Genomic variation and evolution of

Clostridium difficile

Miao He

Darwin College, University of Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy

September 2011

Declaration

This dissertation describes my work undertaken at the Wellcome Trust Sanger Institute between May 2008 and September 2011, under the supervision of Profs. Gordon Dougan and Julian Parkhill in fulfilment of the requirements for the degree of Doctor of Philosophy, at Darwin College, University of Cambridge. This dissertation is identical to that which was examined, except as required by the examiners by way of correction.

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where indicated in the text. The work described here has not been submitted for a degree, diploma or any other qualification at any other university or institution. I confirm that this dissertation does not exceed the page limit specified by the Biology Degree Committee.

Miao He

Cambridge, September 2011

Abstract

Genomic variation and evolution of *Clostridium difficile* Miao He

Clostridium difficile has rapidly emerged, in part through the transcontinental spread of various PCR ribotypes including 001, 017, 027 and 078, as the leading cause of antibiotic-associated diarrheal disease in humans. In particular, a previously rare ribotype 027 was recognized as the underlying cause of a number of hospital outbreaks worldwide. However, the genetic basis of the emergence of *C. difficile* as a human pathogen is unclear.

In this thesis, comparative genomic analysis was used to identify genetic variation within the *C. difficile* population and to further understand the evolution of this organism. Genome comparison between isolates belonging to different ribotypes revealed *C. difficile* is a genetically diverse species, which is estimated to have evolved within the last 1.1–85 million years. Disease-causing isolates have arisen from multiple lineages, suggesting that virulence can evolve independently. Horizontal gene transfer and large-scale recombination of core genes have shaped the *C. difficile* genome over both short and long time scales.

Ribotype 027 isolates have a highly similar genomic backbone. To understand the genetic characteristics driving the emergence of this group and identify the genetic relationships between pre-epidemic and recent 027s, whole genome sequencing was applied to a global collection of 339 isolates spanning 25 years. Phylogenetic analysis based on SNPs identified within the core genome discriminated between >100 distinct genotypes and identified two distinct epidemic lineages that have acquired fluoroquinolone resistance independently. One of these lineages has spread more widely and contains descendants from Canada, the USA, the UK, and Australia. Further antibiotic resistance mutations and potential signatures of immune selection were also identified. Strikingly, even among these isolates, which share a highly similar core genome, there is evidence that large-scale homologous recombination and horizontal gene transfer are significant.

The global collection also included >100 ribotype 027 isolates sampled from the same English hospital and the associated patient capture areas. Phylogenetic analysis was used to distinguish relapse from re-infection cases within this particular sample set. Additionally, by combining temporal and spatial data, the use of genetic variation analysis to study local hospital transmission was explored.

Publications

Publications associated with the work described in this thesis:

He, M., Miyajima F., *et al*. Two independent fluoroquinolone-resistant lineages of epidemic *Clostridium difficile* 027/BI/NAP1 emerged in North America and spread globally. Submitted.

Castillo-Ramirez, S., Harris, S.R., Holden, M.T., **He, M.**, Parkhill, J., Bentley, S.D., and Feil, E.J. (2011). The impact of recombination on dN/dS within recently emerged bacterial clones. PLoS Pathog 7, e1002129.

He, M., Sebaihia, M., Lawley, T.D., Stabler, R.A., Dawson, L.F., Martin, M.J., Holt, K.E., Seth-Smith, H.M., Quail, M.A., Rance, R., *et al.* (2010). Evolutionary dynamics of *Clostridium difficile* over short and long time scales. Proc Natl Acad Sci U S A 107, 7527-7532.

Stabler, R.A., Valiente, E., Dawson, L.F., **He, M.**, Parkhill, J., and Wren, B.W. (2010). In-depth genetic analysis of *Clostridium difficile* PCR-ribotype 027 strains reveals high genome fluidity including point mutations and inversions. Gut Microbes 1, 269-276.

Stabler, R.A., **He, M**., Dawson, L., Martin, M., Valiente, E., Corton, C., Lawley, T.D., Sebaihia, M., Quail, M.A., Rose, G., *et al.* (2009). Comparative genome and phenotypic analysis of *Clostridium difficile* 027 strains provides insight into the evolution of a hypervirulent bacterium. Genome Biol 10, R102.

Croucher NJ, Fookes MC, Perkins TT, Turner DJ, Marguerat SB, Keane T, Quail MA, **He M**, Assefa S, Bähler J, Kingsley RA, Parkhill J, Bentley SD, Dougan G, Thomson NR. (2009) A simple method for directional transcriptome sequencing using Illumina technology. Nucleic Acids Res. 37(22):e148.

Perkins TT, Kingsley RA, Fookes MC, Gardner PP, James KD, Yu L, Assefa SA, **He M**, Croucher NJ, Pickard DJ, Maskell DJ, Parkhill J, Choudhary J, Thomson NR, Dougan G. (2009) A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi*. PLoS Genet. 5(7):e1000569.

Acknowledgements

I am deeply indebted to my supervisors, Profs. Gordon Dougan and Julian Parkhill for allowing me to pursue this project, and for their guidance, advice, encouragement, patience and constructive criticism throughout my PhD. I am most grateful to my thesis committee advisors Drs. Alex Bateman, Chris Tyler-Smith, and Prof. James Wood for ideas and discussions. My gratitude also goes to the Wellcome Trust for providing for my research and living expenses.

I am very thankful to Dr. Trevor Lawley for numerous discussions on *C. difficile*, and for generously sharing his time and expertise. The mouse colonization experiment was conducted by Drs. Trevor Lawley and Simon Clare. It is also through Dr. Lawley's effort in obtaining a comprehensive strain collection that I was able to carry out the study of global BI/NAP1/027 microevolution.

I am hugely indebted to WTSI sequencing and finishing teams. Michael Quail, Richard Rance, David Harris, Elizabeth Gibson, Craig and Nicola Corton, Hilary Browne, Graham Rose, Karen Brooks, Christine Burrows, Louise Clark, Vicky Murray, Scott Thurston, Andries van Tonder, and Danielle Walker have all participated in generating the data used for this analysis. I thank all of them very much.

So many people in Teams 15 and 81 at WTSI have provided invaluable help and support to this project. The genomic DNA used for sequencing was prepared by Derek Pickard, Louise Ellison and Claire Raisen. I am very grateful to Kathryn Holt, who kindly helped me with the SNP analyses and provided ideas and suggestions when I first started. I am also thankful to Mohammed Sebaihia and Stephen Bentley for teaching me the basics of Artemis and comparative genomic analyses, and to Helena Seth-Smith, who helped to validate strain identities. A very big thank you to Simon Harris and Thomas Connor, who had generously shared their time and expertise in phylogenetics and population genetics, and to the Pathogen informatics team, including Thomas Dan Otto, Sammy Assefa and Jason Tsai, who have contributed analysis tools for sequencing data.

This project would not have been possible without the help of many researchers around the world who generously provided strains to be sequenced. Their names and affiliations are listed in Appendix A. Many thanks to Fabio Miyajima and Paul Roberts at the Liverpool University Hospital, and Brendan Wren, Richard Stabler, Lisa Dawson and Melissa Martin at London School of Hygiene & Tropical Medicine, who not only provided isolates but also shared valuable insights and information important to the project.

I also want to thank Sophie Palmer, Annabel Smith and Christina Hedberg-Delouka for looking after me while I'm writing and keeping me in check. I wouldn't be here without the support of my family and friends, whose words and presence keep me positive and sane.

To my parents

"The secret… I guess you've just got to find something you love to do and then… do it for the rest of your life."

Glossary

ACT	Artemis comparison tool
CDI	*C. difficile* infection
CDS	Coding sequence
CGH	Comparative genomic hybridization
contig	Contiguous sequence from overlapping reads
GA	Billion years ago
GTR	General time reversible substitution model
homosplasy	Similarity through convergent evolution but not by descent
ICE	Integrative and conjugative elements
IS	Insertion sequence
JC	Jukes and Cantor substitution model
MLST	Multi-locus sequence typing
MLVA	Multiple-locus variable-number tandem-repeat analysis
MRCA	Most recent common ancestor
PFGE	Pulsed-field gel electrophoresis
PMC	psedomembraneous colitis
REA	Restriction endonuclease analysis
RFLP	Restriction fragment length polymorphism
tMRCA	Time to most recent common ancestor
WTSI	Wellcome Trust Sanger Institute

List of Figures

List of Tables

Chapter 1

Introduction

Seventy-six years after its discovery, *Clostridium difficile* is now a leading cause of antibiotic-associated infections in hospitals worldwide. Major outbreaks have occurred in healthcare facilities since 2002, including many in North America (Loo *et al.*, 2005; McDonald *et al.*, 2005), the United Kingdom (O'Connor *et al.*, 2009) and continental Europe (Kuijper *et al.*, 2008). It is unclear how *C. difficile* emerged as a human pathogen. In this thesis, genome sequences of close to 400 *C. difficile* isolates were analyzed to investigate the reasons behind the emergence. This introduction reviews the background knowledge of the species *C. difficile* and explores its role as an opportunistic pathogen. The analysis exploits and develops methods for studying bacterial populations using genome sequences.

1.1 *C. difficile*

1.1.1 The bacterial species *C. difficile*

C. difficile is a Gram-positive, spore-forming, anaerobic bacterium that commonly resides in the large intestine of human and other mammals (Bartlett, 1994). About three percent of healthy adults and 20 - 40 percent of hospitalized patients are normally colonized with *C. difficile* (Bartlett and Perl, 2005). In healthy individuals this bacterial species is present as a component of normal intestinal microbiota and in the form of dormant spores (Bartlett and Perl, 2005). However, when a carrier is subject to antibiotic treatment, *C. difficile* can rapidly expand within the gastrointestinal tract and produce toxins that can impact on the clinical well-being of the host (Burdon *et al.*, 1981;

George *et al.*, 1982). Indeed, antibiotic treatment is considered to be an important cause for *C. difficile*-associated diseases. During and after antibiotic treatment, *C. difficile* can also sporulate, potentially resulting in an increased level of *C. difficile* spores in patients' stools. The resilient nature of *C. difficile* spores makes them highly transmissible. Risk factors for *C. difficile* infections include long-term hospital residency, advanced age, a compromised immune system, and the use of antimicrobial drugs, in particular receipt of clindamycin, cephalosporins, and fluoroquinolones (Barbut and Petit, 2001; Bartlett and Perl, 2005; Bignardi, 1998). However, in recent years an increase in *C. difficile* clinical disease has been observed in younger populations and people who have received no antibiotics weeks prior to the emergence of *C. difficile*-associated disease symptoms (Rupnik *et al.*, 2009). Reports have also indicated an increasing risk of *C. difficile* infections in children (Kim *et al.*, 2008) and pregnant women (Rouphael *et al.*, 2008).

1.1.1.1 Classification

C. difficile belongs to the class Clostridia under the phylum Firmicutes (Table 1.1). The genus *Clostridium* is a group of bacteria which are extremely diverse phylogenetically (Collins *et al.*, 1994; Kalia *et al.*, 2011). According to calibrations based on chemical compounds (2-methylhopanoids) found in cyanobacterial membranes, the *Clostridium* group diverged around 2.34 billion years ago (Sheridan *et al.*, 2003). A phylogenetic tree based on 16S rRNA (Figure 1.1) indicates the evolutionary position of *C. difficile* among close relatives.

Phylum	Firmicutes
Class	Clostridia
Order	Clostridiales
Family	Clostridiaceae
Genus	*Clostridium*
Species	*Clostridium difficile*

Table 1.1: Classification of *Clostridium difficile.*

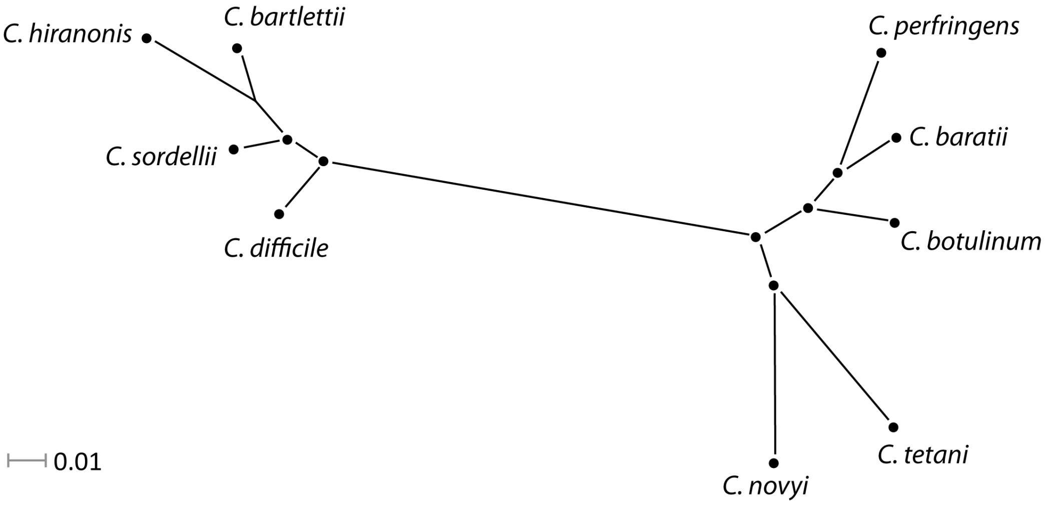


Figure 1.1: Phylogenetic tree of *C. difficile* and other *Clostridium* species. Created using PHYML (Guindon *et al.*, 2010) with 16S rRNA gene sequences retrieved from RDP Project (Cole *et al.*, 2009). Scale bar indicates substitutions per site.

1.1.1.2 History of *C. difficile* discovery and research

C. difficile was first identified by Hall and O’Toole in 1935 (Hall and O'Toole, 1935). They described this species as a normal inhabitant of the intestinal microbiota in infants (Hall and O'Toole, 1935). They also noticed that the supernatant of broth culture could cause death in a number of different experimental animal species (Hall and O'Toole, 1935). However, it was only in the late 1970s that the disease-causing nature of *C. difficile* in humans was reported. In 1974, Tedesco *et al* (Tedesco *et al.*, 1974) reported that 41 (21%) out of 200 patients treated with clindamycin developed diarrhoea and of these 20 (10%) had developed psedomembraneous colitis (PMC). This was the first report of *C. difficile* infection associated with clindamycin treatment. Bartlett *et al* suggested a toxin-producing Clostridia was responsible for antibiotic-associated PMC in 1978 (Bartlett *et al.*, 1978). This report was among the first that established the link between disease and the organism (Bartlett, 1994).

1.1.2 *C. difficile* infection

The term *'C. difficile* infection' (CDI) refers to clinical diseases associated with *C. difficile* outgrowth in humans and animals. Clinical features can range from asymptomatic carriage, moderate antibiotic-associated diarrhoea, severe PMC (Kuijper *et al.*, 2007) and even death (Vonberg *et al.*, 2008).

1.1.2.1 Symptoms

Diarrhoea is a common feature of *C. difficile* associated disease following hospitalization and antibiotic administration. It has been estimated that *C. difficile* infection underlies 10% - 25% of all antibiotic-associated diarrhoea cases (Bartlett and Gerding, 2008). Diarrhoea is usually the only symptom associated with so-called mild *C. difficile* infection (Bartlett, 2010). Other clinical presentations of CDI include abdominal cramps, fever, hypoalbuminemia (excessively low blood albumin) and leucocytosis; some patients also have faecal leucocytes (Bartlett and Gerding, 2008). These symptoms are more commonly found in patients with moderate to severe CDI, but are less likely in mildly-diseased patients (Bartlett, 2010). Fever occurs in ~28% of cases, leucocytosis in ~50%, and abdominal pain in ~22% of cases (Bartlett and Gerding, 2008).

The most advanced form of *C. difficile* associated disease is PMC. Although *C. difficile* is not the only cause of PMC, it accounts for the majority of the cases (Bartlett and Gerding, 2008). "*C. difficile* colitis" has been used as an alternative name for PMC (Dallal *et al.*, 2002). This or a similar disease was first described in 1893 (Bartlett, 1994; Rupnik *et al.*, 2009). The patient was a 22-year-woman who developed diarrhoea after a surgical removal of a gastric tumor (Bartlett, 1994). Her diarrhoea became increasingly severe and she died 15 days after the surgery (Bartlett, 1994). PMC is a severe infection which results in changes in the inner surface of the colon. Following damage, potentially precipitated by *C. difficile* toxins, immune responses are triggered and leucocytes are attracted to the gut. In severe inflammation, a mixture of dead intestinal cells, leucocytes and bacteria form yellow patches ("pseudomembrane") in the infected area (Rupnik *et al.*, 2009). Clinical presentation often includes high fever, diffuse abdominal pain and distension, and can lead to toxic dilatation (toxic megacolon) or perforation of the colon in extreme cases; this condition results in mortality 25% - 40% of the time (Kuijper *et al.*, 2007). It should also be noted that significant asymptomatic *C. difficile* colonization exists in patients. For example, toxigenic *C. difficile* strains can be carried by up to 50% of patients in nursing home facilities without causing obvious symptoms (Riggs *et al.*, 2007; Rupnik *et al.*, 2009).

1.1.2.2 Diagnostics

Suspicion of CDI in healthcare facilities was traditionally based on diarrhoea following antibiotic treatment, in addition to the foul odour of a patient's stool. However, this is hardly sufficient for diagnosis (Rupnik *et al.*, 2009). Until recently the prime diagnostic measure or "gold standard" for detecting CDI has been cytotoxin neutralization assay using *C. sordellii* or *C. difficile* antitoxin on diluted faecal samples (Bartlett, 2010; Keessen *et al.*, 2011; Wilcox *et al.*, 2010). This assay detects a cytotoxic phenotype in tissue culture (Wilcox *et al.*, 2010). An alternative enzyme immunoassay method (EIA) for *C. difficile* toxins was reported in 1984 (Laughon *et al.*, 1984), which gradually replaced the cytotoxin assay because it is rapid, inexpensive and convenient (Bartlett, 2010). An EIA test for glutamine dehydrogenase (GDH) is also being used (Bartlett, 2010). Various PCR-based assays have also been developed (Belanger *et al.*, 2003; Peterson *et al.*, 2007; Stamper *et al.*, 2009).

Among these diagnostic tests, EIA for both toxins or toxin B alone have been used most widely (Schmidt and Gilligan, 2009) but this test is relatively low in terms of sensitivity (Bartlett, 2010; Schmidt and Gilligan, 2009). Although EIA test for GDH, a cell-wall antigen almost exclusively produced by *C. difficile* proved to be more sensitive, it lacks specificity (Bartlett, 2010; Schmidt and Gilligan, 2009). Anaerobic toxigenic culture is high in terms of sensitivity but it takes too long (3-5 days) and is technically complex (Bartlett, 2010), thus it is not favored by most clinicians. However, this culture method is useful for organism characterization during an epidemic and for evaluating new methods for genotyping or phenotyping *C. difficile* (Schmidt and Gilligan, 2009). PCR-based assays may be favorable in the longer term as they are potentially more rapid and sensitive, but they are also costly and require molecular expertise within the test laboratory (Schmidt and Gilligan, 2009). A future gold standard for CDI diagnostics would be a combination of two or more of the above tests, perhaps in the form of a combined assay (Bartlett, 2010; Schmidt and Gilligan, 2009).

1.1.2.3 Epidemiology

From 2000 the number of patients with CDI during hospital stay in the USA increased from <150,000 cases to over 300,000 cases in 2006 (Rupnik *et al.*, 2009). Currently, it is estimated that CDI is responsible for 15,000 to 20,000 deaths in the USA each year (Rupnik *et al.*, 2009).

The increase in CDI incidence and severity over the past decade was first reported from the University of Pittsburgh Medical Centre (Dallal *et al.*, 2002; Rupnik *et al.*, 2009). Dallal *et al* noted that the monthly incidence of PMC had increased from ~10 cases in 1993 to >30 cases in 2000 (Dallal *et al.*, 2002). Similar increases in CDI cases were observed in many hospitals in the USA, including examples in Pennsylvania, New Jersey, Maine, Illinois, Georgia, and Oregon (McDonald *et al.*, 2005; O'Connor *et al.*, 2009). The first report from Canada emerged in 2004 (Pepin *et al.*, 2004). In Sherbrooke Hospital in Quebec province the frequency of CDI rose from 35.6 per 100,000 in 1991 to 156.3 per 100,000 population in 2003 (Pepin *et al.*, 2004). In particular, the incidence increased 7-fold for individuals over 65 (Pepin *et al.*, 2004). Rises in CDI cases were also documented in Europe (Kuijper *et al.*, 2008). Multiple outbreaks have occurred in healthcare facilities worldwide, notably in Quebec, Canada and Stoke Mandeville hospital in the UK, where outbreaks resulted in a total of 334 CDI cases and 38 deaths between 2004 and 2005 (OConnor *et al.*, 2009).

Several studies published in 2005 identified a new variant of *C. difficile* responsible for a large number of infections and epidemics in North America and Europe (Loo *et al.*, 2005; McDonald *et al.*, 2005; Warny *et al.*, 2005). This group of *C. difficile* isolates, designated as BI/NAP1/027, were found to be highly similar according to a number of typing methods: they are assigned to a single BI group when typed by restriction endonuclease analysis (REA), designated as NAP1 (North American pulse-field type I) by pulsed-field gel electrophoresis (PFGE), and characterized as ribotype 027 and toxintype III by ribotyping and toxintyping respectively (Loo *et al.*, 2005; McDonald *et al.*, 2005). Subsequently *C. difficile* infection cases attributed to BI/NAP1/027 were found in 16 countries in Europe (Figure 1.2) (Kuijper *et al.*, 2008; OConnor *et al.*, 2009), 40 states in the USA and all provinces in Canada (Figure 1.3). As of the end of 2008, BI/NAP1/027 *C. difficile* accounted for >40% of the CDI cases in UK hospitals (Brazier *et al.*, 2008).

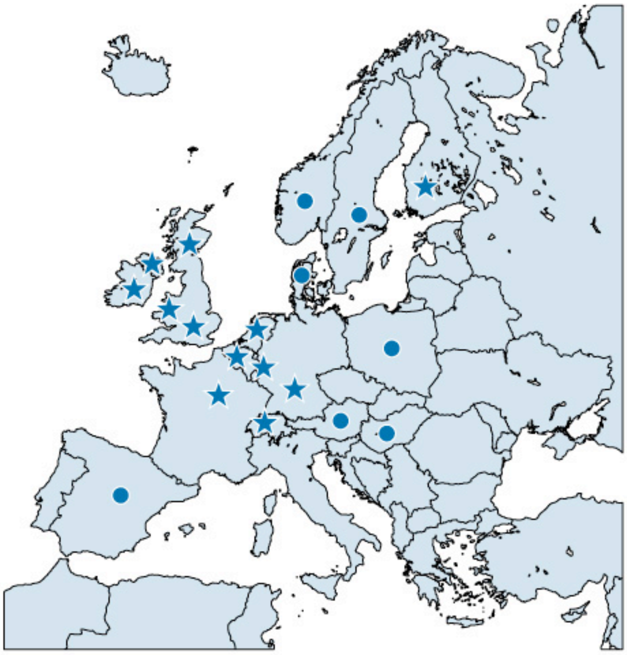


Figure 1.2: Distribution of BI/NAP1/027 *C. difficile* in European countries as of June 2008. Reproduced from (Kuijper *et al.*, 2008). Stars denote countries where outbreaks due to BI/NAP1/027 have been reported. Countries reporting sporadic cases of BI/NAP1/027 are shown by circles.

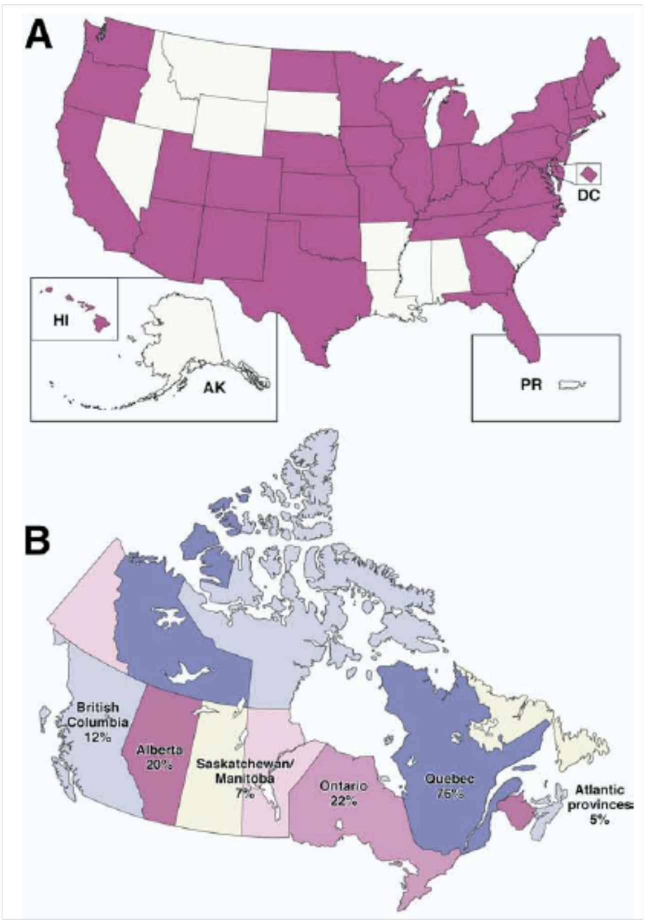


Figure 1.3: Distribution of BI/NAP1/027 *C. difficile* in the USA and Canada. Reproduced from (OConnor *et al.*, 2009). (*A*) States in the USA with >1 hospital that has reported CDI associated with BI/NAP1/027 as of October 2008 (shown in red). (*B*) Percentage of BI/NAP1/027 *C. difficile* isolates in Canadian provinces in 2005.

Prior to their epidemic spread, BI/NAP1/027 *C. difficile* were already present in the USA but they were associated with sporadic *C. difficile* disease (OConnor *et al.*, 2009). One notable phenotypic difference is the emergence of resistance to fluoroquinolone antibiotics in many of the later isolates (Loo *et al.*, 2005; McDonald *et al.*, 2005). Fluoroquinolone-resistant BI/NAP1/027 *C. difficile* variants are now prevalent in many European countries (Bauer *et al.*, 2011) and they have recently caused several outbreaks in Australia (Richards *et al.*, 2011). It is unclear what has driven the success of this lineage. While some studies have shown that BI/NAP1/027 *C. difficile* are capable of producing increased levels of toxins (Warny *et al.*, 2005) and spores (Merrigan *et al.*, 2010), other studies found no significant difference in toxin

(Akerlund *et al.*, 2008; Merrigan *et al.*, 2010) or spore production (Burns *et al.*, 2010).

The number of reported CDI cases due to BI/NAP1/027 has declined recently (Duerden, 2010). At the same time, reports have increased for CDI caused by *C. difficile* ribotypes 001, 106 and 078 (Rupnik *et al.*, 2009). In the Netherlands, the incidence of CDI associated with ribotype 078 increased from 3% to 13% from 2005 to 2008 (Goorhuis *et al.*, 2008). Ribotype 078 is currently the third most common ribotype in Europe, the first and second being 014 and 001; while 027 now only accounts for 5% of the total number (Bauer *et al.*, 2011). The population infected with 078 are younger compared to those infected with BI/NAP1/027 and the infections are more likely community-associated (Goorhuis *et al.*, 2008).

1.1.2.4 Community-associated CDI

The majority of CDI cases are found in hospitalized patients. However, community-associated CDI is an emerging issue that warrants attention. According to estimates based on data from Philadelphia (USA) and the surrounding counties, the community CDI rate is 7.6 cases per 100,000 population per year (Chernak, 2005). Moreover, community-associated CDI was found in individuals with no recent exposure to health-care settings or antibiotics; it also affects population previously considered to be of low risk, such as young children and pregnant women (Chernak, 2005; Rupnik *et al.*, 2009). The sources for community-associated *C. difficile* infections are unclear. Both hospital infections and environmental reservoir are possible contributors. There is evidence for a correlation between the presence of a BI/NAP1/027 in a hospital and the presence of similar strains in nearby communities (Rupnik *et al.*, 2009).

1.1.2.5 Environmental and zoonotic *C. difficile*

According to an early study, sources of *C. difficile* include water, soil, farm animals and vegetables (al Saif and Brazier, 1996). The highest yield of *C. difficile* has been reported from river waters, with 87.5% of the samples positive, while carriage in farm animals was as low as 1% (al Saif and Brazier, 1996). A more recent study from Austria showed that three (3%) out of 100 meat samples contain *C. difficile* (Jobstl *et al.*, 2010). However, currently there is no conclusive proof that environmental *C. difficile* contamination leads to infections in human patients (Rupnik *et al.*, 2009).

Besides humans, *C. difficile* also infects animals. CDI has been described in a number of animals, including pigs, cattle, horses, dogs, hamsters, guinea pigs, even wildlife species such as ostriches and elephants (Keessen *et al.*, 2011). The number of ribotypes of *C. difficile* from animals (approximately 30-50) is not as great as those isolated from humans (approximately 190), though the diversities recovered from both sources are relatively large (Rupnik *et al.*). The dominant ribotype found in pigs and cows is 078, although ribotype 027 was also identified (Keessen *et al.*, 2011; O'Connor *et al.*, 2009). There is no conclusive evidence for *C. difficile* transmission from animals to humans, though this has been implicated (Keessen *et al.*, 2011; Rupnik *et al.*). The presence of *C. difficile* in meat products and vegetables varies from 2.5% in Sweden to 42% in the USA (Songer *et al.*, 2007; Von Abercron *et al.*, 2009). However, no food-borne *C. difficile*-associated outbreaks have been reported (Keessen *et al.*, 2011).

1.1.2.6 Treatment, antibiotic use and resistance

C. difficile infection symptoms are a consequence of perturbation or eradication of the normal intestinal microbiota. Many antimicrobial drugs have been associated with *C. difficile* infection in humans (Bartlett, 2010). In particular, clindamycin, cephalosporins, and β-lactams (including penicillin, ampicillin and amoxicillin-clavulanic acid) confer high risks (Bartlett, 2008). Historically, the majority of human CDI agents are clindamycin-resistant (Johnson *et al.*, 1999). The persistence of CDI varies depending in part on the antimicrobial drug administered. Clindamycin-treated hamsters experience longer periods of infection; while diseased individuals overall recover rapidly following treatment with cephalosporins (Rupnik *et al.*, 2009). *C. difficile* isolates resistant to clindamycin and erythromycin have been associated with outbreaks (O'Connor *et al.*, 2009). More recently administration of fluoroquinolone antibiotics has been implicated as a prominent risk factor for CDI. All fluoroquinolones, including gatifloxacin, moxifloxacin, ciprofloxacin and levofloxacin, have been associated with subsequent CDI in humans (Rupnik *et al.*, 2009). This is very likely attributed at least in part to the rising resistance to fluoroquinolones in BI/NAP1/027 *C. difficile* (McDonald *et al.*, 2005; Rupnik *et al.*, 2009). Studies have also reported resistance to macrolide, lincosamide, streptogramin, tetracycline, chloramphenicol and rifampin and rifaximin (Dridi *et al.*, 2002; OConnor *et al.*, 2008; OConnor *et al.*, 2009; Roberts *et al.*, 1994; Spigaglia *et al.*, 2005).

The two most commonly prescribed drugs for CDI are vancomycin and metronidazole but resistance to these drugs has not yet become a serious problem (Rupnik *et al.*, 2009). Oral vancomycin was established as an agreed form of treatment for PMC in 1959 (Bartlett, 1994). This treatment was approved by FDA in 1978 and vancomycin remains the only FDA-approved drug for CDI (Bartlett, 2010) until May 2011, when fidaxomicin was approved. Reports have shown vancomycin to be highly effective for *C. difficile*-related colitis and diarrhoea (Silva *et al.*, 1981). Oral vancomycin treatment has several advantages, including few side effects by oral administration, and poor absorption, therefore high effective doses of the drug can remain in the infective site within the colon lumen (Bartlett, 1994). Fidaxomicin was very recently shown to be at least equally effective as vancomycin (Louie *et al.*, 2011).

Compared to vancomycin, metronidazole is less expensive and sometimes favoured as an alternative (Bartlett, 2010). When oral metronidazole treatment was first tested in 1978, it proved to be highly effective (Mogg *et al.*, 1979). Later reports showed that metronidazole is as equally effective as vancomycin (Teasley *et al.*, 1983). Metronidazole is almost always effective in acting against *C. difficile in vitro*, but it is absorbed efficiently and the levels remaining in the colon are extremely low (Bartlett, 1994). A recent study suggested that vancomycin and metronidazole are equally effective for treating mild *C. difficile* infection, but vancomycin is superior for severe disease (Zar *et al.*, 2007). However, no highly effective treatment has been found for the most severe cases of CDI, where orally administered vancomycin may not be able to reach diseased areas in the colon (Bartlett, 2010). Management measures in these cases include increasing doses of orally administered drugs, intravenous metronidazole, and vancomycin enemas. For patients who are severely ill and unresponsive to any treatment, colectomy (surgical removal of the colon) can be the only remaining procedure (Bartlett, 2010; Rupnik *et al.*, 2009).

Recurrent CDI is fairly typical and can be found in 5%-35% of patients (Bakken, 2009). Rates of recurrent infection for patients treated with vancomycin or metronidazole are comparable (Teasley *et al.*, 1983). The symptoms of relapse cases are nearly identical to previous infection scenarios (Bartlett, 2010). It was estimated that 20%-50% of the reoccurring infections are caused by a new *C. difficile* organism, suggesting re-infection rather than relapse (Johnson *et al.*, 1989; Rupnik *et al.*, 2009). Most relapse cases are traditionally dealt with by prolonged administration of vancomycin (Rupnik *et al.*, 2009).

Other antibiotics administered for CDI include bacitracin, teicoplanin, nitazoxanide, and fusidic acid (Bartlett, 1994, 2010). Rifaximin is also occasionally prescribed, and increasing resistance to this drug has been reported recently, especially in BI/NAP1/027 isolates (O'Connor *et al.*, 2008).

Non-antibiotic treatment for CDI includes probiotics and faecal replacement therapy; the latter recently emerged as an effective and promising treatment for recurrent CDI (Bartlett, 2010). In one study, faecal therapy successfully resolved 89% of the refractory relapsing CDI cases (Bakken, 2009). In another study, all 12 patients demonstrated immediate and durable responses after treatment (Yoon and Brandt, 2010).

1.1.2.7 Typing schemes for *C. difficile*

A number of typing schemes have been developed to discriminate between *C. difficile* isolates. These schemes, ranging from phenotypic identification to molecular methods based on genetic polymorphism, have evolved over the years and some have contributed to our understanding of *C. difficile* epidemiology and pathogenesis (Cohen *et al.*, 2001). Phenotypic typing schemes came to prominence in the 1980s and are mainly based on monitoring phenotypic traits such as antibiotic or bacteriophage susceptibility patterns (Cohen *et al.*, 2001; Killgore *et al.*, 2008). Among these methods, serotyping and immunoblotting were used more widely (Cohen *et al.*, 2001). However, these schemes have problems with reproducibility in general and have low discrimination and are being replaced by molecular typing methods.

Molecular typing schemes for *C. difficile* were introduced in the mid-1980s. These methods include restriction REA, RFLP, PFGE and PCR-based methods (Cohen *et al.*, 2001; Killgore *et al.*, 2008; Stubbs *et al.*, 1999). Ribotyping is the more widely used PCR-based method. This approach discriminates between isolates through analyzing polymorphisms in the DNA sequence of the 16S-23S intergenic spacer region (Cohen *et al.*, 2001; O'Neill *et al.*, 1996; Stubbs *et al.*, 1999). *C. difficile* isolates in the UK were assigned into 116 distinct PCR ribotypes in 1999 (Stubbs *et al.*, 1999). A toxintyping scheme specific to *C. difficile* was developed in 1998 (Rupnik *et al.*, 1998). This molecular typing method utilizes sequence variation in the toxin locus and separates isolates into ~10 groups (toxintypes I to X) (Rupnik *et al.*, 1998).

The disadvantage of these molecular typing schemes lies in the reliability and reproducibility of gel patterns, which makes them error-prone and difficult to compare between laboratories (Maiden *et al.*, 1998). A multilocus sequence typing (MLST) method was developed for *C. difficile* to overcome some of these disadvantages (Lemee *et al.*, 2004). This typing method differentiates microbial strains into sequence types according to allele pattern in nucleotide sequences of housekeeping gene fragments (Maiden *et al.*, 1998). By sequencing 7 housekeeping gene fragments of 72 isolates, Lemee *et al* identified 34 sequence types (STs) (Lemee *et al.*, 2004). A more recent MLST analysis of 1290 clinical isolates from Oxfordshire (UK) yielded 69 STs, bringing the total number of *C. difficile* sequence types to 78 (Dingle *et al.*, 2011). The discriminatory power of MLST and PCR ribotyping is roughly comparable (Griffiths *et al.*, 2009). According to a comparative study of various *C. difficile* typing methods, the most discriminatory scheme is multiple-locus variable-number tandem-repeat analysis (MLVA) (Killgore *et al.*, 2008). MLVA typing for *C. difficile* was developed in 2007 (van den Berg *et al.*, 2007) and was shown to be effective in discriminating subtypes of ribotype 027 within a single faecal specimen (Tanner *et al.*, 2010).

