However, no left inverted repeat (LIR) or right inverted

repeat (RIR) were identified. Cdi3 was located 64 bp upstream of CD1616/CDR20291_1514/CD196_1539 (**Fig. 6**). Cdi3 was inverted in both R20291 and CD196 compared to 630. The presence of these inversions indicates the possibility that phase variation is an important mode of genetic regulation. The absence of similarity between the LIR/RIR of the two inversions suggests that at least two invertases are responsible for these inversions and potentially a different mechanism for Cdi2. In addition to CD1167, the recombinase responsible for inverting *cwpV*,[11] and a number of unconserved transposon- or phage-related recombinases, there are at least three other tyrosine recombinases conserved in the three genomes (CD1222, CD1333 and CD1932). Tyrosine recombinases have previously been shown to be associated with phase-variable inversions in *Bacteroides fragilis*.[12]

Discussion

Genetic changes, such as inversions and point mutations, are key mechanisms for genetic variability in bacteria. Often these mechanisms for genetic variation are under estimated, due to difficulty in detection, compared to horizontal gene transfer and gain-of-trait functions. We show that inversions are located upstream of genes encoding for putative signaling and cell surface proteins. The Cdi1 inversion is located downstream of the promoter (P_{cwpV}), suggesting that phase variation in this instance is based on intrinsic terminator formation, resulting in switching off transcription.[11] For Cdi3 the position of the promoter for the upstream CDS is unknown, therefore this inversion may affect transcription of the downstream genes in a number of different ways: it could function in the classical way to flip a promoter contained within the LIR/RIR, such as in *E. coli* (reviewed in ref. 13) or may function in a similar way to *cwpV* by forming a transcriptional terminator.[11]

Closer analysis of the genomes has revealed putative disruptions of many genes through small mutations that have resulted in either frame shifts or premature stop codons. Opal stop codons may alternatively encode for the insertion of selenocysteine and therefore may result in fully functional proteins; however this also requires a selenocysteine incorporation sequence (SECIS) element in close proximity and our analysis did not identify any SECIS signatures. In *E. coli*, selenocysteine requires constitutively expressed *selABCD*, with *selC* encoding a unique tRNA species.[14,15] DNA BLAST using *selC* did not identify any homologues in the three *C. difficile* strains. However, *selABD* are present (e.g., CD2495, CD2493 & CD2496 respectively in 630), and all opal stop codons are conserved in all three strains, suggesting that these are indeed functional proteins with the insertion of selenocysteine. This indicates that the *selC* homologue and SECIS signatures in *C. difficile* are too divergent from the *E. coli* sequences to be recognized by DNA comparisons. The majority of interruptions are due to frame shifts (23/39) with 11 occurring in homopolymeric adenosine tracts. Although both 027 sequences were derived

