Chapter 2

Genomic variation of *C. difficile* over short and long time scales

2.1 Introduction

Although *C. difficile* was first discovered in 1935 (Hall and O'Toole, 1935), the species was recognized as having pathogenic potential only three decades ago (Bartlett *et al.*, 1978) Since then, a number of emergent PCR ribotypes have been responsible for outbreaks worldwide (Bartlett, 2006), with different PCR ribotypes dominating both temporally and geographically (Brazier *et al.*, 2008; Cheknis *et al.*, 2009).

Major outbreaks occurred in Canada, the USA and the UK around 2003 (Loo *et al.*, 2005; McDonald *et al.*, 2005), caused by a previously rare PCR ribotype 027. The earliest recorded PCR-ribotype 027 isolate was CD196 in 1985, which was from a sporadic case involving a single patient with CDI in a Parisian hospital (Popoff *et al.*, 1988). Subsequently, several studies have shown that patients infected with PCR-ribotype 027 strains have more severe diarrhoea, higher mortality and more recurrences than similar patients infected with other ribotypes (Hubert *et al.*, 2007; Loo *et al.*, 2005; Redelings *et al.*, 2007). The 027 ribotype has spread globally and currently accounts for ~50% of isolates in UK and North American hospitals (Brazier *et al.*, 2008; Goorhuis *et al.*, 2008; Joseph *et al.*, 2005). The CDI outbreak at the Stoke Mandeville hospital, Buckinghamshire, which marked the arrival of the epidemic 027 isolates to the UK, resulted in a total of 334 CDI cases and 38 deaths between 2004 and 2005 (O'Connor *et al.*, 2009).

Other ribotypes, including 001, 017, and 078 (Brazier *et al.*, 2008; Cheknis *et al.*, 2009; Drudy *et al.*, 2007b; Goorhuis *et al.*, 2008; Huang *et al.*, 2009; Kim *et al.*, 2008) have emerged recently, suggesting an evolutionary trend associated with *C. difficile* in terms of adaptation to the modern healthcare environment. It is unclear what genetic characteristics differentiate ribotype 027 from other ribotypes (aside from sequence differences in the ribosomal RNA locus) that apparently make ribotype 027 more virulent. Also unclear are the additional genetic changes which underlie the emergence of modern day ribotype 027. A comparative genomic hybridization (CGH) approach (Stabler *et al.*, 2006) was used to study isolates representative of the *C. difficile* population and this analysis revealed the phylogeny consists of four major clades, including B1/NAP1/027.

MLST has also been used to study the population structure of *C. difficile* (Lemee *et al.*, 2004) and this analysis indicated that isolates recovered from severe infection cases do not cluster into distinct lineages, and thus no particular lineage is associated with increased virulence. The study also proposed that the population structure of *C. difficile* is clonal, although recombination events do occur (Dingle *et al.*, 2011; Lemee *et al.*, 2004). It was estimated that a single allele in *C. difficile* is 8- to 10-fold more likely to be altered by point mutation than by recombination (Lemee *et al.*, 2004). A more recent MLST analysis suggested that the effect of homologous recombination is four times lower than mutation on genetic diversification (Dingle *et al.*, 2011). However, MLST analysis is based on a limited number of loci in the genome and it was not clear whether the same findings would be supported if the analysis was based on whole genome sequences.

This chapter presents an analysis of whole genome sequences for a collection of eight *C. difficile* human and animal isolates generated using combined 454 (Roche) and Sanger sequencing. These sequences were used to explore macroevolution within the *C. difficile* genome. In addition, 21 isolates representing the hypervirulent clade identified by Stabler *et al.* (Stabler *et al.*, 2006) were sequenced with Illumina (Solexa) to investigate

microevolution within this group. A three-way genome comparison was also undertaken that included a 'historic' non-epidemic 027 *C. difficile* (CD196), a recent epidemic and hypervirulent 027 (R20291) and the previously published PCR-ribotype 012 strain (630). The aims of this study were: -

- to infer phylogenetic relationships between different ribotypes and within ribotype 027
- to assess the genetic diversity of *C. difficile*,
- to identify key mechanisms of *C. difficile* genome changes, and the relative impact of these mechanisms,
- to identify genetic difference between 027 and other ribotypes, between historic and more recent 027 isolates, and to assess aspects of the functional impact of these differences

2.2 Materials and methods

2.2.1 Bacterial isolates

Isolates were provided by the following individuals: Dale Gerding, Hines VA Hospital, IL (CF5 and BI isolates); Jon Brazier, Anaerobe Reference Laboratory, Cardiff, UK (R20291); Michel Popoff, Institut Pasteur, Paris, France (CD196); Denise Drudy, Centre for Food Safety, University College Dublin, Ireland (M68 and M120); Peter Mullany, Eastman Dental Institute, London, UK (630); and Glenn Songer, Department of Veterinary Science and Microbiology, University of Arizona (all other isolates). *C. difficile* 630 (Wust *et al.*, 1982) was isolated from a patient with PMC in Zurich, 1982 and has been fully sequenced by the Wellcome Trust Sanger Institute (WTSI) (Sebaihia *et al.*, 2006). 027 CD196 is a non-epidemic strain isolated from a patient with PMC in Paris, 1985. The hypervirulent 027 R20291 was isolated during a recent outbreak in Stoke Mandeville, UK. Selected details of the isolates are provided in Tables 2.1 and 2.2.

2.2.2 DNA sequencing and assembly

Genomic DNA was prepared according to Wren *et al.* (Wren and Tabaqchali, 1987) by Dr. Trevor Lawley at WTSI. Isolates were sequenced using 454 Life Sciences GS-20 sequencer (Roche) (R20291), 454 Life Sciences GS-FLX sequencer (Roche) (all other isolates in Table 2.1), and Illumina (Solexa) Genome Analyzer System (isolates in Table 2.2) with a multiplexed protocol according to the manufacturer's specifications. Shotgun capillary reads were also generated for R20291 and CD196 with ABI 3730xl analyzers. Paired-end reads were generated for all isolates except CF5, M68, M120, BI-1, 2007855, R20291, and CD196, for which single-end reads were produced. 454 reads were assembled *de novo* into contigs using newbler (Roche). For isolates with capillary data available, 454 contigs were shredded into reads of comparable length to capillary reads, and assemblies were created using data from both platforms using Phrap (Gordon *et al.*, 2001).

To further correct homopolymer tract errors inherent in early 454 sequencing data, Solexa (Illumina) sequence data were generated for isolate R20291. The Illumina sequences were assembled *de novo* using Velvet (Zerbino and Birney, 2008) and the resulting contigs were incorporated with the combined 454 and capillary assembly. Closing gaps between contigs for both CD196 and R20291 was either by primer walking on subclones from the capillary shotgun or by sequencing PCR products covering gaps between adjacent contigs. The final contiguous sequence for CD196 was mostly from combined data but small regions were covered with only 454 data (a total of less than 2.6% of the sequence) or with only capillary reads, giving a consensus confidence of < 41 (< 0.3% of the sequence). All regions of the final finished R20291 assembly are covered by high quality capillary reads or by combinations of data from at least two sequencing technologies, although three gaps remain where ribosomal rRNA operons have not been bridged by read-pairs. Sequencing and assembly described above were carried out by WTSI Sequencing and Finishing teams. The order of contigs was estimated by comparison with 630 genomic sequence using the MUMmer package

(nucmer program) (Kurtz *et al.*, 2004), implemented in ABACAS (Assefa *et al.*, 2009). Although some manual error checking was performed, these should still be considered to be draft genomes. Reads from Illumina (Solexa) were directly mapped back to a reference sequence (CD196) using MAQ version 0.7.1 (Li *et al.*, 2008).

2.2.3 Genome annotation, comparison, identification of orthologues and unique genomic regions

Genome annotation of *C. difficile* strains was based on previously published annotations of *C. difficile* 630 (Sebaihia *et al.*, 2006). The genomic sequences of CD196 and R20291 were compared against the database of 630 proteins by blastx, and a CDS feature in the query genome was created when a hit of over 90% identity was found. Glimmer3 (Delcher *et al.*, 1999) was used to predict CDSs in genomic regions where no significant hits were found. Any unique genomic regions left were examined and annotated manually in Artemis (Rutherford *et al.*, 2000). The genome comparisons were visualized in Artemis and ACT (Artemis Comparison Tool) (Carver *et al.*, 2005).

The reciprocal-best-hit fasta search algorithm was used to identify orthologues among 630, CD196 and R20291. All CDSs in the query genome were searched in the database of subject CDSs by FASTA (Pearson and Lipman, 1988). When a hit of over 30% identity and over 80% length was found, the hit CDS in the subject genome was searched again in the database of query CDSs in a similar fashion. If the top hit in the second search was the same as the original query CDS, the two CDSs were considered as orthologues by this method. These identified orthologues were manually curated to take into account inaccuracies caused by inserted elements, frameshifts and pseudogenes.

2.2.4 Phylogenetic analyses

The genomic sequences of nine *C. difficile* isolates (630, BI-9, CF5, M68, M120, CD196, BI-1, 2007855, and R20291) were aligned with ProgressiveMauve (Darling *et al.*, 2010). Consensus alignments were used to build a maximum likelihood tree with RAxML (Stamatakis, 2006) with 100 re-samplings of alignment data. To root the phylogenetic tree, the genomic sequences of *Clostridium bartlettii* strain DSM 16795 and *Clostridium hiranonis* DSM 13275 were used. Phylogenetic relationships between hypervirulent isolates were inferred using PHYML (Guindon and Gascuel, 2003) and the split decomposition method implemented in SplitsTree4 (Huson and Bryant, 2006) with 100 bootstraps.