1.1.3 Prominent virulence factors and transmission agent

C. difficile toxins are perhaps the major virulence factors of the organism. The toxins *per se* and cell responses to toxins have been extensively studied since their discovery. Additionally, resilient spores make *C. difficile* potentially highly transmissible. *C. difficile* spores and toxins will be discussed in more detail in the following sections. Aside from these two factors, *C. difficile* also produce a surface layer (S-layer) covering the vegetative cell of the bacterium. The S-layer predominantly consists of two proteins and can act as an adhesin promoting interactions between host cells and the bacterium (Calabi *et al.*, 2001; Fagan *et al.*, 2009).

1.1.3.1 *C. difficile* toxins

Two major *C. difficile* virulence associated factors are toxin A and toxin B. Since the discovery of *C. difficile* in 1935, there has been a consensus that *C. difficile* produces a cytotoxin. However, it was not until 1981 that both toxins were characterized (Taylor *et al.*, 1981). The genes encoding these toxins, known as *tcdA* and *tcdB,* are situated in the same pathogenicity locus (PaLoc) in the *C. difficile* genome, surrounded by three additional regulatory CDSs, *tcdC*, *tcdD*, and *tcdE* (Voth and Ballard, 2005) (Figure 1.4). The toxins are glucosyltransferases that share a degree of sequence similarity (66%) and have similar domain organizations: both consist of an enzymatic domain, receptor-binding domain and a putative translocation domain (Voth and Ballard, 2005) (Figure 1.4). The toxins are detectable during the late log and stationary phase of growth; this expression appeared to be inhibited by the presence of glucose or other rapidly metabolizable sugars in the medium (Dupuy and Sonenshein, 1998). The gene *tcdC* encodes a negative regulator for toxin production (Matamouros *et al.*, 2007) and a truncated TcdC can lead to increased toxin levels (Matamouros *et al.*, 2007; Spigaglia and Mastrantonio, 2002)

The link between *C. difficile* toxins and CDI has been the subject of extensive study. *C. difficile* toxins are more likely to be detected in patients with severe symptoms (Bartlett, 1994). A study also showed that production of toxin A and toxin B is ~16 and ~23 times higher respectively in some epidemic BI/NAP1/027 strains compared to isolates of other ribotypes (Warny *et al.*, 2005).

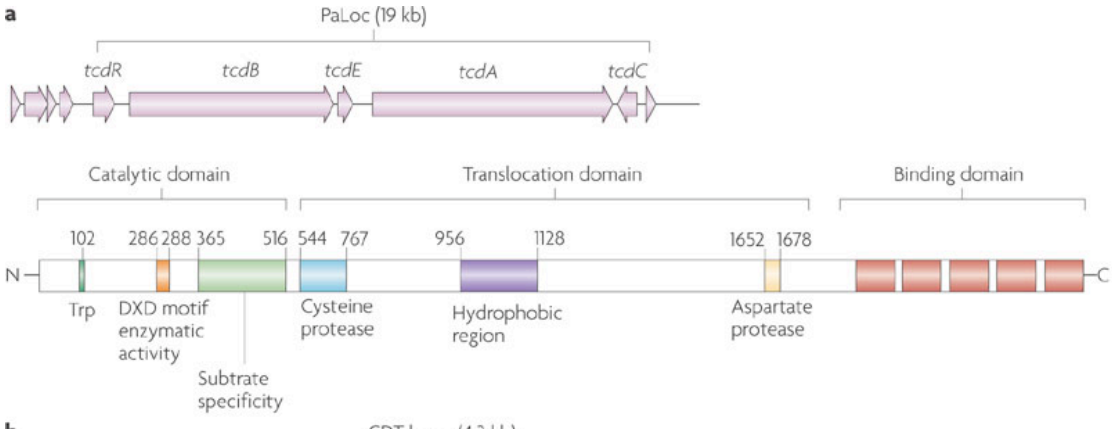


Figure 1.4: *C. difficile* pathogenicity locus (PaLoc) and domain organization of toxin B (below). Reproduced from (Rupnik *et al.*, 2009).

There has been a debate in the literature about which of these toxins is essential for CDI (Carter *et al.*, 2011; Carter *et al.*, 2010). Initially it was postulated that toxin A and toxin B act synergistically but that toxin A is the more important virulence factor (Lyerly *et al.*, 1985). However, a recent study using *tcdA* and *tcdB* mutants showed that toxin B, but not toxin A, was essential for virulence in hamsters (Carter *et al.*, 2010; Lyras *et al.*, 2009). However, an apparently conflicting study indicated that both *tcdA* and *tcdB* mutants are capable of causing disease in hamsters (Kuehne *et al.*, 2010). Natural TcdA-deficient *C. difficile* isolates do exist, particularly strains belonging to PCR-ribotype 017 (Drudy *et al.*, 2007a). These isolates, also called TcdA-TcdB+ *C. difficile*, harbour deletions in the *tcdA* gene and produce a truncated form of toxin A but an intact toxin B (Drudy *et al.*, 2007a). The disease symptoms caused by TcdA-TcdB+ strains are very similar to those due to TcdA+TcdB+ isolates (Carter *et al.*, 2010; Johnson *et al.*, 2001). TcdA-deficient *C. difficile* have also been shown to cause nosocomial outbreaks (Alfa *et al.*, 2000).

In addition to toxin A and toxin B, a small proportion of *C. difficile* harbor the genes *cdtA* and *cdtB*, which encode a two-component ADP-ribosyltransferase known as binary toxin (Voth and Ballard, 2005). TcdA+TcdB+ *C. difficile* that express binary toxins do not apparently exhibit increased virulence (Voth and Ballard, 2005). It is perhaps also worth noting that non-toxigenic *C. difficile* make up 20-25% of the total *C. difficile* population (Schmidt and Gilligan, 2009). In non-toxigenic *C. difficile* strains, the PaLoc region is substituted by a 115 bp sequence (Rupnik *et al.*, 2009).

1.1.3.2 *C. difficile* spores, spore formation and germination

C. difficile can exist in two living states: as spores and in the vegetative form. Spores are critical for *C. difficile* transmission in most healthcare facilities (Savage and Alford, 1983), while the existence of *C. difficile* vegetative cells requires an anaerobic environment such as the intestinal tracts of humans and other mammals. Spores are excreted by diseased patients colonized with *C. difficile* (Paredes *et al.*, 2005). Under traditional routine cleaning regimes, the spores could potentially persist in the environment for months while maintaining their transmissible nature (Gerding *et al.*, 2008). They are also potentially highly contagious. Indeed, only <7 spores per cm^2 is required to establish an infection in mice (Lawley *et al.*, 2009). Dormant spores can tolerate very harsh environments. For example, *Bacillus* spores can persist under extreme heat, radiation, pH and in the presence of toxic chemicals (Setlow, 2003). *C. difficile* spores are also resistant to heat and commonly used disinfectants. After remaining in a 60 °C environment for 24 hours, only <1% of the spores are inactivated (Lawley *et al.*, 2009). The viability of spores is not detectably affected by 70% ethanol (Lawley *et al.*, 2009). When tested with other common hospital cleaning agents, it was revealed that although working concentrations of all agents inhibit *C. difficile* growth in culture, only chlorine-based germicides and Virkon are effective in inactivating *C. difficile* spores (Fawley *et al.*, 2007; Lawley *et al.*, 2009). Upon ingestion and/or in a favorable environment, germination occurs and *C. difficile* spores can resume vegetative growth. Germination of *Bacillus* spores is triggered by a variety of chemical nutrients or non-nutrients, also called germinants (Moir, 2006; Setlow, 2003). These germinants include amino acids, potassium ions, and carbohydrates (Moir, 2006; Setlow, 2003). *C. difficile* spore germination is induced most effectively by the presence of cholate-derivatives, a component of bile (Lawley *et al.*, 2009). It was estimated that only 0.1% - 1% of the spores germinate routinely under normal conditions, but germination rate increased 100- to 1000-fold when cholate derivatives were supplied (Lawley *et al.*, 2009). Saxton *et al* showed that fluoroquinolones can trigger high levels of toxin production and spore germination for BI/NAP1/027 and ribotype 001 strains (Saxton *et al.*, 2009).

The conditions that trigger *C. difficile* sporulation have not been very well characterized. In general, sporulation occurs when environments do not favor normal vegetative growth. For example, *Bacillus* species produce spores as a response to starvation (Setlow, 2003). Spores may contribute to the survival and persistence of *C. difficile* within hosts during antibiotic treatment (Underwood *et al.*, 2009; Walters *et al.*, 1983). The gene *spo0A*, which encodes sporulation stage 0 protein A, plays an essential role in the initiation of sporulation; inactivation of *spo0A* results in complete deficiency in spore formation (Underwood *et al.*, 2009). The protein Spo0A is a response regulation transcription factor which requires phosphorylation to be activated. Upon activation, Spo0A can bind to specific target sequences near or in the promoters of genes under its direct regulation (Hatt and Youngman, 2000). Spo0A is directly activated by histine kinases through phosphorylation in *C. difficile*, as opposed to a relay of a series of response regulators including SpoB and SpoF, as in *Bacillus* species (Underwood *et al.*, 2009).

1.1.4 *C. difficile* genomics and genetic diversity

Significant progress has been made in the area of *C. difficile* genomics. The first sequenced *C. difficile* genome was from 630, a multidrug-resistant isolate from Switzerland in 1982 (Sebaihia *et al.*, 2006). This study revealed that the *C. difficile* genome is highly dynamic, containing many mobile genetic elements, including seven putative conjugative transposons and two highly similar prophages (Sebaihia *et al.*, 2006). These mobile elements make up a large proportion (11%) of the genome, which in the case of 630 is 4,290,252 bp in size for the chromosome (Sebaihia *et al.*, 2006). Mobile elements *C. difficile* have made a significant contribution to the acquisition of genes involved in antibiotic resistance, virulence and surface structure (Sebaihia *et al.*, 2006).

Elucidating the genomic sequence of *C. difficile* can provide a valuable resource for subsequent studies into *C. difficile* phylogeny and genetic diversity. Stabler *et al* designed DNA microarrays based on the 3,688 predicted CDSs from strain 630 and analyzed the gene expression profiles of 75 isolates (Stabler *et al.*, 2006). Using a Bayesian algorithm, they identified four distinct statistically supported clusters, including a “hypervirulent” clade, which comprised 20 of 21 BI/NAP1/027 isolates (Stabler *et al.*, 2006). The remaining clades are a toxin A-B+ clade, and two clades both incorporating human and animal isolates (Stabler *et al.*, 2006). Only 19.7% of the genes were deemed to be shared by all strains (Stabler *et al.*, 2006). However, a different comparative genome hybridization analysis of 73 *C. difficile* isolates from humans, cattle, horses and pigs indicated that approximately 16% of the genes in strain 630 are conserved across all strains (Janvilisri *et al.*, 2009). A more recent analysis of genome conservation based on 167 isolates estimated that *C. difficile* core-genome consists of 947 to 1,033 genes (25% - 28% of strain 630 genome), with a pan-genome comprised of 9,640 genes (Scaria *et al.*, 2010). Importantly, all three studies support the concept of a highly variable *C. difficile* genome.

The first analysis using MLST (Lemee *et al.*, 2004) indicated that there is limited correlation between genotype and geographical affiliation and that animal isolates do not occupy a distinct lineage but are scattered among human isolates in the clustering dendrogram (Lemee *et al.*, 2004). In addition, 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). Based on these analyses the population structure of *C. difficile* was proposed to be clonal, although recombination events do occur (Dingle *et al.*, 2011; Lemee *et al.*, 2004). Lemee *et al* further 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 based on 1,290 clinical isolates determined that the population exists as five distinct genetic clades, and that the effect of homologous recombination is four times lower than mutation on genetic diversification (Dingle *et al.*, 2011).

1.2 Genetic variation and evolution of bacterial populations

Individuals in a population are constantly evolving, generating genetic variation, which is shaped by selection forces imposed by the environment. In bacterial pathogens, genetic changes are likely to reflect interactions between microbes and hosts or aspects of infectious disease. The study of genetic variation in bacteria can improve our knowledge of bacterial pathogenesis and potentially help us design better public health measures (Maiden and Urwin, 2006). Understanding how bacterial pathogens evolve can also inform studies on epidemiology. One of the aims of epidemiology studies is to find out how isolates underlying a local or recent epidemic are related to strains recovered globally or that were previously circulating in the same area (Spratt and Maiden, 1999). The details of the approaches to address these questions need to vary depending on the organism in focus (Spratt and Maiden, 1999).

1.2.1 Genetic diversity and evolution of bacterial populations

Microbial species boundaries can be regarded as “fuzzy” and lacking a universally acknowledged definition (Achtman and Wagner, 2008; Fraser *et al.*, 2009). The definition for eukaryotic species is based on their ability to interbreed and their physical or morphological properties, but such a rationale does not translate well for bacteria (Fraser *et al.*, 2009). Our existing definitions of microbial species are based on practical methods that characterize their genotypic or phenotypic properties (Achtman and Wagner, 2008; Fraser *et al.*, 2009). In recent years there is an increasing call for combining information on genetic diversity and ecological niches to better define microbial species (Fraser *et al.*, 2009). ‘Genetic distance’ has been used to measure relatedness between bacterial isolates. Isolates are considered to belong to the same species if they show 70% or more similarity by DNA hybridization (Hanage *et al.*, 2006). However, there is no universal cut off for genetic similarity when it comes to species definition (Fraser *et al.*, 2009). The level of genetic diversity within any given bacterial species is far from uniform. Some bacterial pathogens are genetically diverse, such as *Helicobacter pylori* (Falush *et al.*, 2001) and *Neisseria meningitidis* (Jolley *et al.*, 2000; Maiden and Urwin, 2006). In contrast, the genetic diversity of other pathogens such as *Bacillus anthracis* (Keim and Smith, 2002), *Yersinia pestis* (Achtman *et al.*, 2004), and *Mycobacterium tuberculosis* (Sreevatsan *et al.*,

1997) is relatively low. The population structure of different bacterial species can vary hugely (Achtman and Wagner, 2008). This situation can be attributed to genetic mechanisms in the organisms themselves, but also external environmental factors or evolutionary processes. Varying population structure and genetic diversity is a result of a balance between evolutionary forces (Barrick and Lenski, 2009). The following sections discuss these areas in more detail.

1.2.2 Mechanisms that generate genetic diversity in bacteria

There are a number of mechanisms through which bacterial genomes change, with the key ones being point mutation, horizontal gene transfer and recombination. These events generate new variants, which are shaped by selection forces, demographic processes and chance events such as population bottlenecks (Figure 1.6) (Gupta and Maiden, 2001; Maiden and Urwin, 2006). The following sections are dedicated to introducing these mechanisms and the evolutionary consequences for each.

1.2.2.1 Nucleotide substitution

Point mutation, or single base substitution has been regarded as a driving force of genome evolution since DNA sequences became available. Although bacteria also evolve by acquisition and loss of genetic sequences, the impact of nucleotide substitution is primary and universal, as it is perhaps the original process that generates genetic variation (Lawrence, 2006). Mutation may arise when a wrong base is incorporated during DNA replication. Based on the nature of the base change, point mutations can be classified as transitions and transversions. Transition refers to replacement of a purine base with another purine or replacement of a pyrimidine base with another pyrimidine; transversion entails replacement of a purine with pirimidine or *vice versa*. Based on functional consequence, nucleotide substitutions in genes are categorized as synonymous substitutions (also called silent substitutions) and non-synonymous substitutions. Synonymous substitution does not result in a change of amino acid residue, therefore the protein sequence is not altered; non-synonymous substitutions lead to changes in protein sequences and may result in gene products that are truncated or functionally impaired.

The terms mutation and substitution are not interchangeable in the strict sense. Mutation refers to change in nucleotide sequence in an individual, while substitution suggests this mutation is fixed and carried by at least a proportion of the individuals in the whole population. Therefore substitution can be considered as a long-term consequence for a subset of mutations that have arisen, as some mutations are deleterious to the organism and purged from the population. The causes that lead to fixation of a mutation can be complex. A novel variant that results in significant advantage (reproduction success, more often termed “fitness”) to the organism can be selected for and therefore quickly spread within the population. This process is known as positive selection (Maiden and Urwin, 2006) or molecular adaptation (Yang and Bielawski, 2000). In contrast, deleterious mutations are subject to purifying selection and removed from the population (also known as negative selection or selection constraints) (Bielawski and Yang, 2003; Yang and Bielawski, 2000). The neutral theory of molecular evolution (Kimura, 1983) maintains that the majority of the genetic variation we observe is a result of random fixation of selectively neutral mutations (Yang and Bielawski, 2000); and although cases of positive selection exist, they occur relatively rarely (Endo *et al.*, 1996). In this sense neutral theory attributes a larger role to stochastic events such as population bottlenecks and genetic drift.

A number of methods have been proposed to identify sequence under positive selection. The most simple and widely used is to calculate the relative ratio of non-synonymous substitution rate (d_N) and synonymous substitution rate (d_S) in protein-coding DNA sequences (Nei and Gojobori, 1986; Yang and Bielawski, 2000). If $d_N/d_S = 1$, this indicates the amino acid change is neutral, as non-synonymous substitution is fixed at the same rate as synonymous substitution. If the change is deleterious, purifying selection should prevent it from being fixed, which leads to a $d_N/d_S < 1$. In contrast, if the change confers significant advantage, it will be promoted to fixation by positive selection, thus $d_N/d_S > 1$ signifies positive selection (Bielawski and Yang, 2003; Yang and Bielawski, 2000).

This method leads to numerous case discoveries of molecular adaptation, which more often than not for pathogens reveal genes associated with immune adaptation or reproduction (Yang and Bielawski, 2000). However, the test of d_N/d_S has limitations and can be influenced by multiple factors, including gene length, sampling of species and the amount of sequence divergence (Yang and Bielawski, 2000). As the ratio is an average over time, it may only identify positive selection acting in a suitable time frame, and thus is ineffective when applied to sequences that are either too similar or too divergent (Yang and Bielawski, 2000). Over a long evolutionary period, it is highly likely that signatures of purifying selection dominate, resulting in a d_N/d_S smaller than 1. Towards the other end of the spectrum, it was shown that d_N/d_S ratio between closely related bacterial genomes is dependent on time, therefore it is only meaningful to examine the trajectory of d_N/d_S in relation to time since divergence (Rocha *et al.*, 2006). It was also suggested that for individuals sampled from a single population, $d_N/d_S > 1$ cannot be interpreted as a signature of positive selection, as in such cases allele differences may represent segregating polymorphisms instead of fixed substitutions between species (Kryazhimskiy and Plotkin, 2008). In addition to the factor of time frame, d_N/d_S calculation averages over sites; thus its power is reduced if adaptive selection only affects a limited number of sites (Yang and Bielawski, 2000).

1.2.2.2 Horizontal gene transfer

The term “horizontal gene transfer” or “lateral gene transfer” refers to bacteria acquiring genetic material from organisms other than the immediate ancestor. This is one of the most important mechanisms contributing to bacterial genetic diversity (Gogarten and Townsend, 2005; Nakamura *et al.*, 2004; Thomas and

Nielsen, 2005). A study based on nucleotide composition analysis revealed that ~14% of the genes in 116 prokaryotic genomes can be attributed to recent horizontal gene transfer (Nakamura *et al.*, 2004). Analysis of the *Streptococcus agalactiae* pan-genome (total gene set for a species (Medini *et al.*, 2005)) suggested each added *S. agalactiae* genome brings 27 novel genes to the gene pool (Tettelin *et al.*, 2005). A comparative genomics study on *E. coli* showed that approximately 40%-50% of the genes are common to all strains examined, while the rest result from potential horizontal gene transfer events and are largely arranged into pathogenicity islands (Lloyd *et al.*, 2007; Welch *et al.*, 2002). Pathogenicity islands refer to horizontally acquired genomic regions that encode multiple genes contributing to virulence. Horizontal gene transfer appears to be extensive throughout microbial evolution, and has influenced the distribution of virtually all classes of genes, including rRNA operons (Gogarten and Townsend, 2005). Transferred genes or regions can be different in G+C content and overall sequence composition when compared to the genomic background of the recipient cell (Vernikos and Parkhill, 2006). A number of methods have been developed to detect regions in the genome that have previously undergone horizontal gene transfer. These methods are often based on a combined analysis involving G+C content, codon bias and/or nucleotide composition (Koski *et al.*, 2001; Vernikos and Parkhill, 2006), but because not all horizontally acquired DNA exhibits compositional bias, the power of these methods is still limited (Koski *et al.*, 2001).

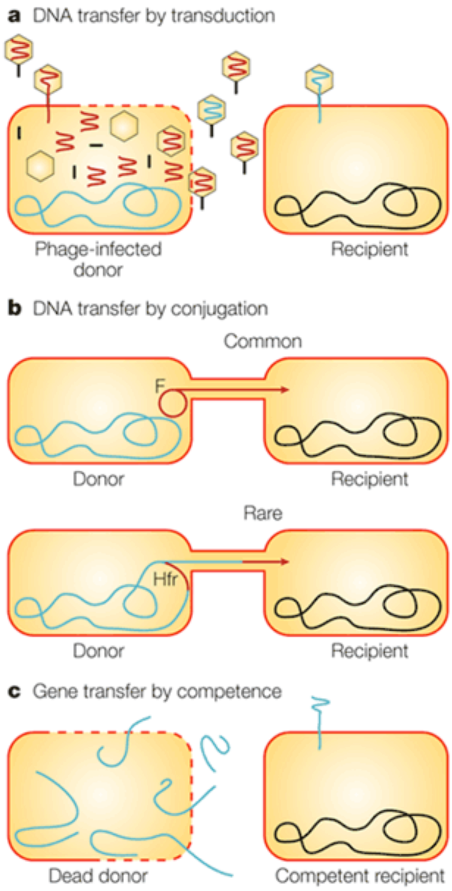


Figure 1.5: Mechanisms for DNA transfer in bacteria. Reproduced from (Redfield, 2001). a) Phage-mediated transduction. DNA from the host chromosome is accidentally packaged into the phage particle and transferred to the recipient cell. b) Conjugation by F plasmid or Hfr-mediated conjugation. c) After host cells die and decay, free DNA is released into the environment and taken by a competent recipient cell.

Prior to horizontal gene transfer, genetic material has to be transferred to a different bacterial cell. The transfer of genetic material in bacteria is unidirectional and involves a donor and a recipient (Redfield, 2001). Three main mechanisms are responsible for the process: transformation, in which a competent recipient cell uptakes free DNA from the environment; transduction, in which genetic material is transferred with the aid of bacteriophages; and conjugation, in which the recipient cell acquires DNA directly from the donor cell, with help from plasmids or other conjugative elements (Redfield, 2001) (Figure 1.5).

Transformation relies on uptake of extracellular DNA and integrating foreign DNA into the host genome, which allows transferred DNA to replicate and persist (Thomas and Nielsen, 2005). DNA is released into the environment by dead and decaying cells or by living cells through active excretion (Thomas and Nielsen, 2005). Extracellular DNA normally has to bind to specific sites on the cell surface before entering competent cells through translocation, at the same time being converted to single-stranded DNA (Thomas and Nielsen, 2005). Once within the cell, DNA is integrated into the host genome through homologous or illegitimate recombination, with the former occurring much more frequently than the latter (Thomas and Nielsen, 2005). Homologous recombination has been predicted to require a region in the DNA fragment at least 25-200 bp long with high similarity to the recipient DNA molecule (Lovett *et al.*, 2002; Thomas and Nielsen, 2005). The rates for such events are low. Recombination rates in *E. coli*, *Streptococcus* spp. and *Neisseria* spp. are in the same order of magnitude as mutation rates (Feil and Spratt, 2001). This frequency drops sharply as sequence divergence increases (Gogarten and Townsend, 2005; Thomas and Nielsen, 2005); but a maximum sequence divergence of 25% is limiting if homologous recombination is to take place (Majewski *et al.*, 2000). In some cases recombination can result in additive integration, where additional genetic material is incorporated into the recipient genome (Thomas and Nielsen, 2005).

Compared to transformation, conjugative transfer requires donor and recipient cells to be in close contact with each other. During the process, a cell-to-cell junction is formed that enables DNA to pass to the new host (Thomas and Nielsen, 2005). Most characterized conjugative systems involve plasmids, probably because they allow a set of genes to be transferred together quickly and efficiently (Thomas and Nielsen, 2005). Integration of self-transmissible elements into the recipient genome may require DNA sequence homology, which can be provided by repeated sequences in the genome, such as IS (insertion sequence) elements (Thomas and Nielsen, 2005). On the other hand`, integrative and conjugative elements (ICEs) are capable of site-specific integration in addition to excision and conjugation. These elements can spread among a range of diverse species (Skippington and Ragan, 2011; Wozniak and Waldor, 2010). Reports have shown conjugative transposons are capable of transferring between *C. difficile* and *B. subtilis* (Mullany *et al.*, 1990), also between *C. difficile* and *Enterococcus faecalis* (Jasni *et al.*, 2010). In particular, the *Tn916* family mobile elements have a very broad host range and have been identified in several phyla, including Actinobacteria, Firmicutes and Proteobacteria (Roberts and Mullany, 2009).

Horizontal gene transfer has a profound impact on microbial evolution. Bacterial genomes evolve through acquisition and loss of genes. Multiple genes associated with the same function or process are often organized into genomic islands; transfer of such elements can introduce novel functions for the recipient in a single event (Achtman and Wagner, 2008; Wozniak and Waldor, 2010). Characterized mobile elements have been found to encode functions related to antibiotic resistance, virulence, metabolism and regulation (Dobrindt *et al.*, 2004). For example, a single conjugative transposon identified from the genome of *S. pneumoniae* serotype 14 carries genes conferring resistance to erythromycin, streptothricin, kanamycin and chloramphenicol (Ding *et al.*, 2009). It appears that genes involved in DNA replication, transcription and translation are less frequently transferred than genes participating in other housekeeping functions (Jain *et al.*, 1999). In some cases genes transferred bring significant advantages to the recipient bacteria and allow rapid adaptation to new ecological niches (Gogarten and Townsend, 2005). However, adaptation to a new niche can also be achieved through gene inactivation for bacterial pathogens (Maurelli, 2007). Through insertion, point mutation or deletion, bacterial pathogens selectively discard genes or genomic regions incompatible with their new living style (Maurelli, 2007). For example, it was proposed that *Burkholderia mallei*, a horse and human pathogen lost multiple genetic loci while evolving from *Burkholderia pseudomallei*, a soil organism capable of causing infection in a number of animal hosts (Maurelli, 2007).

Despite numerous cases of adaptive traits and increased fitness resulting from horizontal gene transfer, it is still unclear what selection force is acting on the transferred DNA (Didelot and Maiden, 2010). Study of acquired genes in *E. coli* and *Salmonella enterica* indicates such genes are often still under purifying selection pressure, although the selection is weak compared to core genes (Daubin and Ochman, 2004). It is also possible that many of these transferred genes are selectively neutral (Gogarten and Townsend, 2005). A conceivable scenario is that a majority of horizontal gene transfer events are neutral or even deleterious to the recipient organism, particularly if the transfer occurs across large phylogenetic distances (Didelot and Maiden, 2010; Gogarten and Townsend, 2005). Very rarely, transferred genetic materials are advantageous to the recipient, therefore a selective sweep leads this particular subtype to spread or become fixed in the population, which is the outcome we often observe in comparative genomic analyses (Didelot and Maiden, 2010; Gogarten and Townsend, 2005). Selective sweeps promoted by the prescription of antimicrobial drugs are probably the most studied and can have a rapid impact on a population. However, horizontally acquired antibiotic-resistance genes or mutations conferring resistance can also impose fitness costs for the host organism (Gagneux *et al.*, 2006; Skippington and Ragan, 2011). As suggested by Gagneux *et al* for *M. tuberculosis*, compensatory mutations can arise to amend the fitness cost of the initial mutation; in the long term a subpopulation with lower fitness cost can become dominant (Gagneux *et al.*, 2006).

1.2.2.3 Recombination

Recombination refers to the exchange of nucleotide sequences between two similar DNA molecules. This process is ubiquitous in eukaryotic reproduction. However, as bacteria reproduce by binary fission, it was believed that this asexual process results in daughter cells genetically identical to their mother cell (Gupta and Maiden, 2001), that genetic information is vertically inherited from one generation to the next. Mutations arising endogenously are only carried by the descendents of the cells in which they occur (Gupta and Maiden, 2001; Maiden and Urwin, 2006). Consequently, genetic diversity within bacterial population is limited, as proved by an early study on *E. coli*

(Selander and Levin, 1980). In this sense, selection forces result in sequential replacements of clones by new clones of higher fitness, a process termed “periodic selection” (Figure 1.6) (Levin, 1981). This eventually leads to a population structure consisting of limited number of genetically distinct lineages (Levin, 1981; Maiden and Urwin, 2006; Spratt and Maiden, 1999). Although homologous recombination between sequences in closely related bacterial species exists, rates of such events were considered to be low (Selander and Levin, 1980).

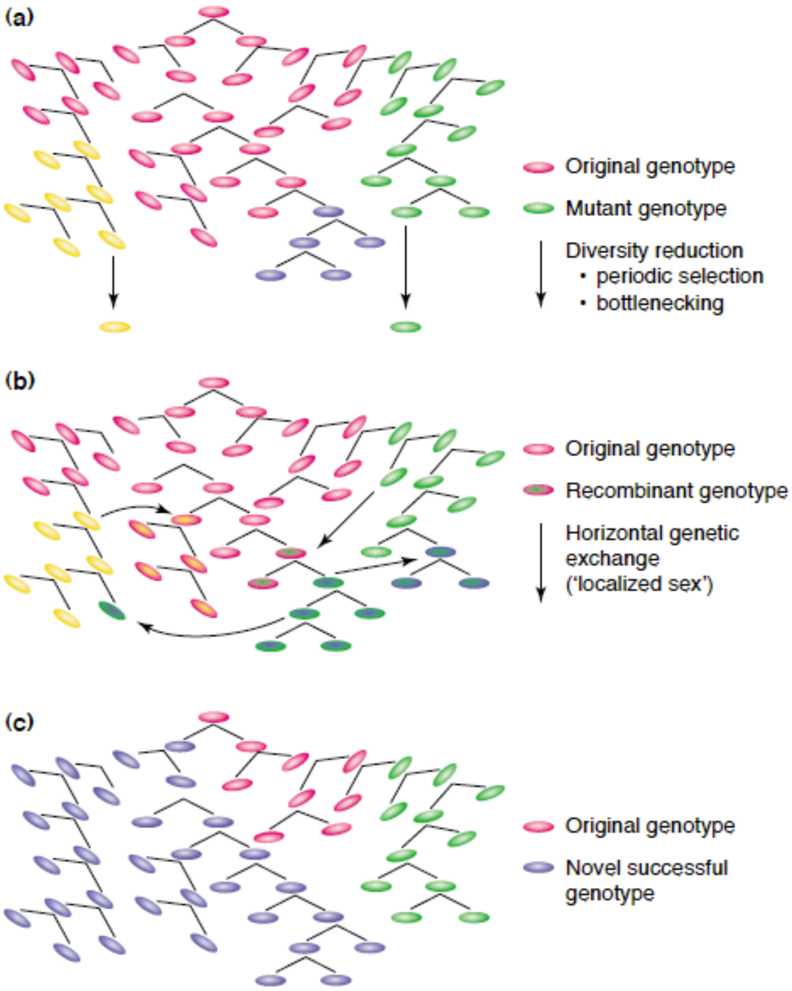


Figure 1.6: Models for bacterial population structure shaped by selection and demographic processes: (a) a clonal population with mutation and stochastic events causing diversity reduction; (b) a population with mutation, horizontal genetic exchange and recombination; and (c) dominance by a successful genotype resulting from a rapid clonal expansion. Reproduced from (Gupta and Maiden, 2001).

As the population structures of more bacterial species were revealed, conflicts with this perception arise. For example, the population structure of *Salmonella* Typhi (Roumagnac *et al.*, 2006) is characterized by abundant extant haplotypes, which argues against periodic selection (Achtman and Wagner, 2008). In particular, previous notions on the impact of recombination were called into question in the 1990s when it was shown by Guttman *et al* that gene phylogenies in *E. coli* are not congruent (Guttman and Dykhuizen, 1994). Their study argued that recombination instead of mutation is the dominant force in sequence diversification for *E. coli* (Guttman and Dykhuizen, 1994). Maynard Smith *et al* (Smith *et al.*, 1993) raised the question “How clonal are bacteria?” and showed allele associations between genes at different loci vary depending on the organism. Since then it has been widely accepted that homologous recombination does influence the population structures of bacterial species (Feil and Spratt, 2001; Spratt *et al.*, 2001; Spratt and Maiden, 1999). Its impact was shown to be extensive for *E. coli* (Wirth *et al.*, 2006), which had previously been considered as largely clonal (Levin, 1981; Selander and Levin, 1980).

The level of effect homologous recombination imposes on population structure varies hugely among bacterial species, resulting in species ranging from clonal to panmictic (Spratt, 2004). For example, little evidence of recombination can be found in pathogens such as *S.* Typhi (Roumagnac *et al.*, 2006), *Y. pestis* (Achtman *et al.*, 2004), *Bacillus anthracis* (Vogler *et al.*, 2002), and *Mycobacterium bovis* (Smith *et al.*, 2006). As for non-clonal bacteria species, an extreme example would be *H. pylori*. *H. pylori* has a long history of colonizing the gastric tracts of humans. Estimates of mutation rate, recombination rate, and average recombinant DNA size indicate that recombination is the primary driving force for *H. pylori* genetic diversification (Falush *et al.*, 2001). Imported DNA fragments have an average size of ~417 bp, and an *H. pylori* genome undergoes approximately 60 imports spanning 25,000 bp per year (Falush *et al.*, 2001). The level of recombination is so extensive that 40-2000 years is sufficient to replace 50% of the genome (Falush *et al.*, 2001). For *H. pylori*, virtually all phylogenetic signatures between isolates are obliterated, thus traditional approach used to elucidate relationships between strains (e.g. phylogenetic tree construction) is no longer appropriate (Didelot and Maiden, 2010). Aside from being fully clonal or panmictic, most bacterial species are intermediate between the two, pathogens such as *S. pneumoniae* (Croucher *et al.*, 2011; Hanage *et al.*, 2009)*, Staphylococcus aureus* (Feil *et al.*, 2003; Robinson and Enright, 2004) and *N. meningitidis* (Feil *et al.*, 2000; Feil *et al.*; Holmes *et al.*, 1999) all fall into this class. By examining the congruence between gene trees for six microbial species, Feil *et al.* concluded that recombination significantly impacts the population structures of *S. pneumoniae*, *S. aureus, N. meningitidis* and *S. pyogenes*, as the gene trees of each species showed little or no congruence (Feil *et al.*, 2001). In contrast, gene trees of *Haemophilus influenzae* and pathogenic *E. coli* isolates exhibit higher levels of congruence, suggesting they are less affected by recombination (Feil *et al.*, 2001). Statistical analysis revealed recombination rates vary extensively across bacterial species (Perez-Losada *et al.*, 2006).

To quantify the impact of recombination, a recombination/mutation (r/m) parameter was calculated to represent the relative probability that an individual nucleotide is changed by recombination or mutation (Feil *et al.*, 2001). Estimated r/m values imply recombination plays a more important role than mutation for *N. meningitidis*, *S. pneumoniae*, and *S. aureus* (Feil *et al.*, 2001). Although the impact of recombination varies for different organisms, in many cases it is sufficient to mask or distort true phylogenetic relationship (Feil *et al.*, 2001; Wirth *et al.*, 2006). More recently a comparative study of homologous recombination rates in multiple bacterial and archaeal species was undertaken (Vos and Didelot, 2009). Vos *et al.* calculated the ratio of r/m using published MLST datasets for 48 microbial species and showed that it ranges from 0.02 (95% confidence interval is 0.0-0.1) for *Leptospira interrogans* to 63.6 (95% confidence interval is 32.8–82.8) for *Flavobacterium psychrophilum* (Vos and Didelot, 2009). These estimates from different studies do not always agree (Didelot and Maiden, 2010) but some of these conflicts can be attributed to analysing methods and sampling strategies (Didelot and Maiden, 2010); both are discussed in 1.2.4.