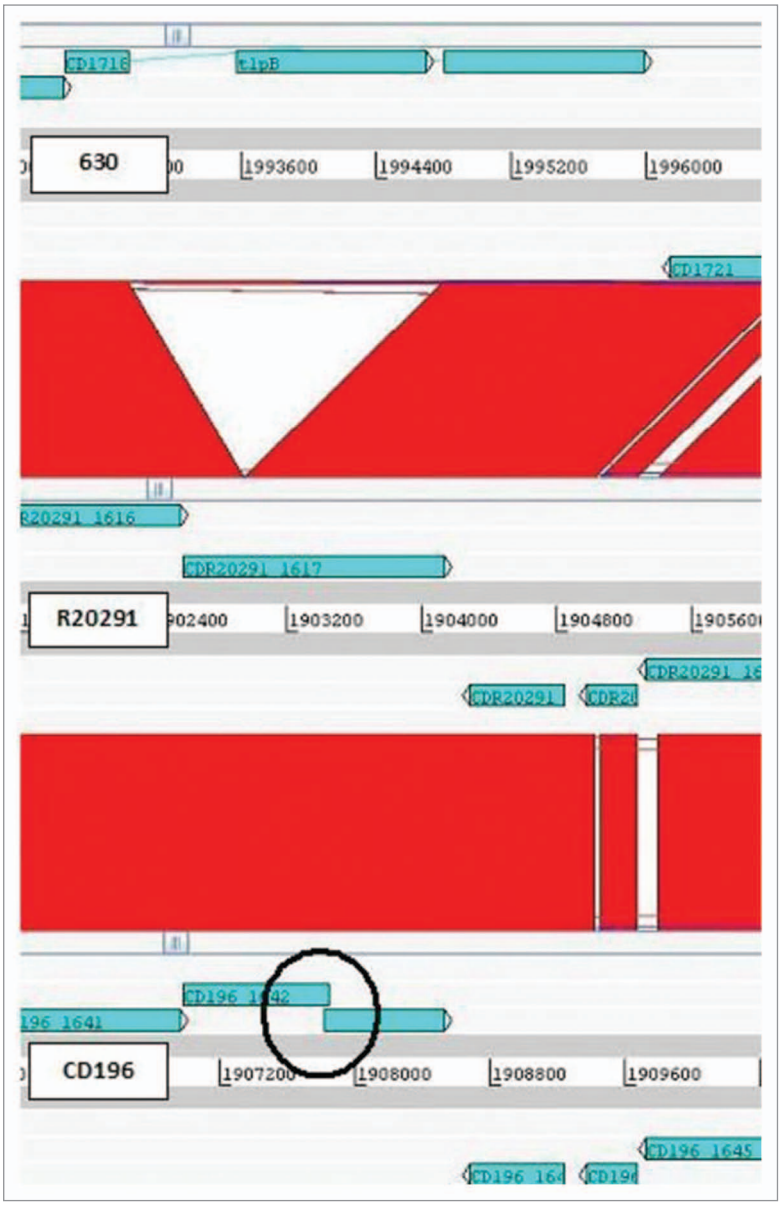

Figure 3. *C. difficile* hypervirulent 027 retains gene function. *C. difficile* R20291 CDS CDR20291_1617 encodes a functional putative hydantoinase but the orthologue in CD196 (CD196_1642) contains a frame shift due to an additional adenosine and has been interrupted in 630 (CD1718) by a transposase-like protein B (*tlpB*). Red bars indicates ≥99% homology between DNA sequences.