2.2.5 SNPs detection and CDS alignments

SNP calling between isolates of different ribotypes was performed using the nucmer program in the MUMmer package (Kurtz *et al.*, 2004). Default settings were used. SNP calls within 20 bp of a contig end were removed. SNPs called within 2 bp of another SNP were excluded, as they may be due to mis-alignment or mis-assembly of sequences. All primary SNPs were further checked by FASTA (Pearson and Lipman, 1988) in all isolates to validate alleles. Orthologous CDSs were retrieved from the whole genome alignment of nine *C. difficile* isolates generated using progressiveMauve (Darling *et al.*, 2010), with the guide of the fully annotated reference strain 630. The genome was inspected manually to exclude CDSs in bacteriophages, transposons and other mobile elements, as these regions are generally repetitive and can confound SNP finding programs. A multiple sequence alignment for each CDS was obtained by aligning SNP alleles against the 630 sequence.

The program MAQ (Li *et al.*, 2008) was used for the analysis between hypervirulent isolates to align Solexa reads to the reference genome CD196 and to call SNPs. A minimum mapping quality of 30 was specified, therefore disallowing ambiguous mapping of most repetitive regions. The SNPfilter algorithm implemented in MAQ was also used and SNPs covered by too few (<3) or too many (>250) reads were removed. The identities of hypervirulent

isolates were validated by PCR. Preliminary SNPs were confirmed for each isolate in all sequencing reads. Only SNP alleles supported by all reads were included in downstream analysis. As a final filtering process repetitive regions in the genome were identified and SNPs called within these regions were excluded. Repetitive regions were identified by: (i) manually marking up prophages, transposons, and other mobile elements in Artemis (Rutherford *et al.*, 2000); and (ii) using repeat-finding programs REPuter (Kurtz *et al.*, 2001), nucmer, and repeat-match in the MUMmer package (Kurtz *et al.*, 2004). A multiple sequence alignment formed by the concatenated variable positions from all isolates was then used for phylogenetic analysis.

2.2.6 Recombination and selection analysis

The program CLONALFRAME (Didelot and Falush, 2007) was used to infer recombination events and calculate r/m ratio within the deep-branching phylogeny. Pairwise dN/dS for concatenated orthologous CDSs was calculated using the method of Nei and Gojobori (Nei and Gojobori, 1986). Site models M1a and M2a implemented in PAML (Yang, 2007) were used to identify genes under positive selection, and individual gene trees built with RAxML (Stamatakis, 2006) were used to correct for homologous recombination. M1a assumes neutral evolution, and M2a allows positive selection. Likelihoods from the two models were compared by a likelihood ratio test. Bayes empirical Bayes (BEB) analysis was used to identify sites under positive selection if the likelihood ratio test is significant.

2.2.7 Estimates of age and population size

Two methods were used to calculate the age of *C. difficile*. The first method follows the formula:

$$Age = \frac{d_s}{rate \times 2}$$

where d_s is the mean synonymous substitutions per site calculated from concatenated non-recombining core CDSs after Jukes-Cantor correction (Jukes and Cantor, 1969). The rate represents a synonymous molecular clock rate of $2.5 \times 10^{-9} – 1.5 \times 10^{-8}$ per site per year, which is equivalent to a universal mutation rate of 0.0001 - 0.0002 per genome per generation proposed by Ochman *et al.* (Ochman *et al.*, 1999). Here 100 - 300 generations per year are assumed (Gibbons and Kapsimalis, 1967).

As a second approach, the program BEAST (Drummond and Rambaut, 2007) was used to estimate the age of the whole *C. difficile* collection. To obtain an independent estimate of the molecular clock rate, orthologues between *C. difficile* and *C. tetani* were identified and their sequence divergence was calculated following the model of Jukes-Cantor (Jukes and Cantor, 1969). The age of the *Clostridium* lineage was previously estimated to be 2.34 billion years (Sheridan *et al.*, 2003), and this was taken to be a maximum divergence time for these two species. This gave rise to a molecular clock rate of 1.15×10^{-10} per site per year. The population history of hypervirulent isolates was inferred using Bayesian skyline plot (Drummond *et al.*, 2005).

2.2.8 Identifying non-recombining core coding sequence

To obtain gene sets that had not undergone homologous recombination, a stringent measure was adopted: If a CDS contains any base position for which a posterior recombination probability of more than 0.2 was inferred by CLONALFRAME (Didelot and Falush, 2007) in any of the genomes, this CDS was excluded from the gene set. This method resulted in 622 non-recombining core CDSs.

2.3 Results

2.3.1 Macroevolution of the *C. difficile* species

2.3.1.1 The deep-branching phylogeny

Previous phylogenomic analysis identified four genetically distinct *C. difficile* clades (Stabler *et al.*, 2006); based on this phylogeny, eight strains were selected to cover the broad genetic diversity of *C. difficile* and DNA prepared from these was subjected to whole genome sequencing as described in Methods (Table 2.1).

Isolate	Year	Country	Source	Ribotype	Coverage		
					454	**Sanger**	**Solexa**
630	1982	Switzerland	Human	012	Published (Sebaihia *et al.*, 2006)		
M68	2006	Ireland	Human	017	16.7×	5.3×	-
CF5	1995	Belgium	Human	017	11.8×	5.6×	-
M120	2007	UK	Human	078	11.1×	5.1×	-
BI-9	2001	USA	Human	001	11.0×	14.8×	>200×
BI-1	1988	USA	Human	027	16.2×	5.1×	92×
R20291	2006	UK	Human	027	14.8×	5.7×	132.5×
CD196	1985	France	Human	027	13.6×	5.1×	-
2007855	2007	USA	Bovine	027	17.5×	5.4×	-

Table 2.1: Sequence coverage of the broad collection of *C. difficile* isolates.

Genome assemblies were created based on combined 454 (Roche) and Sanger sequencing to increase accuracy and coverage. A consensus whole-genome alignment was used to build a maximum likelihood phylogenetic tree that also included the published genome of *C. difficile* 630 (ribotype 012) (Figure 2.1A). To minimize the impact of recombined sequences on the tree building, a concatenated alignment of non-recombining core CDSs was also used to build a tree of six isolates representing the deep-branching phylogeny, resulting in the same topology (Figure 2.2). See section 2.2.8 for details of the methods for identifying non-recombining core coding sequences. The resulting phylogeny recapitulates the four major lineages based upon microarray analysis (Stabler *et al.*, 2006) but provides much more depth.

The phylogenetic tree reveals that the broad genetic diversity of *C. difficile* is predominantly reflected in the ribotyping scheme. For example, the four 027 ribotype sequences occupy a single lineage (unresolved in Figure 2.1A but separated in Figure 2.1B). Strains CF5 and M68, which are historic and recent representatives respectively of the 017 ribotype, occupy a distinct lineage. Isolate BI-9, verified as ribotype 001, is more closely related to isolate 630 (ribotype 012). Isolate M120 (ribotype 078) appears to be highly divergent as indicated by its long branch length. The average sequence divergence between M120 and the other isolates is 2.4%, indicating that *C. difficile* is an old species.