The size of imported regions in most bacteria range from several to several hundred thousand base pairs (Didelot and Maiden, 2010; Falush *et al.*, 2001). The exchange of large chromosomal regions is thought to be rare, although cases of large scale recombination, probably by conjugation, have been identified in *S. aureus* (Robinson and Enright, 2004) and *S. agalactiae* (Brochet *et al.*, 2008). Large imports may be more likely to be selected against due to imposed reduction in fitness (Didelot and Maiden, 2010). It appears that certain genomic regions are influenced to a greater extent by homologous recombination than others. Although recombination spreads variant alleles among isolates, its effect on different loci is not uniform (Milkman, 1997). A comparative study on house-keeping genes in several bacterial species claimed that recombination rates vary hugely among loci (Perez-Losada *et al.*, 2006), but there is no apparent “hot spot” or “cold spot” for recombination (Didelot and Maiden, 2010).

A key question is what roles these recombined regions play in bacterial genome evolution. A comparative genomics study of 26 *Streptococcus* genomes suggests that a large proportion of genes under positive selection have also been shaped by recombination (Lefebure and Stanhope, 2007). An analysis of genetic diversity in *E. coli* also showed there is an association between frequent recombination and virulence (Wirth *et al.*, 2006). Wirth *et al.* postulates that more frequently recombined subpopulations are selected for their increased virulence (Wirth *et al.*, 2006). An alternative view was offered to explain association between recombination frequency and selectively advantageous traits. The view maintains that because most deleterious imports are purged from the population, what we are more likely to observe are the beneficial ones (Vos, 2009).

1.2.3 Impact on speciation

The mechanisms of genetic material exchange have a profound impact for bacterial speciation (Fraser *et al.*, 2009; Fraser *et al.*, 2007; Lawrence, 2002). Horizontal gene transfer introduces novel gene sets to the recipient organism, thus increases their ability to adapt to new ecological niches different from their old habitat, while homologous recombination acts as a force that ameliorates the genetic diversity among isolates (Lawrence, 2002). The complementary effects of mutation, horizontal gene transfer, homologous recombination and selection forces could in the long term lead to the formation of genetically isolated, distinct lineages; or as otherwise stated, new species (Lawrence, 2002). During this long evolutionary period, homologous recombination does not affect all loci uniformly, as it is bounded by selection for individuals that are more fit, thus resulting in a "grey zone" where species boundaries are "fuzzy" (Falush *et al.*, 2006; Hanage *et al.*, 2005; Lawrence, 2002). Eventually reproductive isolation can be achieved due to a large number of sequence mismatches, which act as a barrier for recombination (Falush *et al.*, 2006). This theory gained support from a computer simulation study, which showed that biological species can be created based on homologous recombination and the barrier from DNA mismatches even under a neutral model (Falush *et al.*, 2006).

Although increasing sequence divergence limits the recombination between relatively distantly related species, this process cannot be fully eliminated and has impact on speciation over long timescales (Fraser *et al.*, 2009; Fraser *et al.*, 2007). This effect was again shown by computer simulation, which suggests that homologous recombination between sequence clusters can result in either clonal divergence or sexual speciation, depending on recombination rate and sequence divergence between clusters in each case (Fraser *et al.*, 2009; Fraser *et al.*, 2007). When genetic distance between two bacterial populations is low enough and recombination between the two is sufficiently frequent, the two clusters will eventually converge into a single cluster, or species (Fraser *et al.*, 2007). This process requires the two populations to occupy the same ecological niche for genetic exchange to occur. An analysis of genetic variation in *Campylobacter jejuni* and *Campylobacter coli* proposes that the two species, which both have broad host ranges but share environments in common, are in the process of merging into one species (Sheppard *et al.*, 2008). Interestingly, S. Typhi and Paratyphi A, two serovars of the species *S. enterica*, have exchanged a quarter of their genomes historically, probably through phage-mediated exchange (Didelot *et al.*, 2007), but it appears that such broad genetic exchange has not reoccurred since then (Holt *et al.*, 2008; Roumagnac *et al.*, 2006).

1.2.4 Considerations in studying bacterial populations

In order to reveal true phylogenetic relationships between isolates, and to obtain real understanding of the entire bacterial population, two considerations are of great importance: the genetic loci examined, which are largely determined by the typing schemes we choose, and the individuals we sample from a population.

1.2.4.1 Typing schemes and choice of genetic loci

Traditionally, researchers classified bacterial pathogens based on phenotypic characteristics such as capsular and protein serotypes (Medini *et al.*, 2008). Advances in nucleic acid sequencing technologies prompted the use of 16S ribosomal RNA (rRNA) gene sequence as a marker. Indeed, this approach is still in wide use today, with 98.7%-99% sequence identity regarded as signifying the borders of different species (Medini *et al.*, 2008; Stackebrandt, 2006). However, the use of 16S rRNA is not appropriate for studying sub-populations within a given species due to a paucity of variation (Medini *et al.*, 2008).

A satisfactory typing scheme should be able to identify strains unambiguously, be easy to perform and interpret, and have good discriminatory power and reproducibility (Cohen *et al.*, 2001). Early molecular typing schemes such as ribotyping, PFGE and RFLP involve gene fragment amplification using specific primers or digesting DNA fragments with restriction enzymes and comparing gel bands following electrophoresis. These methods are easy to perform and relatively discriminatory for epidemiological purposes. PFGE has been more widely used and was employed in epidemiological studies of a range of bacterial species, including *C. difficile* (Loo *et al.*, 2005; McDonald *et al.*, 2005). Another typing method that improved our early understanding of bacterial populations is MLEE (multi-locus electrophoresis), which discriminates isolates based on differential mobilities of cellular enzymes during electrophoresis (Maiden *et al.*, 1998; Selander and Levin, 1980). MLEE was used to produce large amounts of statistically meaningful data that aided early studies on bacterial populations (Smith *et al.*, 1993). However, a major drawback with these early methods is the difficulty in comparing typing results from different laboratories (Maiden *et al.*, 1998).

MLST, which was developed to overcome this limitation, is one of the current gold standard for typing bacterial populations(Maiden *et al.*, 1998). This scheme made typing results comparable and accessible to the scientific community by recording the actual sequences of house-keeping gene fragments and storing them in publicly available databases. For example, PubMLST (http://pubmlst.org/) hosts MLST data for close to 50 bacterial species. In addition, standard phylogenetic and evolutionary analyses can be applied to DNA sequences. This enables research into the population history of sampled isolates. A number of programs have been designed to investigate bacterial population structure using MLST data, including eBurst (Feil *et al.*, 2004), START (Jolley *et al.*, 2001) and ClonalFrame (Didelot and Falush, 2007). ClonalFrame was designed to suit both MLST data and a limited number of whole genomes (Didelot and Falush, 2007; Didelot and Maiden, 2010).

For genetically uniform (monomorphic) bacterial pathogens, MLST reveals too little variation, and is less favourable compared to SNP typing (Achtman, 2008). These studies in bacteria usually utilize sets of SNPs (<10). Recent examples include the analysis of *M. tuberculosis* (Bouakaze *et al.*, 2010), *L. monocytogenes* (Ward *et al.*, 2008)*, E. faecalis* and *Enterococcus faecium* (Rathnayake *et al.*, 2011). Large scale genotyping studies in bacteria have been carried out in *Mycobacterium leprae* (>100 SNPs) (Monot *et al.*, 2009), *Y. pestis* (933 SNPs) (Morelli *et al.*, 2010) and *S.* Typhi (1500 SNPs) (Holt *et* *al.*, 2010). Recently more studies are combining genotyping and whole-genome sequencing. Comparative genomic analysis is conducted to identify polymorphic sites, which are then used to design SNP typing assays. In these cases the isolates used for initial SNP discovery have to be chosen carefully, because using biased evolutionary markers can lead to “branch collapse” during phylogenetic tree construction (Pearson *et al.*, 2004), where secondary branching is eliminated and some taxa present as a single node in the tree, although accurate positions of the node can be retained (Pearson *et al.*, 2004).

Homologous recombination can also distort true phylogenetic relationships in bacterial populations that recombine (Doolittle and Papke, 2006; Smith *et al.*, 1993). Sampling more loci from the genome (or using the entire genome) is more reliable for obtaining true phylogenies (Doolittle, 1999; Doolittle and Papke, 2006).

1.2.4.2 Sampling of bacterial pathogens

A crucial factor for studying bacterial pathogens is sampling of isolates. In every population study, the sample collection directly influences the conclusions that can be sensibly made by analyzing them (Maiden and Urwin, 2006). The variability of different pathogens determines sampling strategies for each and there is no “one size fits all” approach (Maiden and Urwin, 2006). If the goal is to study evolutionary forces acting on the entire population, then the strains sampled should be representative of the natural population of the bacteria under investigation (Maiden and Urwin, 2006). One should consider many factors when designing population sampling strategies, such as host range, natural history of the bacteria, level of genetic diversity, and measures of isolation; sometimes multiple studies are needed (Maiden and Urwin, 2006). The majority of the studies on bacterial populations rely on isolation of genomic DNA from pure cultures, however, for some pathogens only a fraction of the bacteria resume growth in culture conditions; this introduces bias.

There is extensive variability in the lifestyles of pathogens and their abilities in causing disease. “Obligate pathogens” such as *M. tuberculosis* require host environments for survival and are dependent on disease processes in order to transmit. “Opportunistic pathogens” on the other hand can be transmitted between hosts without causing disease, but become disease agents when host immune systems are weak or damaged. *C. difficile* falls into this category. There are also “accidental” pathogens which only cause disease by chance (Maiden and Urwin, 2006). It is perhaps unfortunate but not surprising that most studies on bacterial pathogens are biased towards the more “virulent” isolates in human hosts (Didelot and Maiden, 2010; Gupta and Maiden, 2001; Maiden and Urwin, 2006). Since most pathogens do not require host infections for their long term survival, such sampling strategies result in poor representation of the entire population (Gupta and Maiden, 2001). In addition, although environmental reservoirs play an important part for many non-obligate pathogens, they are often overlooked. Previous cases have demonstrated that improper sampling can lead to inaccurate and even misleading conclusions. For example, early studies on the *E. coli* population structure based on limited strain collections concluded that it is largely clonal (Selander and Levin, 1980). However, by analyzing a collection of highly diverse isolates from multiple geographical locations and hosts including both healthy and diseased individuals, Wirth *et al.* revealed that homologous recombination is frequent enough to disrupt the clonal framework of *E. coli* house-keeping genes (Wirth *et al.*, 2006). A similar case involves *N. meningitidis*. An early study based on MLEE suggested it has a clonal population structure (Caugant *et al.*, 1987) but it was later discovered that the sampling of the initial analysis is biased towards more virulent isolates, thus resulted in an over-representation of certain genotypes (Smith *et al.*, 1993).

1.3 Genome sequencing of bacterial pathogens

Before the advent of next-generation sequencing technologies, people studied genetic variation in bacteria by focusing on particular genes, or a set of housekeeping genes (Lawrence, 2006; Lemee *et al.*, 2004; Reid *et al.*, 2000). These analyses provided valuable knowledge of the genetic diversity of bacterial populations. However, the level of genetic variation uncovered by sequencing fragments of house-keeping genes can be low; this limits the discriminatory power of MLST (Medini *et al.*, 2008). In particular, some bacterial pathogens have very little sequence diversity (referred to as "genetically monomorphic") and yield only a handful of polymorphisms or even none when sequencing several genes (Achtman, 2008). Prominent examples of monomorphic bacterial pathogens include *B.anthracis* (Keim and Smith, 2002), *S.* Typhi (Holt *et al.*, 2008; Roumagnac *et al.*, 2006), *Y. pestis* (Achtman *et al.*, 1999) and *M. tuberculosis* (Sreevatsan *et al.*, 1997). Isolates belonging to these species will appear to be almost uniform when examined with MLST. In addition, studies focused on particular genes associated with virulence may not be informative in revealing genetic diversity or real ancestry for the population, as these genes are typically subject to constant selection pressure, including that of the human immune system; the conclusions drawn from these genes are hardly representative of the rest of the genome (Medini *et al.*, 2008).

Until half a decade ago, the gold standard for DNA sequencing has been the chain-termination method developed by Frederick Sanger (Sanger and Coulson, 1975) in the 1970s. Using this method, Sanger and colleagues determined the sequence of bacteriophage phiX174 in 1977, which marks the first time researchers sequenced a complete genome (Sanger *et al.*, 1977). This technique was later automated so that data can be directly recorded into a computer (Hutchison, 2007; Smith *et al.*, 1986). Automated Sanger sequencing (also called "capillary sequencing" as it more recently included a capillary electrophoresis step) was used to determine the genome sequences of several important organisms. Most notably, the published sequence of the first human genome marks a milestone in biological science research (Rubin *et al.*, 2004). The first whole-genome sequence of a bacterium was published in 1995 and is that of *H. influenzae* strain Rd (Fleischmann *et al.*, 1995). Since then the number of complete genomes has increased exponentially (Figure 1.7). As of July 2011, there are 9,133 genome projects as recorded by

Genomes Online Database (http://www.genomesonline.org/) (Bernal *et al.*, 2001), of which 7,400 (81%) are bacterial genome projects. Out of all bacterial genome projects, about 40% are for human pathogens (Fraser-Liggett, 2005). This rapid progress in genome sequencing was primarily made possible by the developments in next-generation sequencing technologies.

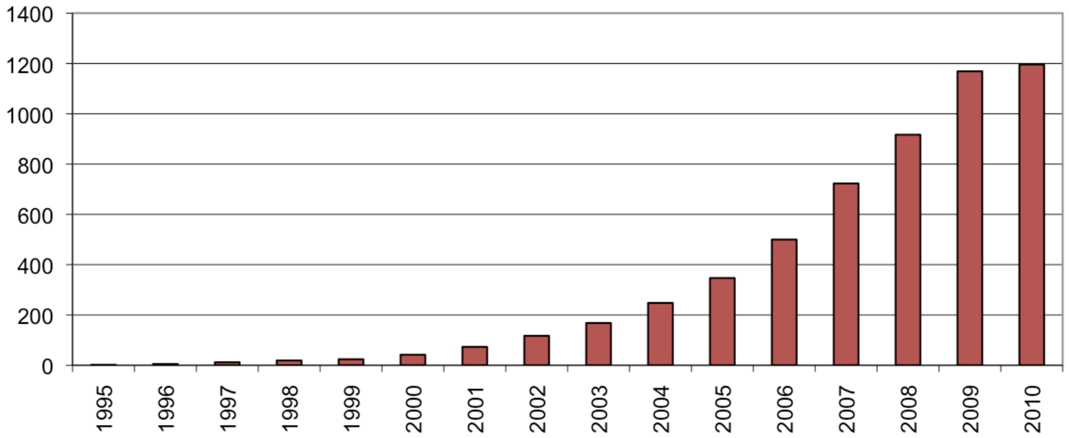


Figure 1.7: Number of sequenced complete genomes in each year (y-axis) from 1995 to 2010 (Genbank). Data sourced from Genomes Online Database.

1.3.1 Next generation sequencing

Next-generation sequencing is an umbrella term used to refer to new technologies in sequencing. Compared to conventional Sanger sequencing (now "the first generation of sequencing"), the new technologies have greater through-put (Figure 1.8), but are reduced in time and cost. Traditional Sanger capillary sequencing involves a cloning step, which is both labour-intensive and speed-limiting. It usually takes several years to finish a bacterial genome by capillary sequencing. Next generation sequencing technologies circumvent some of these steps and allow large numbers of DNA fragments to be sequenced in parallel (Shendure and Ji, 2008). One main feature of the new technologies lies in the data they produce. Next generation sequencing platforms produce far more sequences of shorter read length (30 bp-300 bp) and lower raw accuracy (Shendure and Ji, 2008). This brings new challenges to genome assembly and annotation (Pop and Salzberg, 2008). The following sections will be dedicated to discussing the technologies *per se*, the accompanying analysis methods and their applications in bacterial genomics and population genetics research.

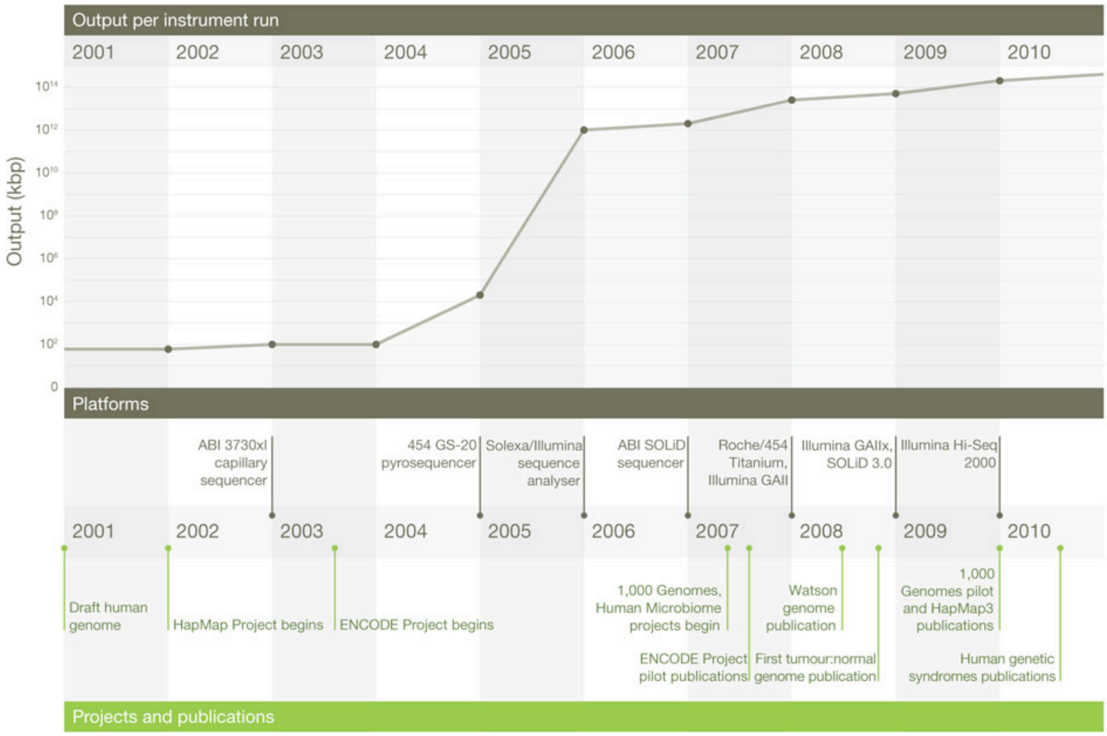


Figure 1.8: Increase in sequencing capacity during the first decade in 21st century. Reproduced from (Mardis, 2011). Top: amounts of data per run per instrument; middle: sequencing platforms; bottom: major genome projects.

1.3.1.1 The new sequencing technologies

Currently, three next-generation sequencing technologies are probably more widely used than others: namely, 454 Life Sciences/Roche, Illumina/Solexa (or Illumina Hi-seq 2000 more recently), and Applied Biosystems/SOLiD (Metzker, 2010; Shendure and Ji, 2008). Each has its own specifics in terms of sample preparation, run time, data yield, read length and accuracy (Table 1.2). However, the principles underlying these new technologies are similar. Basically, next generation sequencing involves random fragmentation of genomic DNA, ligation of adaptor sequences, PCR amplification, and the actual sequencing process, which is alternating cycles of enzyme-driven chemical reactions and data recording based on captured images (Shendure and Ji, 2008). Differences between platforms are mainly reflected in DNA amplification, the chemistry and the method for base detection.

Roche/454 Life Sciences technique involves a sequencing by synthesis method and pyrosequencing on DNA beads in tiny wells (Margulies *et al.*, 2005). The amplification process is achieved by PCR in a water-oil emulsion (Margulies *et al.*, 2005). In pyrosequencing, incorporation of a nucleotide triggers a reaction cascade involving ATP sulfurylase and luciferase, which leads to emission of a pulse of light (Margulies *et al.*, 2005). One major limitation of this platform are errors due to homopolymer tracts, which are stretches of sequence consisting of the same base (eg. “AAAAAA”). Because only one type of nucleotide is added at one single time, and no termination procedure is in place, multiple nucleotides of the same kind can be added consecutively; and the length of homopolymer is determined based on signal intensity (Shendure and Ji, 2008). This intrinsic feature makes 454 technology more prone to insertion and deletion errors (Huse *et al.*, 2007; Shendure and Ji, 2008). However, Roche/454 platform produces longer reads (200–400 bp) than other new sequencing technologies and thus is preferred for *de novo* assembly (without the guide from a reference sequence).

Platform	**Throughput**	**Read length (bp)**	**Cost**	**Accuracy base read**
Sanger/capillary (ABI 3730xl)	115 kb/day	500 – 1,000	$500/Mb	>99%
Roche/454 FLX	400 – 600 Mb/run (8h)	200 - 400	$60/Mb	>99.5%
Illumina/Solexa GAII	95 Gb/run (9 d)	150	$2.0/Mb	99.8%
Illumina HiSeq 2000	600 Gb/run (11 d)	100	$0.1/Mb	

Table 1.2: Comparison of sequencing platforms. Data source: Roche/454 FLX from (Gupta *et al.*, 2010; Metzker, 2010; Shendure and Ji, 2008); Illumina from http://www.illumina.com/systems/sequencing.ilmn; Sanger/capillary from (Mardis,

2011; Shendure and Ji, 2008); accuracy per read sourced from (Tettelin and Feldblyum, 2009).

The Illumina/Solexa sequencing technology is also a sequencing by synthesis method. It is currently more cost-effective compared to Roche/454 platform and is also the most widely used platform in the field (Metzker, 2010). The amplification step is achieved by a process called “bridge PCR”. Prior to this process, templates linked with adaptors are attached to a solid surface, usually a glass slide (called “flowcell”), which is already densely covered by covalently bound primers (Turcatti *et al.*, 2008). Clone clusters of the same DNA fragment are formed through amplification of the static template with nearby primers (Bentley *et al.*, 2008; Turcatti *et al.*, 2008). In terms of sequencing biochemistry, Illumina/Solexa utilizes a reversible termination technique, which is essentially based on the same principle as the Sanger method, the difference being in Solexa sequencing DNA synthesis is momentarily terminated after incorporation of each base (Bentley *et al.*, 2008; Turcatti *et al.*, 2008). The nucleotides used in Solexa sequencing are a set of four “reversible terminators”, each labelled with a blocker and a fluorophore corresponding to the base it carries; both the blocker and the fluorophore are cleavable (Bennett, 2004; Bentley *et al.*, 2008). The fluorophore on an incorporated base is excited by a laser and releases a coloured light, which is captured by the imaging system and used to determine base identity (Bentley *et al.*, 2008). After each cycle both the blocker and the fluorescent label are removed; this allows the next cycle of incorporation (Turcatti *et al.*, 2008). The Illumina/Solexa platform is much less prone to homopolymer tract errors compared to Roche/454. Instead, substitution is the dominant error. It was revealed by an independent study that an inaccurate base call is more likely to occur to a base immediately following base “G” (Dohm *et al.*, 2008). Also, low sequence coverage tends to fall in AT-rich, repetitive regions (Harismendy *et al.*, 2009). Increasing sequencing depth is able to compensate for these errors (Dohm *et al.*, 2008). Improvements to Illumina protocols have been made to achieve better results (Quail *et al.*, 2008).

The standard library preparation protocols can be modified to perform paired-end sequencing and multiplexed (indexed) sequencing. In paired-end sequencing, two sets of sequencing primers are used; this allows sequencing reads to be generated from both ends of each DNA fragment (Roach *et al.*, 1995). Because DNA fragments are size-selected prior to cluster formation, a read pair contains not only sequence information but also approximate length of the insert between them. Paired-end sequencing is particularly useful for *de novo* assembly and identifying structural variants or chromosomal rearrangements (Campbell *et al.*, 2008; Korbel *et al.*, 2007). A multiplexed library is prepared by introducing a set of “index tags” to fragmented DNA. Illumina/Solexa sequencing currently allows up to 12 samples per lane, or 96 samples per flowcell. In the analysis stage, each read can be traced back to individual sample based on the unique index it carries. This modification greatly enhances sequencing throughput.

1.3.1.2 Next generation sequencing bioinformatics

Large numbers of short reads with less accurate base calls pose challenges for bioinformatic analysis (Pop and Salzberg, 2008). Mapping reads to a reference sequence and *de novo* assembly are the two most frequently used strategies for short read data. In genome assembly, overlapping short sequence reads are identified and joined to form a contiguous sequence or “contig”. With read pair information, contigs can be further joined together to form “scaffolds” or “super contigs”, which are larger segments of the genome. Gaps could remain between the ends of two adjacent contigs. Targeted PCR amplification is required to fill the gaps and stitch contigs or scaffolds together.

Whole genome assembly using short sequencing reads is computationally challenging due to sequence repeats, differential coverage and base call error (Horner *et al.*, 2010; Miller *et al.*, 2010; Nagarajan and Pop, 2010). Assembly programs intended for Sanger sequencing reads (such as PHRAP (Gordon *et al.*, 2001)) are not suitable when dealing with large datasets produced by new sequencing platforms (Horner *et al.*, 2010). The longer reads from Roche/454 platform are preferred for *de novo* assembly, as shown with several bacterial genome projects (Chaisson and Pevzner, 2008); although Illumina data can also be used (Hernandez *et al.*, 2008; Studholme *et al.*, 2009). The Newbler program distributed with 454 machines has been used in several sequencing projects. Paired-end data allows more accurate placements of sequencing reads and is favourable. A common indicator of assembly quality is N50 contig length, which refers to "the length of the smallest contig in the set that contains the fewest (largest) contigs whose combined length represents at least 50% of the assembly" (Miller *et al.*, 2010). Currently almost all assembly programs are based on a mathematical graph approach (MacLean *et al.*, 2009; Nagarajan and Pop, 2010). In particular, the Velvet assembler (Zerbino and Birney, 2008) utilizes a de Brujin graph-based approach and works with short sequence reads. It also takes read pair information into account (Zerbino and Birney, 2008). Other assemblers specifically designed for short Illumina reads include SOAPdenovo (Li *et al.*, 2010) and ALLPATHS (Butler *et al.*, 2008).

Some assemblers allow genome assembly using mixed data from multiple platforms, such as Celera (http://sourceforge.net/projects/wgs-assembler/) (Nagarajan and Pop, 2010). It appears that mixed-platform data can be used to generate good quality assemblies (Aury *et al.*, 2008; Goldberg *et al.*, 2006). However, the performance of many genome assemblers strictly relies on input parameters and data quality (MacLean *et al.*, 2009). Programs aiming at identifying best parameter options were developed, such as VelvetOptimiser (http://bioinformatics.net.au/software.velvetoptimiser.shtml). Due to the changes brought by new sequencing technologies, most genome projects are not brought to a finished standard but instead towards a reasonably accurate "draft" sequence (Chain *et al.*, 2009).

In many cases scientists are interested in genetic polymorphisms within a known species, rather than the sequence of a new organism. Genome re-sequencing and variation detection based on read alignment (Figure 1.9) is more suitable for this purpose. Again, mapping programs should be able to accommodate sequencing errors while handling millions of short reads

(Horner *et al.*, 2010; Miller *et al.*, 2010). Phred score (Ewing *et al.*, 1998), a quality indicator initially developed for Sanger sequencing, is also used for next generation sequencing data. It gives the logarithmic probability that a given base is incorrectly called (Ewing and Green, 1998). The program Mapping and Assembly with Quality (MAQ) (Li *et al.*, 2008) is specifically designed to take base quality scores from Illumina data into account. Other mapping programs include BWA (Li and Durbin, 2009), SSAHA (Ning *et al.*, 2001) and Bowtie (Langmead *et al.*, 2009). In general there is a trade-off between mapping accuracy and efficiency, as reflected from these programs (Nagarajan and Pop, 2010). Illumina data are probably the norm for variant detection, although longer reads from Roche/454 or Sanger sequencing can also be processed with software such as MUMmer (Kurtz *et al.*, 2004).

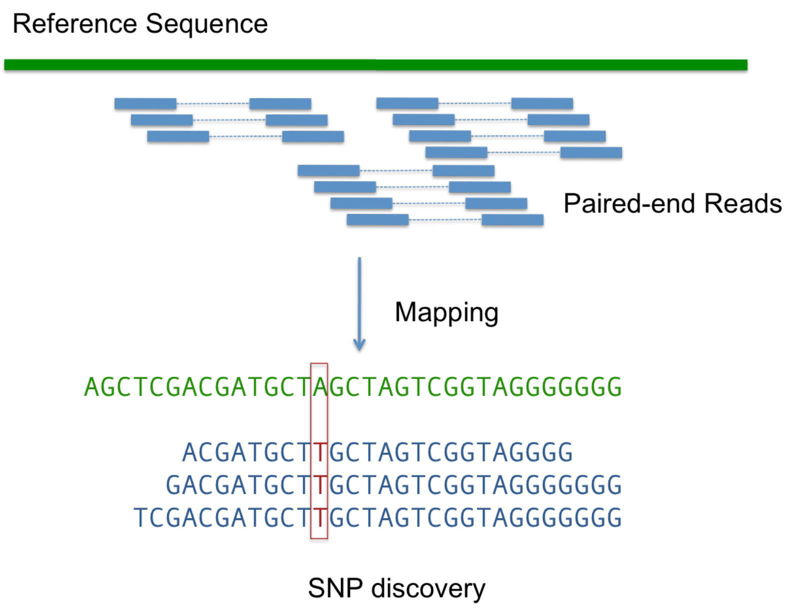

Reference Sequence
Paired-end Reads
Mapping
AGCTCGACGATGCTAGCTAGTCGGTAGGGGGGG
ACGATGCTTGCTAGTCGGTAGGGG
GACGATGCTTGCTAGTCGGTAGGGGGGG
TCGACGATGCTTGCTAGTCGGTAGGGGGGG
SNP discovery

Figure 1.9: Mapping paired-end reads to a reference sequence and identify SNPs. Reproduced from (Nagarajan and Pop, 2010).

1.3.2 Studying bacterial populations using next-generation technologies

Next generating sequencing brought major changes to the study of bacterial pathogens. The ability to sequence more bacterial genomes rapidly and cheaply impacts on the type of questions that can be addressed and the ways to address them. With the new technology, advances were made in many areas in just a few years. The studies transformed by new sequencing technologies include, but are not limited to, comparative genomics, population genetics, evolution, epidemiology, metagenomics, and gene expression.

Bacterial genomics was transformed from sequencing one single isolate at a time to large-scale comparative genomics (Brinkman and Parkhill, 2008). An early comparative study of pathogenic *E. coli* O157:H7 and non-pathogenic isolate K12 identified 1,387 genes unique to O157:H7 (Perna *et al.*, 2001). This result revealed the surprising variability among isolates of the same species. Later studies in this area focused on core-genomes and pan-genomes of a number of bacterial species, including *S. agalactiae* (Tettelin *et al.*, 2005), *H. influenzae* (Hogg *et al.*, 2007), *S. pneumoniae* (Donati *et al.*, 2010; Hiller *et al.*, 2007), and *C. difficile* (Scaria *et al.*, 2010). Analysis of core and pan-genomes was extended to compare several *Streptococcal* species (Lefebure and Stanhope, 2007). Previously researchers studied bacterial genome evolution by comparing a limited number of genomes (Bentley and Parkhill, 2004) or by sampling a limited number of loci in the genome (Achtman, 2008). However, for monomorphic bacterial pathogens, whole-genome sequencing may be the only way to investigate their evolution, transmission and epidemiology (Parkhill, 2008) and allow us to link phenotypes and genotypes through association studies (Falush, 2009; Falush and Bowden, 2006). The power of these analyses is exemplified in studies in *S.* Typhi (Holt *et al.*, 2008), *S. aureus* (Harris *et al.*, 2010), and *S. pneumoniae* (Croucher *et al.*, 2011). Through genome sequencing of multiple isolates, scientists are able to trace their trans-continental spread, as well as transmission within and between local health-care facilities (Beres *et al.*, 2010; Lewis *et al.*, 2010; Pallen *et al.*, 2010). In addition to studies of natural populations, genome sequencing has also been applied to study evolution of *E. coli* (Barrick *et al.*, 2009) and *B. subtilis* (Srivatsan *et al.*, 2008) under laboratory conditions.

New sequencing technologies also prompted developments in metagenomics, which generally refers to studying microbial communities directly from their natural habitats without the culturing process (Handelsman, 2004; Medini *et al.*, 2008). Studies in metagenomics broadened our knowledge of the microbial diversity present in communities, which had been largely unappreciated. Previous studies predominantly focused on microorganisms that could be cultured. However, it appears that only <1% of environmental bacteria can be readily cultured in laboratories (Handelsman, 2004). With short read sequencing, researchers have begun to study microorganisms in both natural and extreme environments (Tyson *et al.*, 2004; Venter *et al.*, 2004). The microbial community in human guts is also a subject of intense interest (Dethlefsen *et al.*, 2007; Gill *et al.*, 2006; Turnbaugh *et al.*, 2006), as it is closely linked to our health. In addition, the new technologies promote our understanding of pathogen biology. By sequencing transposon mutant libraries, one can identify genes required for survival or increased fitness under various conditions (Gawronski *et al.*, 2009; Langridge *et al.*, 2009). Sequencing cDNA libraries also provides a novel and unbiased way to study the pathogen transcriptome (Albrecht *et al.*, 2010; Croucher *et al.*, 2009; Perkins *et al.*, 2009; Sharma *et al.*, 2010).

1.4 Thesis outline

In this thesis, whole genome sequencing technologies were used to study the variation of *C. difficile* from genomic and evolutionary perspectives. The data generated were analysed using both phylogenetic and comparative genomic approaches. Chapter 2 investigates *C. difficile* genetic diversity through the analysis of eight isolates belonging to different ribotypes, and 25 isolates within a single ribotype - 027, which emerged recently and is associated with hospital outbreaks. Homologous recombination and horizontal gene transfer were identified as the two primary mechanisms underlying *C. difficile* genetic diversity. Selective pressures acting on the genome and on individual genes were assessed. To identify the genes unique to ribotype 027, particularly genes unique to very recent 027 isolates, a three way genomic comparison was carried out between two 027 isolates (one modern and one historic) and a non-027 isolate.

Chapters 3 and 4 continue with the investigation of genetic variation and evolution of ribotype 027. The analyses in these two chapters can be seen as representing both ends of the spectrum in terms of spatial distribution of samples. The study in Chapter 3 is based on a global collection of 339 ribotype 027 isolates. Illumina sequencing was used to identify SNPs from core genomes of these isolates which have highly similar genomic backbones. The analysis provides insights into the emergence and global transmission of modern day *C. difficile* 027 and highlights mutations, genes and genomic regions that potentially underlie this emergence. Finally, Chapter 4 focuses on dissecting the genetic variation between ribotype 027 isolates sampled from human patients in a local hospital area within a limited time frame. The study explores the use of whole genome sequencing in investigating local epidemiology.

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.

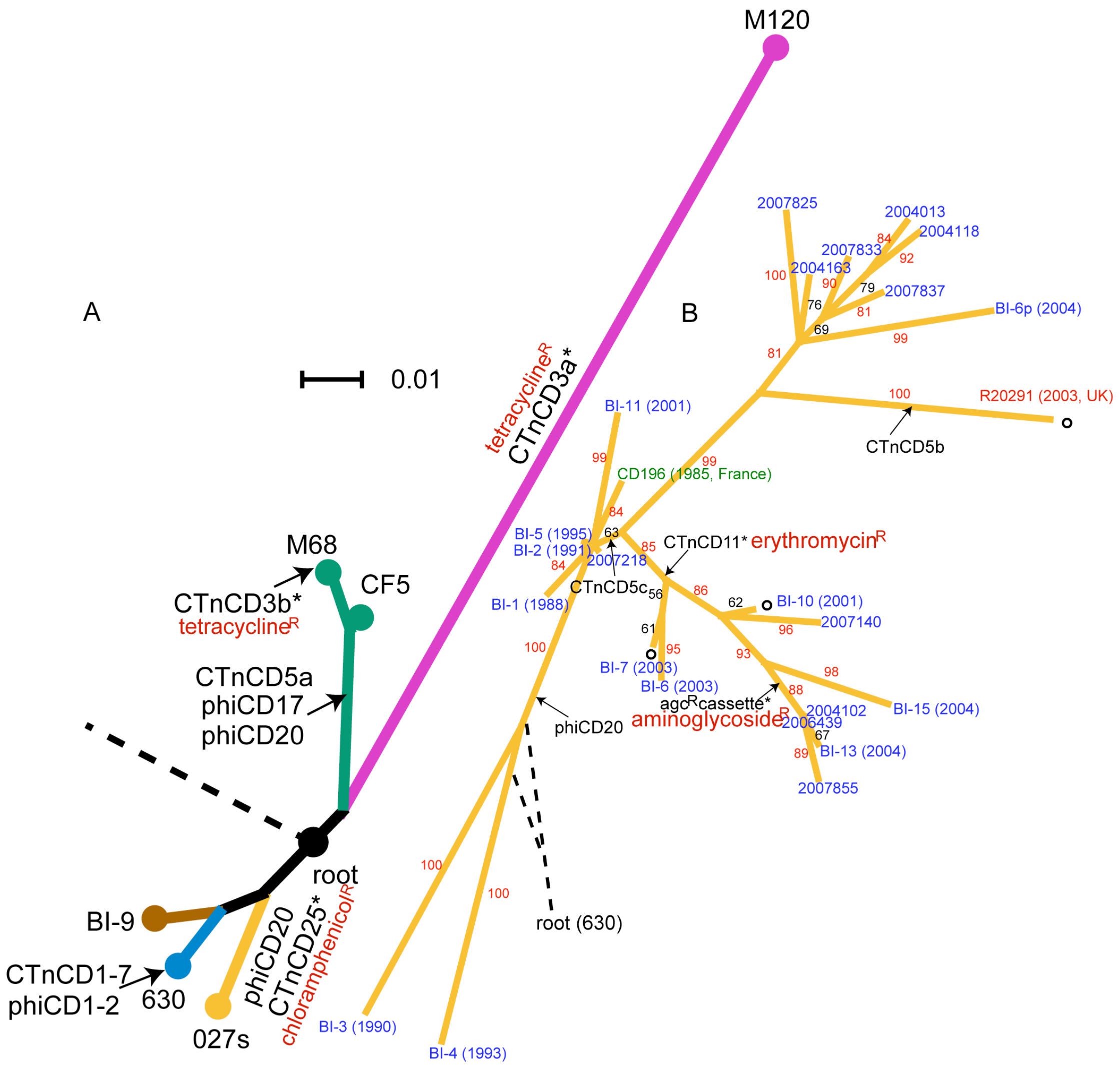


Figure 2.1: Phylogenetic trees of *C. difficile* based on whole-genome sequences. Arrows and unfilled circles denote insertion and deletion events, respectively. Genomic islands carrying drug resistance genes are shown with asterisks. (A) Deep-branching phylogeny that illustrates the relationships between 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.