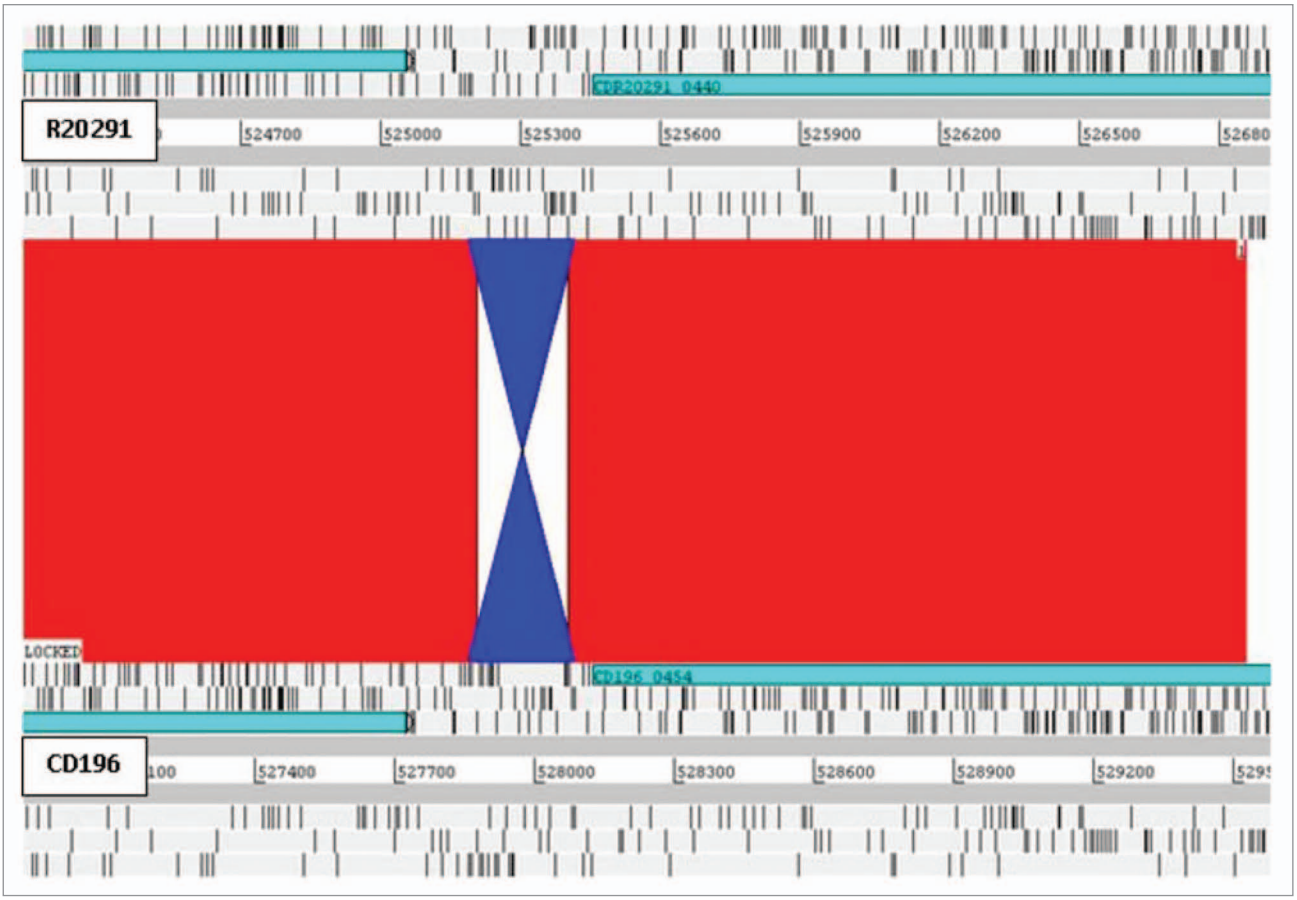

Figure 4. *C. difficile* inversion 1. Inversion (blue) in CD196 of 231 bp, 40 bp upstream of cell surface protein (*cwpV*).[11] Red bars indicate ≥99% homology between R20291 and CD196 DNA sequence. 630 Cdi1 was in the same orientation as R20291.

Table 2. Putative inversion sites

Name	Gene ID 630	Gene ID R20291	Gene ID CD196	LIR/RIR	Spacer	Gene function	Inversion
Cdi1	CD0514	CDR20291_0440	CD196_0454	5'-TTTTAATTCTAAAGGcTACTT 5'-AAGTAtCCTTTAGAATTA-GAA	195 bp	Cell surface protein (*cwpV*)	Inverted in CD196
Cdi2	CD0757	CDR20291_0685	CD196_0704	none	178 bp	Putative signaling protein	Inverted in CD196
Cdi3	CD1616	CDR20291_1514	CD196_1539	5'-CATTTCTTGTAAAATGGA-TAGTTT 5'-AAACTATCCATTTACAA-GAAATG	215 bp	Putative signaling protein	Inverted in R20291 & CD196

Analysis of three *C. difficile* genomes identified three conserved intergenic inversions. Two of the three inversion sites are flanked by inverted repeats, Cdi1 has LIR (left inverted repeat) and RIR (right inverted repeat) as described by Emerson et al. found upstream of *cwpV*. Cdi3 contains a novel set of inverted repeats, but Cdi2 has no identifiable repeats. Inversion = orientation relative to *C. difficile* 630 genome sequence.

using 454 sequencing technology, which can introduce errors in homopolymeric tracts this sequence data was confirmed using Solexa sequencing technology, which usually negates the problems associated with 454 sequencing technology.[16] Furthermore, seven of the frame shifts identified in homopolymeric tracts occur in 630, that was sequenced using the dye terminator method, which does not have this problem with polymeric tracts. Given the dual approach to sequencing these genomes, this suggests that these variations maybe genuine and that *C. difficile* may undergo phase variation by slipped-strand mispairing. However, experimental validation is required.