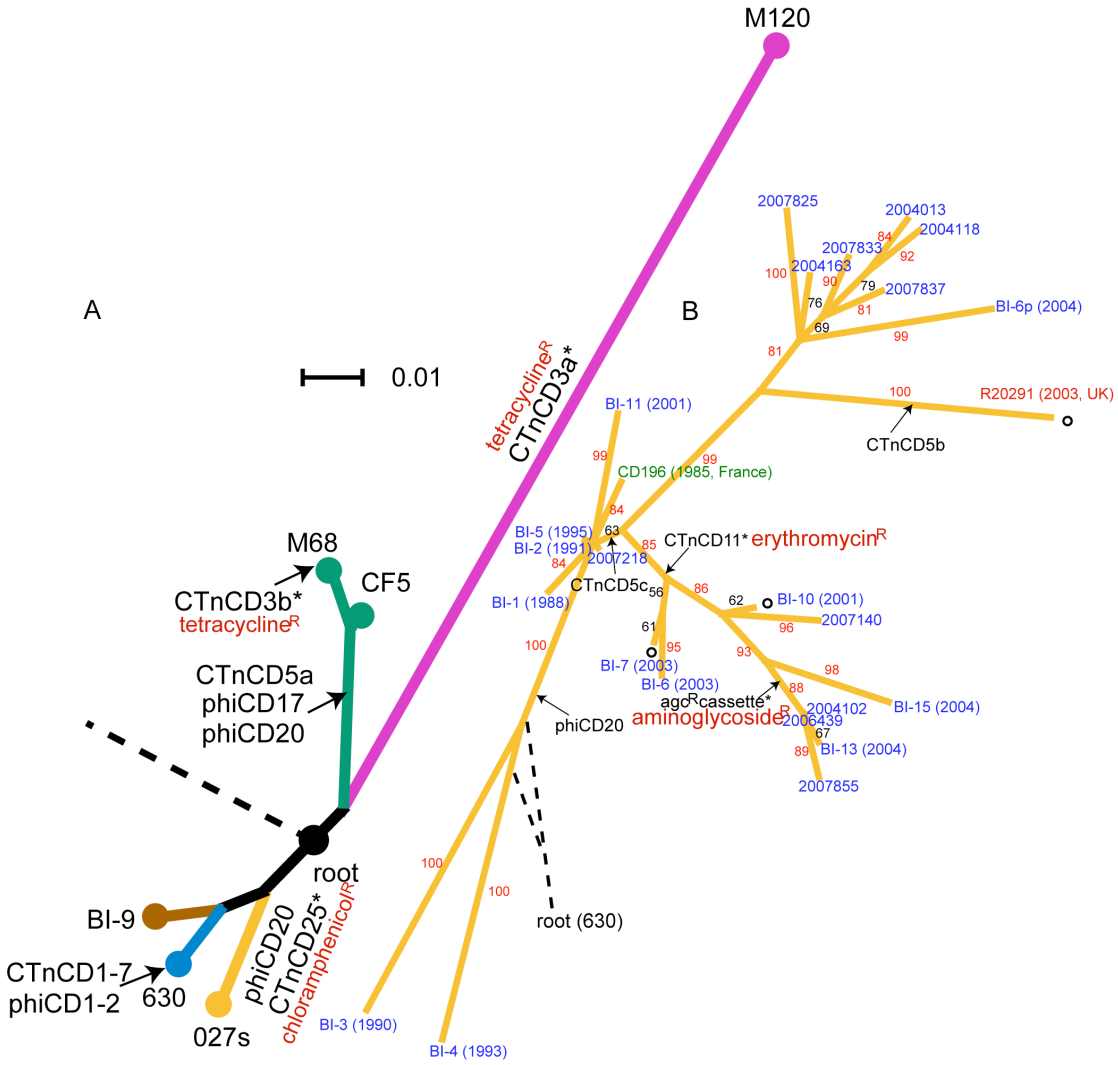

Figure 2.1: Phylogenetic trees of *C. difficile* based on whole-genome sequences. Arrows and unfilled circles denote insertion and deletion events, respectively. Genomic islands carrying drug resistance genes are shown with asterisks. (A) Deep-branching phylogeny that illustrates the relationships between different lineages/ribotypes (shown by different colours). The four 027 ribotype isolates are collectively represented as node “027s”. Scale bar indicates number of substitutions per site. The root connects to *C. bartlettii* and *C. hiranonis*. (B) Split decomposition network indicating microevolution within the hypervirulent lineage. Strain names are coloured according to countries of isolation (blue, USA; red, UK; green, France). Bootstrap values are labelled along branches. The root connects to strain 630.

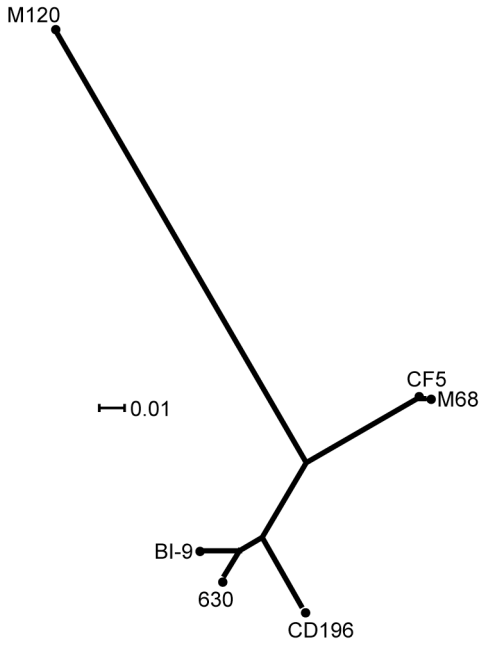

Figure 2.2: Phylogenetic tree of a diverse collection of *C. difficile* isolates based on concatenated non-recombining core CDSs. Scale bar indicates number of substitutions per site.

2.3.1.2 The age of the *C. difficile* species

The sequence data were used to estimate the age of the *C. difficile* species. Dating methods for microbes are imperfect and the subject of controversy, so to find the range of possibilities, two independent methods based on different underlying assumptions were used. One is based on an average synonymous substitution per site (dS) of 0.032 in concatenated non-recombining core CDSs and a synonymous substitution rate of $2.5 \times 10^{-9} – 1.5 \times 10^{-8}$ per site per year (See 2.2.7 for details). This indicates an age of 1.1 - 6.4 million years before present. As an alternative, the software BEAST (Drummond and Rambaut, 2007) was used; a calibrated molecular clock rate of 1.15×10^{-10} was specified in the analysis (see 2.2.7 for details). This was calculated based on a sequence divergence of 0.54 between orthologous CDSs of *C. difficile* and *C. tetani*, and the hypothesis that the *Clostridium* lineage diverged 2.34 billion years ago (Ga) (Sheridan *et al.*, 2003). Therefore, although *C. difficile* and *C. tetani* diverged relatively early in the *Clostridium* lineage, the

divergence time between them should not exceed 2.34 Ga. This analysis resulted in a divergence time of 85 million years. Clearly these methods produce highly divergent estimates, indicating high levels of uncertainly, but which could be viewed as maximum and minimum boundaries.

2.3.2 Microevolution within the hypervirulent clade.

To study microevolution within the hypervirulent clade and recent ribotype 027 isolates, a collection of 25 isolates spanning 1985 - 2007 (Tables 2.1 and 2.2) were sequenced using multiplexed Illumina (Solexa) or a combination of 454 (Roche) and Sanger sequencing technologies.

Isolate	Year	Country	Source	PCR Ribotype	Solexa Coverage
BI-2	1991	USA	Human	027	29×
BI-3	1990	USA	Human	027	59×
BI-4	1993	USA	Human	027	104×
BI-5	1995	USA	Human	027	74×
BI-6	2003	USA	Human	-	38×
BI-6p	2004	USA	Human	027	60×
BI-7	2003	USA	Human	027	49×
BI-10	2001	USA	Human	027	16×
BI-11	2001	USA	Human	-	51×
BI-13	2004	USA	Human	027	62×
BI-15	2004	USA	Human	027	90×
2004013	2004	USA	Human	027	37×
2004163	2004	USA	Human	027	32×
2004102	2004	USA	Human	027	50×

2004118	2004	USA	Human	027	30×
2006439	2006	USA	Food	027	38×
2007140	2007	USA	Human	027	29×
2007837	2007	USA	Human	027	53×
2007833	2007	USA	Human	027	9×
2007825	2007	USA	Human	027	27×
2007218	2007	USA	Food	027	29×

Table 2.2: Details of the hypervirulent isolates included in this chapter.

SNPs were detected by comparing the sequence of each isolate with the early 027 ribotype isolate CD196. A total of 1847 SNP differences were discovered among 25 isolates; however, 1670 (90.4%) of these SNPs appear in tight clusters and are present only in isolate BI-4 or BI-11 (Figure 2.3), indicating that they could have resulted from recent recombination events. These SNPs were excluded from phylogenetic analyses as they could mask the true phylogenetic signal. A split-decomposition network based on the remaining SNPs is shown in Figure 2.1B. No conflict between placements of branches was identified by split decomposition analysis. A lack of bipartitions in certain parts of the lineage and low bootstrap values can possibly be explained by the scarcity of genetic variation between isolates, and some recombinant sites potentially remaining in the analysis. The placement of the root for this lineage cannot be uniquely determined. This also suggests recombination between isolates sitting at the basal branches of this phylogeny and those outside this group. Interestingly, a Bayesian skyline plot analysis (Drummond *et al.*, 2005) suggests this hypervirulent group has undergone a population expansion around the start of the century (Figure 2.4), which coincides with the time when hospital outbreaks caused by this *C. difficile* ribotype were first reported.