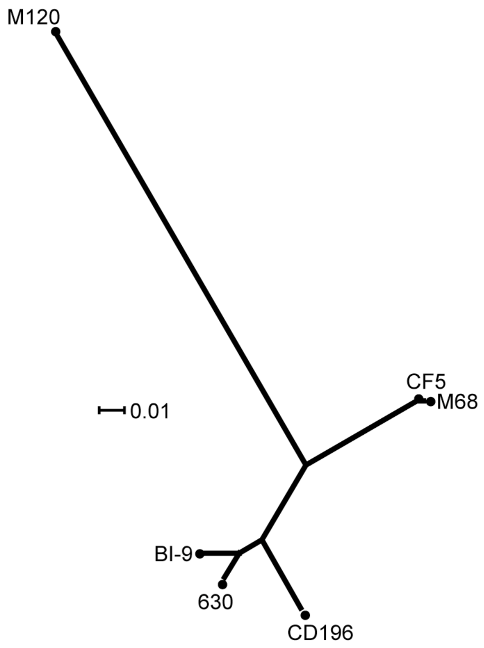


Figure 2.2: Phylogenetic tree of a diverse collection of *C. difficile* isolates based on concatenated non-recombining core CDSs. Scale bar indicates number of substitutions per site.

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 2.5×10^{-9} – 1.5×10^{-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.

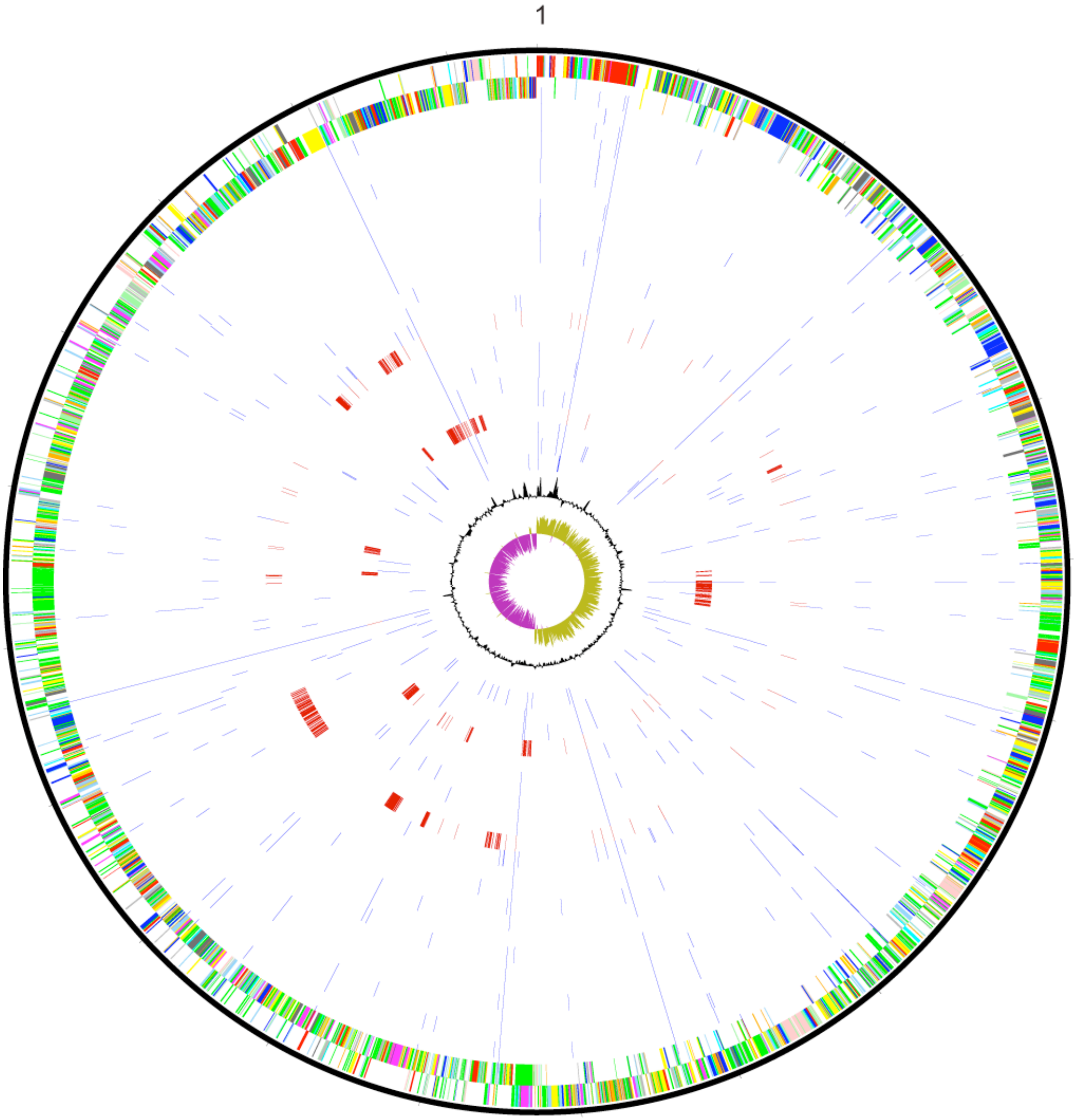


Figure 2.3: SNPs between CD196 and 24 other hypervirulent *C. difficile* isolates. Outer circle: CDSs of *C. difficile* CD196 genome, shown on a pair of concentric rings representing both coding strands; two inner circles: G+C% content plot and GC deviation plot (>0% olive, <0% purple); in between: SNPs (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.

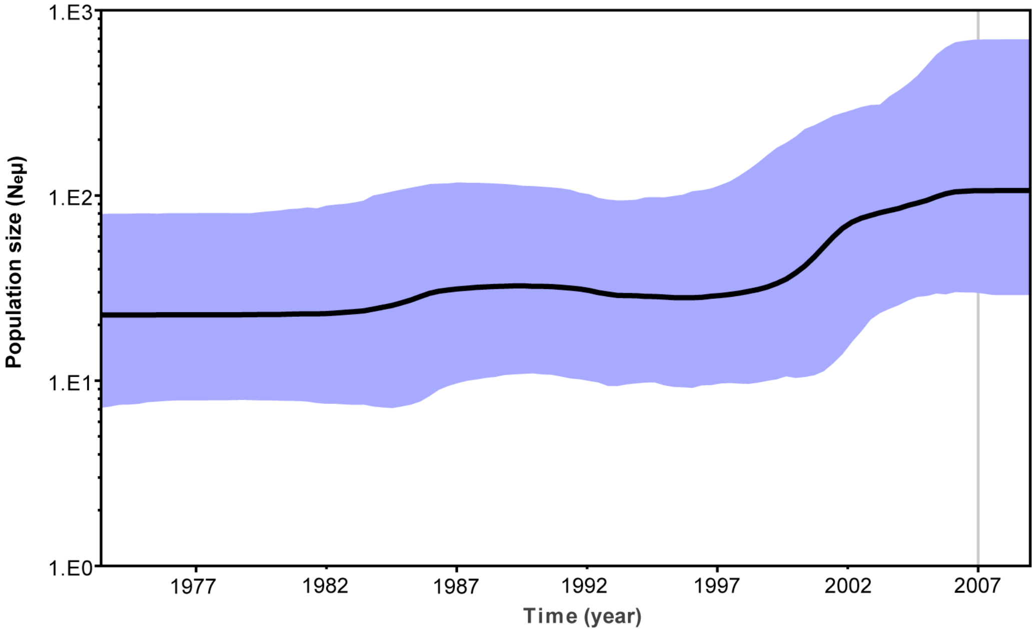


Figure 2.4: Bayesian skyline plot (group number = 10) shows a recent population expansion of the hypervirulent group. The x axis gives units of years, and the y axis is equal to N_eμ (product of effective population size and generation length in years). Thick solid line indicates median estimate, and 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).

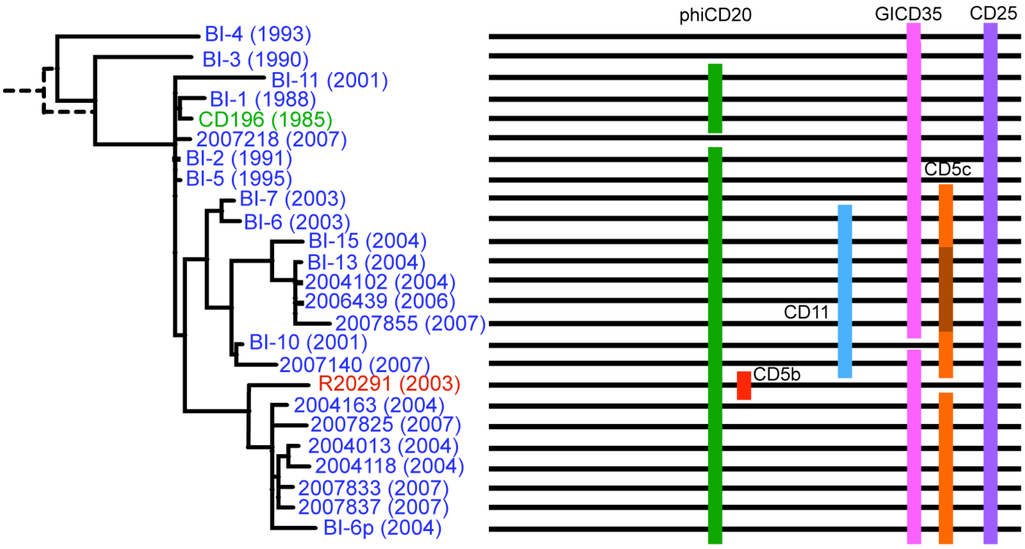


Figure 2.5: Genomic islands within isolates of the hypervirulent clade. Colour scheme of strain names is the same as in Figure 2.1 The presence of each genomic island is shown by a coloured box. Dark brown boxes denote copies of CTn*CD5c* with an aminoglycoside resistance cassette insertion.

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.

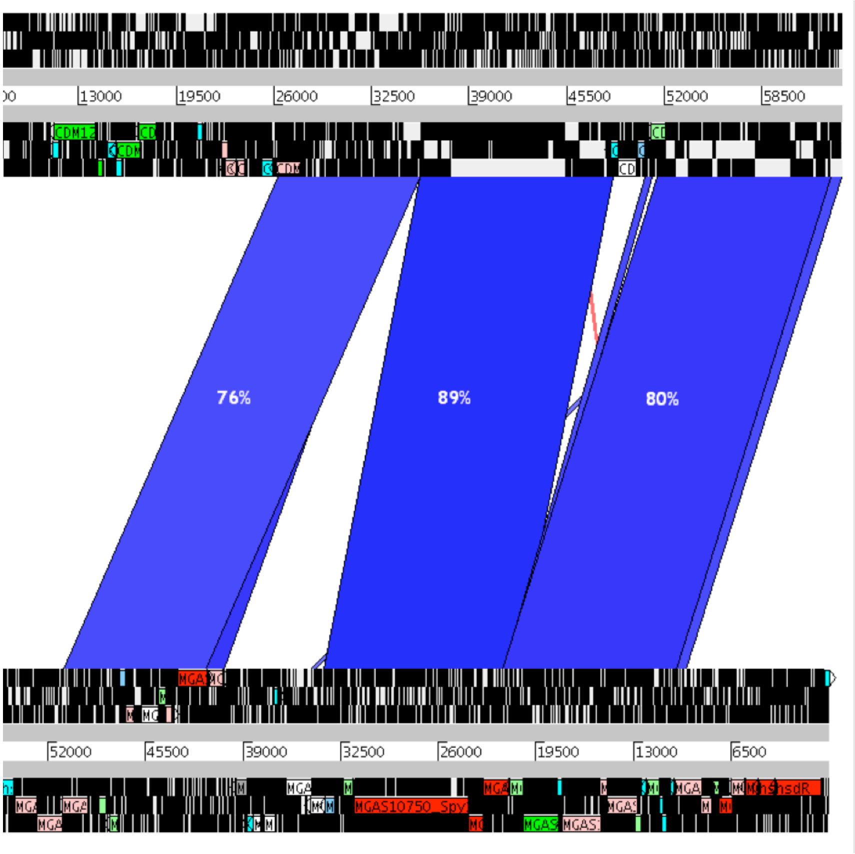


Figure 2.6: Comparison between parts of the *C. difficile* M120 genome (top) and the genome of *S. pyogenes* MGAS10750 (bottom). Each pair of black and white boxes represents both strands of a sequence. Coloured boxes present annotated CDSs; un-annotated parts of the sequence are left blank. Blue blocks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

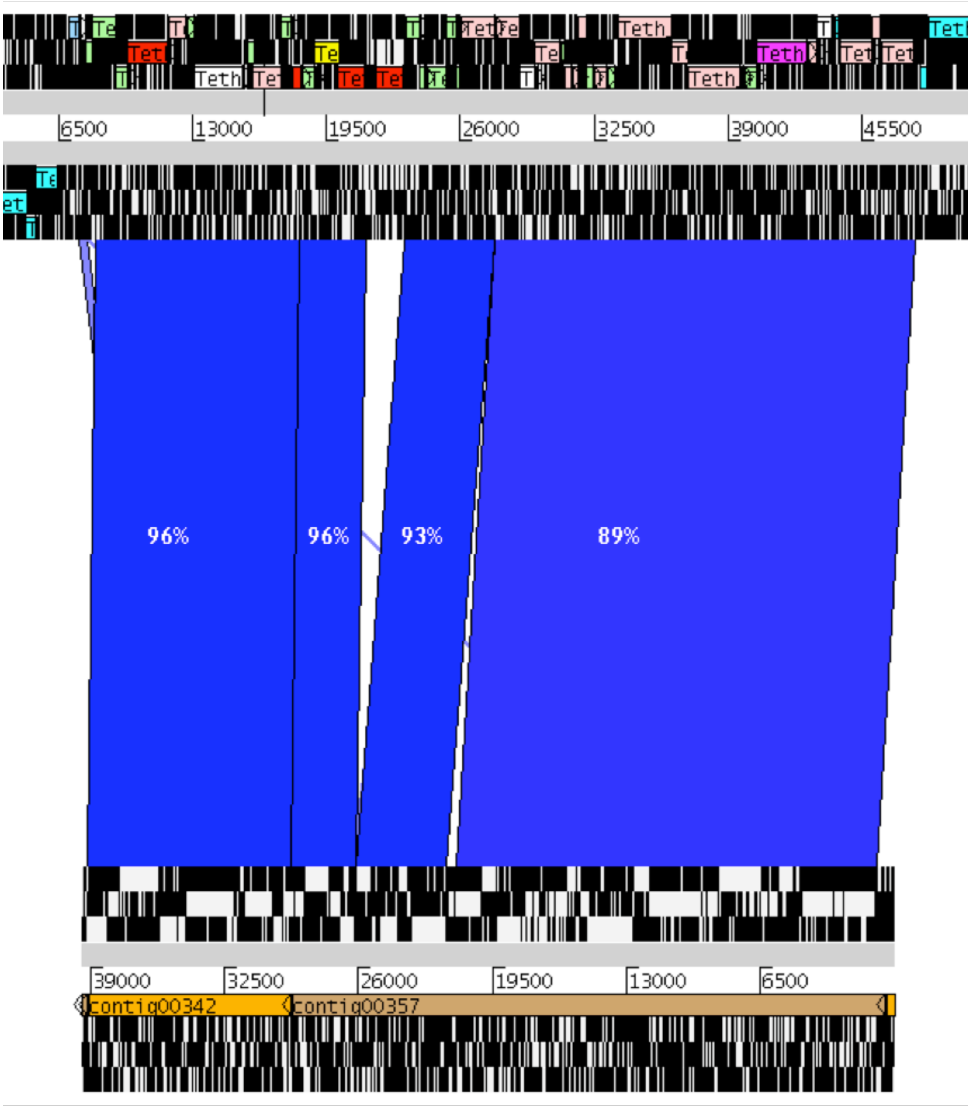


Figure 2.7: Comparison between parts of the *C. difficile* M120 genome (bottom) and the genome of *Thermoanaerobacter* sp.X514 (top). Other descriptions for this figure are the same as for Figure 2.6.

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.

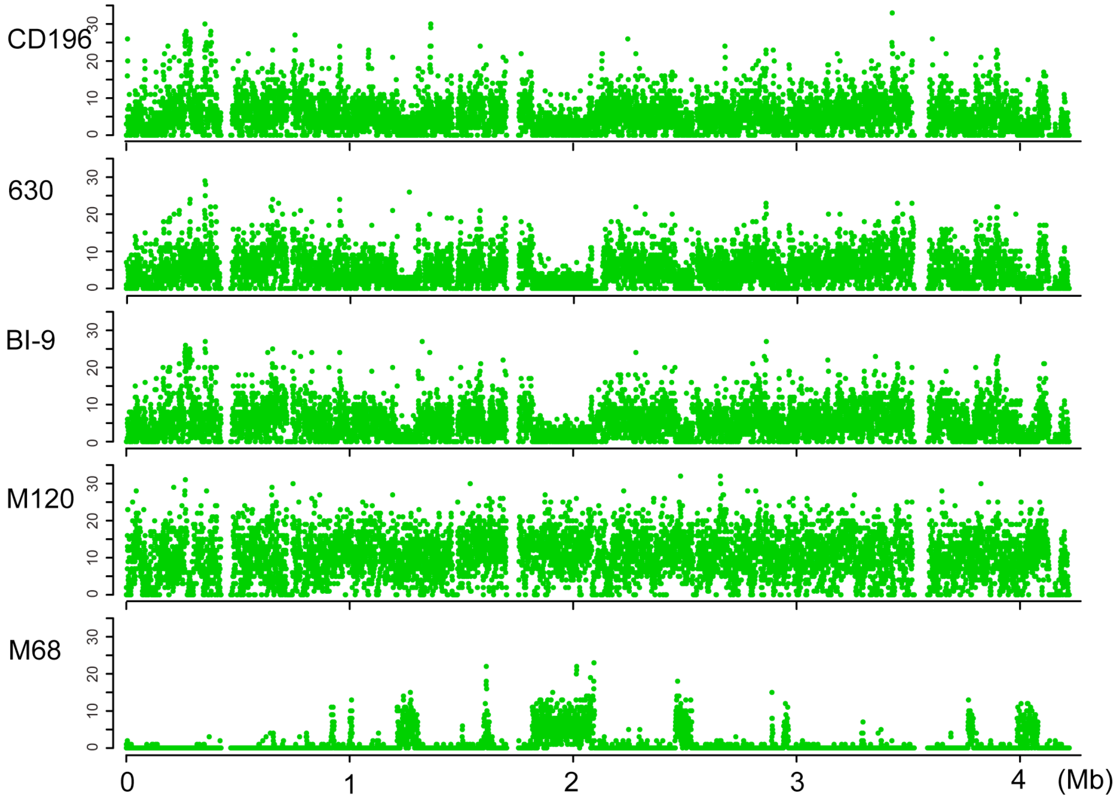


Figure 2.8: Signature of recombination in the deep-branching phylogeny. The genome-wide distribution of SNPs is shown for each strain against the core genome (excluding repetitive sequences) of strain CF5, which is indicated along the x axis. The y axis gives the number of SNPs in each 500-bp window.

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.

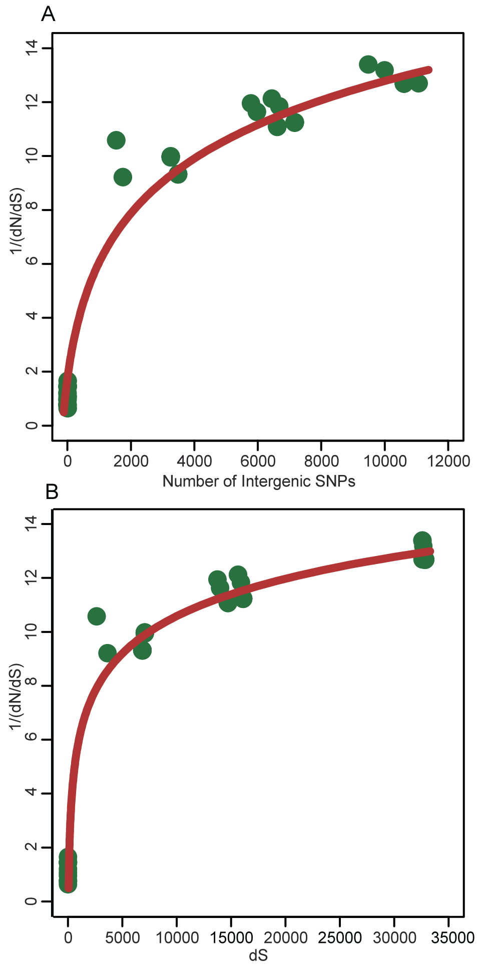


Figure 2.9: Trajectory of 1/(dN/dS) within the *C. difficile* phylogeny over time. The number of intergenic SNPs (A) and synonymous changes (B) serve as measures of time since divergence.

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).

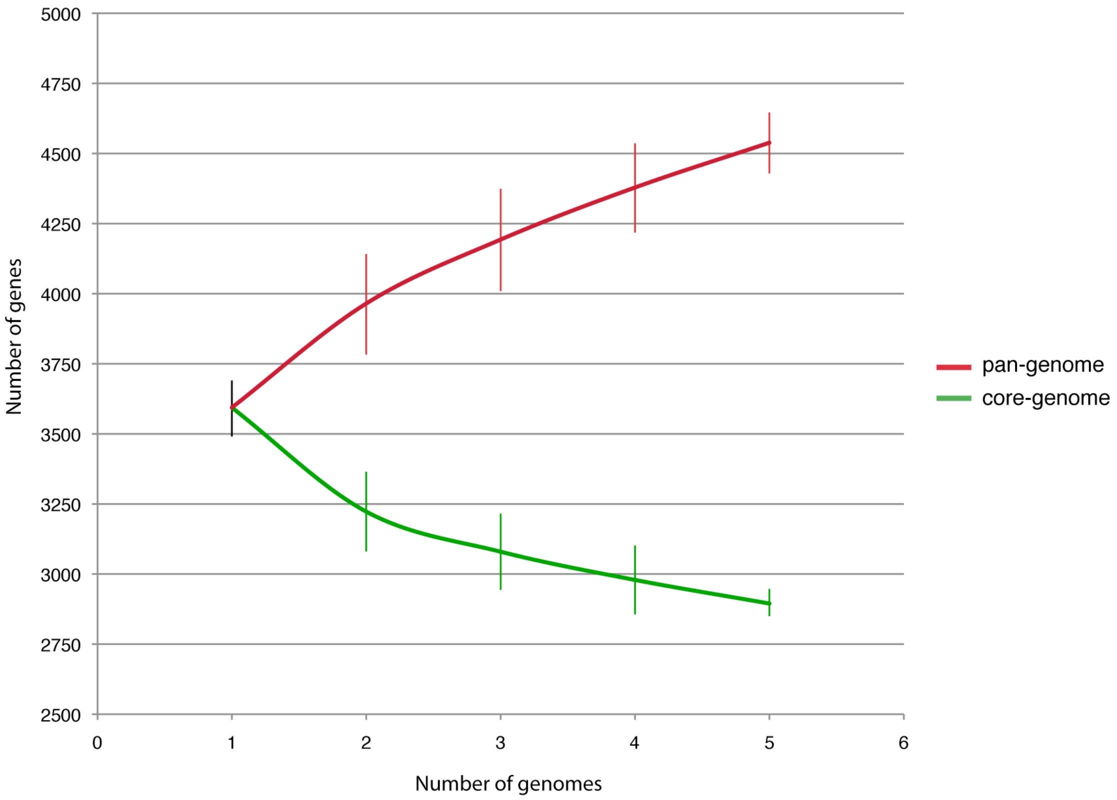


Figure 2.10: Changes in core- and pan-genome size of *C. difficile* in relation to the number of genomes. Red and green lines give mean estimates of numbers of genes in the pan- and core-genome, respectively. Coloured bars represent ranges of the estimates for all possible permutation of genome order.

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.

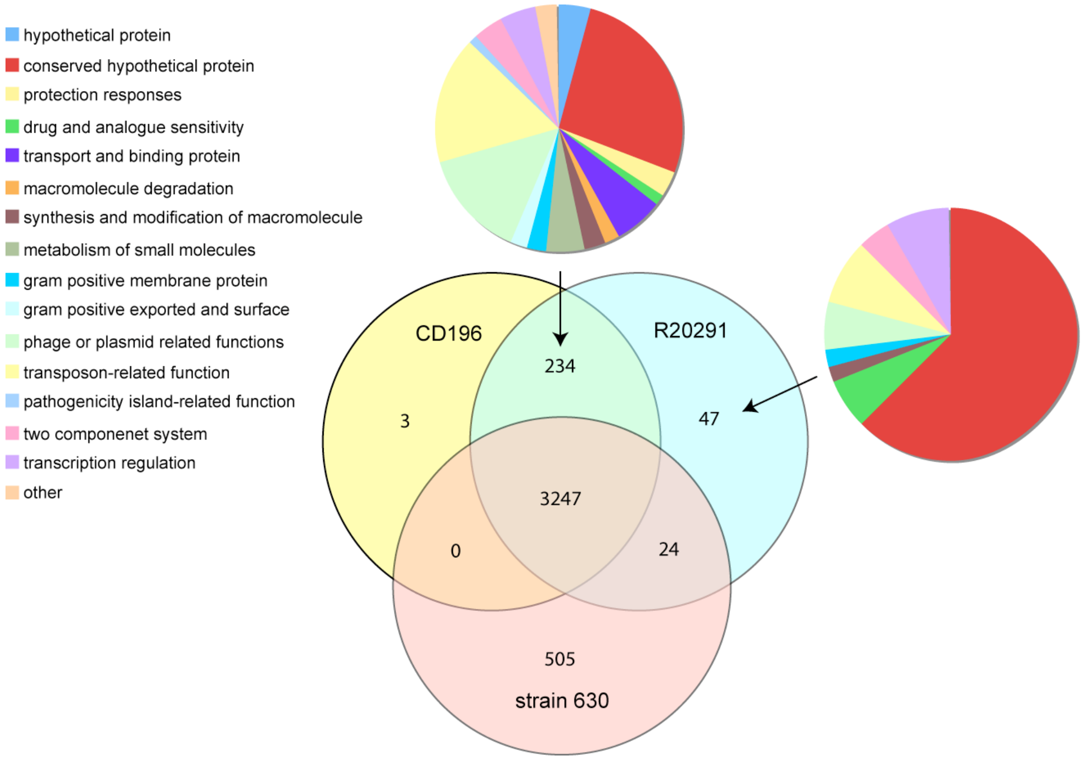


Figure 2.11: Distribution of orthologous CDSs in *C. difficile* strains 630, CD196 and R20291. The Venn diagram shows the number of genes unique, shared or core between the three strains. The associated pie charts show the breakdown of the functional categories assigned to these CDS.

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.

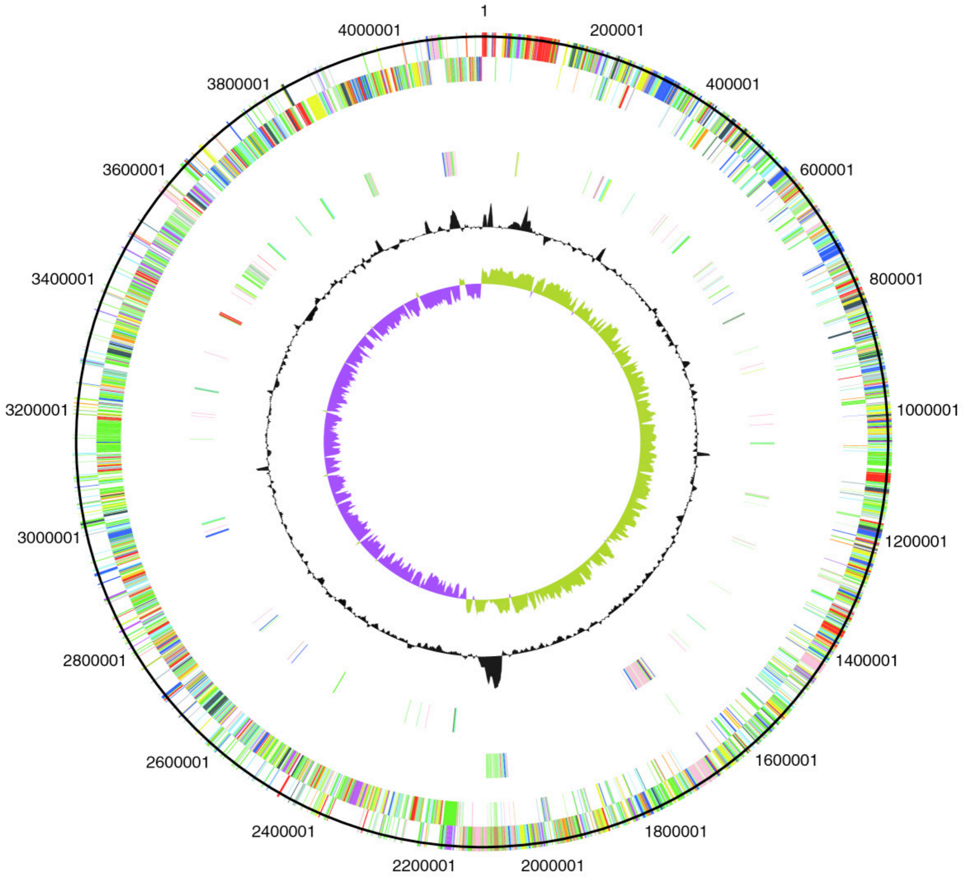


Figure 2.12: Circular representations of *C. difficile* chromosomes. From the outside (scale in bp): circles 1 and 2 show the position of R20291 CDS transcribed in a clockwise and anti-clockwise direction coloured according to predicted function; circle 3 shows CDS unique to R20291; circle 4 shows CDS 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.

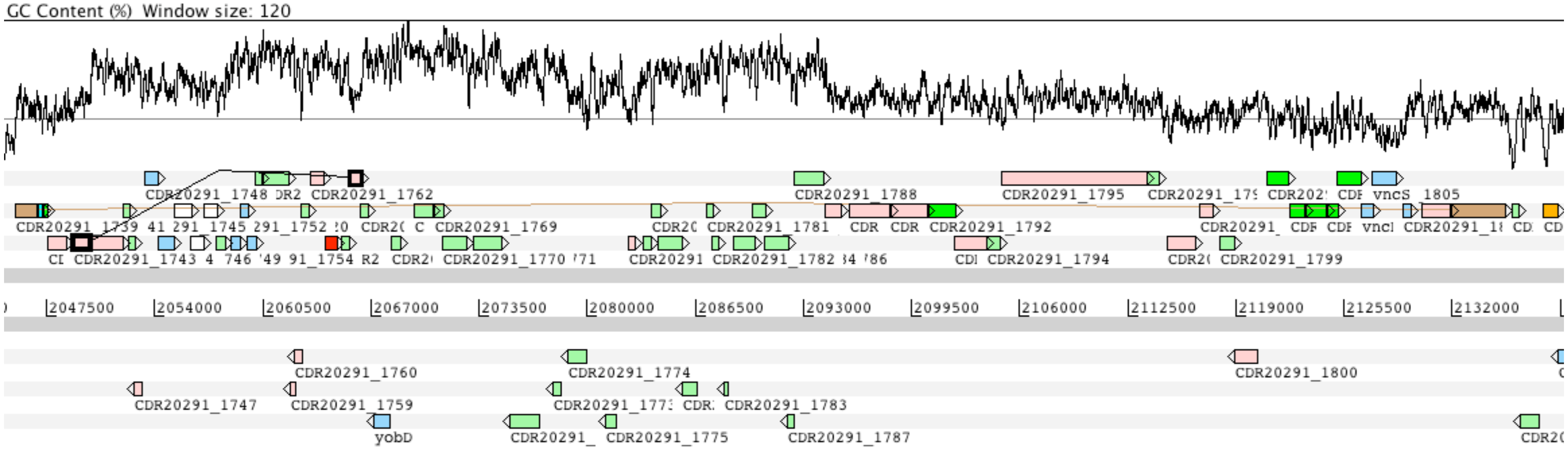


Figure 2.13: A conjugative transposon carried by R20291 but absent from CD196. The graph on top shows G+C content of the corresponding DNA sequence below. Two boxes connected by a line denote a CDS disrupted by the insertion of a genomic island. Colour coding for CDS functions: white, pathogenicity/adaptation; 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.

Chapter 3

Microevolution and global transmission of *C. difficile* BI/NAP1/027

3.1 Introduction

The biological, environmental and genetic characteristics that underlie the success of the *C. difficile* clade BI/NAP1/027 are not known. Prior to the emergence of significant outbreaks, BI/NAP1/027 *C. difficile* were present as a small percentage of cases in the United States and the UK, causing only sporadic infections (O'Connor *et al.*, 2009). However, in the last decade members of this ribotype have rapidly spread to all provinces in Canada, 40 States in the United States, and 16 countries in Europe (Kuijper *et al.*, 2008; OConnor *et al.*, 2009). BI/NAP1/027 *C. difficile* were essentially below detection levels in Canada in 2000, but they have accounted for 72.5% of these strains since 2003 (MacCannell *et al.*, 2006). As of 2008, BI/NAP1/027 *C. difficile* have been responsible for >40% of the CDI cases in UK hospitals (Brazier *et al.*, 2008). One notable phenotypic difference between the historical and modern isolates is the acquisition of resistance to the fluoroquinolone antibiotics in latter isolates (Loo *et al.*, 2005; McDonald *et al.*, 2005). In a surveillance study involving 411 isolates from European hospitals, 83 (20%) exhibited high resistance to moxifloxacin (Spigaglia *et al.*, 2008). Fluoroquinolone antibiotics work by targeting DNA gyrase in bacteria; therefore mutational changes in gyrase-associated genes can lead to resistance. Studies have identified one common amino acid change in *gyrA* (Thr82Ile) and four substitutions in *gyrB* (Ser416Ala, Asp426Asn, Asp426Val and Arg447Lys) in moxifloxacin-resistant *C. difficile* isolates, Thr82Ile and Ser416Ala being the most common ones; however Ser416Ala does not appear to be specifically associated with fluoroquinolone resistance, since it was also detected among susceptible isolates (Spigaglia *et al.*, 2008). The antibiotics moxifloxacin and levofloxacin have been shown to select for fluoroquinolone resistance mediated by changes in *gyrA* and *gyrB in vitro* (Spigaglia *et al.*, 2009).

Fluoroquinolone-resistant BI/NAP1/027 *C. difficile* variants are now prevalent in many European countries (Bauer *et al.*, 2011) and they have recently emerged in South Korea (Kim *et al.*, 2011) and Australia, where they have caused several outbreaks (Richards *et al.*, 2011). However, the details and pathways behind this global spread are unknown. It is also unclear whether the emergence of fluoroquinolone-resistance occurred once or several times independently. Interestingly, resistance to rifaximin and rifampicin has also been observed in these isolates indicating significant selection mediated by antibiotic usage. In a recent study, O'Connor *et al.* found 17.5% of the clinical isolates are resistant to both drugs; and BI/NAP1/027 *C. difficile* make up 64.3% of such resistant strains (O'Connor *et al.*, 2008). Further analysis revealed seven distinct amino acid substitutions in the *rpoB* gene of independent isolates, which encodes RNA polymerase β subunit, and is targeted by rifampicin, suggesting that these changes were independently derived rather than from a clonal expansion (O'Connor *et al.*, 2008) Phylogenetic analysis of the BI/NAP1/027 *C. difficile* lineage is required to confirm and expand on these observations.