The re-analysis of the three genome sequences suggests that *C. difficile* has altered its gene content and functionality

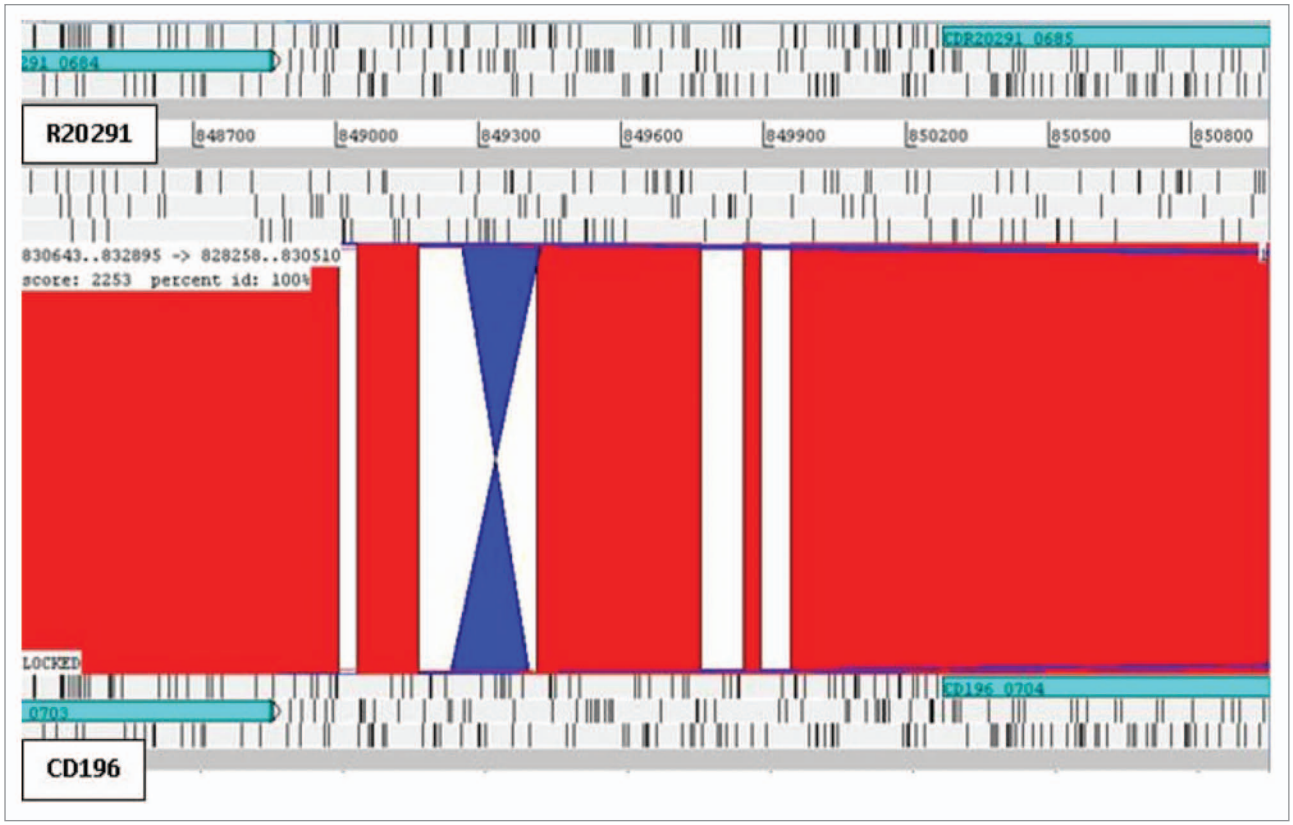

Figure 5. *C. difficile* inversion 2. Inversion (blue) in CD196 of 179 bp, 872 bp upstream of a putative signaling protein. Red bars indicate 100% homology between R20291 and CD196 DNA sequence. 630 Cdi2 was in the same orientation as R20291.

to significantly affect adaptation and emergence of hypervirulent strains.