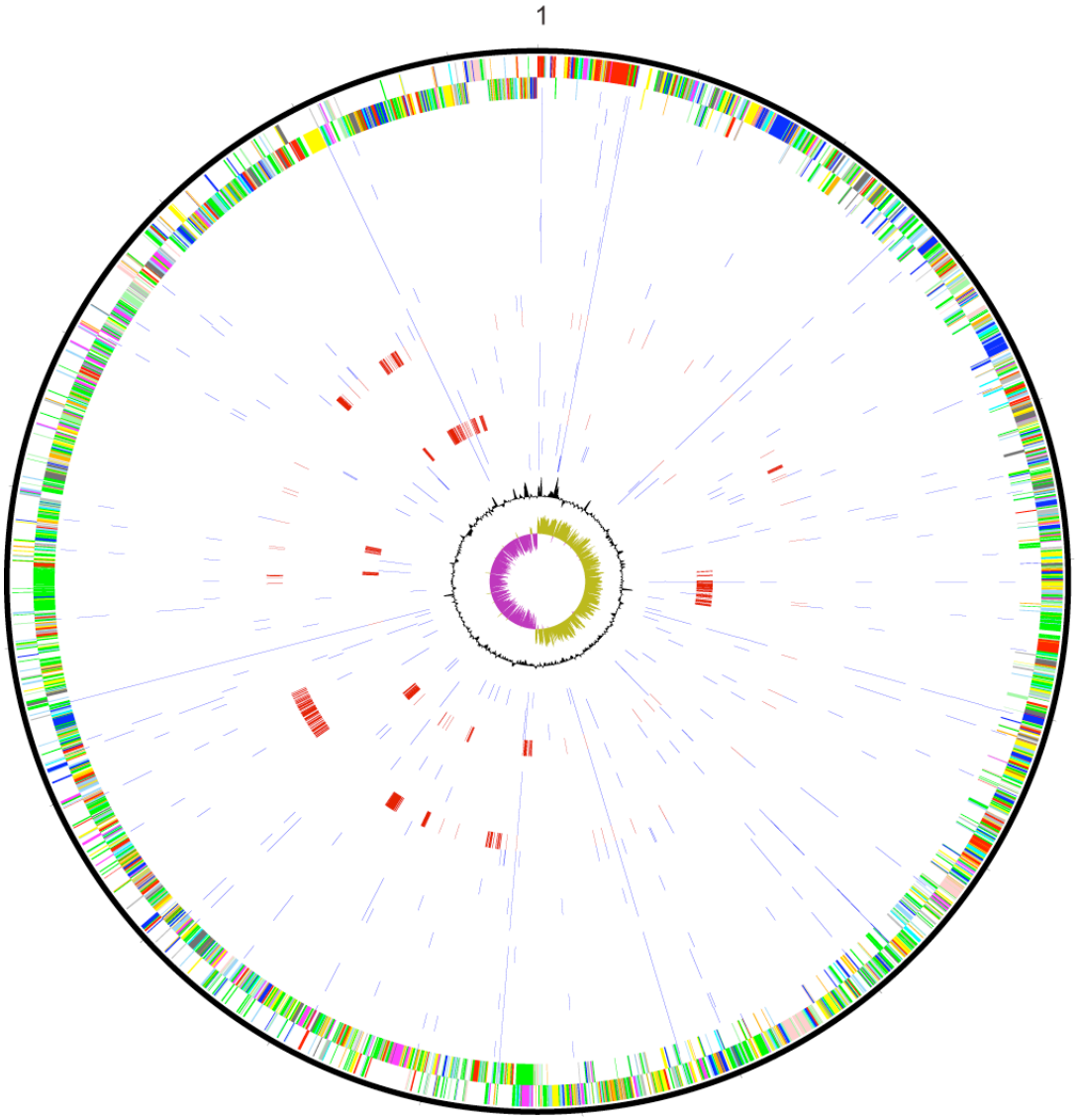

Figure 2.3: SNPs between CD196 and 24 other hypervirulent *C. difficile* isolates. Outer circle: CDSs of *C. difficile* CD196 genome, shown on a pair of concentric rings representing both coding strands; two inner circles: G+C% content plot and GC deviation plot (>0% olive, <0% purple); in between: SNPs (blue and red) between CD196 and other isolates, from outer to inner: 2004013, 2004102, 2004118, 2004163, 2006439, 2007140, 2007218, 2007825, 2007833, 2007837, 2007855, BI-1, BI-2, BI-3, BI-4, BI-5, BI-6, BI-6p, BI-7, BI-10, BI-11, BI-13, BI-15, and R20291. The rings representing isolates with large homologous recombination blocks (BI-4 and BI-11) are shown in red.

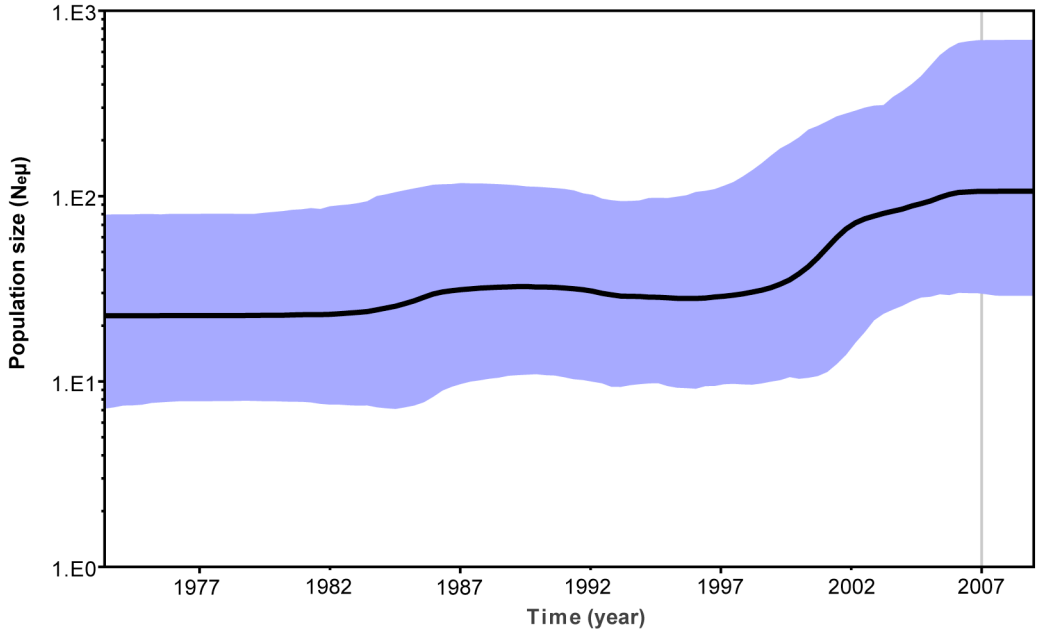

Figure 2.4: Bayesian skyline plot (group number = 10) shows a recent population expansion of the hypervirulent group. The x axis gives units of years, and the y axis is equal to $N_e\mu$ (product of effective population size and generation length in years). Thick solid line indicates median estimate, and purple areas indicate its 95% confidence interval.

2.3.3 Extensive role of horizontal gene transfer in *C. difficile* evolution

The establishment of a rooted phylogeny for *C. difficile* facilitated the tracing of various evolutionary signatures such as genomic insertions and deletions back to where they occurred in the phylogenetic tree.

Putative conjugative transposons and bacteriophages account for a large proportion of the mobile elements present within the *C. difficile* genomes. Many of these mobile elements code for a variety of antibiotic resistance genes (Figure 2.1), suggesting a significant role for horizontal gene transfer in resistance acquisition. In isolate 630, CTn*CD1*, CTn*CD3*, CTn*CD6*, and CTn*CD7* are closely related to Tn*916* in *Enterococcus faecalis* (Sebaihia *et al.*, 2006). Here the similarity was also found to extend to CTn*CD25* and

CTn*CD11*, but these carry different drug-resistance determinants (CTn*CD3*, tetracycline; CTn*CD25*, chloramphenicol; CTn*CD11*, erythromycin). CTn*CD25* is a conjugative transposon carried by all hypervirulent isolates in this collection, while CTn*CD11* was only found in 8 isolates, which occupy a sub-lineage within the tree.

In both the deep-branching phylogeny and the lineage of hypervirulent isolates, evidence was detected for the same genomic island entering different parts of the phylogenetic tree. CTn*CD3*, previously characterized as Tn*5397* (Wang *et al.*, 2000), is a conjugative transposon carrying *tetM* (a tetracycline resistance gene), which appeared to have entered 630, M68, and M120 independently, as indicated by completely different locations in each genome. A high level of similarity was also observed between structural genes in prophage 1 and 2 in isolate 630 and phi*CD20* found in 22 of the 25 hypervirulent isolates in this collection (Figure 2.5). Almost all hypervirulent isolates sampled after 2001 harbour the same conjugative transposon CTn*CD5c*, except R20291 and 2007218. However, a variant of this island was found to have inserted at a different location in R20291 (Figure 2.5).

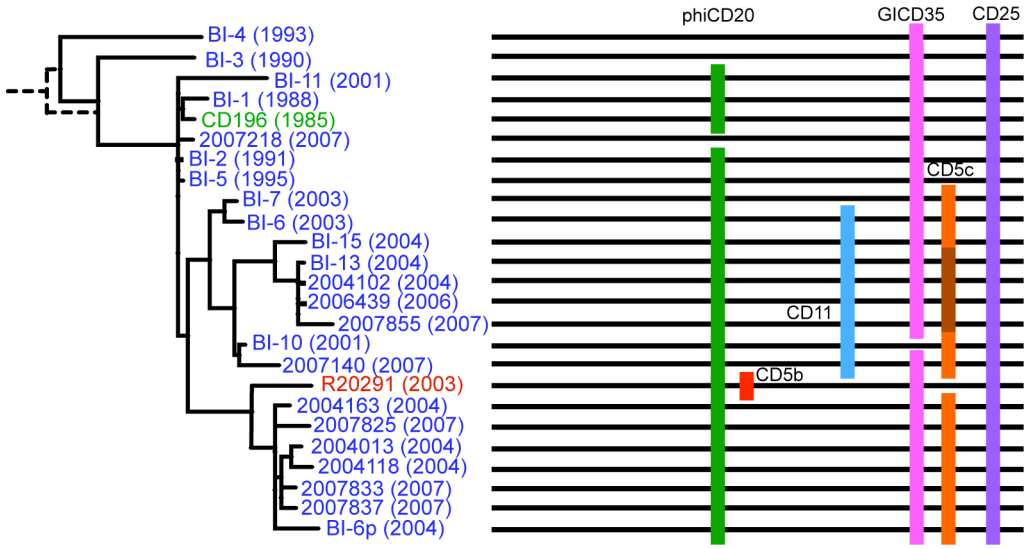

Figure 2.5: Genomic islands within isolates of the hypervirulent clade. Colour scheme of strain names is the same as in Figure 2.1 The presence of each genomic island is shown by a coloured box. Dark brown boxes denote copies of CTn*CD5c* with an aminoglycoside resistance cassette insertion.

There are also cases of new insertions occurring within existing genomic islands. For example, copies of the conjugative transposon CTn*CD5* found in 2004102, 2006439, 2007855 and BI-13 all contain an extra 7.5-kb cassette (Figure 2.1B and Figure 2.5). This region harbours CDSs encoding a DNA recombinase and aminoglycoside resistance genes *aph(2')-Ib* and *aac(6')-Im*. Combining the information from the phylogenetic tree, this suggests that the insertion event within CTn*CD5* occurred in the common ancestor of these isolates. Isolate M120, which is divergent from the other isolates, harbours a number of unique genomic regions, two of which exhibit ~80% sequence similarity to *Streptococcus pyogenes* (Figure 2.6) and a *Thermoanaerobacter* species (Figure 2.7), respectively, suggesting gene transfer across very large phylogenetic distances.