Aside from derived antimicrobial resistance, studies on BI/NAP1/027 *C. difficile* have also focused on spore and toxin production. One study showed that BI/NAP1/027 *C. difficile* are capable of producing elevated levels of toxins A and B (Warny *et al.*, 2005), whereas other researchers found no significant difference in toxin production (Akerlund *et al.*, 2008; Merrigan *et al.*, 2010). BI/NAP1/027 *C. difficile* can also produce a novel binary toxin and can harbour an 18-bp deletion in the *tcdC* gene, which encodes a negative regulator of toxins A and B (Loo *et al.*, 2005; McDonald *et al.*, 2005).

However, according to reports the increased toxin production reported in BI/NAP1/027 isolates is not directly linked to the 18-bp deletion but to a single base deletion in *tcdC* (Matamouros *et al.*, 2007). Another study showed that epidemic BI/NAP1/027 isolates produce more spores than some other *C. difficile*, a property which may underlie an increased ability to transmit (Merrigan *et al.*, 2010). However, a different study argues that sporulation rate of BI/NAP1/027 is not significantly higher, and suggests that such variability is not associated with strain type (Burns *et al.*, 2010). The relationship between antimicrobial drugs and sporulation was also studied, indicating that fluoroquinolones trigger high levels of toxin production and spore germination in BI/NAP1/027 isolates (Saxton *et al.*, 2009).

This chapter presents the details of a comprehensive whole-genome sequence analysis of *C. difficile* isolates belonging or similar to ribotype 027 based on Illumina short read data. Whole-genome analysis is necessary to provide sufficient resolution to study the BI/NAP1/027 lineage in a discriminatory manner, as these isolates share a highly similar genomic backbone. During the analysis, the accuracy of the analysis methods employed for this short read data was assessed. Various parameters were tested, and an optimal 'cutoff' in SNP calling was chosen for the actual study. The aims of this study were:-

- to gain insights into the details of the global spread of this lineage;
- to examine recently derived variants in this lineage and their functional consequences; and
- to investigate the reasons behind the emergence.

3.2 Materials and Methods

3.2.1 Bacterial isolates

A total of 339 *C. difficile* isolates spanning 1985-2010 were included in the study. The isolates were previously genotyped as either PCR-ribotype 027, REA type BI or PFGE type NAP1. The typing was performed in individual contributing laboratories. In order to assess the coherence of data accumulated over time, the hypervirulent isolates used in the study described in chapter 2 were also included. Three *C. difficile* ribotype 176 isolates (lon004, lon005, lon006) were also included since they were suspected of being highly-related to the *C. difficile* 027/BI/NAP1 genotype (Nyc *et al.*, 2011). The isolates were obtained from the United States (45), Canada (13), Australia (6), Singapore (3), Korea (6), France (1), and the UK (261). The isolates are predominantly from hospital patients infected with *C. difficile* except for 8 United States isolates, which were from farm animals and food sources. This collection includes historic strains (isolated before 2000) from France (1985) (Stabler *et al.*, 2009), the USA (1988, 1990, 1991, 1993, 1995) (McDonald *et al.*, 2005) and the UK (1998), isolates from notable outbreaks in Montreal, Canada (2003) (Loo *et al.*, 2005), North Eastern United States (2001-2003) (McDonald *et al.*, 2005), London, UK (2005) (Stabler *et al.*, 2009), Melbourne, Australia (2010) (Richards *et al.*, 2011), and modern disease-causing isolates from Korea (Kim *et al.*, 2011). This collection also included a focused sampling of 111 *C. difficile* 027 isolates at a single hospital (Royal Liverpool Hospital) in the UK. The Liverpool isolates were collected between July 2008 - May 2010. Information on the isolates used in this study is summarized in Appendix A.

3.2.2 Sequencing, mapping, and SNP detection

Sample DNA was prepared and sequenced on the Illumina GAII platform according to protocols described in Harris *et al.* (Harris *et al.*, 2010). Paired-end multiplex libraries were created with a 200 bp insertion size. The read length was 54 bp for samples Liv1-Liv21, 108 bp for samples Gla001-Gla022, and 76 bp for the rest. All isolates were sequenced to a minimum coverage of 9-fold, with an average coverage of 110-fold across all isolates. New sequencing data of more satisfactory coverage were generated for five isolates in the study described in chapter 2 (BI-1, BI-2, BI-3, BI-7, and BI-10). These five isolates were assigned slightly different names as BI-1a, BI-2a, BI-

3a, BI-7_L22 and BI-10a respectively to differentiate them from the earlier sequencing data.

Sequencing reads were aligned with BWA (Li and Durbin, 2009) against the genome sequence of the ribotype 027 reference strain R20291. SNPs were identified with SAMtools (Li *et al.*, 2009). A coverage cut off of >5-fold and < three times the average coverage was set for each individual isolate during SNP detection. Repetitive regions in the reference genome sequence were characterized by using REPuter (Kurtz *et al.*, 2001) and the repeat finder functions in the MUMmer package (Kurtz *et al.*, 2004). The boundaries of repetitive regions were extended to include the mobile elements in R20291. SNPs falling within these repetitive regions were excluded. To confirm the alleles for each variant position, SNPs were checked at each position in all sequencing reads in all isolates. An allele is only considered to be valid if supported by all reads (if 5<coverage<=40) or >92.5% of the total reads (if coverage >40) covering the position; otherwise it was treated as missing data. SNPs within 2,000 bp of each other were considered as potentially affected by recombination and excluded from phylogeny construction.

3.2.3 Assessing accuracy of SNP detection

A simulation approach was used to assess the accuracy of short read mapping and variant detection. A pseudo sequence was made by introducing artificial variants (single base substitutions, insertions and deletions) into the genome of R20291. The software INDELible (Fletcher and Yang, 2009) was used for this purpose. Default options were implemented, including a JC model (Jukes and Cantor, 1969) and insertion and deletion rates both of 0.1 relative to substitution rate. Two pseudo-genomes (named “scale 1” and “scale 2” for simplicity) with different levels of divergence were created. Scale 1 genome differs from R20291 by 74 SNPs; scale 2 genome differs from R20291 by 869 SNPs. Paired-end Illumina reads generated for R20291 were aligned to pseudo references. This step was performed multiple times, each time with data of a different coverage. The range of the tested data coverage is 8.5-fold to 100-fold. For variant detection, four different SNP filtering and validation measures were tested: (I) use default settings in BWA (Li and Durbin, 2009) and specify a coverage cut off of >5-fold and < three times the average coverage; (II) by excluding SNPs within repetitive regions following the measure stated in (I); (III) validate SNP alleles by checking at all variant positions in all sequencing reads following the measures stated in (II) and only consider a SNP allele true if it is supported by all sequencing reads; and (IV) validate SNP alleles by checking at all variant positions in all sequencing reads following the measures stated in (II) and only consider a SNP allele true if a) it is supported by all sequencing reads and b) the depth for this position is no less than 40, or a) if it is supported by > 92.5% of the reads and b) the depth for this position is >40. The numbers of false positives and false negatives were calculated in each case. A flow chart of the whole process is given in Figure 3.1.

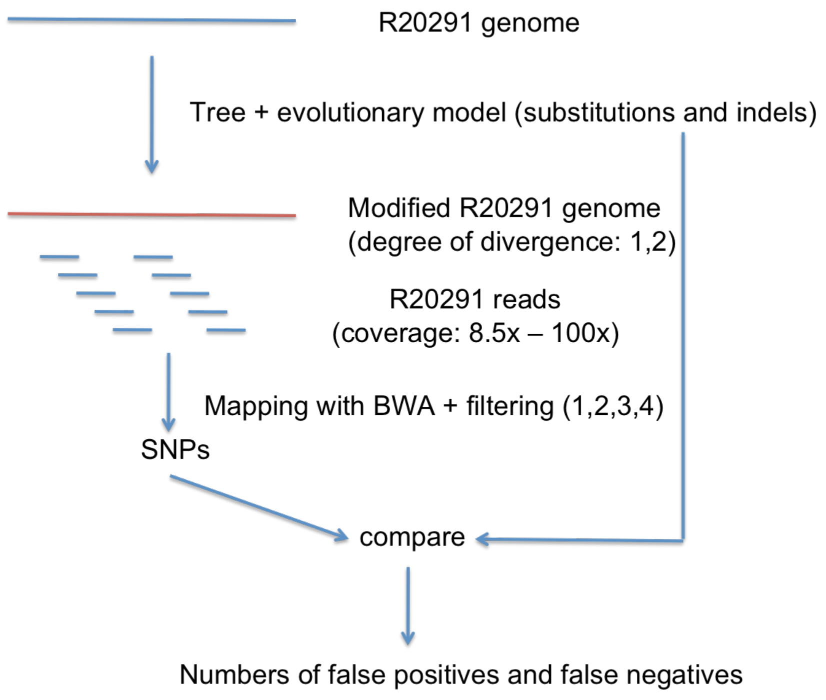

R20291 genome
Tree + evolutionary model (substitutions and indels)
Modified R20291 genome
(degree of divergence: 1,2)
R20291 reads
(coverage: 8.5x – 100x)
Mapping with BWA + filtering (1,2,3,4)
SNPs
compare
Numbers of false positives and false negatives

Figure 3.1: Methods for assessing accuracy in short reads mapping and SNP detection.

3.2.4 Phylogenetic analysis

An appropriate evolutionary model (simple GTR) was determined using jModelTest 0.1.1 (Posada, 2008). Phylogenetic relationships were inferred with three methods: 1), split-decomposition and neighbour-net methods in SplitsTree4 (Huson and Bryant, 2006); 2), the program PHYML (Guindon and Gascuel, 2003); and 3), the program BEAST (Drummond and Rambaut, 2007). In the first two cases, a simple GTR model was used. The model assumes all sites evolve at the same rate, with no invariable site. Neighbour-joining trees were also constructed with PHYML, and the results were compared.

In the BEAST analysis, two clock models (relaxed lognormal and relaxed exponential) and two datasets were tested initially (see 3.2.6 for details of both datasets). The relaxed exponential clock model was determined as more suitable based on Bayes Factor calculations (Suchard *et al.*, 2001) and was used for later BEAST runs. The program was specified to estimate tMRCA (time to the most recent common ancestor) of taxon groupings. All other parameters were set to default. These analyses were carried out with a chain length of 200,000,000 states, and re-sampling every 10,000 states. The phylogeograpic history was also inferred with BEAST using a Bayesian method as described in (Lemey *et al.*, 2009).

3.2.5 Core and accessory genome

For each isolate, the unaligned sequencing reads were assembled using Velvet (Zerbino and Birney, 2008). To assess whether the resulting contigs were unique, each contig with a length >1kb was searched using BLASTN against the current pan-genome, which was made by concatenating the draft genome sequence of M7404 and already determined unique contigs. Any unique contigs were added to the pan-genome. If the match was of >80% identity and covered >40% of the contig length, this contig was not considered to be unique, and was not added to the current pan-genome. The resulting unique contigs were individually searched against the NCBI bacteria genome database to check for contamination. The filtered set of unique contigs were added to the genome sequence of M7404 to create a pan-genome. Finally, the sequencing reads from each strain were aligned against the constructed pan-genome to assess for the presence and absence of genomic regions in each isolate.

3.2.6 Identification of homoplasic characters and homologous recombination

Homoplasic SNPs were identified by examining the SNP allele pattern across all isolates in relation to the phylogenetic tree. A SNP was considered homoplasic if the allele pattern did not agree with the tree topology. Genomic regions affected by homologous recombination were identified by a) clusters of SNPs within 2,000 bp windows and b) the iterative method to eliminate recombination sites described in (Croucher *et al.*, 2011). The identified homologous recombination blocks were excluded from phylogenetic and population genetic analysis. These methods result in two datasets of 604 SNPs and 852 SNPs respectively.

3.2.7 Population history and mutation rate

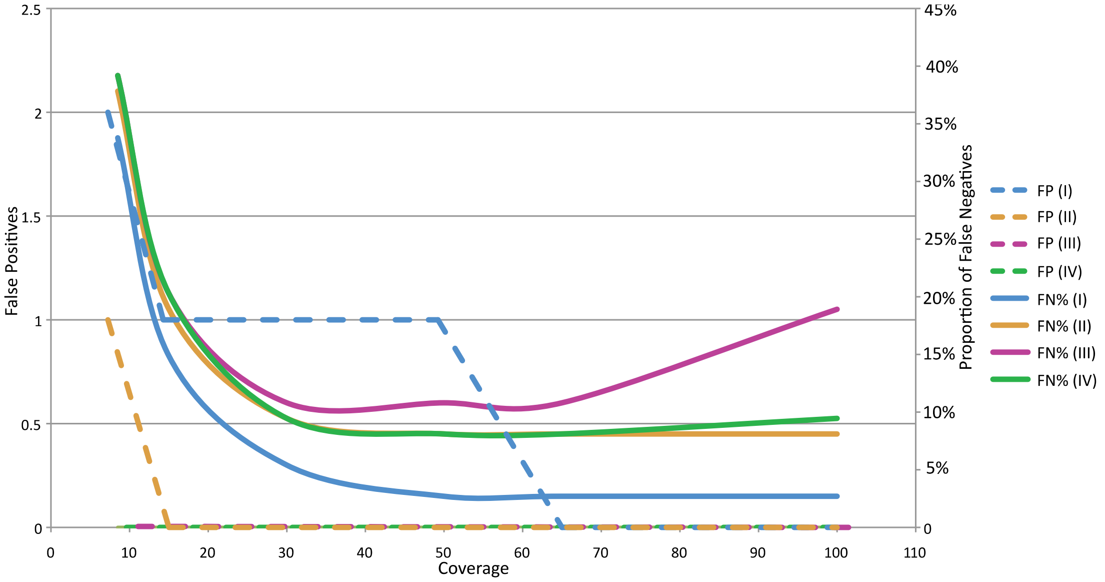
The apparent mutation rate was estimated using two methods: 1), A full maximum likelihood model assuming a rapid expansion which results in perfect star genealogies, implemented in an R script (Morelli *et al.*, 2010); and 2), the program BEAST (Drummond and Rambaut, 2007). BEAST analyses were carried out as stated in 3.2.4. A final mutation rate was determined by combining median estimates from both methods. Bayesian skyline plot analysis were also performed using BEAST by specifying the skyline population model (Drummond *et al.*, 2005).

3.3 Results

3.3.1 Assessment of SNP detection method

Illumina reads of R20291 were aligned to two pseudo genome sequences to identify SNPs. The accuracy of SNP detection was assessed. The numbers of false positive and false negative SNPs are influenced by sequencing data coverage and the measure used for SNP filtering and validation, as shown in Figure 3.2. The number of false positive SNPs decreases as data coverage increases, and a minimum of 15-fold coverage is necessary to achieve a result of no false positive SNPs using measures (II), (III) and (IV). The proportion of false negative SNPs also decreases as data coverage increases, except for SNP validation measure (III). The overly stringent validation criterion of (III) rejects more SNPs when data coverage is higher, as this method only considers a SNP allele correct if it is supported by all sequencing reads. A significant number of true SNPs were therefore missed due to a few sequencing errors in abundant reads covering the variant position, despite the majority of the reads indicating the correct allele. Method IV is an improvement with respect to this situation, as shown by a false negative rate comparable to method II, which does not include a SNP validation step. After comparing four SNP validation methods, method IV was selected for analyzing the actual sequencing data of BI/NAP1/027 *C. difficile* isolates. This method allows for no false positive SNPs and a false negative rate of 7%-10% when data coverage is above 30-fold, depending on the similarity between subject sequence and reference sequence.

Scale 1 (74 SNPs)



Scale 2 (869 SNPs)

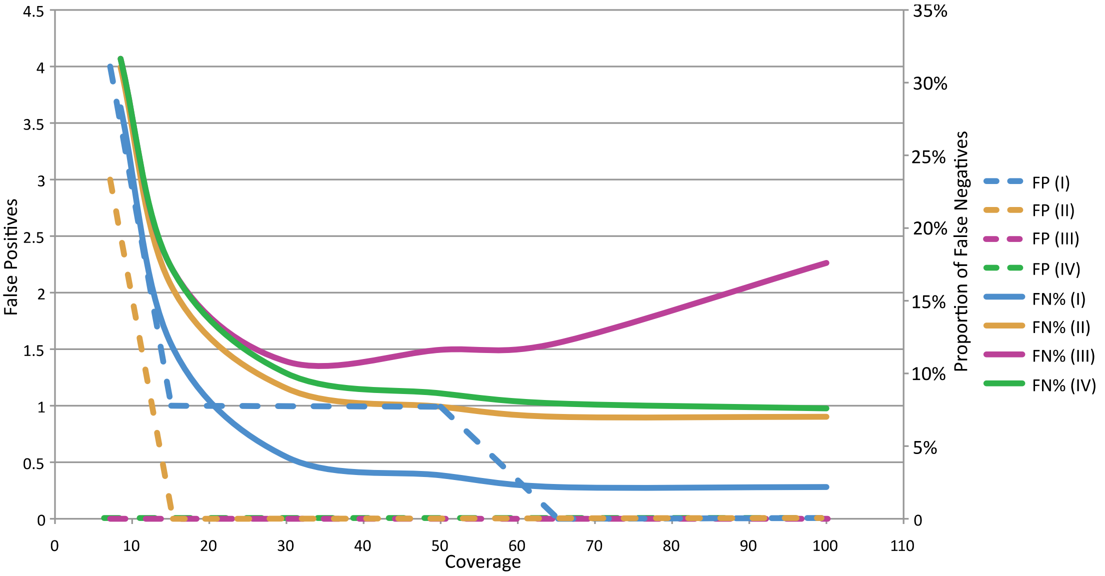


Figure 3.2: Numbers of false positive SNPs (left axis) and percentage of false negative SNPs (right axis) in relation to sequencing data coverage (x-axis) and different SNP filtering and validation measures (coloured lines). Dashed lines represent the absolute numbers of false positive SNPs; solid lines represent the proportions of false negative SNPs. Scale 1 (top) and scale 2 (bottom) indicate two scenarios with different level of divergence, as shown by the numbers of SNPs in brackets. SNP filtering and validation measures I – IV correspond to what stated in 3.2.3.

3.3.2 Phylogenetic relationship

3.3.2.1 Maximum-likelihood phylogeny and phylogenetic networks

Using the method identified in 3.3.1, a total of 3,580 SNPs were discovered from the global collection of isolates of the *C. difficile* BI/NAP1/027 lineage. However, 2,943 (83.3%) of these SNPs are clustered and private to 8 individual strains (kor001, BI-3, BI-4, BI-10, BI-11, Can001, Can007, lon004), which suggests that these are due to recombination events involving the acquisition of DNA from donors outside of the 027 lineage. We therefore removed these from downstream analysis. These regions will be discussed in more detail in section 3.3.4. After this analysis, a total of 604 SNPs remained from the core-genome. A maximum likelihood phylogeny based on these variable positions was constructed using a simple GTR model (Figure 3.3). The root of this phylogeny was determined by incorporating 630 and CF5, two *C. difficile* isolates outside the BI/NAP1/027 lineage. The long branches leading to isolates BI-3 and BI-4 could be a consequence of potential recombination sites remaining in the dataset. The dataset was also analyzed using an iterative method (Croucher *et al.*, 2011) aiming at identifying sites affected by homologous recombination, but the long branches of BI-3 and BI-4 remained. The topology of the entire phylogeny is also supported by the neighbour-joining algorithm. Network analysis was carried out with split-decomposition and neighbour-net algorithms (Huson and Bryant, 2006), and the results confirmed that the dataset is very much tree-like. In particular, the split-decomposition analysis resulted in a reproducible tree. The fit values for split-decomposition and neighbour networks were 72.35 and 98.975 respectively, indicating the latter is more suitable for the dataset. NeighbourNet network is shown in Figure 3.4. The four main clades in the maximum likelihood tree are supported by both networks.

Overall, the SNP phylogeny can discriminate between >100 distinct genotypes within the *C. difficile* BI/NAP1/027 collection that are clustered into several clades. The global location, where each *C. difficile* BI/NAP1/027 was isolated, is indicated with colour in Figure 3.3 and this demonstrates a general lack of geographical clustering at the global and UK levels (clades 2 and 3) and strong, but incomplete, clustering at the hospital level (clade 4).

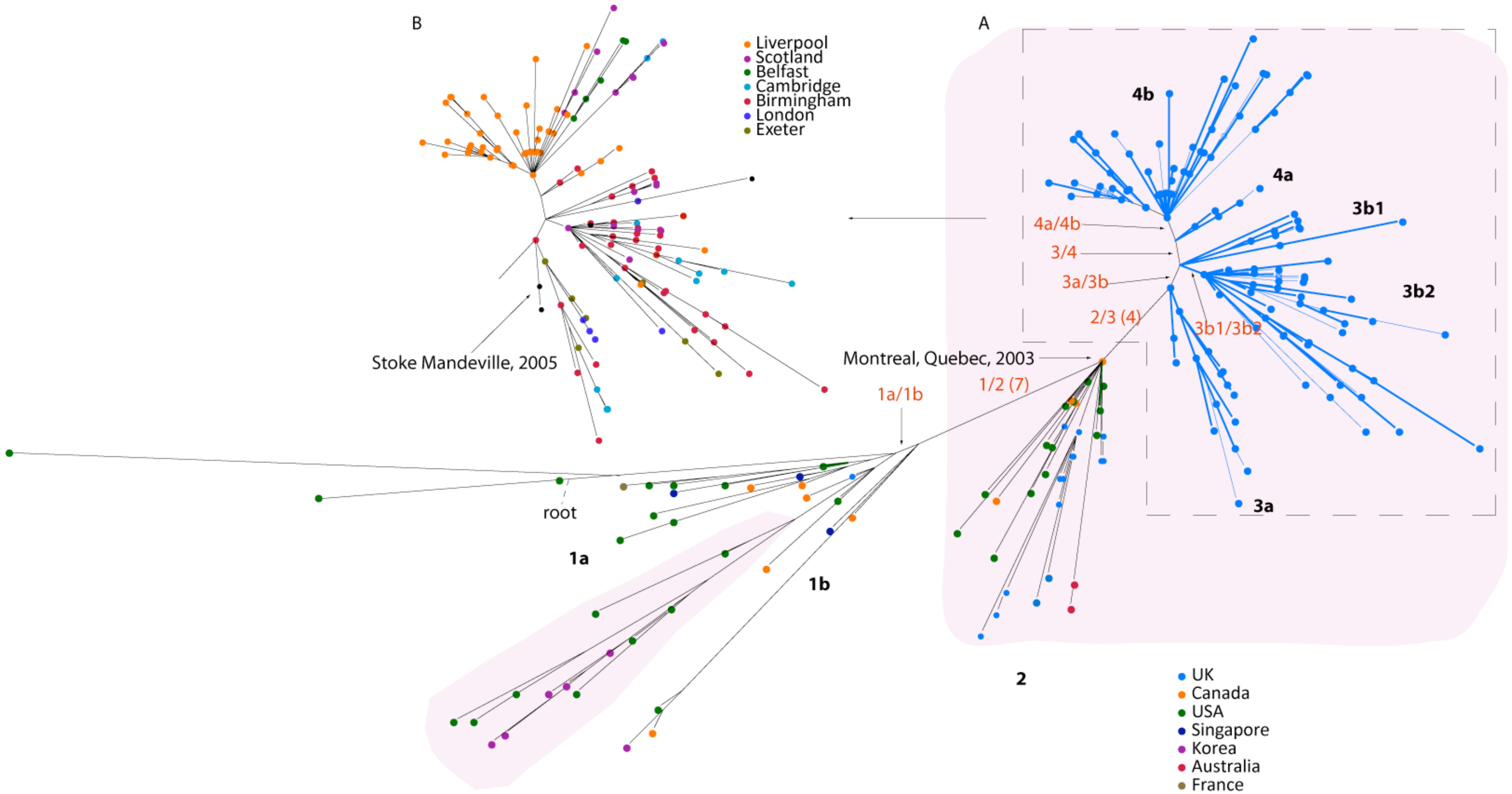


Figure 3.3: Global phylogeny of *C. difficile* BI/NAP1/027. A). Phylogenetic tree of a global collection of 339 isolates based on core genome SNPs. Nodes are coloured according to origin of isolate. Letters in bold denote the clade names (1a, 1b, 2, 3a, 3b1, 3b2, 4a and 4b). Orange text indicates SNPs that differentiate between clades, with SNP numbers given in brackets (1 SNP if unlabeled). Lineages harbouring the *gyrA* mutation are shown by pink shaded areas. B). Expansion of the part of UK isolates circled by a dashed line in A), with nodes coloured according to origins of isolates. Isolates from Inverness, Dundee, Dumbarton, Glasgow, Edinburgh, Ayrshire, and Dumfries are collected noted as from Scotland.

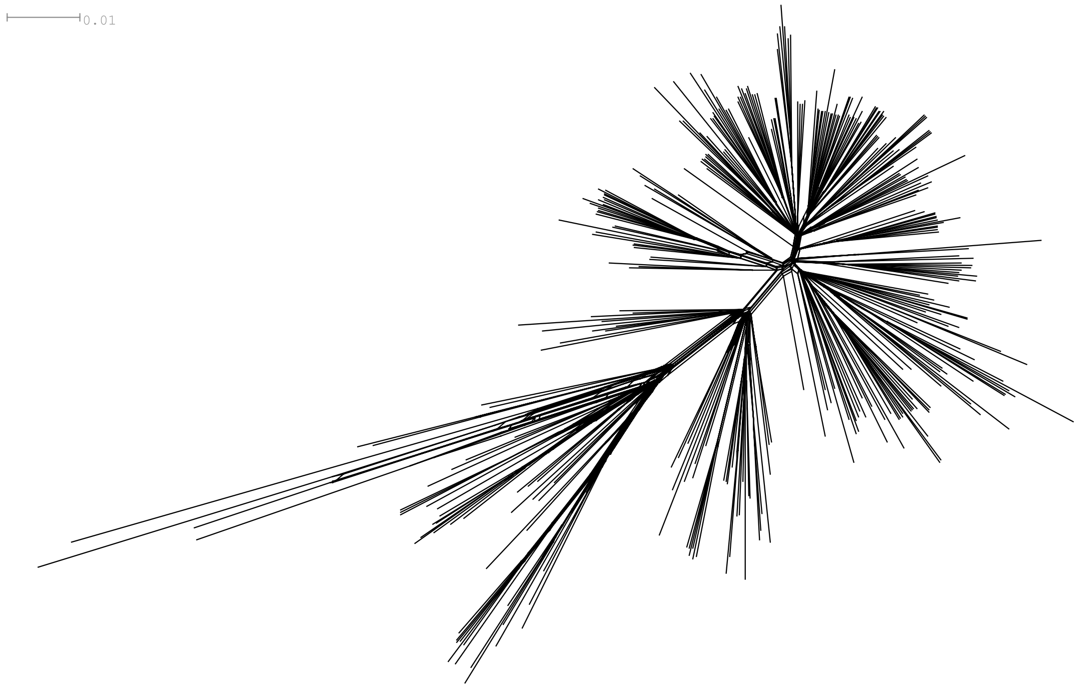


Figure 3.4. NeighbourNet network of the *C. difficile* BI/NAP1/027 lineage. Label indicates genetic distance per site.

All historical, farm animal and food isolates are located within clade 1 (Figure 3.5). In particular, multiple isolates from food sources in Arizona were derived from a historical Arizona human isolate (BI-2 from Tucson, 1991). This analysis suggests transmission of *C. difficile* through the food chain, although more data would be needed to confirm this. Also within clade 1 is a distinct lineage that contains epidemic isolates associated with healthcare outbreaks in the USA (McDonald *et al.*, 2005) and sporadic infections in South Korea (irregular shaded areas in Figures 3.3 and 3.5). This lineage carries the mutation (Thr82Ile) in DNA gyrase subunit A (*gyrA*), which has been shown to result in an increased level of fluoroquinolone resistance (Spigaglia *et al.*, 2010).

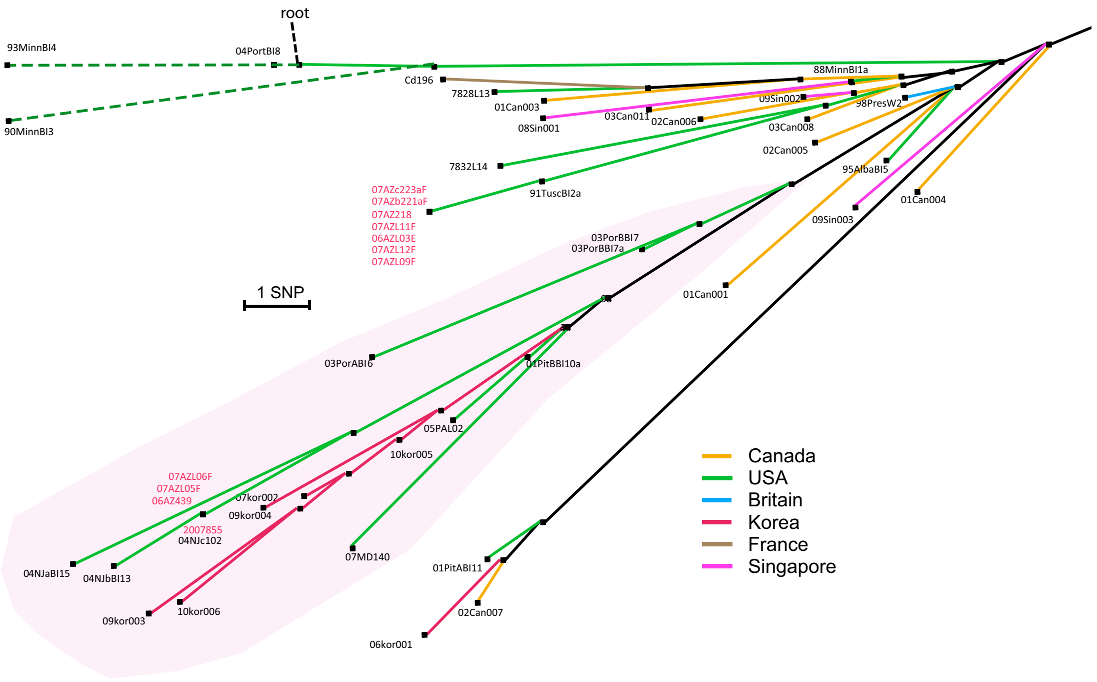


Figure 3.5: Expansion of clade 1 from Figure 3.3. Branches are coloured according to sampling locations of the isolates the branches directly lead to. Isolate names are shown in black (human isolates) or red (isolates from animals and food sources). The lineage harbouring the *gyrA* mutation is shown by a pink shaded area. The branches leading to BI-3 and BI-4 are shown with dashed lines as these branch lengths were artificially shortened to fit the graph.

The earliest isolate in this lineage was from Pittsburgh, Pennsylvania in 2001, consistent with the fact that the initial report of an increase in the incidence of 027-associated *C. difficile* infections came from Pittsburgh (Dallal *et al.*, 2002; Muto *et al.*, 2005). Other isolates in this lineage are from various states in the USA, including Oregon (2003), New Jersey (2004), Arizona (2006 and 2007) and Maryland (2007). The earlier isolates from Oregon, New Jersey and Pennsylvania are all associated with healthcare facility outbreaks (McDonald *et al.*, 2005). Also found within this lineage are isolates from five South Korean patients between 2007 and 2010. The only other isolate from South Korea in this collection is from 2006 and is fluoroquinolone-sensitive, implying fluoroquinolone-resistant BI/NAP1/027 *C. difficile* was first introduced to South Korea before 2007, possibly from the USA.

It appears that the same mutation in *gyrA* occurred twice independently in our dataset, resulting in two distinct fluoroquinolone-resistant lineages (Figure 3.3); both lineages contain isolates underlying outbreaks. The other fluoroquinolone-resistant lineage includes clades 2, 3 and 4 and covers the majority of isolates in this collection. The most striking feature of the phylogeny is a star-like topology in clade 2, implying a population expansion event (Figures 3.3 and 3.6). More interestingly, an isolate associated with the 2003 Quebec outbreak (Can010 from Montreal) (MacCannell *et al.*, 2006) sits on the node at the base of this star-like topology (arrow A in Figure 3.6), implying the founding nature of this isolate or isolates with similar or identical genotype. Interestingly, all Canadian isolates found in clade 2 were from Montreal, Quebec, while the Canadian isolates found in clade 1 were all from Calgary. These findings suggest that the two fluoroquinolone-resistant lineages consist of genetically different BI/NAP1/027 variants. Descendents of the expansion event in clade 2 include isolates from the USA, Canada, the UK and Australia, the most recent being from 2010. This suggests that the isolate underlying the outbreaks in Quebec in 2003 has subsequently spread to several continents. This part of the lineage includes epidemic isolates underlying healthcare outbreaks in Australia and the UK. In particular, the isolate from the outbreak in Maidstone, UK (2004) (arrow B in Figure 3.6) appears to have been derived from the expansion event in Quebec, and to have subsequently spread to London and Cambridge (Figure 3.6). According to the phylogeny, BI/NAP1/027 *C. difficile* arrived the UK in at least four independent events, as suggested by isolates from Exeter (2008), Maidstone (2004) and Ayrshire (2008) in clade 2, and Birmingham (2002) in clade 3.

According to this phylogeny the isolate from Birmingham (2002) or an isolate with similar or identical genotype appears to be the ancestor of a significant number of UK BI/NAP1/027 *C. difficile* (arrow A in Figure 3.7). This also suggests a rapid transatlantic transmission event following the Quebec outbreak. The descendents of this genotype include very recent isolates from Exeter, Birmingham, Cambridge, London, Liverpool, and multiple locations in Scotland (Figure 3.7). The isolate R20291 underlying the Stoke Mandeville outbreak is also an early variant of this genotype (arrow B in Figure 3.7), consistent with the fact that R20291 was among the first epidemic BI/NAP1/027 *C. difficile* to have been reported in the UK.

There is generally a lack of clear geographical structure in the UK lineage except for the clade of Liverpool isolates (clade 4). The ancestors of clade 3b2 are five isolates of the same genotype sampled from three different locations (Glasgow, Dundee, Inverness) in Scotland (arrow C in Figure 3.7). The descendents in this lineage consist of isolates from almost all sampled regions in the UK, except Northern Ireland (Figures 3.3B and 3.7).