Acknowledgements

We acknowledge the Wellcome Trust for funding this research.

Note

Supplementary materials can be found at: www.landesbioscience.com/supplement/StablerGUT1-4-Sup.pdf

References

1. Bartlett JG. *Clostridium difficile*: history of its role as an enteric pathogen and the current state of knowledge about the organism. Clin Infect Dis 1994; 18:265-72.
2. O'Connor JR, Johnson S, Gerding DN. *Clostridium difficile* infection caused by the epidemic BI/NAP1/027 strain. Gastroenterology 2009; 136:1913-24.
3. Goorhuis A, Van der Kooi T, Vaessen N, Dekker FW, Van den Berg R, Harmanus C, et al. Spread and epidemiology of *Clostridium difficile* polymerase chain reaction ribotype 027/toxinotype III in The Netherlands. Clin Infect Dis 2007; 45:695-703.
4. Hubert B, Loo VG, Bourgault AM, Poirier L, Dascal A, Fortin E, et al. A portrait of the geographic dissemination of the *Clostridium difficile* North American pulsed-field type 1 strain and the epidemiology of *C. difficile*-associated disease in Quebec. Clin Infect Dis 2007; 44:238-44.
5. Loo VG, Poirier L, Miller MA, Oughton M, Libman MD, Michaud S, et al. A predominantly clonal multi-institutional outbreak of *Clostridium difficile*-associated diarrhea with high morbidity and mortality. N Engl J Med 2005; 353:2442-9.
6. Mooney H. Annual incidence of MRSA falls in England, but *C. difficile* continues to rise. Bmj 2007; 335:958.
7. Redelings MD, Sorvillo F, Mascola L. Increase in *Clostridium difficile*-related mortality rates, United States, 1999–2004. Emerg Infect Dis 2007; 13:1417-9.
8. Sebaihia M, Wren BW, Mullany P, Fairweather NF, Minton N, Stabler R, et al. The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. Nat Genet 2006; 38:779-86.
9. Pallen MJ, Wren BW. Bacterial pathogenomics. Nature 2007; 449:835-42.
10. Drudy D, Kyne L, O'Mahony R, Fanning S. gyrA mutations in fluoroquinolone-resistant *Clostridium difficile* PCR-027. Emerg Infect Dis 2007; 13:504-5.
11. Emerson JE, Reynolds CB, Fagan RP, Shaw HA, Goulding D, Fairweather NF. A novel genetic switch controls phase variable expression of CwpV, a *Clostridium difficile* cell wall protein. Mol Microbiol 2009; 74:541-56.
12. Weinacht KG, Roche H, Krinos CM, Coyne MJ, Parkhill J, Comstock LE. Tyrosine site-specific recombinases mediate DNA inversions affecting the expression of outer surface proteins of *Bacteroides fragilis*. Mol Microbiol 2004; 53:1319-30.
13. Henderson IR, Owen P, Nataro JP. Molecular switches—the ON and OFF of bacterial phase variation. Mol Microbiol 1999; 33:919-32.
14. Sandman KE, Tardiff DF, Neely LA, Noren CJ. Revised *Escherichia coli* selenocysteine insertion requirements determined by in vivo screening of combinatorial libraries of SECIS variants. Nucleic Acids Res 2003; 31:2234-41.
15. Sandman KE, Noren CJ. The efficiency of *Escherichia coli* selenocysteine insertion is influenced by the immediate downstream nucleotide. Nucleic Acids Res 2000; 28:755-61.
16. Aury JM, Cruaud C, Barbe V, Rogier O, Mangenot S, Samson G, et al. High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. BMC Genomics 2008; 9:603.