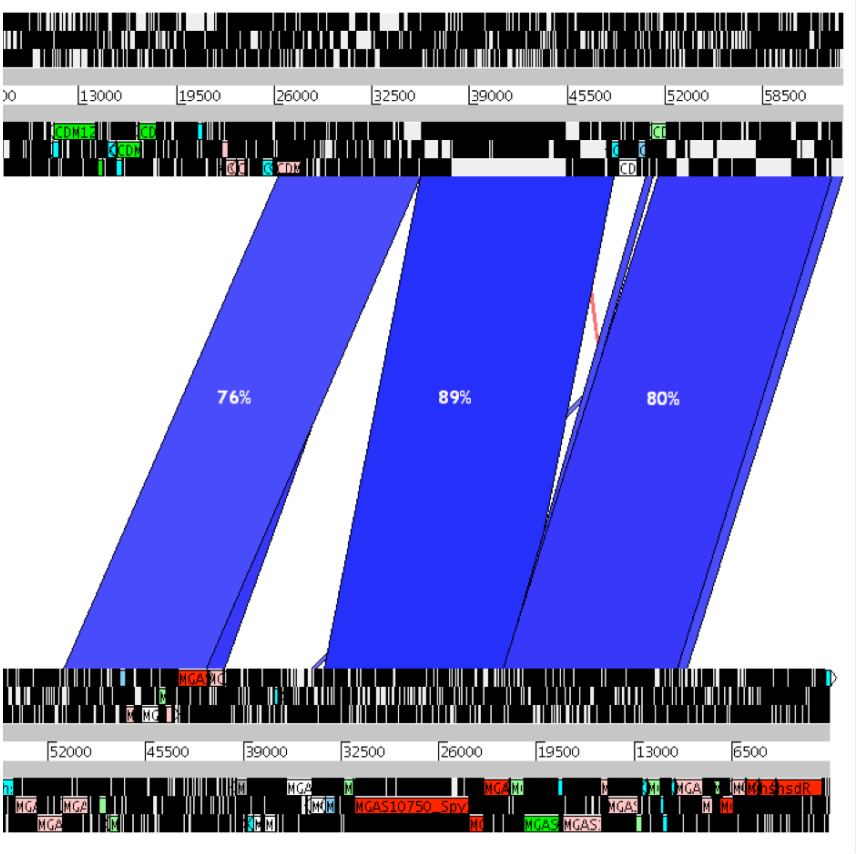

Figure 2.6: Comparison between parts of the *C. difficile* M120 genome (top) and the genome of *S. pyogenes* MGAS10750 (bottom). Each pair of black and white boxes represents both strands of a sequence. Coloured boxes present annotated CDSs; un-annotated parts of the sequence are left blank. Blue blocks indicate sequence similarity. Percent sequence identity is labelled onto each matching block in white.

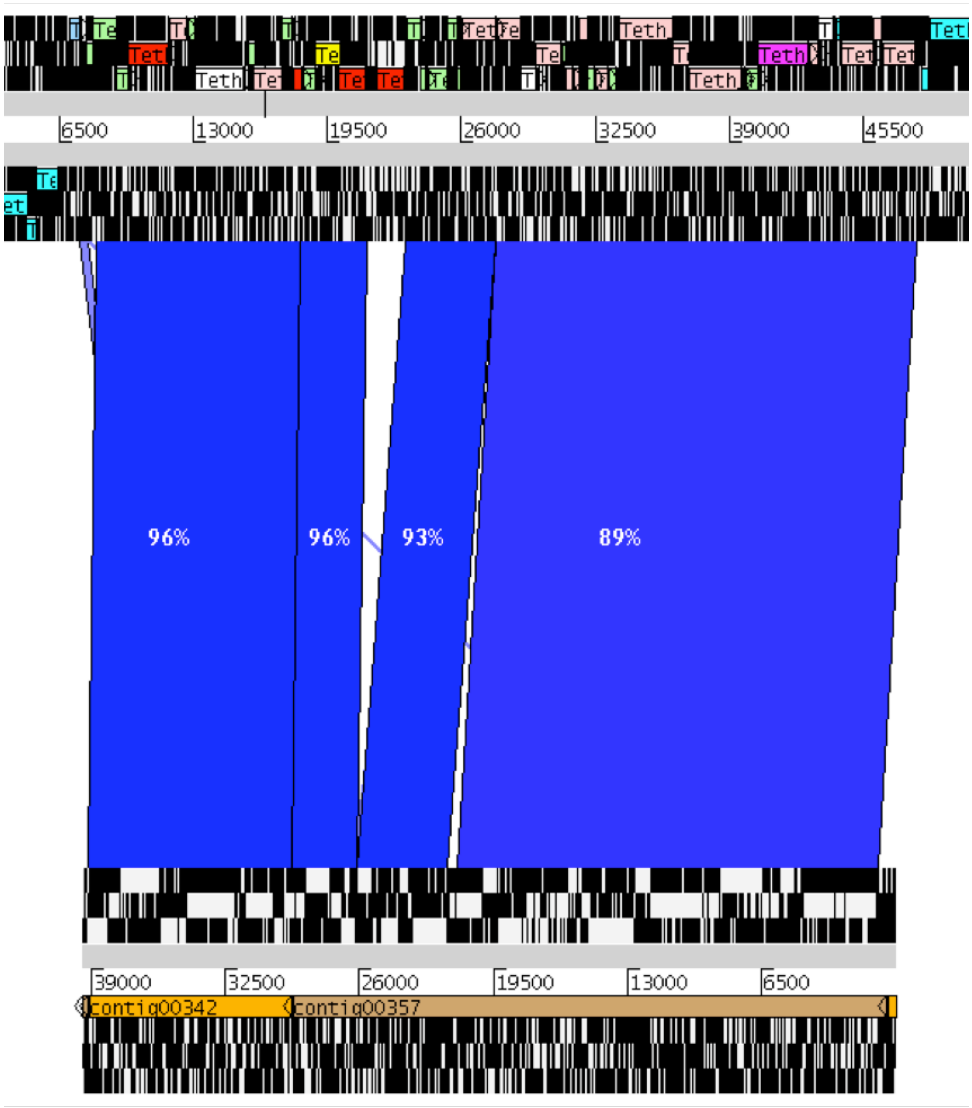

Figure 2.7: Comparison between parts of the *C. difficile* M120 genome (bottom) and the genome of *Thermoanaerobacter* sp.X514 (top). Other descriptions for this figure are the same as for Figure 2.6.

2.3.4 Large chromosomal regions exchanged by homologous recombination.

Besides horizontal transfer of mobile elements, bacteria can also evolve through exchange of common chromosomal segments by homologous recombination (Smith *et al.*, 1991). To test for signatures of recombination, the distribution of SNPs along the conserved *C. difficile* genome was examined.

Strikingly, this identified strong evidence for exchange of very large chromosomal regions both within the deep-branching phylogeny and the recent hypervirulent group.

The distribution of SNPs along the genomic backbone of the hypervirulent isolates demonstrates dense SNP clusters in BI-4 and BI-11, suggesting imports from different phylogenetic backgrounds into these isolates (Figure 2.3). The sizes of these regions range from 9 kb to 170 kb. In an attempt to identify potential donors for recombined regions within the hypervirulent group, these regions were compared to genomic sequences of known *C. difficile* isolates outside the hypervirulent group with BLAST to assess sequence similarity. No hit was found with a higher percentage identity than that between the strain in question and CD196, which implies the donors for these recombined regions are not closely related to sequenced *C. difficile* isolates in this collection.

To identify recombination between ribotypes, SNPs in non-repetitive regions of the genome were identified between all pair-wise combinations of the six isolates (CD196, 630, BI-9, M120, CF5, and M68). Figure 2.8 shows, as an example, the distribution of SNPs between isolate CF5 and each of the others. A very low level of diversity was observed between isolates CF5 and M68 across the entire chromosome except for several discrete regions, characterized by significantly increased SNP numbers with clear-cut boundaries, which suggests that these regions were acquired through homologous recombination events. The largest of these regions was around 300 kb. The complementary pattern of SNP peaks and valleys between M68 and BI-9, CD196 and 630 indicates a donor similar to BI-9, CD196, and 630 in these chromosomal regions, suggesting this recombination has occurred across a relatively large phylogenetic distance. The total size of imported sequence is ~640 kb (15% of the CF5 genome). Similar large blocks of homologous recombination were recently identified in *Streptococcus agalactiae*, where it was suggested that they may arise from Hfr-like conjugation driven by origins of transfer in mobile genomic islands (Brochet *et*

al., 2008). This is also a possible explanation in *C. difficile*, given the large numbers of mobile elements in the chromosome.