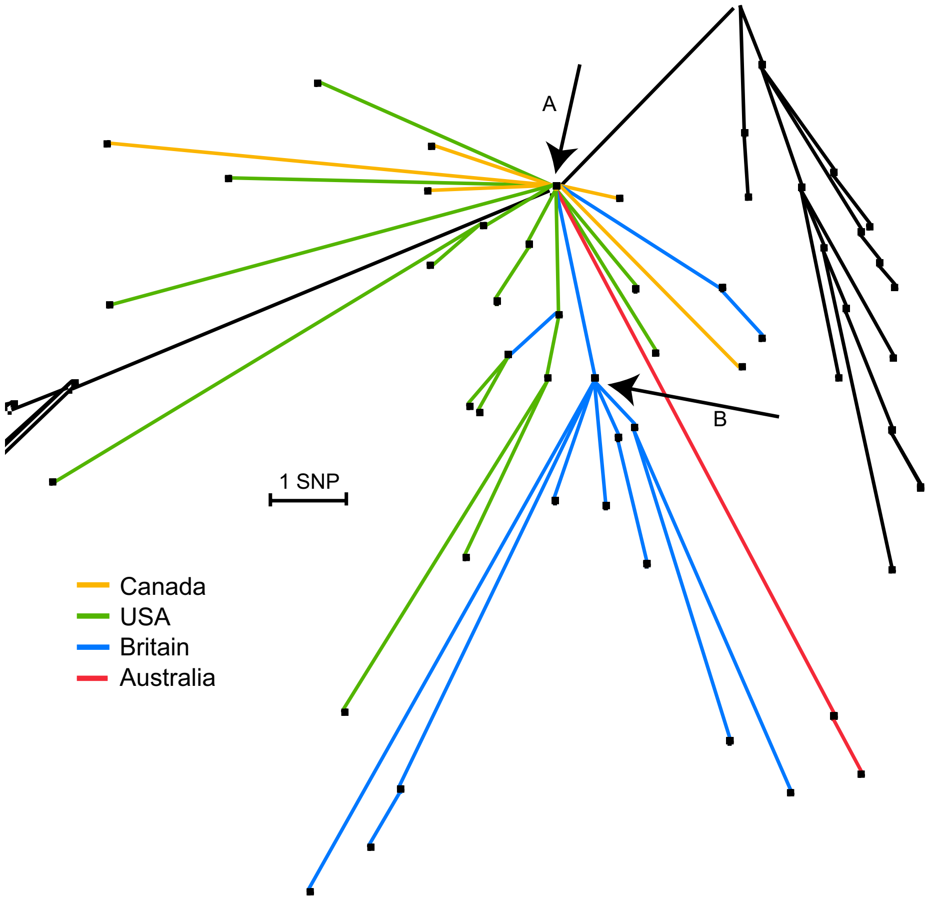
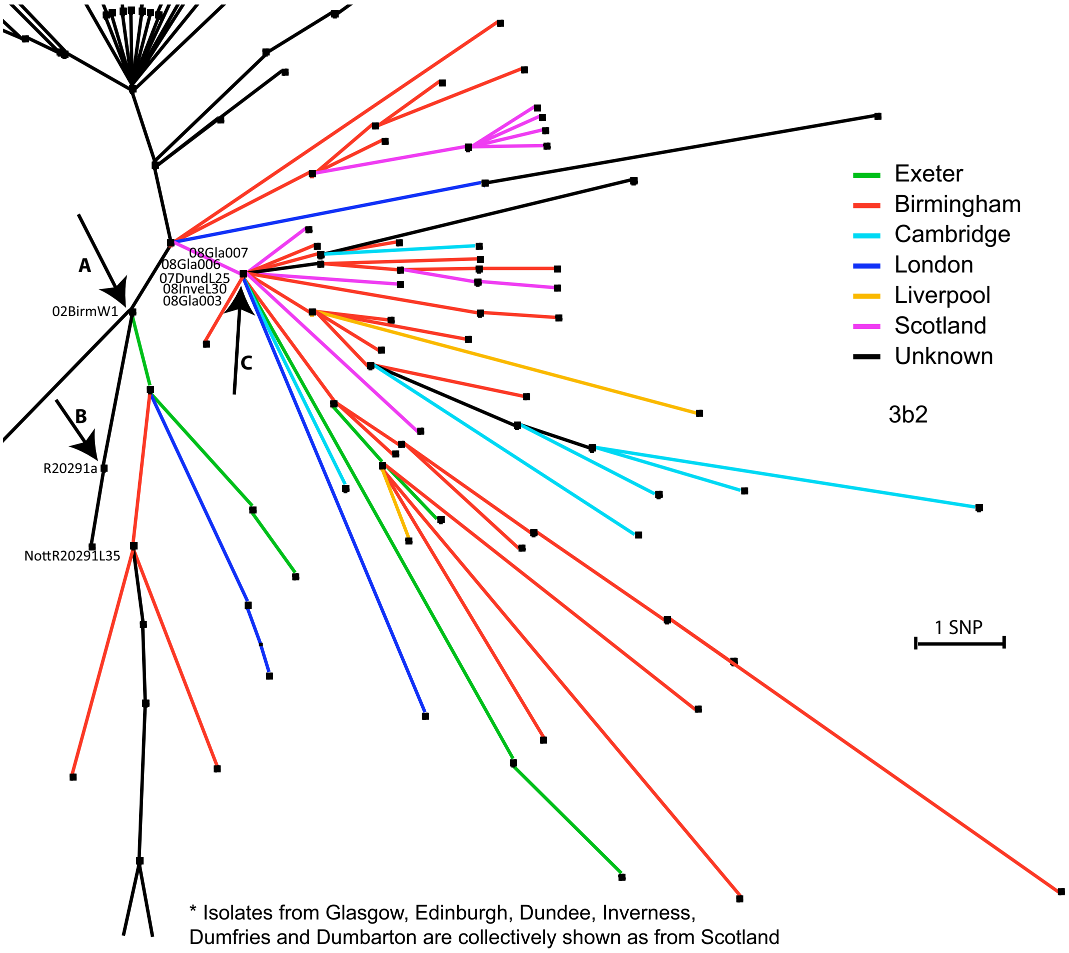


Figure 3.6: Expansion of clade 2 from Figure 3.3. Branches are coloured according to sampling locations of the strains the branches directly lead to. Arrows point to the isolate or lineage that are mentioned in the text.

This suggests that a significant number of UK BI/NAP1/027 *C. difficile* isolates could be traced back to an introduction event in Scotland. More recent parts of this lineage show that isolates from Liverpool, Birmingham and Exeter all share the same recent common ancestor. This data highlights the impact of rapid transmission on the structure of the tree. It should be noted that there is evidence of locally evolving clusters in Birmingham and Cambridge (Figure 3.7). The star-like topology of the UK isolates mirrors the global expansion represented by clade 2, implying an expansion of a smaller scale in the UK in the last four years. All isolates sampled from Belfast, Northern Ireland appear to have been derived from the Liverpool-associated genotype (Figures 3.3B and 3.8). This observation could be linked with frequent transport between Liverpool and Belfast, providing increased chances of transmission between the two cities.



* Isolates from Glasgow, Edinburgh, Dundee, Inverness, Dumfries and Dumbarton are collectively shown as from Scotland

Figure 3.7: Expansion of clade 3 from Figure 3.3. Branches are coloured according to sampling locations of the strains the branches directly lead to. Arrows point to the isolates that are mentioned in the text.

The majority of the Liverpool hospital isolates cluster into clade 4 although several are also located in clade 3. The three ribotype 176 isolates from London (lon004, lon005 and lon006) are found among BI/NAP1/027 *C. difficile* isolates in clades 2 and 3 and are indistinguishable at the whole genome level, confirming that *C. difficile* ribotype 176 is actually a variant of BI/NAP1/027 with an altered PCR-ribotype pattern (Nyc *et al.*, 2011).

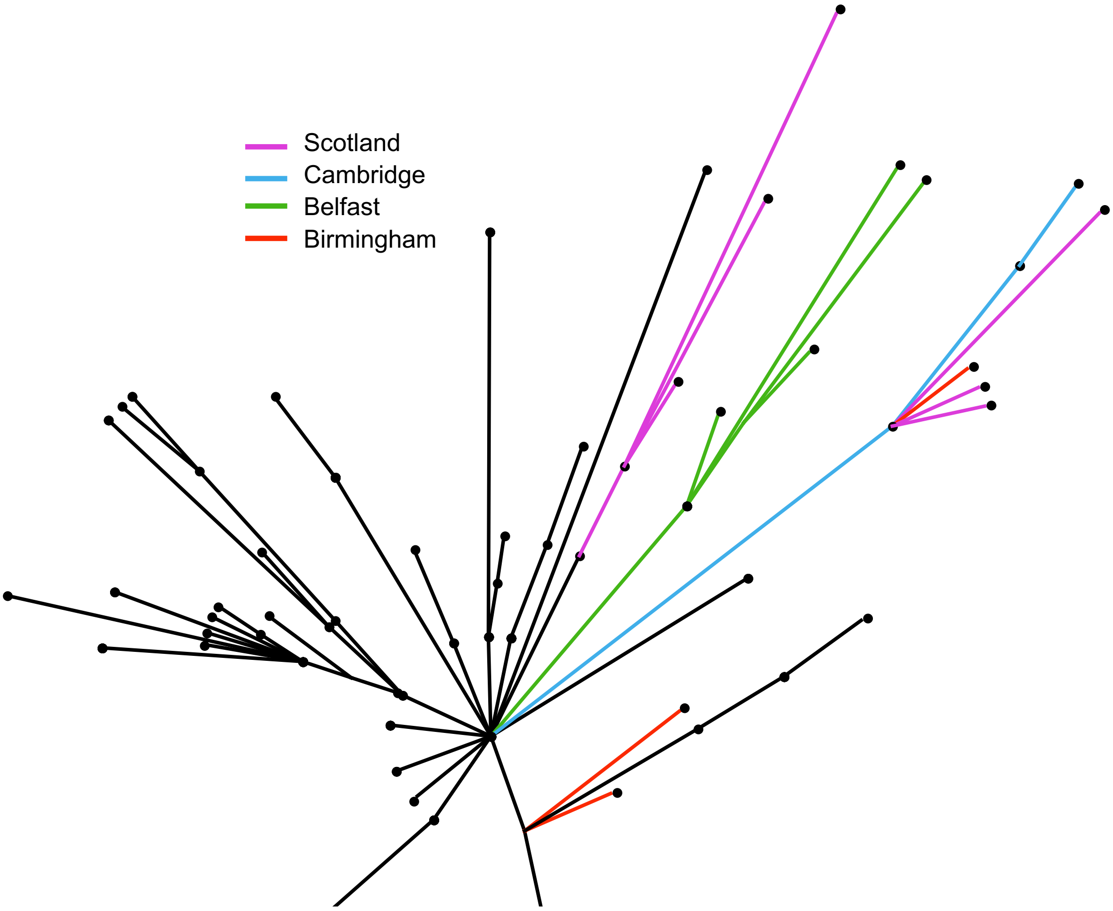


Figure 3.8: Expansion of clade 4 from Figure 3.3. Branches are coloured according to sampling locations of the strains the branches directly lead to, except isolates from Liverpool (shown in black).

3.3.2.2 Phylogenetic relationships inferred with Bayesian analysis

Phylogenetic relationships were also inferred for the same dataset using the program BEAST (Drummond and Rambaut, 2007) to check for agreement. Bayesian methods were developed for phylogenetic and population history analyses by sampling large numbers of trees (Drummond and Rambaut, 2007). Multiple methods, which are all contained in the same BEAST program, serve different analysis purposes, including re-constructing phylogenies, dating lineages and inferring phylogeographic histories (Lemey *et al.*, 2009). Overall, the four clades identified in the maximum likelihood tree are all strongly supported by the Bayesian method, with posterior probabilities of 0.97, 1, 1, and 0.98, respectively (asterisks in Figure 3.9). Although the Bayesian tree appears to resolve the star-like phylogeny (clade 2) into clear bifurcations, the bifurcations were very poorly supported. The three branching lineages in clade 2 received posterior probabilities of only 0.019, 0.031, 0.001 respectively (branches pointed to by arrows A to C in Figure 3.9). This un-resolved branching order can be explained by a rapid expansion, which warrants the star-like genealogies in the maximum likelihood tree. Bayesian phylogeographic analysis also provided no conclusive inference with respect to the geographic affiliation of the most common ancestor (MRCA) of these lineages.

The time to the most recent common ancestor (tMRCA) was estimated for each node in the Bayesian tree. The age of the entire sampled BI/NAP1/027 collection was estimated to be 78 years (1933) with a 95% confidence interval (CI) between 36 to 142 years (1879 – 1975) (Figure 3.9). However, two isolates BI-3 and BI-4 appeared to be outliers in our collection; both have long branches in the tree. The common ancestor of the remaining isolates was estimated to have emerged 44 years ago (1967). The Bayesian tree supports independent acquisition of the *gyrA* mutation and implies the two *gyrA* mutant lineages associated with fluoroquinolone resistance appeared in 1994 (95% CI from 1989 to 1998) and 1992 (95% CI from 1987 to 1997) respectively. The *gyrA* mutant lineage in clade 1 consists of isolates from the USA near the base and South Korean isolates at the tip, supporting the suggestion that the fluoroquinolone-resistant South Korean BI/NAP1/027 *C. difficile* was derived from the USA. Bayesian phylogeographic analysis also indicates the MRCA of this *gyrA* mutant lineage is from the USA (supported by >99% probability).

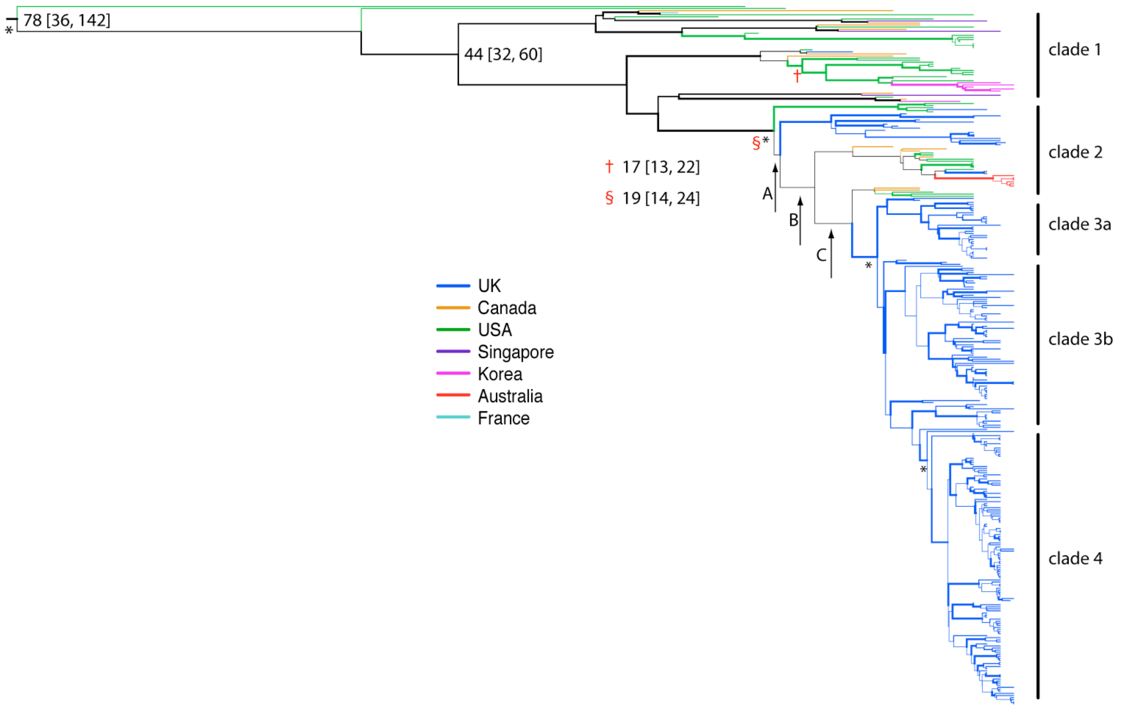


Figure 3.9: A summarized tree based on sampled trees from a BEAST run with a relaxed log exponential clock model. Labels on the right correspond to clade names in Figure 3.3. Branches are coloured according to origins of isolates. Line widths of internal branches indicate how well-supported the branches are. The bolder a branch is, the better supported it is. The three poorly supported bifurcations in clade 2 are pointed to by arrows A to C. The numbers labelled near the nodes indicate estimates (median and 95% CI) of tMRCA of the isolates above that node. Red symbols indicate lineages associated with fluoroquinolone resistance and ages inferred for them.

3.3.3 Discriminatory SNPs and potential functional consequences

The level of genetic divergence is low within the BI/NAP1/027 *C. difficile* lineage. For example, only 7 SNPs differentiate between clade 1 and clade 2; this includes the *gyrA* mutation mentioned above. Table 3.1 summarizes the

SNPs that discriminate among clades and their predicted effects. One SNP that differentiates clade 3a and clade 3b is a non-synonymous mutation in a penicillin-binding protein. This mutation can result in resistance to penicillin or other beta-lactam class antibiotics. The SNPs that were fixed along the branch leading to the expansion in clade 2 are of particular interest, as they could potentially lead to an increased level of fitness, which could contribute to the sudden expansion. Among these SNPs is an amino acid change (A240D) in a probable transporter. This amino acid change is within the transmembrane helix domain of the protein, which belongs to the PFAM Nramp (PF01566) family (natural resistance-associated macrophage protein family) (Finn *et al.*, 2008). This family of proteins normally acts as cation transporters and have been shown to be involved on both ‘sides’ in interactions between intracellular microbial pathogens and their hosts (Govoni and Gros, 1998; Pinner *et al.*, 1997). However, there is no evidence that this protein could be involved in host interactions in an extracellular pathogen such as *C. difficile*, and the nature of the amino acid change would suggest impaired protein function in clade 2.

Two SNPs are found within the coding sequence of R20291_1052 (*spo0A*), which is essential for spore formation. However, these mutations were only detected in two isolates (BI-15 and Liv131, one in each). It is unclear whether the mutations in this gene have significant impact on sporulation. Overall, there is little evidence that a significant change in phenotype could result from any of the SNPs that differentiate between clade 1 and clade 2, except perhaps the fluoroquinolone resistance mutation itself. However, a novel genomic region was gained along the branch leading to clade 2; this will be discussed in 3.3.8.

position	separates clades	Type	Product	reference residue	alternative residue	amino acid change	Predicted effect
2656051	1a/1b	Nonsyn	quinolinate synthetase A	L	I	conservative	None
6310	1/2	Nonsyn	DNA gyrase subunit A	I	T		Fluoroquinolone resistance
118571	1/2	Nonsyn	putative ribosomal protein	D	N	conservative	None
1239212	1/2	Nonsyn	conserved hypothetical protein, DUF_177 family	T	N	conservative	None
1466990	1/2	Nonsyn	Probable cation transporter, Nramp homolog	D	A	non-conservative	Possible transmembrane helix disruption
2938388	1/2	Nonsyn	hypothetical protein (Small, very Histidine-rich protein, unique to *Clostridia*)	F	L	non-conservative	Unknown
3118366	1/2	synonymous	phosphoenolpyruvate-protein phosphotransferase	P	P	none	None
3507157	1/2	Intergenic				none	200 bp upstream of conserved hypothetical protein
120932	2/3	synonymous	DNA-directed RNA polymerase alpha chain	I	I		
2304160	2/3	Nonsyn	conserved hypothetical protein, DUF162 family	E	G	non-conservative	Unknown
2983263	2/3	Nonsyn	UDP-N-acetylmuramoylalanine--D-glutamate ligase	T	I	non-conservative	Unknown; not within any Pfam domain
3538081	2/3	Nonsyn	PTS system, IIabc component	V	I	conservative	None
886105	3a/3b	Nonsyn	penicillin-binding protein	T	I	non-conservative	Transpeptidase, could potentially result in penicillin resistance
1374216	3b1/3b2	Nonsyn	putative mannosyl-glycoprotein endo-beta-N-acetylglucosamidase	L	S	non-conservative	Unknown; not within a functional domain
1163835	3b/4a	Nonsyn	putative membrane protein	T	I	non-conservative	Unknown
4150703	4a/4b	Nonsyn	putative transcription antiterminator	M	I	conservative	Unknown

Table 3.1: Discriminatory SNPs and predicted functional effects within the BI/NAP1/027 *C. difficile* lineage. nonsyn – nonsynonymous.

3.3.4 Signatures of recombination

Section 2.3.4 revealed that *C. difficile* is capable of exchanging large chromosomal regions through homologous recombination. In this dataset, eight isolates (kor001, BI-3, BI-4, BI-10, BI-11, Can001, Can007, lon004) exhibit clusters of SNPs when compared to R20291, suggesting imported genomic regions from outside the BI/NAP1/027 lineage (Figure 3.10).

The largest homologous recombination blocks are found in isolates BI-4 (123 kb, approximately at 2.8 Mb in Figure 3.10), BI-11, kor001 and Can007 (134 kb and 147 kb, approximately at 1.1 Mb and 3.9 Mb locations in Figure 3.10 respectively). Interestingly, BI-11, kor001 and Can007 demonstrate homologous recombination blocks of an almost identical pattern. This is consistent with the knowledge that the three isolates occupy the same lineage in the phylogenetic tree (Figure 3.5), implying the recombination event affected their common ancestor. A putative phage element was found adjacent to the 147 kb blocks in BI-11, kor001 and Can007. However, no mobile element was found near the other two blocks.

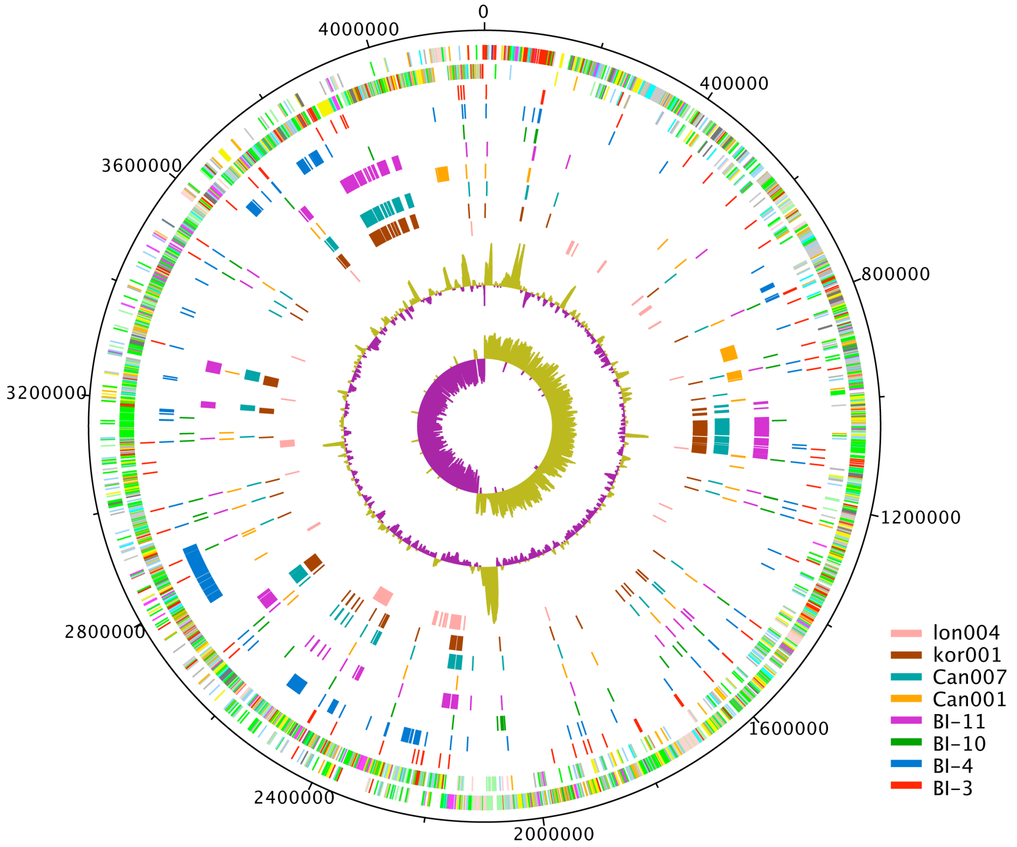


Figure 3.10: SNPs between R20291 and eight *C. difficile* isolates showing homologous recombination blocks. Outer circle: CDSs of *C. difficile* R20291 genome, shown on a pair of concentric rings representing both coding strands; two inner circles: G+C% content plot and GC deviation plot (>0% olive, <0% purple); in between: SNPs between R20291 and eight isolates (from outer to inner: BI-3, BI-4, BI-10, BI-11, Can001, Can007, kor001, lon004), coloured according to legend.

3.3.5 Homoplasic SNPs and convergent evolution

In order to detect signals of convergent evolution driven by selection, the 604 SNPs from the non-recombining core genome were checked to identify those that are in conflict with the phylogenetic tree (homoplasic SNPs). This approach detected 13 homoplasic SNPs (2% of the total number), displayed in Table 3.2. It appears that homologous recombination between isolates within the BI/NAP1/027 *C. difficile* lineage has not played a major role in shaping the phylogeny. However, it should also be noted that in a bacterial population with such highly similar genomic backbones as BI/NAP1/027 *C. difficile*, it is very difficult to detect recombination between the isolates.

An examination of the functional consequences of the homoplasic SNPs highlights the significant impact of antimicrobial drugs and potential immune selection in BI/NAP1/027 *C. difficile* microevolution. Besides the mutation in *gyrA* discussed above, DNA gyrase subunit B was also found to harbour an amino acid substitution (Asp426Asn), which is associated with increased fluoroquinolone resistance (Spigaglia *et al.*, 2008). Mutations associated with rifampicin and fusidic acid resistance are also apparent (Table 3.2). Both substitutions in *rpoB* gene (His502Asn and Arg505Lys) have been reported in *C. difficile* (O'Connor *et al.*, 2008) to be associated with resistance to rifampin and rifaximin; while the two substitutions in *fusA* gene were not identified in a previous study (Noren *et al.*, 2007).

The isolates carrying homoplasic substitutions that result in antibiotic resistance are shown in Figure 3.11. Interestingly, the resistance to both rifampicin and fusidic acid occurred only in the fluoroquinolone-resistant lineages. It is unclear whether other antibiotic resistances are more likely to develop on a fluoroquinolone background. It is possible that since the fluoroquinolone-resistant isolates are more numerous and more recent, further mutations are more likely to occur in them. There is no evidence for a specific multidrug resistant strain or lineage, as none of the isolates are resistant to all three antibiotics. The earliest isolates in our collection that developed resistance to rifampicin and fusidic acid are from the USA (2004) and Canada (2003) respectively; while resistance to fluoroquinolones may have developed even earlier, as it was discovered in two 2001 isolates from the USA and Canada.

Beyond drug resistance, among the list of genes affected by homoplasic SNPs are a pair of two-component regulatory system genes and two cell surface proteins, implying that changes driven by other factors such as environmental or immune selection pressure could also be detected through this approach.

Position	Gene	Product	Reference allele	Alternative allele	Amino acid change	Functional Impact
5420	CDR20291_3546	DNA gyrase subunit B	G	A	Asp426Asn	Fluoroquinolone^{R}
6310	CDR20291_3547	DNA gyrase subunit A	T	C	Thr82Ile	Fluoroquinolone^{R}
95412	CDR20291_0060	DNA-directed RNA polymerase beta chain	C	A	His502Asn	Rifampicin^{R}
95422	CDR20291_0060	DNA-directed RNA polymerase beta chain	G	A	Arg505Lys	Rifampicin^{R}
103867	CDR20291_0064	translation elongation factor G	C	A/T	His455Asn / His455Tyr	Fusidic acid^{R}
104117	CDR20291_0064	translation elongation factor G	C	T	Pro538Leu	Fusidic acid^{R}
1681194	CDR20291_1418	putative phage-related protein	C	T	synonymous	
1681261	CDR20291_1418	putative phage-related protein	A	G	synonymous	
1681354	CDR20291_1418	putative phage-related protein	A	G	synonymous	
1800920	CDR20291_1522	two-component response regulator	G	A	Glu -> Lys	
1802086	CDR20291_1523	two-component sensor histidine kinase	C	T	Thr -> Ile	
3170481	CDR20291_2682	cell surface protein (S-layer precursor protein)	G	A/T	Pro -> Leu/Gln	
3938789	CDR20291_3294	putative membrane protein	A	C	Tyr -> stop	

Table 3.2. Homoplasic SNPs and associated gene products.

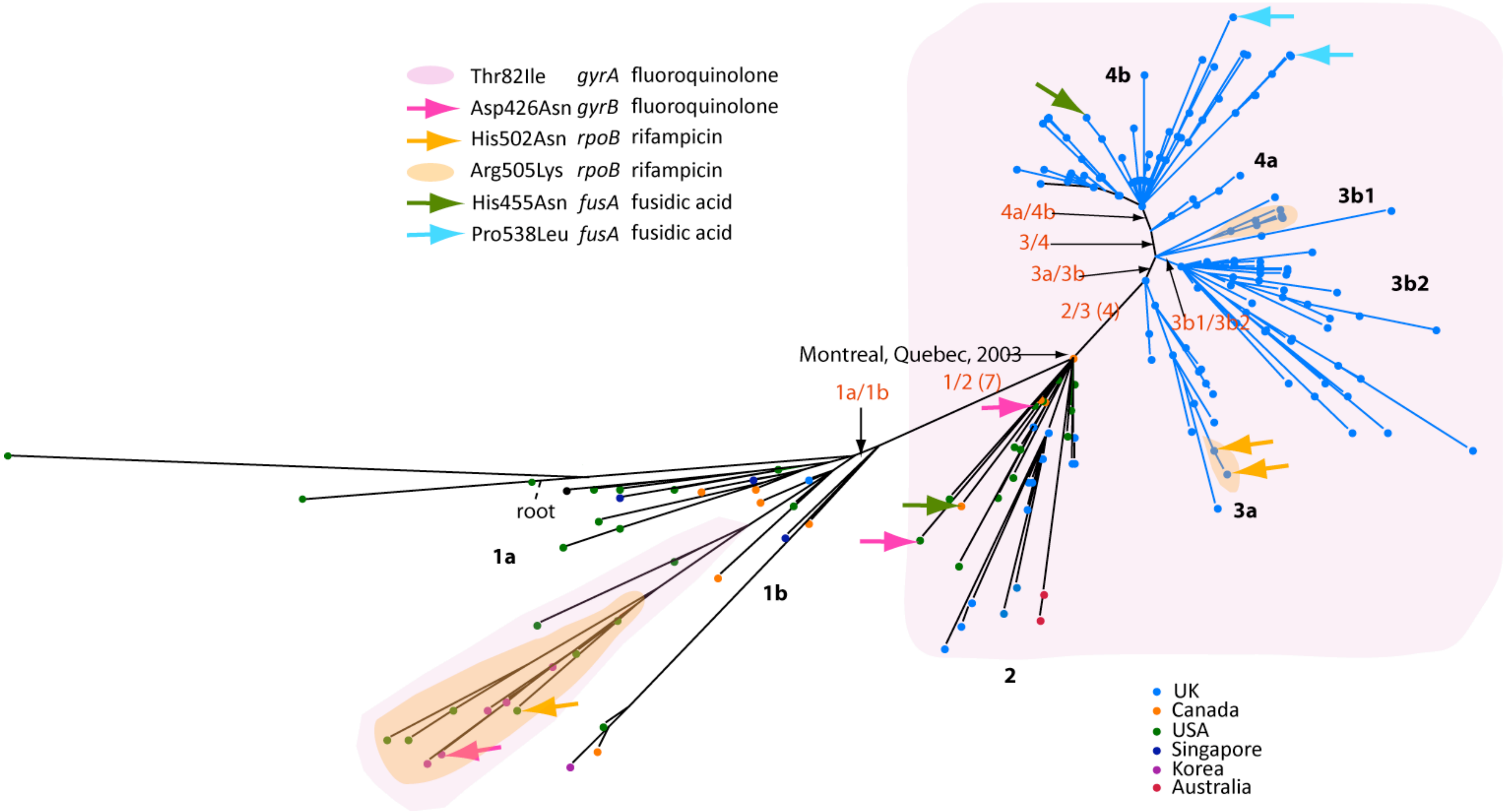

Thr82Ile gyrA fluoroquinolone
Asp426Asn gyrB fluoroquinolone
His502Asn rpoB rifampicin
Arg505Lys rpoB rifampicin
His455Asn fusA fusidic acid
Pro538Leu fusA fusidic acid
4b
4a
3b1
3b2
3a
2
1a
1b
root
4a/4b
3/4
3a/3b
2/3 (4)
3b1/3b2
Montreal, Quebec, 2003
1/2 (7)
1a/1b
UK
Canada
USA
Singapore
Korea
Australia

Figure 3.11: Isolates and lineages carrying substitutions conferring resistance to antimicrobial drugs. Strains carrying different substitutions are pointed out by either shaded areas (Thr82Ile in *gyrA* and Arg505Lys in *rpoB*) or coloured arrows (the rest). The legends on top left indicate amino acid substitutions, the genes affected, and the resulting resistance. Descriptions for the phylogenetic tree are the same as given in Figure 3.3.

Three other synonymous SNPs were found within a single gene encoding a putative phage-related protein. These same isolates (CD196, LSTM013 and LSTM014) carry these alternate alleles, indicating these SNPs were likely acquired through a single gene transfer event.

3.3.6 Estimating mutation rate

The finding that this sample collection, which spans a 25-year-period, is only differentiated by a few hundred SNPs is striking. The mutation rate for this group of *C. difficile* was estimated based on dates of isolation and the number of mutations accumulated using a full maximum likelihood model. This model assumes a rapid expansion that results in perfect star genealogies (Morelli *et al.*, 2010). This calculation was performed with isolates from clade 2, as the star genealogy is more suitable for this model. The results generated by this method are shown in Table 3.3. The entire dataset (without recombination sites) was also used to estimate mutation rate with BEAST (Drummond and Rambaut, 2007). This yielded comparable results in the range of 1.59×10^{-7} to 1.68×10^{-7} substitutions per site per year. Taking results from both methods together, the mutation rate for the BI/NAP1/027 lineage was estimated to be within the range of 1.59×10^{-7} – 4.41×10^{-7} substitutions per site per year.

Clade	NumStrains	NumLoci	Maximum likelihood	Mutation rate	Lower 95% CI	Upper 95% CI
2	41	1000	-231.72	4.07E-07	3.12E-07	5.28E-07
2	41	1500	-353.46	3.48E-07	2.74E-07	4.33E-07
2	41	2000	-518.85	4.41E-07	3.69E-07	5.28E-07

Table 3.3. Estimated mutation rates using a maximum likelihood expansion model. NumLoci refers to numbers of CDSs randomly selected for each analysis. Mutation rate unit is substitutions per site per year.

This rate is equal to 1-2 mutations per genome per year, and ~10 times slower when compared to other bacteria over equivalent (recent) timescales, such as *Streptococcus pneumoniae* (1.57×10^{-6} substitutions per site per year)

(Croucher *et al.*, 2011) and *Staphylococcus aureus* (3.3×10^{-6} substitutions per site per year) (Harris *et al.*, 2010). The finding is in agreement with the hypothesis that spore-forming bacteria generate genetic variation much slower because they are only actively replicating when in vegetative form, and the time spent in dormant spore form does not contribute to the evolutionary rate (Keim *et al.*, 1997; Pearson *et al.*, 2004). A slow mutation rate is also consistent with the lack of geographical structure we observed in clade 3b. Since the *C. difficile* ribotype 027 genome only changes at a rate of 1-2 mutations per year, and spores in the environment can be carried by people travelling to distant places, our observations can be explained as result of rapid and frequent transmission, rather than isolates transmitting slowly and evolving locally. This finding further underlines the need to control *C. difficile* transmission, particularly by eradicating environmental spores.

3.3.7 Population history inference

The polymorphism data (excluding sites affected by recombination) were used to infer the population history of this collection of BI/NAP1/027 isolates using Bayesian skyline plot (Drummond *et al.*, 2005) (Figure 3.12). This analysis indicates the population size underwent a minor increase around 2010 and a two-step sharp decrease beginning in 2007. The increase slightly predates reports of hospital outbreaks caused by BI/NAP1/027 *C. difficile.* The sharp decrease is possibly due to implementation of more stringent hospital cleaning regimes in recent years. However, the inferred population size still maintains the same order of magnitude overall, and both changes are not significant,

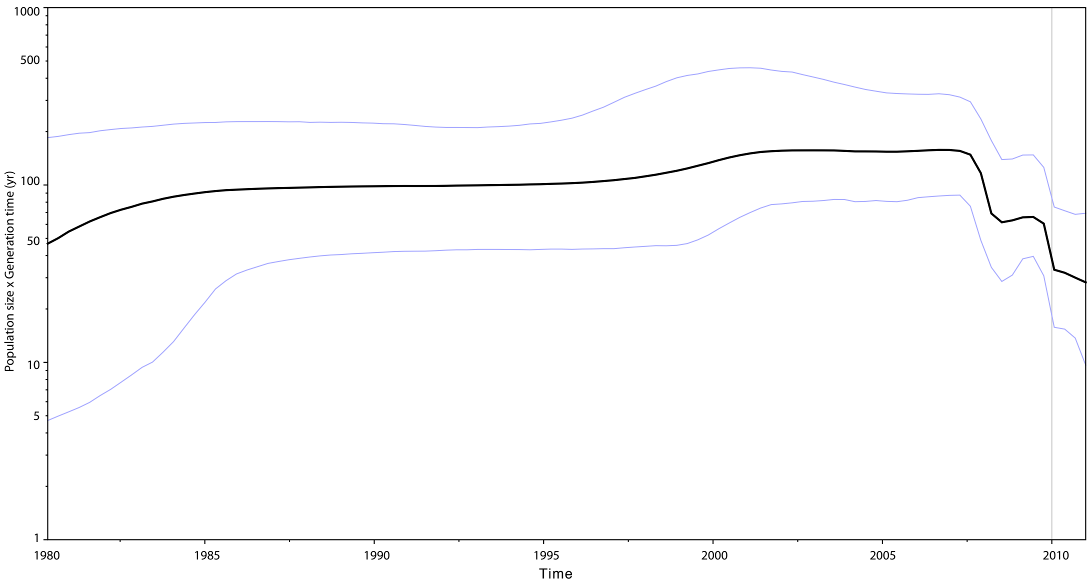


Figure 3.12: Bayesian skyline plot indicating changes in population size of this BI/NAP1/027 sample collection. Time (in years) is given on the x-axis; y-axis shows the product of population size and generation time (in years) in log scale. The black line represents median estimate; purple lines denote 95% CI.