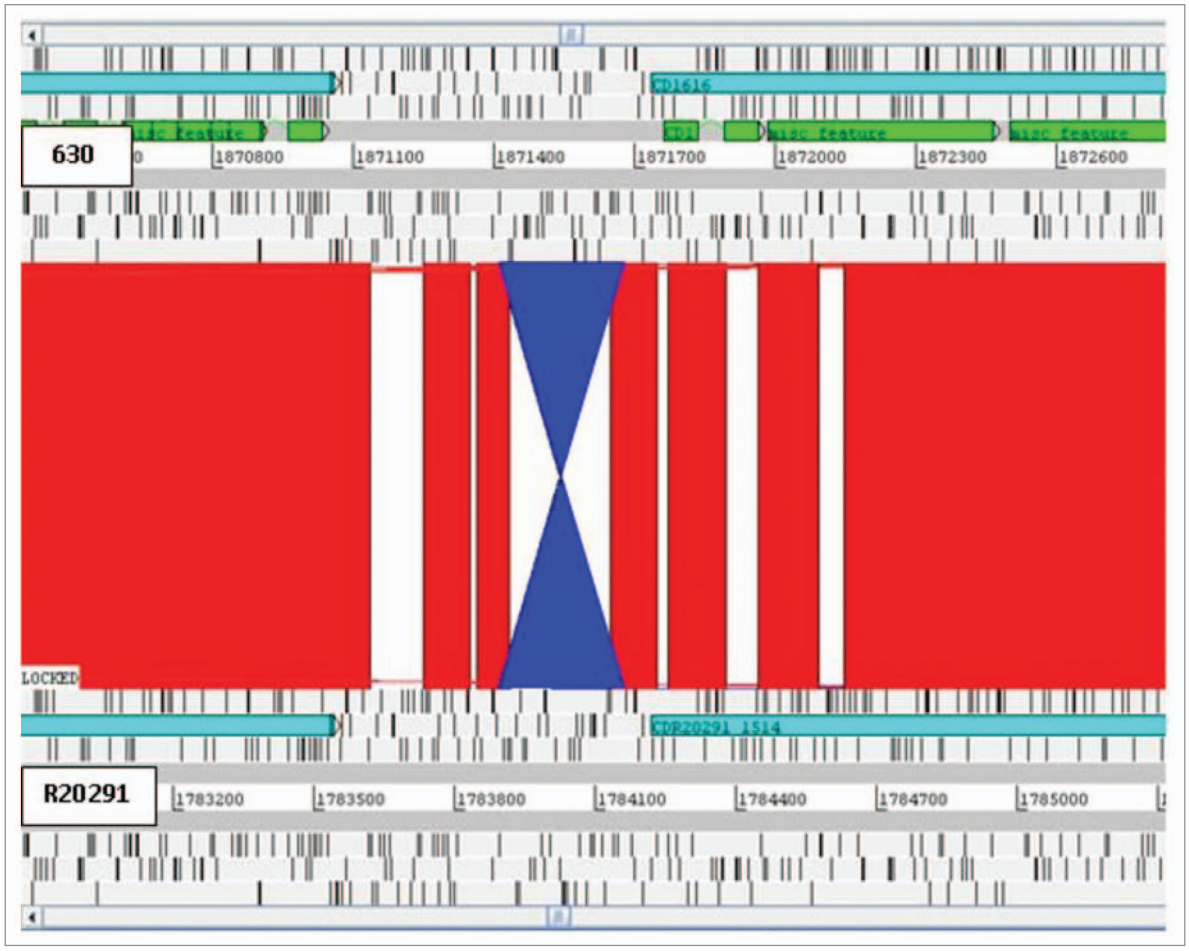

Figure 6. *C. difficile* inversion 3. Inversion (blue) in R20291 of 263 bp, 64 bp upstream of putative signaling protein. Red bars indicate ≥99% homology between R20291 and 630 DNA sequence. CD196 Cdi1 was in the same orientation as R20291.