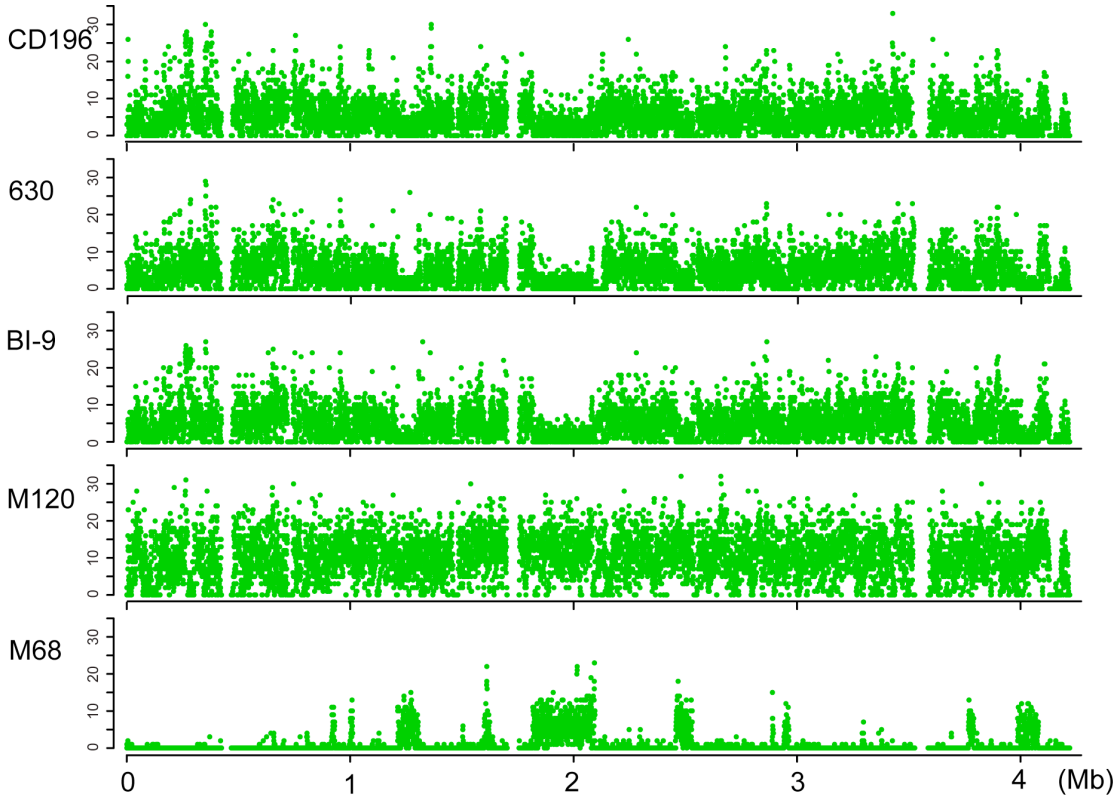

Figure 2.8: Signature of recombination in the deep-branching phylogeny. The genome-wide distribution of SNPs is shown for each strain against the core genome (excluding repetitive sequences) of strain CF5, which is indicated along the x axis. The y axis gives the number of SNPs in each 500-bp window.

The rate of homologous recombination varies hugely among bacterial species (Spratt and Maiden, 1999). To assess the impact of homologous recombination on sequence diversification, the ratio of recombination/ mutation (r/m) was calculated for within the deep-branching phylogeny. This rate gives the relative probability that a nucleotide has changed as the result of recombination relative to point mutation (Guttman and Dykhuizen, 1994; Spratt *et al.*, 2001).

The r/m ratio for *C. difficile* is between 0.63 and 1.13 based on this dataset. Vos *et al.* previously compared r/m for different bacteria based on MLST data (Vos and Didelot, 2009). They calculated this ratio to be 13.6 for *Helicobacter*

pylori, 7.1 for *Neisseria meningitidis*, and 0.1 for *Staphylococcus aureus*. Their estimated r/m for *C. difficile* is 0.2. The difference between this and the current estimates may be due to sampling of loci and strains, as the recombination events detected here seem to be very localized.

2.3.5 Selective forces acting upon the *C. difficile* genome

To investigate the selective forces acting on the *C. difficile* genome, dN/dS, the ratio of non-synonymous vs. synonymous substitution rate was calculated. A ratio significantly smaller than 1 suggests strong purifying selection, whereas a ratio close to 1 is usually taken as indicating a neutral selection pressure. However, it has been shown that for very closely related genomes, dN/dS can be close to 1 (Rocha *et al.*, 2006), either because time has been too short for significant selection to act (Rocha *et al.*, 2006) or because nucleotide substitutions within a species may represent segregating polymorphisms rather than fixed differences (Kryazhimskiy and Plotkin, 2008).

dN/dS was calculated for concatenated alignments of CDSs from the non-repetitive, core genome for each pair-wise combination of 9 isolates (630, BI-9, CF5, M68, M120, CD196, BI-1, 2007855, and R20291). It was previously suggested when effective population size is infinite or sufficiently large; the trajectory of 1/(dN/dS) exhibits a linear trend with time, but when population size decreases, the increase of 1/(dN/dS) with time reaches a plateau (Rocha *et al.*, 2006). dS (number of synonymous substitutions) or dI (number of intergenic SNPs) have been used as a measure of time since divergence, although synonymous changes and intergenic regions could also be subject to selection forces that deviate from neutral. The 1/(dN/dS) trajectory of *C. difficile* appears to be nonlinear, regardless of dS or dI being used as the indicator of time (Figure 2.9). This pattern is similar to the trajectories reported for *S. pyogenes* and the *Bacillus cereus+anthracis+thuringiensis* complex (Rocha *et al.*, 2006).

This data shows that between deeply diverging lineages, there is evidence for strong purifying selection (the average dN/dS between M120 and the rest is ~0.08). However, for recently diverged lineages, dN/dS is very close to 1, in agreement with previous analyses (Rocha *et al.*, 2006). This 1/(dN/dS) trajectory suggests nonsynonymous substitutions were purged less efficiently in *C. difficile* than in *E. coli,* whose trajectory appears to be linear (Rocha *et al.*, 2006), or that the effects of purifying selection on the *C. difficile* genome are somewhat delayed. This could be explained by a relatively small effective population size for *C. difficile* compared with *E. coli*, which has a broader host-range.

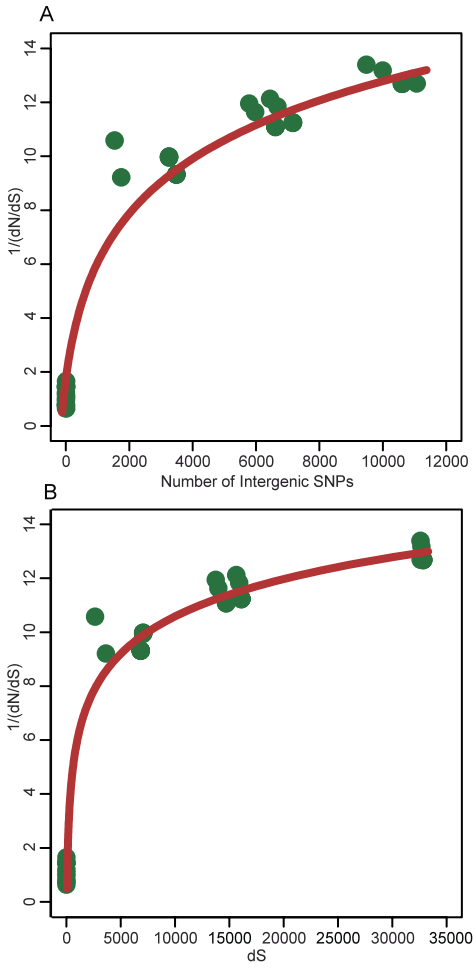

Figure 2.9: Trajectory of 1/(dN/dS) within the *C. difficile* phylogeny over time. The number of intergenic SNPs (A) and synonymous changes (B) serve as measures of time since divergence.

Genes under positive selection were also investigated and 12 potentially positively selected CDSs were identified (Table 2.3), which seems relatively small for such a diverse species. However, among these CDSs are response regulators and surface proteins, including predicted membrane and exported proteins, which are likely candidates for positive selection driven by host factors, such as the immune system or influences in the environment. It is also very likely that many more variable and positively selected genes exist, but that these are part of the accessory gene pool and therefore not captured in this analysis.

Name	Significance	ω_2	Annotation
CD0195	**	40.34	putative membrane protein
CD0707	*	37.60	putative signaling protein
CD1068	*	81.79	putative polysaccharide biosynthesis/sporulation protein
CD1755	**	133.35	putative ABC transporter, permease protein
CD1989	*	113.82	putative membrane protein
CD2022	*	32.06	hypothetical protein
CD2316	*	10.94	two-component response regulator
CD2454	*	3.96	hypothetical protein
CD2468	**	11.30	putative exported protein
CD3094	*	28.28	putative sigma-54-dependent transcriptional regulator
CD3248	*	22.62	putative polysaccharide deacetylase
CD3558	*	74.29	BirA bifunctional protein

Table 2.3: Potentially positively selected genes in *C. difficile.* Gene name refers to systematic identifier in strain 630. Significance level was determined by a likelihood ratio test. * - < 0.05; ** - < 0.01. ω_2 is the approximate mean of the posterior distribution for ω (dN/dS).

2.3.6 Core-genome and pan-genome sizes of *C. difficile*

Analysis of core- and pan-genome of *C. difficile* was conducted using 5 genomes (630, CD196, R20291, CF5, and M120). These isolates (from 4 ribotypes) were chosen because they had more reliable sequence data and they form a relatively diverse collection. Orthologues were identified in a pair-wise way between these genomes (see section 2.2.3 for details of orthologue identification). The mean number of genes contained in core- and pan-genome were determined after calculating both estimates from all possible permutations of genome order. The results show that, based on these 5 genomes, *C. difficile* possess a core-genome of ~2,900 genes and a pan-genome of ~4,550 genes (Figure 2.10); the former is equal to 76% of the 630 genome. Neither core- nor pan-genome size appears to reach a plateau (Figure 2.10).