3.3.8 Horizontal gene transfer in BI/NAP1/027 lineage

The pan-genome for this collection of isolates was identified by assembling un-aligned sequencing reads and sequence comparisons using BLASTN. All sequencing reads were then aligned to the pan-genome to assess the presence and absence of genomic regions. Mobile elements such as conjugative transposons and bacteriophages can make up a large part of the accessory genome. Antibiotic resistance cassettes carried by these mobile elements often confer major advantage to host strains. Genes related to erythromycin, chloramphenicol, tetracycline and aminoglycoside resistance were found in the *C. difficile* BI/NAP1/027 accessory genome. Combining information from the phylogenetic tree and presence or absence of genomic regions enabled us to map specific insertion and deletion events onto the branches (i.e. the time these events occurred).

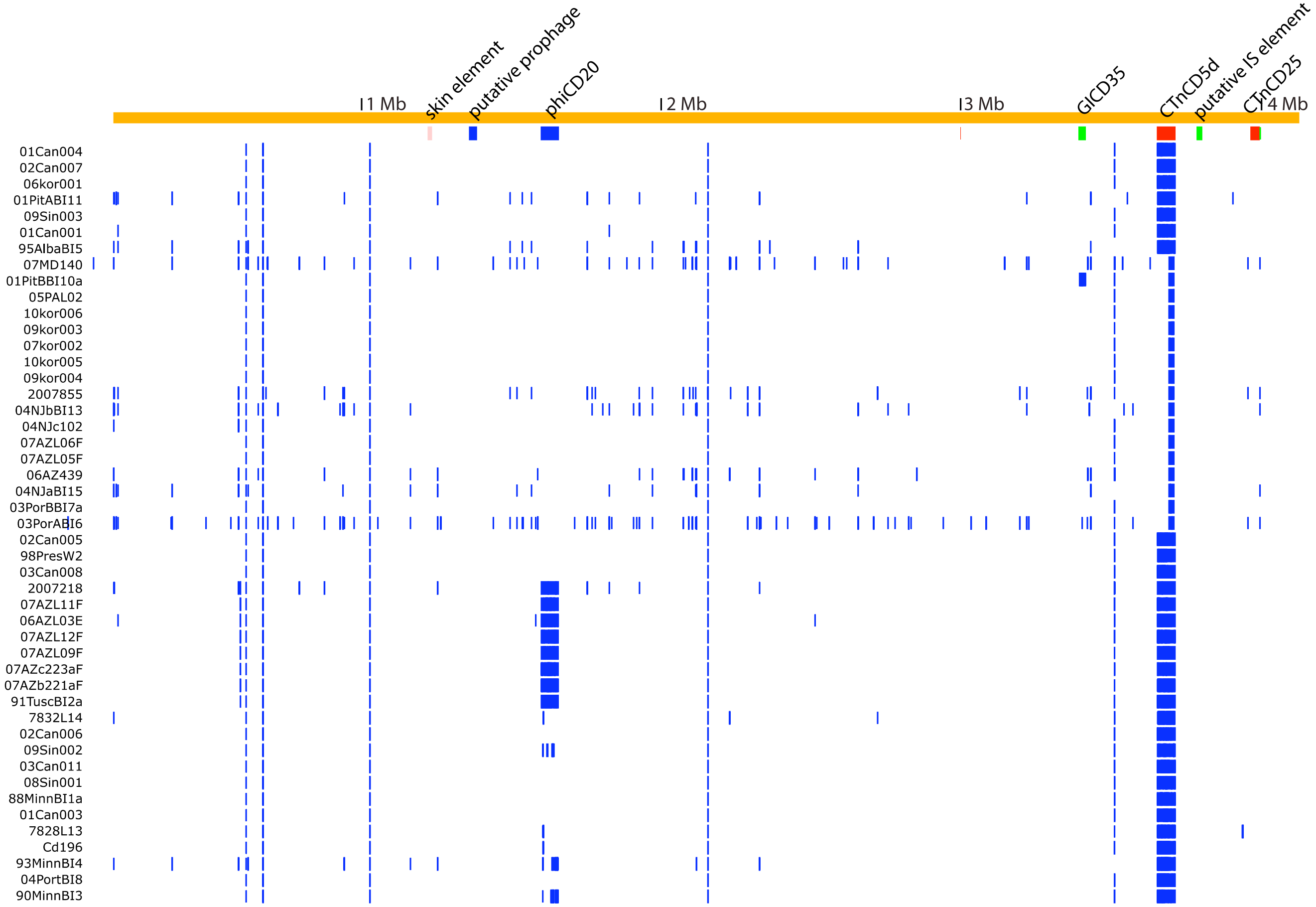


Figure 3.13: Genomic regions that are present in the reference (M7404) but absent (shown by blue boxes) from isolates in clade 1. The genome of M7404 (>4 Mb in size) is represented by an orange bar. The genomic islands carried by M7404 are depicted by coloured boxes beneath it. Names of isolates in clade 1 are given on the left.

For examples, phi*CD20*, the prophage common to ribotype 027 isolates was not present in BI-2 and 7 other isolates from environmental sources, while two outlier isolates BI-3 and BI-4 harbour a different version of phi*CD20* (Figure 3.13). The genomic island GI*CD35* is common to all isolates in the collection except BI-10, which suggests a deletion event.

It is fairly common for *C. difficile* to harbour similar structural backbones and machinery for various mobile elements, which appear in different parts of the tree. For example, at least three versions of the conjugative transposon CTn*CD5* exist in BI/NAP1/027 *C. difficile* lineage, all of which are highly similar to conjugative transposon CTn*5* in 630. A comparison of CTn*CD5b* (in R20291), CTn*CD5d* (in M7404) and CTn*CD5c* (in 2007855) is shown in Figure 3.14. CTn*CD5b* in R20291 was also named Tn*6103* in (Brouwer *et al.*, 2011).

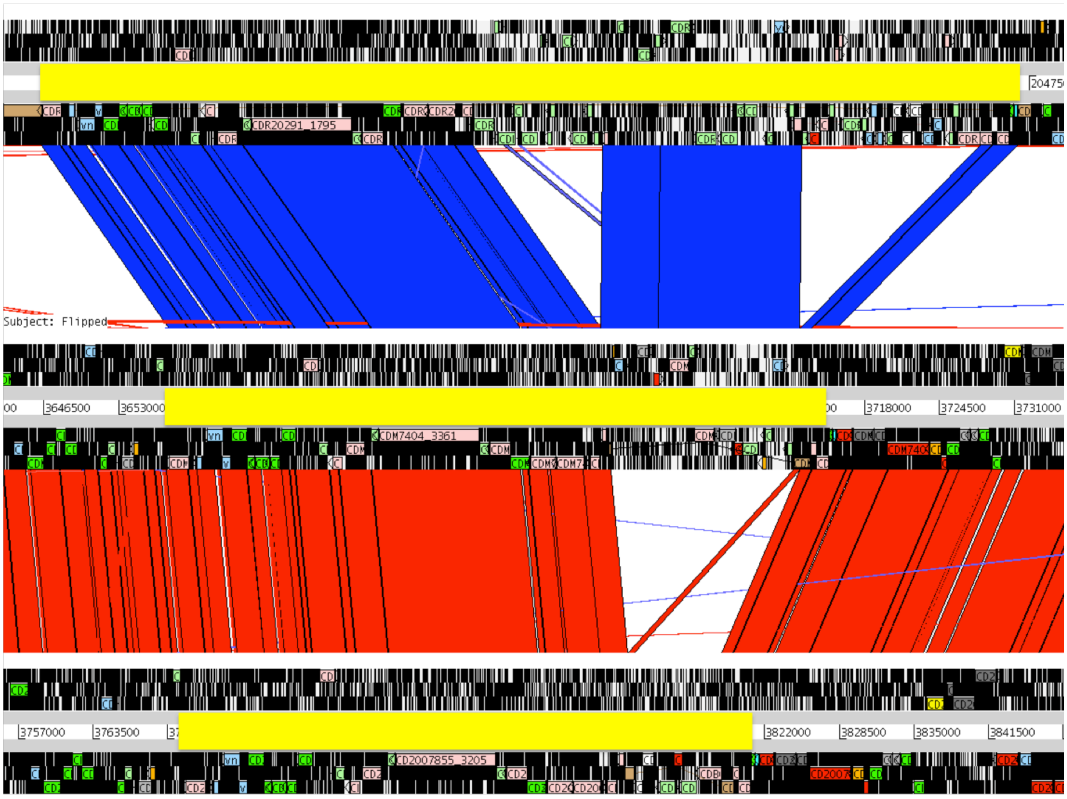


Figure 3.14: Comparison of three versions of CTn*CD5* (indicated by yellow boxes) in strains R20291 (top), M7404 (middle), and 2007855 (bottom).

This comparison also highlights a major difference between isolates in clades 1 and 2. Copies of CTn*CD5* carried by isolates in clade 2 and beyond harbour a contiguous insertion of 15.7 kb, which was also recently named Tn*6105* (Brouwer *et al.*, 2011) (Figure 3.15), containing 14 genes, four of which are predicted to be DNA-binding proteins or transcriptional regulators (CDM7404_3352, CDM7404_3350, CDM7404_3346 and CDM7404_3343) (Table 3.4). This acquisition has the potential to confer significant phenotypic changes on the organism, through a modification of the transcriptome. In addition to this insertion, R20291 also carries an insertion of 16 kb in size (also named Tn*6104*), which has been discussed in section 2.3.7. This small insertion was only found in three other isolates in our collection (LSTM035, Liv14 and Liv16). On the other hand, copies of a different version of CTn*CD5* were found in all isolates in the clade 1 *gyrA* mutant lineage, though at the same location within the genome. Six isolates in this lineage (2007855, BI-13, 2006439, 2004102, 07AZL05F and 07AZL06F) carry the additional agc^R (aminoglycoside resistance) cassette at the 3' end of CTn*CD5* (Figure 3.16), as discussed in section 2.3.3. In addition to CTn*CD5*, all isolates except BI-7 (Portland, 2003) in the clade 1 *gyrA* lineage carry conjugative transposon CTn*CD11*, which implies it may have been deleted from BI-7 (Figure 3.16, part of the sequences from strain Kor005). This conjugative transposon contains a gene *erm(B)* which confers erythromycin resistance.

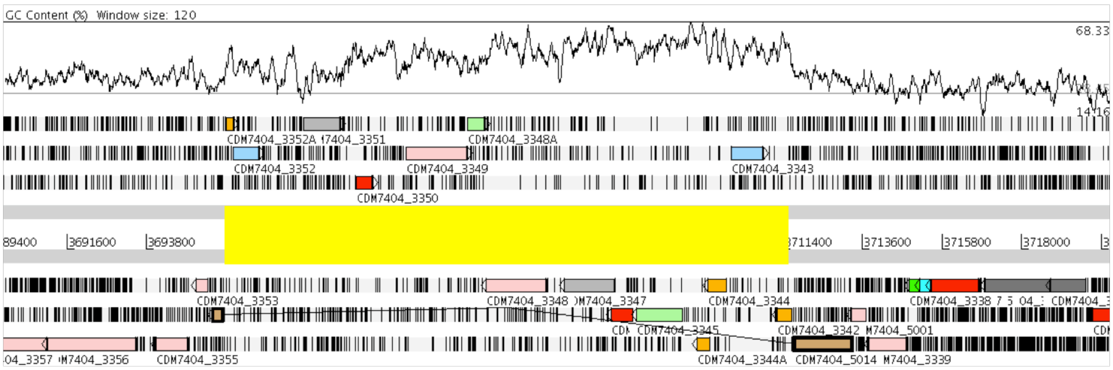


Figure 3.15: Genes gained along the branch connecting clades 1 and 2, shown in the genome of M7404. The graph on top depicts GC content. The extent of the insertion Tn*6105* is indicated by a yellow box.

Figures 3.16 – 3.19 depict additional genomic regions gained by isolates in clade 1 – 4 respectively in relation to M7404. Comparisons with known *C. difficile* genomes revealed that the accessory genome contains sequences from plasmids, phages and conjugative transposons. Four contigs absent from M7404 were identified in the genome assembly of isolate 2007223 (Figure 3.16); the two larger contigs Node_6 and Node_25, which make up 41,648 bp in total, are highly similar to a 40 kb plasmid carried by BI-1. It is likely that the same plasmid is also present in 2007223. It is also probable that more isolates carry this plasmid, such as Aus002, 2007850, Liv189 and a large proportion of the isolates in clade 1 (red boxes in Figures 3.16 – 3.19).

Coding sequence	Product
CDM7404_3352A	conserved hypothetical protein
CDM7404_3352	two-component system regulatory protein
CDM7404_3351	radical SAM enzyme
CDM7404_3350	probable regulator (contains HtH domain from sigma 70)
CDM7404_3349	site-specific recombinase
CDM7404_3348A	hypothetical protein
CDM7404_3348	site-specific recombinase
CDM7404_3347	putative P-loop NTPase
CDM7404_3346	putative DNA-binding protein (contains zinc-finger domain)
CDM7404_3345	putative plasmid mobilisation protein
CDM7404_3344A	conserved hypothetical protein
CDM7404_3344	conserved hypothetical protein
CDM7404_3343	probable transcription regulator (contains HtH domain)
CDM7404_3342	conserved hypothetical protein

Table 3.4: Fourteen coding sequences gained along the branch connecting clades 1 and 2.

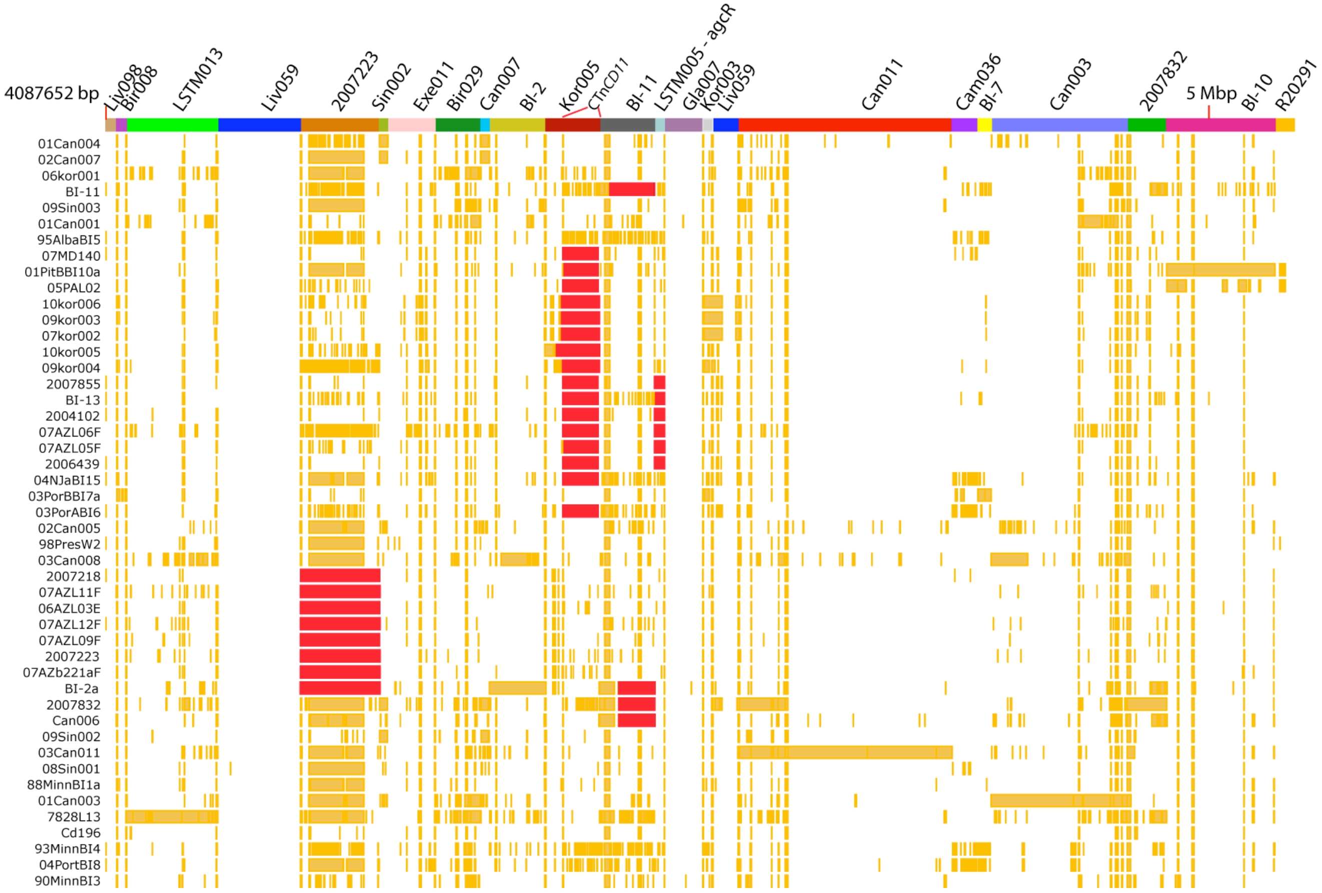


Figure 3.16: Accessory genome components of isolates in clade 1. The coloured boxes on top depict groups of contigs or genomic island sequences from different isolates. Sequence from the same isolate is shown in the same colour, with the isolate name labelled above. Yellow and red boxes depict same genome regions carried by other isolates. Names of isolates are given on the left. Regions shown with red boxes are mentioned in the text.

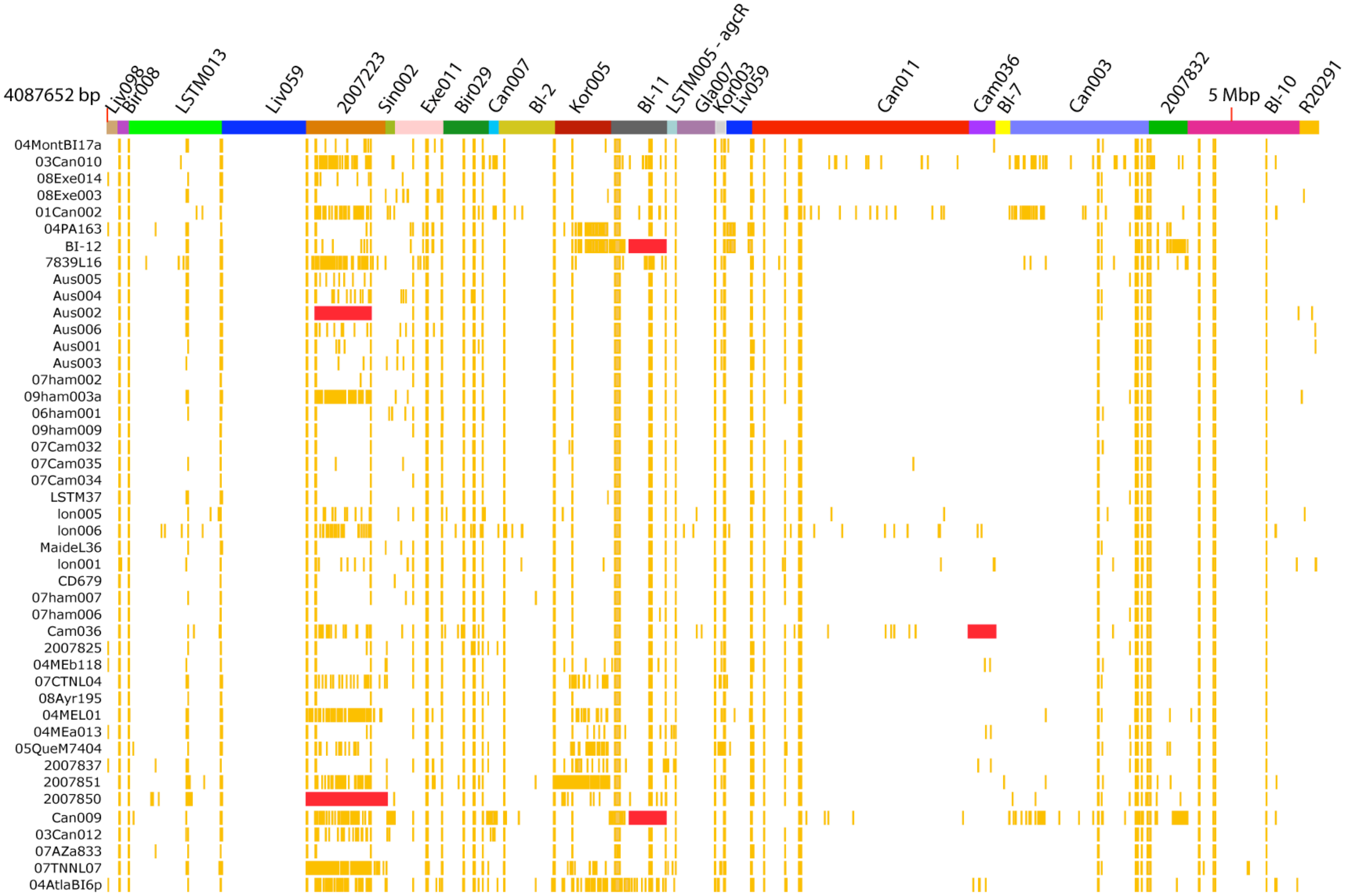


Figure 3.17: Accessory genome components of isolates in clade 2. Other descriptions for this figure are the same as Figure 3.16.

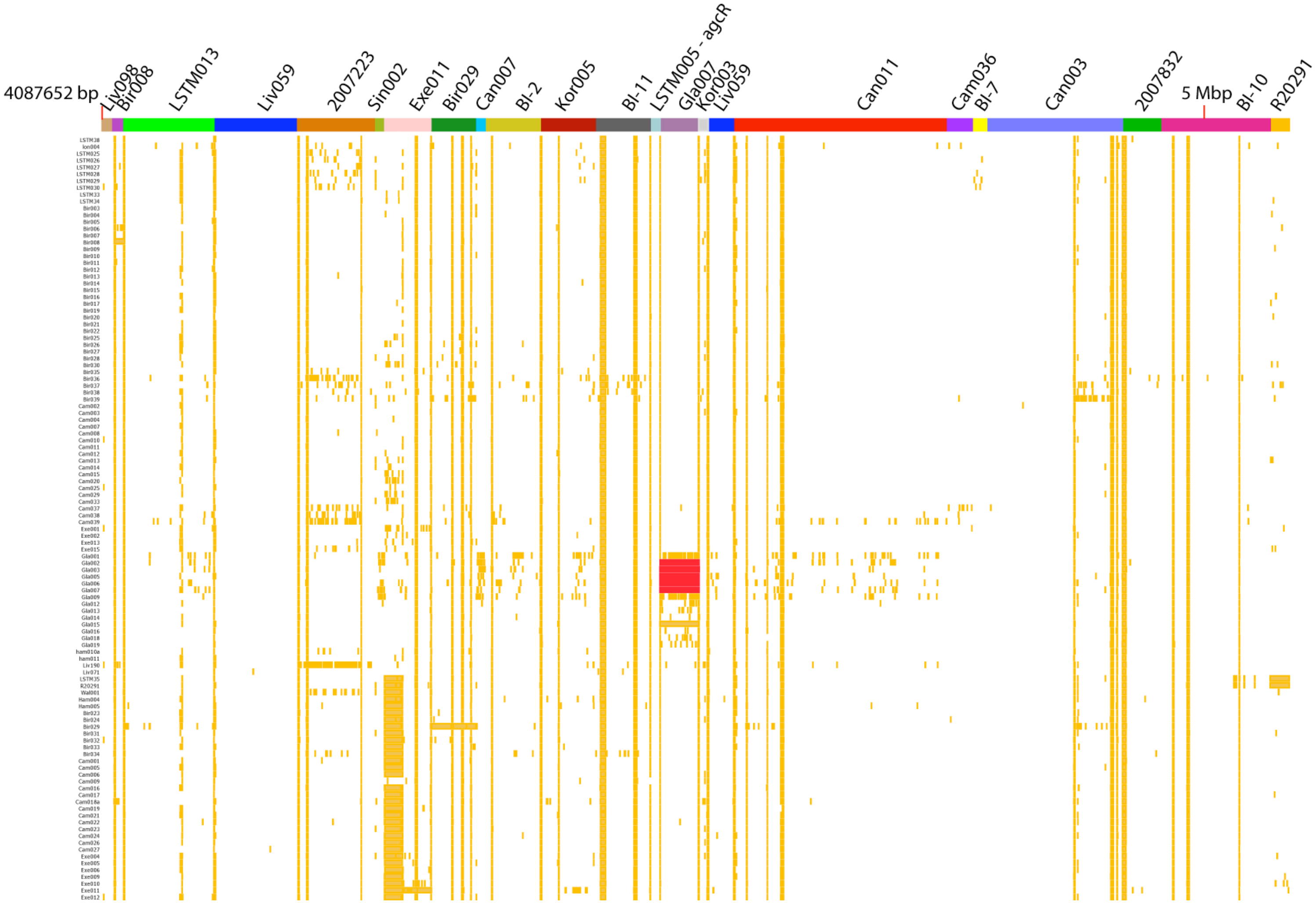


Figure 3.18: Accessory genome components of isolates in clade 3. Other descriptions for this figure are the same as Figure 3.16.

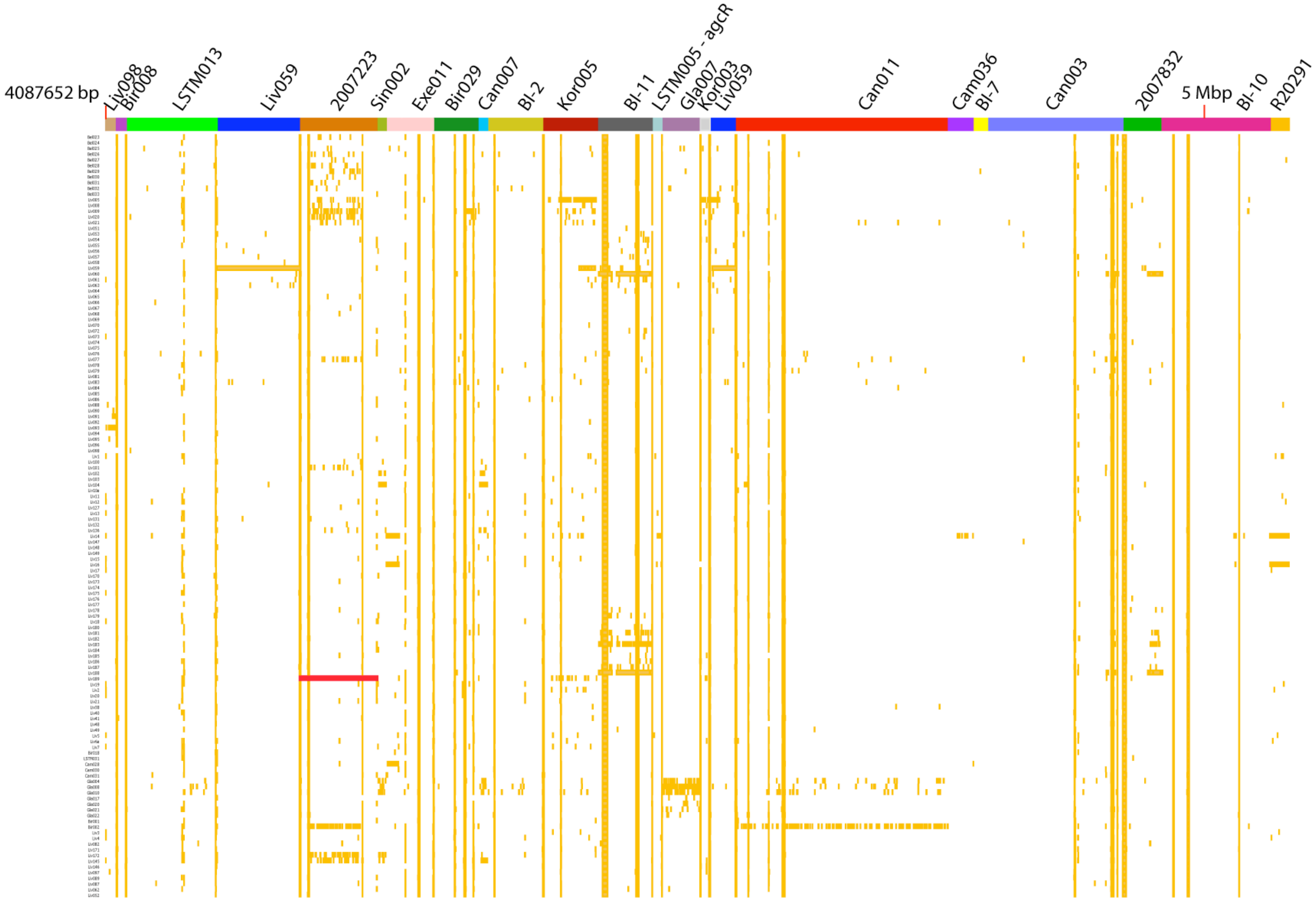


Figure 3.19: Accessory genome components of isolates in clade 4. Other descriptions for this figure are the same as Figure 3.16.

Almost all unique contigs from isolate BI-11 showed 98 – 100 percent similarity to a previously characterized bacteriophage phi*CD38-2* (Sekulovic *et al.*, 2011). The same sequences were also found in BI-2a, BI-12, 2007832, Can006, and Can009, which suggests they may also harbour phi*CD38-2* or a phage highly similar to it. A comparison between phi*CD38-2* and BI-2a is shown in Figure 3.20. However, one gene, which encodes a tail fibre protein in phi*CD38-2*, does not seem to be present in BI-2a (Figure 3.20).

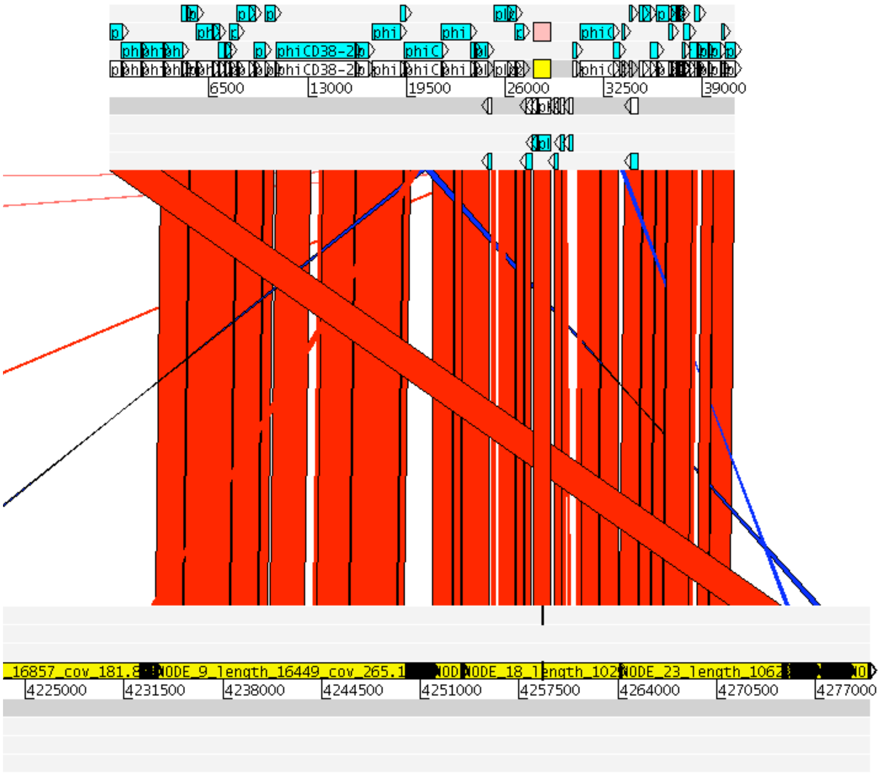


Figure 3.20: Comparison between phi*CD38-2* and unique contigs from isolate BI-2a.

In addition to the CTn*5*-like transposons discussed above, two other conjugative transposons were found in the accessory genome, both showing high similarity to known conjugative transposons in 630. One is only found in 6 isolates from Glasgow (Gla002, Gal003, Gla005, Gla006, Gla007 and Gla015) (Figure 3.18) and is highly similar to CTn*4*, but carries a set of putative antibiotic transporters different from the lantibiotic transporters in CTn*4*. It also contains a putative histidine kinase gene, which is absent from CTn*4* (Figure 3.21).

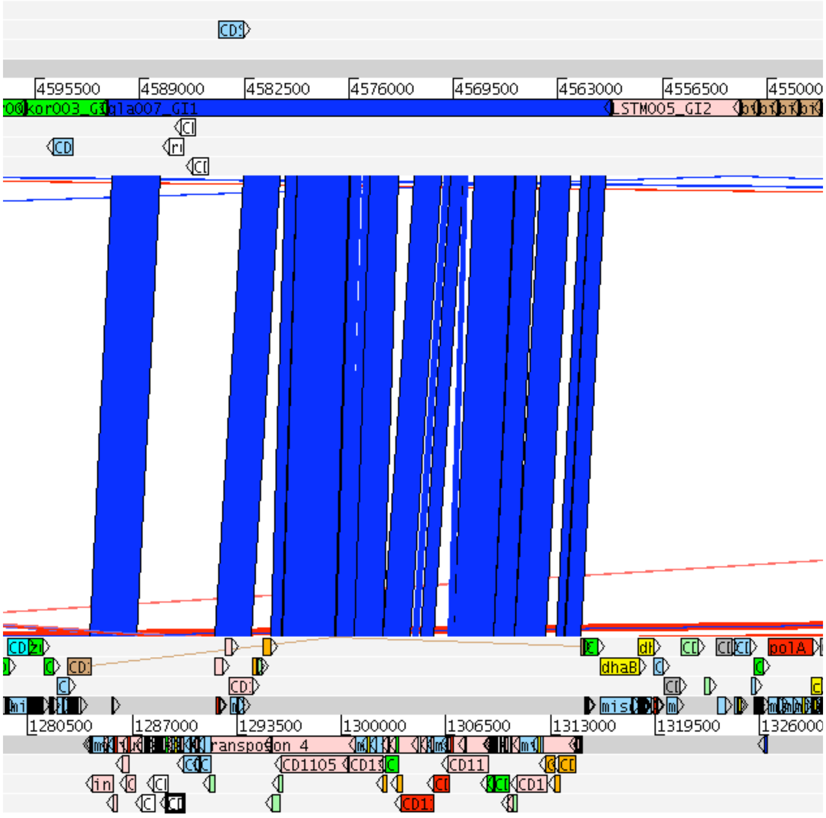


Figure 3.21: Comparison between a contig in Gla007 and CTn*4* in 630.

Another novel transposon, only found in isolate Cam036 (Figure 3.17), is highly similar to CTn*3* (Tn*5397*) in 630 and carries the tetracycline-resistance genes *tetM* and *tetL* (Figure 3.22), while only *tetM* is found in CTn*3* in 630.

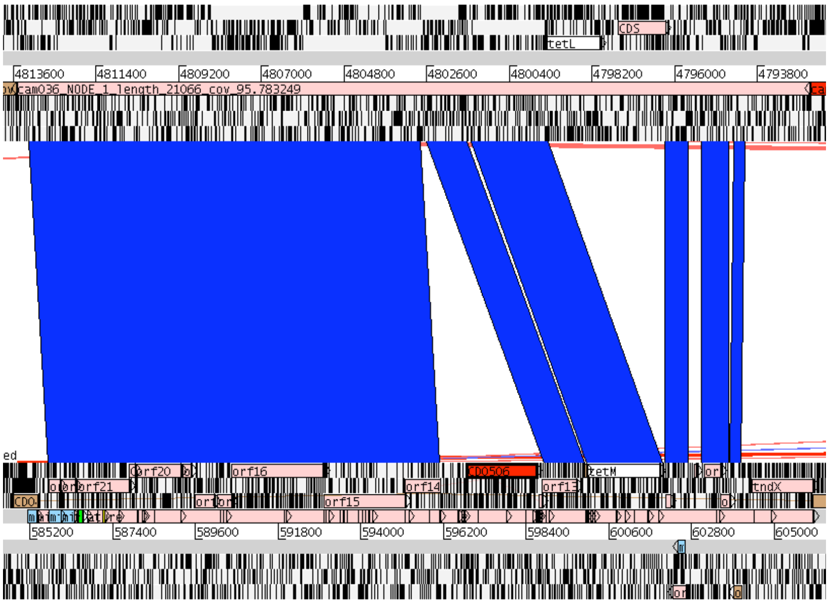


Figure 3.22: Comparison between a contig unique to Cam036 and CTn*3* in 630.

3.3.9 PaLoc region variation

The best-characterized *C. difficile* virulence factors are toxins A and B (encoded by *tcdA* and *tcdB*), which, together with two regulator genes (*tcdC* and *tcdD*) and the holin *tcdE*, form the pathogenicity locus (PaLoc). Perhaps remarkably, only two SNPs were found in the entire 19.6 kb region across 339 027 isolates. One SNP results in a premature stop codon in *tcdB* in isolate 2007825, which could lead to a truncated TcdB that lacks 203 amino acid residues at its C-terminus. Another SNP leads to a residue change (Ser419Ala) in *tcdA* gene in isolate BI-7. Both SNPs are only private to a single isolate. Thus, it is unlikely the genetic changes in PaLoc have had a large functional impact within the BI/NAP1/027 lineage.