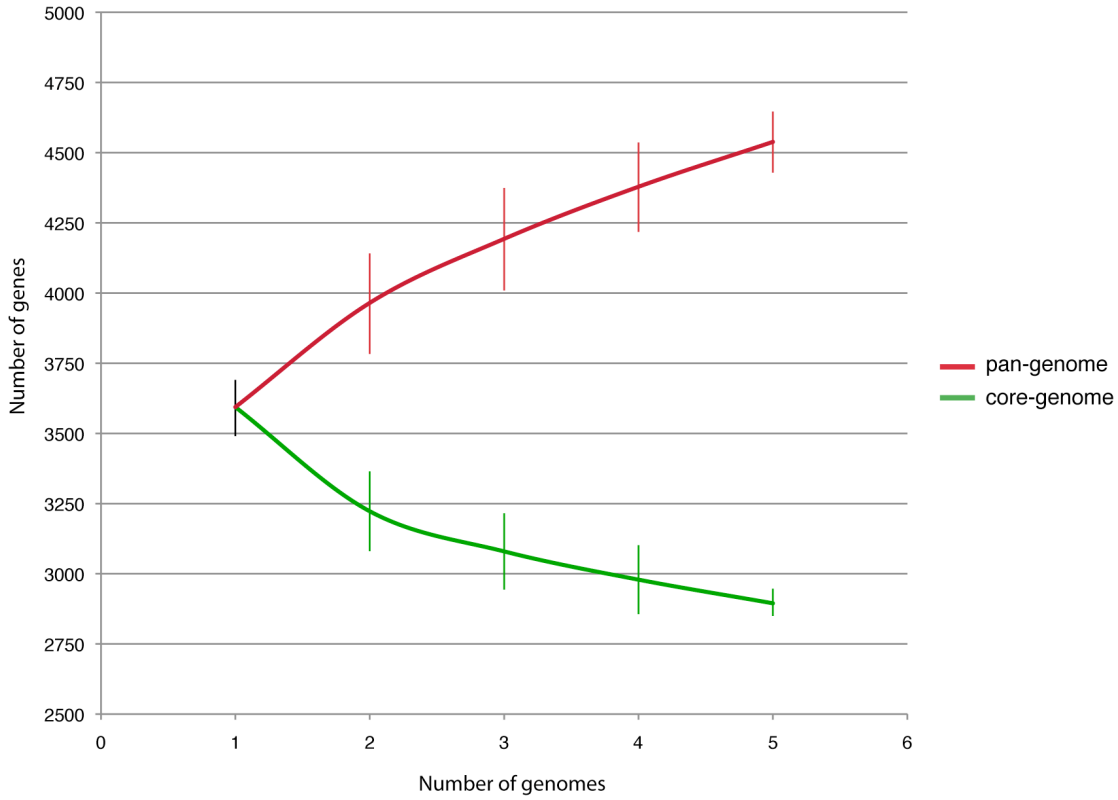

Figure 2.10: Changes in core- and pan-genome size of *C. difficile* in relation to the number of genomes. Red and green lines give mean estimates of numbers of genes in the pan- and core-genome, respectively. Coloured bars represent ranges of the estimates for all possible permutation of genome order.

2.3.7 Genome comparisons between isolates CD630, CD196 and R20291

The *C. difficile* genome is clearly dynamic and undergoes many changes. To gain further insight into the phenotypic consequence of these changes, particularly the changes that underlie recent emergence of epidemic type ribotype 027, a three-way genome comparison was conducted between two ribotype 027 isolates CD196 (historical) and R20291 (modern epidemic), and the ribotype 012 isolate 630. The CDSs unique to both 027 isolates, and to the modern epidemic isolate in particular, could have a functional impact associated with increased transmissibility and virulence. The numbers of CDSs unique to one strain or shared by more than two strains are shown in Figure 2.11.

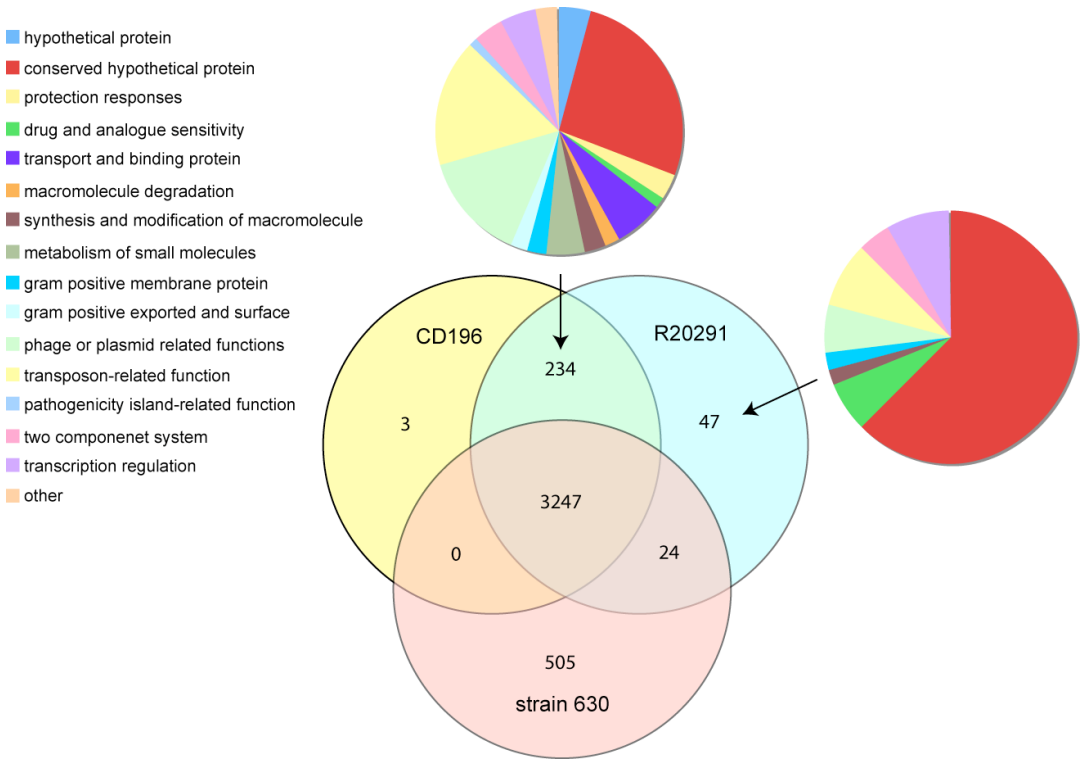

Figure 2.11: Distribution of orthologous CDSs in *C. difficile* strains 630, CD196 and R20291. The Venn diagram shows the number of genes unique, shared or core between the three strains. The associated pie charts show the breakdown of the functional categories assigned to these CDS.

The three strains share 3,247 core genes, including examples encoding determinants important for pathogenesis (Figure 2.11). There are 505 CDSs unique to 630 compared to the two 027 strains, whereas there are 47 CDSs unique to R20291 and three CDSs unique to CD196 (Figure 2.11). The locations of regions of genetic difference between the three strains are highlighted in the concentric circular chromosome representations of the three genomes (Figure 2.12). There are 234 CDSs unique to both ribotype 027 strains spread among at least 50 regions of genetic difference (Figure 2.12). These include a prophage, transposon genes, two-component response regulators, drug resistance genes, and transporter genes. All three genomes have multiple copies of a CDS named *tlpB* (transposase-like protein B). In *C. difficile* 630 there are 10 copies; of which 8 are found within CDSs. In both ribotype 027 strains 17 copies were found, of which only 6 inserted within CDSs. Only three CDSs are interrupted by *tlpB* in all three strains.

Comparison of the toxin locus revealed a single base deletion at position 117 in *tcdC* in both R20291 and CD196; this deletion is absent from 630 and results in truncation of TcdC at the 66th amino acid residue, which has been reported previously (Matamouros *et al.*, 2007). The presence of the 18-bp deletions in both R20291 and CD196 compared to 630 was also confirmed.

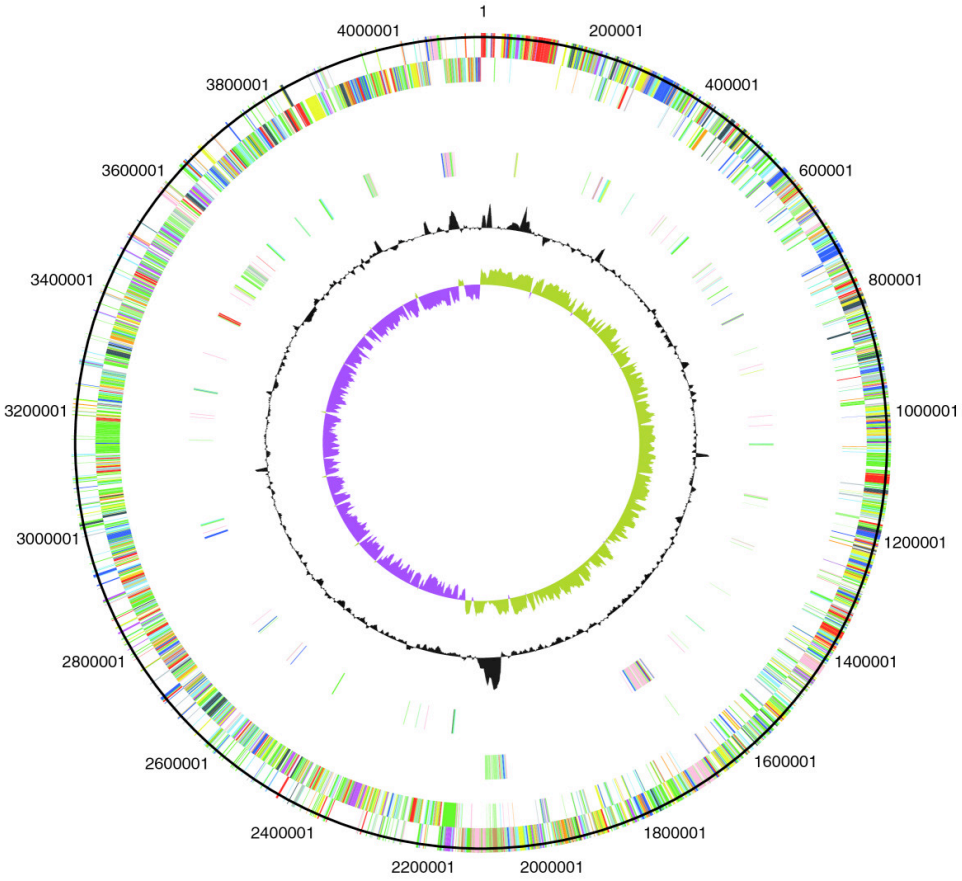

Figure 2.12: Circular representations of *C. difficile* chromosomes. From the outside (scale in bp): circles 1 and 2 show the position of R20291 CDS transcribed in a clockwise and anti-clockwise direction coloured according to predicted function; circle 3 shows CDS unique to R20291; circle 4 shows CDS unique to both R20291 and CD196; circle 5 shows GC content; circle 6 shows GC deviation (> 0%, olive; < 0%, purple). Colour coding for CDS functions: 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 (Insertion Sequence) elements; grey, miscellaneous.

The comparison between the two ribotype 027 strains revealed at least five genomic regions in the epidemic 027 strain (R20291) which were absent from the non-epidemic 027 strain (CD196). Most notably, R20291 harbours a novel 16 kb insertion that exhibits higher G+C DNA content compared to the rest of

the genome (Figure 2.13). This genomic island, named GI-R20291 (also termed Tn*6104* recently by Brouwer *et al.* (Brouwer *et al.*, 2011)), was inserted into a conjugative transposon named CTn*5b*, as it is highly similar to C*Tn5* in 630. The insertion of this genomic island disrupts the CDS CDR20291_1743 in R20291 and carries a number of cargo genes found only in R20291, including a two-component response regulator (CDR20291_1748), a putative lantibiotic ABC transporter (CDR20291_1752), three sigma-like factors (CDR20291_1754 - CDR20291_1756), a putative cell surface protein and a number of hypothetical and conserved hypothetical proteins. GI-R20291 also encodes a toxin-antitoxin system (RelE/StbE family) that is important in maintaining the stability of mobile elements (Hayes, 1998). RelE encodes a stable toxin that inhibits translation by cleaving mRNAs on translating ribosomes (Christensen and Gerdes, 2003). The toxin is inhibited by an unstable anti-toxin (RelB). This toxin-antitoxin system has been linked to translation moderation under amino-acid starvation stress (Christensen and Gerdes, 2003).

Both CD196 and R20291 share the same prophage (prophage phi-*027*, or phi*CD20*), which has integrated between the orthologues of 630 CDSs CD1566-7. The only CD196-specific CDSs in the whole genome are three consecutive CDSs within this phage region. These CDSs encode a putative phage anti-repressor and two putative uncharacterized proteins. In R20291 the three CDSs are replaced with a single putative uncharacterized protein.

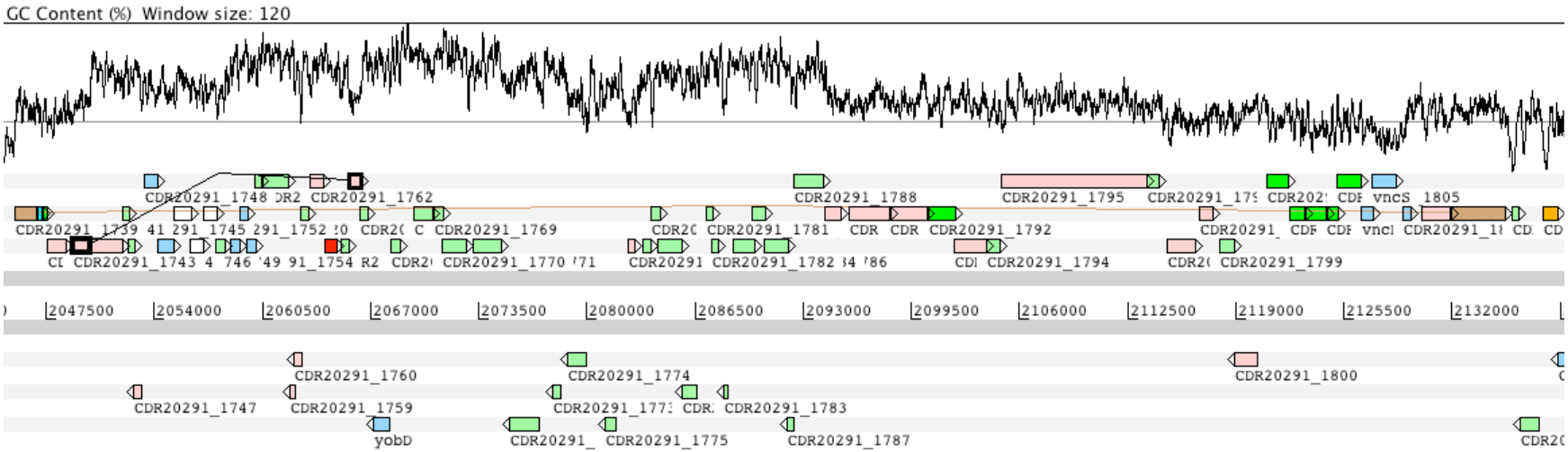

Figure 2.13: A conjugative transposon carried by R20291 but absent from CD196. The graph on top shows G+C content of the corresponding DNA sequence below. Two boxes connected by a line denote a CDS disrupted by the insertion of a genomic island. Colour coding for CDS functions: white, pathogenicity/adaptation; red, information transfer; bright green, surface-associated; pale green, unknown; orange, conserved hypothetical; pale blue, regulators; brown, pseudogene; pink, phage and IS (Insertion Sequence) element.

2.4 Discussion

In this chapter, analyses were carried out at the whole genome level that provided insights into phylogeny, horizontal gene transfer, recombination, and the evolutionary history of *C. difficile*. The findings demonstrated that the species *C. difficile* harbours significant diversity, with disease-associated isolates emerging from multiple lineages. The level of horizontal gene transfer and recombination confirms that *C. difficile* has a highly dynamic genome. However, the effect of horizontal exchange extends beyond mobile genetic elements to include large core chromosomal regions transferred over considerable phylogenetic distances.

2.4.1 Insights from deep-branching phylogeny

In the phylogenetic tree, disease-causing isolates (017s, 027s, and 078) were found in all lineages, contradicting the idea that a single lineage evolved to become pathogenic. This finding agrees with the MLST analysis of Lemee *et al.*, who found that isolates recovered from severe infection cases do not cluster into distinct lineages, thus no particular lineage is associated with increased virulence (Lemee *et al.*, 2004). The phylogeny based on whole genome sequence is also consistent with the clustering analysis from Stabler *et al.* in that all ribotype 027 isolates in our collection grouped closely together, and isolate M120 is connected to the rest of the tree by a long branch, indicating increased divergence from other ribotypes. The finding that the common ancestor of *C. difficile* dates back millions of years, and that pathogenic isolates exist in all lineages has interesting implications for the emergence of *C. difficile* as a human pathogen. It suggests there may be certain genetic elements common to all *C. difficile* strains that underlie virulence. Although *C. difficile* appears to be an ancient species, it was recognized as a pathogen only three decades ago, indicating that besides genetic modifications, changes in interaction between host and pathogen, as well as other factors such as human activity, hospital design, and antibiotic

use, may have contributed to the emergence of *C. difficile* as a major pathogen.

2.4.2 The relative impact of recombination versus mutation

Based on whole genome sequence from 6 isolates in this collection, the relative impact of recombination versus mutation (r/m) to sequence diversification of *C. difficile* is 0.63 - 1.13, indicating the effect of recombination is not negligible. Other studies based on MLST data produced smaller values for this indicator. Lemee *et al* estimated that a single allele in *C. difficile* is 8- to 10-fold more likely to be altered by point mutation than by recombination (Lemee *et al.*, 2004), which suggested an r/m of 0.1 – 0.125. A more recent MLST analysis based on 1,290 clinical isolates proposed that the effect of homologous recombination is four times lower than mutation on genetic diversification (Dingle *et al.*, 2011), which is the same as the estimate of Vos *et al.* (Vos and Didelot, 2009). The reason for this difference between estimates can be attributed to the sampled isolates and genetic loci in the analysis. The recombination blocks detected in this chapter, although large in size, appear to be very localized; and such recombination events were only found in a limited number of isolates in this collection. On the other hand, all MLST analysis mentioned above are based on much larger strain collections but very limited genetic loci for each sampled isolate. It is therefore likely that that the impact of large chromosomal exchange was underestimated in these analysis, especially when the estimate was averaged over a large number of isolates.

2.4.3 Genomic island of potential functional impact

The genetic differences between two ribotype 027 isolates and 630, particularly between R20291 and CD196 may have important functional

implications. The finding that GI-R20291, the genomic island uniquely carried by R20291, encodes many regulatory proteins is intriguing. The presence of this region may potentially have great effect on the *C. difficile* transcriptome, which calls for experimental validation.

2.4.4 Core- and pan-genome sizes

The core- and pan-genome analysis based on 5 genomes was not meant to be conclusive. As neither core- nor pan-genome size appears to reach a plateau, more genomes are needed to obtain more accurate estimates. The estimates achieved in this chapter, however, are comparable to the findings of Scaria *et al.*, whose estimation of core- and pan-genome sizes are 2,300 and 5,300, respectively, at 5 genomes (Scaria *et al.*, 2010). The difference can also be attributed to the methods used in gene prediction and orthologue identification.