3.4 Discussion

3.4.1 Two lineages associated with fluoroquinolone resistance

The phylogeny of this BI/NAP1/027 *C. difficile* collection consists of four well-supported clades. Previously, both studies that documented major epidemics in Canada and the USA (Loo *et al.*, 2005; McDonald *et al.*, 2005) concluded that a new fluoroquinolone-resistant variant of ribotype 027 *C. difficile* was responsible for the epidemics. However, it was not clear whether the same fluoroquinolone-resistant ribotype 027 variant was responsible for both epidemics. Based on the phylogenetic analysis, the isolates associated with these epidemics were found in two separate parts of the tree. Although both lineages contained the same mutation in the *gyrA* gene, they are well-supported, genetically distinct lineages. These findings suggest that the outbreaks were caused by different ribotype 027 variants, and that the resistance to fluoroquinolones has emerged twice independently.

Both the maximum likelihood phylogeny and the Bayesian analysis imply that a sudden expansion of fluoroquinolone-resistant BI/NAP1/027 *C. difficile* occurred; the descendants of this expansion later spread to the USA, Canada, the UK and Australia. It is less clear, however, from which country this expansion originated. Maximum likelihood phylogeny implied it started in Quebec, Canada, while the Bayesian phylogeny provided no well-supported answer to this question. As BEAST analysis is designed to infer the most recent common ancestors for isolates sampled from the present day, isolates are very unlikely to be placed on a node in the tree, which contributes to one of the more important differences between the Bayesian tree and the maximum likelihood phylogeny.

Fluoroquinolone antibiotics were one of the most commonly prescribed antibiotic classes in North America during the late 1990s and early 2000s (Linder *et al.*, 2005). Based on the inferred ages for both fluoroquinolone-resistant lineages from this sample collection, it is difficult to judge whether the *gyrA* mutation occurred prior to or during heavy use of fluoroquinolones. However, it is likely that the mutation was selected for and has spread in response to wide use of this drug.

3.4.2 Other insights drawn from the phylogeny

It is an open question whether environmental *C. difficile* from water, food or meat products leads directly to *C. difficile* infection in humans, or if the process happens in the other direction, where infection is caused by strains circulating among humans, and human *C. difficile* contaminates the environment. Isolates from animals and food sources in our collection appear to have been derived from human *C. difficile*, indicating human activity to be the source for environmental ribotype 027 *C. difficile*. However, since our sampling is biased towards human isolates, the possibility of an environmental reservoir of ribotype 027 *C. difficile* cannot be ruled out.

The observation that ribotype 176 isolates group among ribotype 027 isolates in the phylogenetic tree has important implications for *C. difficile* diagnostic and surveillance laboratories who should consider ribotypes 027 and 176 as the same genome-level variant and therefore of the same virulence potential.

3.4.3 Agreement with earlier analyses based on a smaller sample set

This more recently derived phylogeny of our BI/NAP1/027 collection agrees in most parts with the earlier phylogeny of 26 hypervirulent *C. difficile* isolates in section 2.3.1, except that a short branch leading to BI-6p, 2007837, 2004118, 2004013, 2004163 and 2007825 is present in the early tree but absent from the current one. According to the early dataset, this branch represented 2 SNPs, both with missing allele information in more than 5 isolates. The discrepancy is possibly due to the more complete allele information in the later dataset.

The results of population size inference with the Bayesian skyline plot should be interpreted with caution. However, the current inference agrees with earlier analysis in section 2.3.2 in the order of magnitude. The inferred increase at the beginning of this century is not as apparent as the earlier analysis based on 26 hypervirulent *C. difficile* isolates. In addition, the current analysis implies a two-step decrease in population size after 2007, which was not captured by the earlier analysis. These differences can possibly be explained by sampled collections used in the analyses. Apart from having a large sample size, the current collection contains strains isolated after 2007 whereas the earlier sample set does not.

3.4.4 Antibiotic resistance

Resistance to antibiotics in the BI/NAP1/027 lineage is achieved in two ways: by altering existing genes through mutation and acquiring additional genes through horizontal gene transfer. Both mechanisms were found in this dataset. Two mutations associated with resistance to fluoroquinolones were discovered in genes *gyrA* and *gyrB*, mutations of the same sites have also been found in a previous study of *S. aureus* (Harris *et al.*, 2010) and also in *C. difficile* (Spigaglia *et al.*, 2008). The mutations in the *rpoB* gene that result in resistance to rifampicin and rifaximin were also reported in *C. difficile* (OConnor *et al.*, 2008), but the mutations conferring fusidic acid resistance appear to be novel. Rifampicin and fusidic acid have been used to treat *C. difficile* infections, and reports have shown emerging resistance in *C. difficile* to these drugs (O'Connor *et al.*, 2008). The data also agree with the claim of O'Connor *et al.*, that mutations in *rpoB* were independently derived (OConnor *et al.*, 2008). The resistance to fusidic acid was also developed independently. There is no evidence for a multi-drug resistant lineage, as no isolate possesses the mutations conferring resistance to fluoroquinolones, rifampicin and fusidic acid at the same time.

3.4.5 Genetic changes underlying the success of BI/NAP1/027

It is an intriguing question as to what made BI/NAP1/027 *C. difficile* more successful in spreading around the world and causing epidemics. Two possibilities could account for this. One, this *C. difficile* variant could have become more successful through gaining fitness in a particular genetic trait, be it resistance to antibiotics, ability to evade the host immune system, or increased transmissibility. Two, it is possible that any change in genetic traits itself has not conferred significant advantage, but that a change in the environment has occurred, which has allowed this lineage to spread more rapidly. Of course, these possibilities are not exclusive, and the underlying cause could also be a combination of the above. The fact that there are two outbreak clades that share the same *gyrA* mutation raises the possibility that the increase in incidence and severity of *C. difficile* infection could simply be explained by resistance to fluoroquinolones. It is possible that common use of these antimicrobial drugs, together with refractory *C. difficile* significantly alters the microbial community in the intestine, allowing more *C. difficile* replication without restraint from an inhibitory intestinal microbiota, hence allowing a greater production of toxins and spores. At the same time, the potential impact of the diversifying changes in the accessory genome cannot be overlooked. The insertion of 14 consecutive genes (or Tn*6105*) along the branch leading to clade 2 can potentially have great impact on the biology of this organism by altering the transcription dynamics and changing other phenotypic characteristics. Still more work is required to gain a more complete understanding of the functions encoded by the accessory genome. Even more experimental analysis will be needed to confirm their functions.

3.4.6 Potential mechanisms for large chromosomal region replacement

C. difficile has a highly dynamic genome, as first shown by Sebaihia *et al* (Sebaihia *et al.*, 2006) and later confirmed by other reports (Janvilisri *et al.*, 2009; Scaria *et al.*, 2010). Consistent with findings stated in the previous chapter, large homologous recombination blocks were found in several isolates in this BI/NAP1/027 collection, which otherwise possess a highly similar genomic backbone. *C. difficile* is not known to be naturally transformable. Considering the vast number of mobile elements within the genome and the large sizes of homologous recombination blocks, chromosomal mobilization mediated by mobile elements are therefore the likely mechanisms for genetic material transfer between isolates. Many of the recombination blocks discovered in 3.3.4 are situated adjacent to conjugative transposons, phages or transposes. Hfr-type chromosomal mobilization from multiple sites in the genome was previously suggested for *S. agalactiae* (Brochet *et al.*, 2008), although other mechanism different from Hfr-type DNA transfer can also result in replacements of large chromosomal regions, as has been shown in *Mycobacterium smegmatis* (Wang *et al.*, 2005). It is possible that *C. difficile* adopts similar mechanisms for horizontal DNA transfer.

However, as no putative mobile elements were found adjacent to some of the homologous recombination blocks, other mechanisms remain to be identified.

Chapter 4

Hospital transmission and persistence of *C. difficile* from a whole genome sequencing perspective

4.1 Introduction

The majority of *C. difficile*-associated disease cases are diagnosed within hospitals and healthcare facilities, although community-acquired *C. difficile* disease exist. Indeed, almost all *C. difficile*-associated outbreaks have occurred in hospitals and healthcare facilities. *C. difficile* spores are resistant to heat and commonly used disinfectants, including 70% ethanol (Lawley *et al.*, 2009) and the resilient nature of *C. difficile* spores contributes to their high transmissibility. Spores can potentially persist in the environment for months under traditional routine cleaning regimes, while maintaining their transmissible nature (Gerding *et al.*, 2008). Additionally, *C. difficile* can persist in the spore form in gnotobiotic mice (Onderdonk *et al.*, 1980). It is unclear, however, what fraction of CDI cases arise due to infections mediated by spores directly originating from the environment.

One major difficulty in treating *C. difficile*-associated diseases is the recurrence of infection. Recurrent CDI is fairly typical, it can be found in 5%-35% of patients (Bakken, 2009; Johnson, 2009) and may occur months or years after the initial infection was resolved (Johnson, 2009). The symptoms of recurrent cases are frequently indistinguishable from the previous infection scenarios (Bartlett, 2010). One of the factors behind recurrent CDI has been postulated to be the persistence of spores, either endogenously within the host or in the environment. Theoretically, re-infection can be caused by the same or a different strain independent of the nature of the source (Johnson, 2009; Johnson *et al.*, 1989). Early studies have used REA to type isolates from the same patient at multiple time points in order to discriminate between relapse and re-infection (Johnson *et al.*, 1989; ONeill *et al.*, 1991). These results generally indicate that approximately 50% of re-occurring infections are caused by a new *C. difficile* strain, supporting re-infection rather than relapse (Johnson *et al.*, 1989; ONeill *et al.*, 1991). A recent review summarized five published studies and concluded that 33%-75% of the recurrent cases are due to re-infection with a new strain. However, in these cases the question of whether re-infection was caused by strains from endogenous sources or the environment remained unresolved (ONeill *et al.*, 1991). Additionally, it is unclear whether patients can be colonized simultaneously with multiple *C. difficile* strains. O'Neill *et al.* investigated the possibility of multiple carriage using REA and claimed that such incidents are rare (ONeill *et al.*, 1991).

The accuracy of these studies depends on suitable typing methods. An important question is, how should we define strain types? REA may have sufficient discriminatory power to identify isolates belonging to different ribotypes, but to differentiate isolates within the same ribotype, particularly ribotype 027, which shares a highly similar genomic backbone, a more discriminatory method is needed. The power of whole genome sequencing in discriminating between ribotype 027 isolates was demonstrated by the data outlined in the previous chapter. Here these techniques were used to analyze 027 isolates from the same hospital and its associated areas, including isolates from the same patient. In addition to differentiating relapse from re-infection, a goal of this study was to gain some level of understanding of local transmission of *C. difficile* through an analysis that combined phylogenetic information and spatial/temporal data of isolates. Comparative analysis of multiple samples from the same patient can also potentially provide insights into genetic changes occurring in *C. difficile* over time while being carried by individual patients.

As a comparison, a murine infection model (Lawley *et al.*, 2009) was utilized to monitor the genome change of ribotype 027 *C. difficile* during long-term colonization in mice. In an experiment conducted by Lawley *et al.*, five mice exhibit different outcomes in colonization levels and disease following infection with the same *C. difficile* ribotype 027 strain and subjected to the same antibiotic treatment. The hypothesis is that the differences detected in the genome over time could be attributed to factors such as the host immune system or competition within the intestinal microbiota of individual mice. Selective pressures could be reflected in changes in the *C. difficile* genome before and after colonization.

The aims of the analysis in this chapter were: -

- to differentiate between relapse and re-infection cases in diseased patients;
- to investigate the possibility of carriage of different strains by a single patient;
- to explore the use of whole genome sequencing in understanding local transmission of ribotype 027;
- to assess the level of genome changes in ribotype 027 in mice and humans, and the possible genetic and phenotypic consequences of these changes

4.2 Materials and methods

4.2.1 Bacterial isolates

4.2.1.1 Ribotype 027 isolates from patients

The *C. difficile* hospital collection included 127 *C. difficile* isolates sampled between July 2008 and May 2010 and two retrospective isolates from May and July of 2007. These isolates were sampled from patients at the Royal Liverpool University Hospital (117 isolates) and neighbouring hospitals, including some visiting general practitioners (12 isolates). These isolates were selected and cultured by Paul Roberts and Fabio Miyajima and colleagues at Liverpool University Hospital, in the following steps: Faecal specimens were collected from patients who are suspected of having CDI based on clinical symptoms. The stool specimens were then tested for *C. difficile* toxins A and B using ELISA. *C. difficile* was then cultured from toxin-positive stool samples on Brazier's plates. Five to ten colonies were then inoculated onto fastidious anaerobe agar plates to check for purity. At least one purified colony from a single patient was collected and used to determine PCR ribotype. Only confirmed ribotype 027 isolates were included in this study.

For fourteen patients (named patient A to N), two or three isolates were collected from different infection episodes at an interval of one to eight months. Additionally, five more colonies were sampled from the primary culture at the first infection episodes of patients C, F, I, and the second infection episode of patient H. The full details of these isolates, including sampling dates and hospital locations where the patients are residing, are given in Appendix B.

4.2.1.2 *C. difficile* BI-7 (ribotype 027) isolates obtained over time during a mouse colonization experiment

Colonization with ribotype 027 *C. difficile* strain BI-7 (clindamycin resistant human isolate) was established by infecting C3H/HeN mice via oral gavage containing 10^7 CFU of culture grown organisms. The mice were subsequently treated with clindamycin for seven days. The resulting colonization condition was monitored by culturing *C. difficile* directly from the faeces (Figure 4.1). In this model, mice 1 and 3 became low-level carriers of *C. difficile* ($<10^2$ CFU/gram faeces); mice 2 and 5 exhibit moderate-level carriage (10^4-10^6 CFU/gram faeces), while mouse 4 remained a high-level excretor of *C. difficile* ($>10^8$ CFU/gram faeces) for extended periods. The experiment was carried out over a 90 day period post-infection. Mouse 4 displayed chronic intestinal inflammation, while the other four mice did not exhibit any significant intestinal pathologies.

The samples of *C. difficile* selected for whole genome sequencing were: twelve colonies from the day 0 input inoculum; twelve colonies from each of mouse 3, 4, and 5 immediately after clindamycin treatment, and similar samples collected at 90 days post-infection. This added up to a total of 84 samples. All the *in vivo* work described in 4.2.1.2 was carried out by Drs. Trevor Lawley and Simon Clare at WTSI.

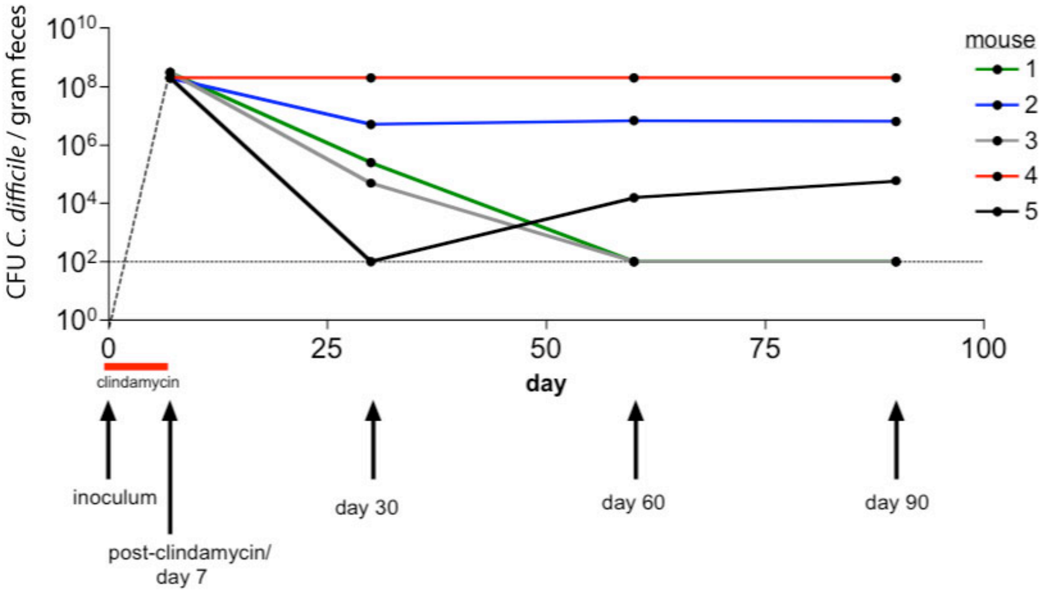


Figure 4.1: The pattern of long-term infection of C3H/HeN mice with *C. difficile* strain BI-7 (ribotype 027). Reproduced from Dr. Trevor Lawley from WTSI.

4.2.2 DNA preparation, sequencing, reads mapping and SNP detection

These steps were carried out as described in section 3.2.2. DNA preparation and sequencing were performed by Louise Ellison, Derek Pickard and the Sequencing Team at WTSI respectively.

4.2.3 Phylogenetic analysis

Phylogenetic relationships were inferred with the program PHYML (Guindon and Gascuel, 2003) with 100 bootstraps. A simple GTR model was used; the model assumes all sites evolve at the same rate, with no invariable sites.

4.3 Results

4.3.1 Genetic diversity and microevolution of hospital ribotype 027

A total of 127 ribotype 027 isolates were sampled from the Royal Liverpool University Hospital and adjacent hospitals between July 2008 and May 2010. Almost all ribotype 027 isolates underlying confirmed CDI cases during this period were included. Two isolates from 2007 were also added to the collection, making the total number 129. All isolates were sampled from patients with confirmed CDI. Ribotype 027 was the most common circulating ribotype in the sampled hospitals within this period, though its relative prevalence has dropped from 50% in 2008 to 31% in 2010 (Table 4.1). The total number of CDI cases increased from 2008 to 2009 but decreased in the later half of 2009.

	027	106	001	002	014/ 020	015	078	005	023	All others	Total
Jun/08 - Nov/08	17	8	0	0	0	0	0	1	1	7	34
Dec/08 - May/09	53	16	7	7	0	0	2	3	0	11	99
Jun/09 - Nov/09	32	14	0	2	3	7	2	3	1	8	72
Dec/09 - May/10	10	5	2	2	2	1	6	0	0	4	32

Table 4.1: Number of *C. difficile* isolates per semester by ribotypes. Data provided by Fabio Miyajima at Royal Liverpool University Hospital.

A total of 70 SNPs from the non-repetitive core genome were discovered among these isolates and an un-rooted maximum likelihood phylogeny was constructed (Figure 4.2). The phylogeny suggests more than 20 distinct genotypes were circulating within the hospitals. Two isolates (Liv071 and Liv190) appear to be outliers in this phylogeny and are differentiated from the other isolates by 2 SNPs (1163835 and 1374216, positions refer to that in R20291 genome). All other isolates can be divided into three groups (A, B, C) based on their genotypes. Based on the phylogeny, group C was derived from group B. There are very few SNPs that differentiate between groups. One SNP (4150703) differentiates between groups B, C and the rest; and one SNP (2979027) differentiates between groups B and C. Despite the paucity of genetic variation that defines each group, the groupings of isolates can be considered reliable, as the divisions are based on SNPs whose alleles have been determined in all the clustered strains, except in the case of SNP 2979027, where the alleles in two isolates (Liv8 and Liv9) can not be determined. In addition, group A, three branches in group B, and four branches in group C all received >80% bootstrap support. These sub-groupings were given the names B10 – B30 and C10 – C40 (Figure 4.2).

There is very little variation within a single genotype group. The numbers of SNPs within each group are 2, 30, and 22 for A, B, and C respectively. Comparing this dataset within the global ribotype 027 dataset in Chapter 3 reveals that group C contains local samples exclusively (only samples from Royal Liverpool University Hospital and neighbouring hospitals), while group B isolates are found in the same genotype cluster as human isolates from Glasgow and Belfast, and group A isolates are found in the same genotype cluster as human isolates from Birmingham. Thus, it is possible that groups A and B reached this region independently, and the emergence of group C occurred locally in Liverpool. Apparently, all three genotypes have persisted and continued to evolve after being introduced into this area.

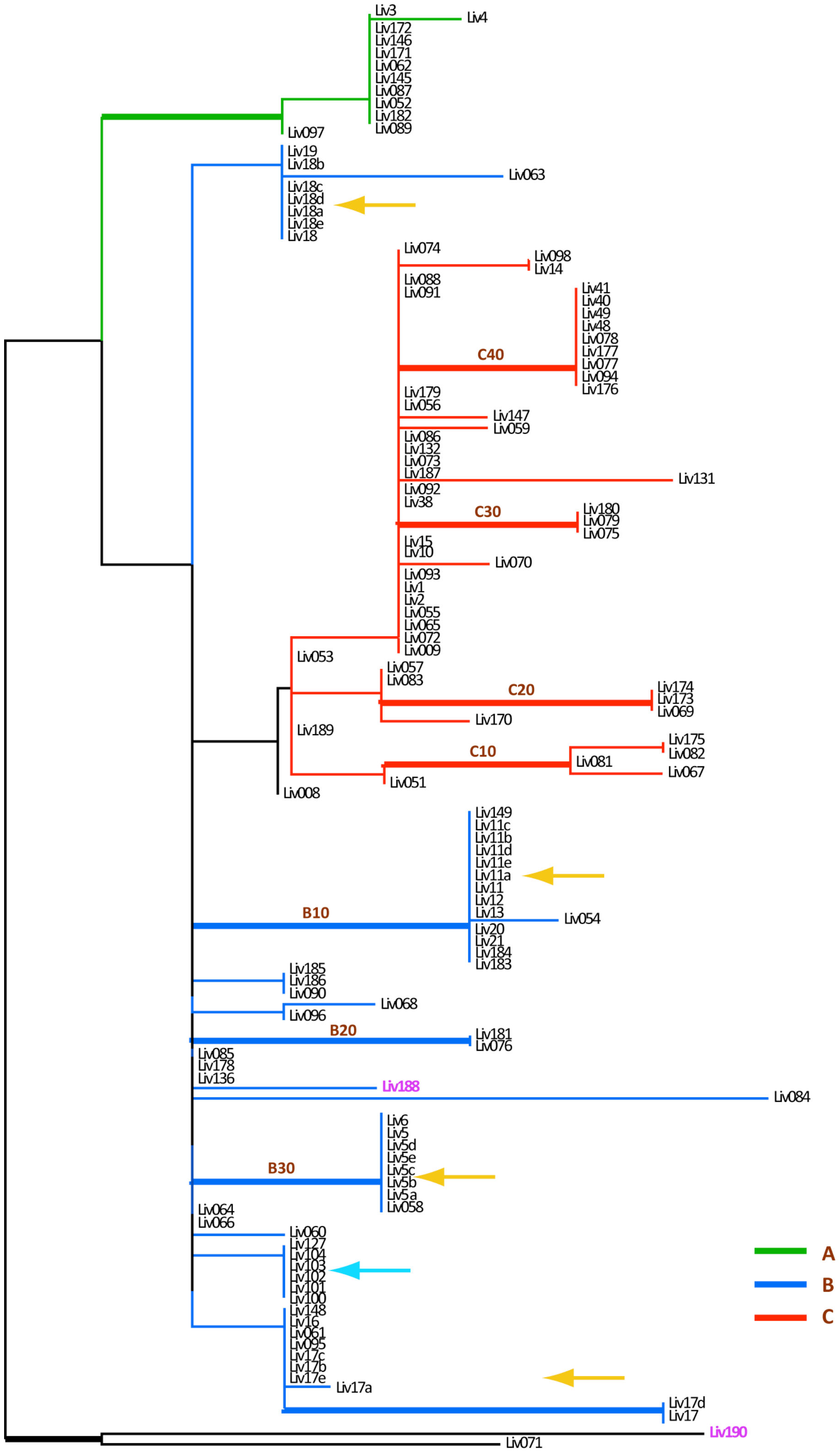


Figure 4.2: Un-rooted maximum likelihood phylogeny of Liverpool isolates. Branches are coloured according to isolate groupings. The groupings supported by > 80% bootstraps are highlighted by thick branches, with associated names assigned for each genotype subgroup labelled in brown. Yellow arrows point to multiple samples taken from the same patient at the same time point. Blue arrow denotes isolates that

exhibit spatial temporal clustering (discussed in 4.3.3). Two isolates from 2007 are shown in pink. Liv008 is labelled in black as its grouping (either B or C) cannot be determined due to missing allele information.

4.3.2 Colonization by *C. difficile* 027 in patients

This study includes samples taken from individual patients sampled at multiple time points. Here the aim was to gain insight into whether recurrent disease is caused by relapse of the same or re-infection with a genotypically distinct *C. difficile* 027 strain. It is worth noting that this collection only includes isolates confirmed to be ribotype 027. During the sampling period, twenty-six patients exhibited at least two separate infection episodes. Among these, two patients were infected with bacteria of a different ribotype at the recurrent infection point (data provided by Fabio Miyajima at Royal Liverpool University Hospital), supporting re-infection with a distinct strain.

Fourteen patients appear to be infected with ribotype 027 *C. difficile* at all monitored infection episodes. Figure 4.3 shows the sampling time points and genotype assignments of the associated *C. difficile*. In all 14 cases, isolates of the same genotype group were found during the subsequent infection period(s) within individual patients except for Patient D, whose case is difficult to determine as Liv8 cannot be assigned to a group with confidence. However, isolates from different infection period(s) of the same patient were not always identical. Sample Liv4 is differentiated from Liv3 by 1 SNP (3283560); these two samples are separated by an interval of 6 months (Figure 4.3). Strikingly, Liv17 is differentiated from Liv16 by 4 SNPs (1767576, 2384792, 2550185 and 3632130). As the two samples were separated by only 5 weeks in isolation date, it could be considered unlikely the 4 SNPs arose during this interval within the same *C. difficile* lineage as it replicated. In addition to Liv8 mentioned above, isolates Liv9, Liv14 and Liv15 also contain missing allele information. This makes it difficult to determine the genetic difference in each sample pair for Patient E (Liv9 and 10) and Patient G (Liv14 and 15).

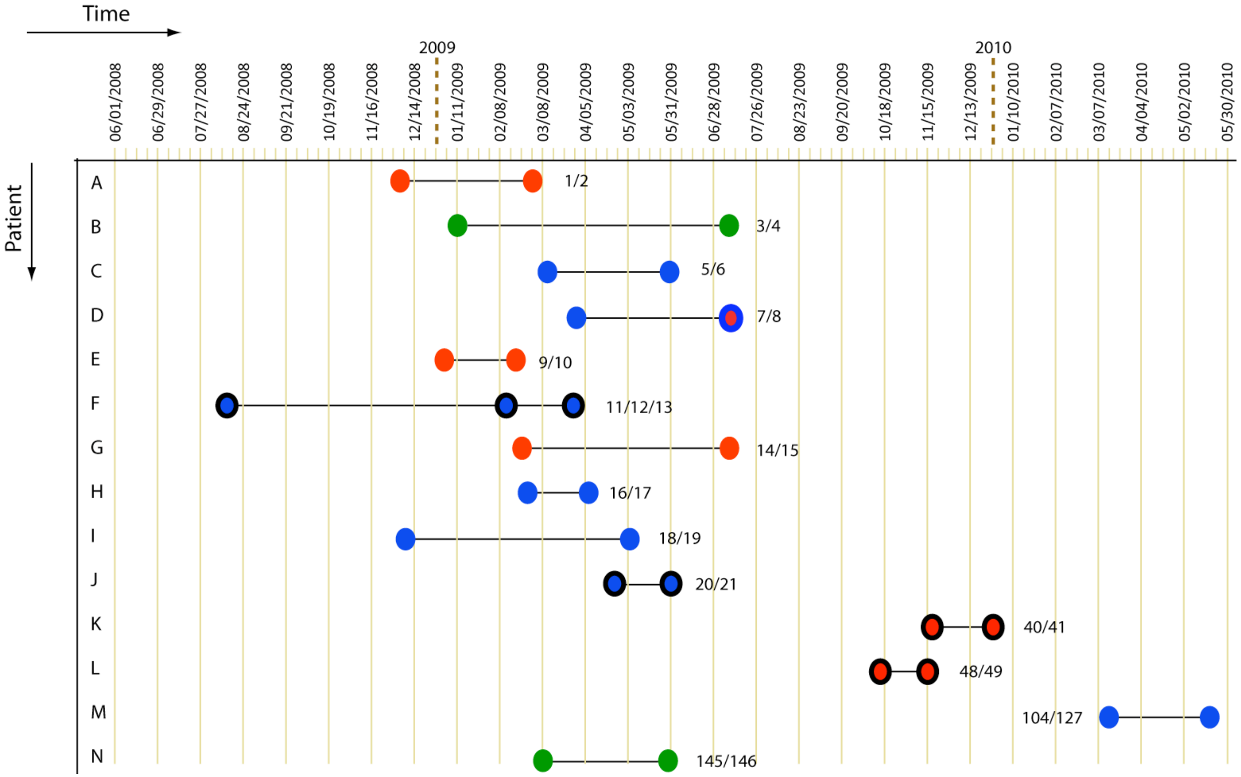


Figure 4.3: Temporal graph of *C. difficile* ribotype 027 genotype assignment from the same patient at multiple infection episodes. Each row depicts two (or three) isolates sampled from the same patient at different time points (dates labelled on top). The isolates are represented by circles, each is coloured according to the genotype group it is associated with, as in Figure 4.2. Circles with black edges represent genotype B10 (blue) and C40 (red) respectively. Liv8 is shown in red and blue as both are possible due to missing allele information. The numbers labelled next to the isolates are sample identifiers.

To investigate the possibility of multiple carriage, five more samples were collected at the first infection episodes of patients C, F, I, and at the second infection episode of patient H; these samples were named Liv5a-e, Liv11a-e, Liv17a-e and Liv18a-e (Figure 4.2), according to the original sample names Liv5, Liv11, Liv17 and Liv18, respectively.

The analysis shows that Liv5a-e, Liv11a-e, and Liv18a-e exhibit an identical genotype to their corresponding original samples. While Liv17a-c and Liv17e are identical, they are differentiated from Liv17 and Liv17d by 4 SNPs. This potentially suggests carriage of multiple lineages, or multiple variants of the same lineage by Patient H. However, a more comprehensive analysis involving a larger sample set of *C. difficile* and patients would be required to firmly resolve the relationship between isolates differentiated by a few SNPs in terms of cross infection between patients compared to evolution within the same patient.

4.3.3 Spatial and temporal distribution of ribotype 027 genotypes

As a next step, the genotypes of Liverpool *C. difficile* isolates were mapped to the date and hospital location where the patient was residing when each sample was isolated. Figure 4.4 illustrates the spatial and temporal information with date of isolation indicated on the x-axis and hospital floors or neighbouring hospitals indicated on the y-axis. Coloured circles (matching Figures 4.2 and 4.3) depict sampled *C. difficile* 027 isolates from infected patients and the associated genotypes. The two isolates from 2007 (Liv188 and Liv190) were not included in Figure 4.4.

The spatial temporal graph reveals a general lack of clustering between genotype and hospital location, implying frequent transmission within the hospital and between neighbouring hospitals. However, persistence of the same genotype group can be seen within certain time periods and spatial locations. For example, during the entire sampling period, Floors 7 (Diabetes/Endocrinology/Cardiology/Haematology) and 9 (Renal transplant/Nephrology/Rheumatology/Urology/Breast) were found to harbour group B genotypes almost exclusively (except Liv182, which belongs to group A). In addition, only group A isolates were found in wards 5XY (Gastroenterology) between November 2008 and February 2009. Group C genotypes appear to dominate between October and December of 2009 (pink shaded area in Figure 4.4). A tight cluster of C40 genotype was found in wards 2XY throughout July 2009 (pointed by an arrow in Figure 4.4). Interestingly, Liv100-102, Liv104, and Liv127, which share an identical genotype and occupy a single lineage in the phylogeny (blue arrow in Figure 4.2), form a cluster between March and May 2010 (yellow shaded area in Figure 4.4). Another observation is that isolates that were derived later in terms of their phylogeny (on branch tips) were not necessarily found at a later time in the sampling period. Isolate Liv188 from 2007 (shown in pink in Figure 4.2) is found at a branch tip in Group B. These findings seem to imply that the emergence of genotypes discovered in the collection predates the sampling period. Although genotypes B and C are more common than A in our collection, it is unclear whether this phenomenon is random or a result of difference in fitness.

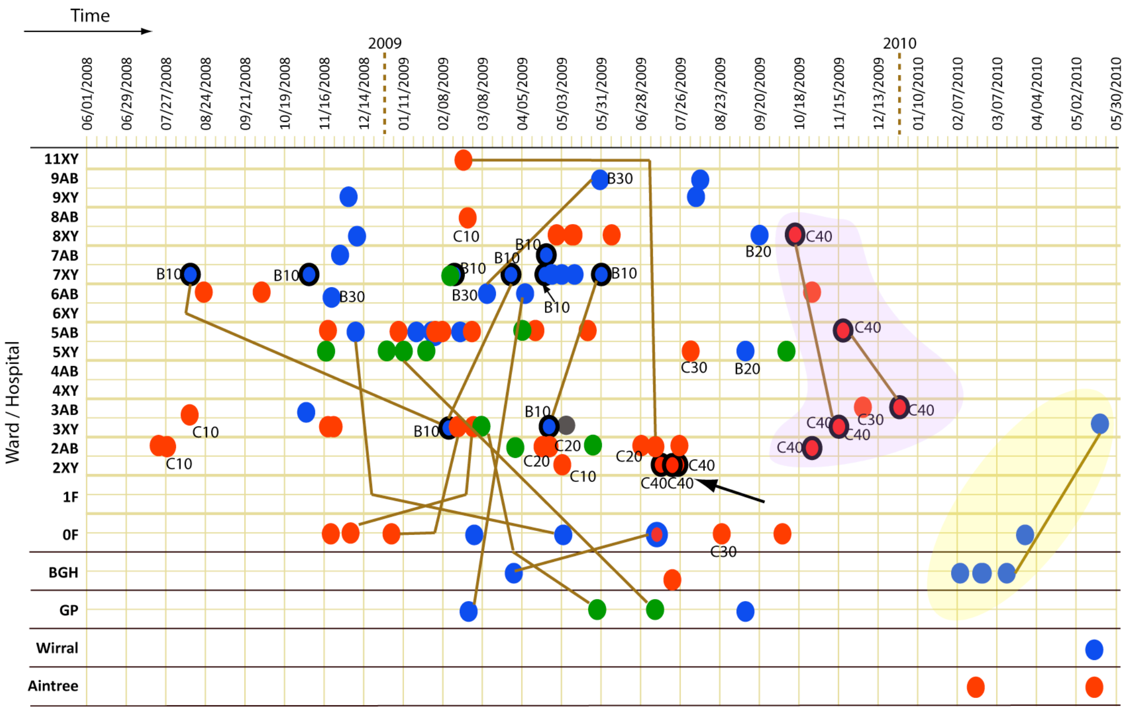


Figure 4.4: Temporal and spatial graph of isolates from local hospitals in Liverpool region. Rows depict hospitals (BGH, Wirral, Aintree), visiting general practitioners (GP) or wards from a single large hospital (0F – 11XY). Time is depicted on the x-axis, with dates labelled on top. Every four weeks is represented by a major division, shown with yellow lines. Minor divisions on the x-axis represent 1 week. The isolates are represented by circles, each is coloured according to the genotype group it is associated with, as in Figure 4.2. The uncoloured circle represents Liv071, which belongs to none of groups A-C. The numbers labelled next to the isolates match genotype names from Figure 4.2. Circles with black edges represent genotype B10 (blue) and C40 (red) respectively. The isolate shown in red and blue is Liv8, which cannot be assigned with confidence to either B or C group due to missing allele information. Isolates from the same patient are connected by brown lines. Shaded areas and the arrow highlight three genotype clusters.

4.3.4 Genetic diversity of ribotype 027 during colonization of mice

An experiment to study aspects of the persistence of ribotype 027 colonization in mice was carried out by Drs. Trevor Lawley and Simon Clare at WTSI. These results showed that five individual mice exhibited different outcomes in terms of their respective colonization levels and disease symptoms following infection with the same dose of *C. difficile* ribotype 027 strain (BI-7) and the same antibiotic treatment (Figure 4.1). This experiment provided an opportunity to sample the genomes of individual *C. difficile* lineages as they persisted *in vivo* in the mouse over time. How quickly does the genome change, if at all? These data could also be used to compare with the hospital patient dataset. Consequently, *C. difficile* samples were collected at three time points: the input inoculum (Day 0), immediately after clindamycin treatment (Day 7), and at 90 days post-infection (Day 90). For Day 7 and Day 90, samples were collected from each of mouse 3, 4, and 5 (more details in 4.2.1.2). Twelve colonies were collected for each time point for each mouse, resulting in 84 isolates in total. The samples were named Day0-1 to 12; Day7-M3/4/5-1 to 12, and Day90-M3/4/5-1 to 12.

Sample Day90-M5-7 yielded little data and was excluded from the analysis. Five samples (Day7-M3-1, Day7-M3-2, Day7-M5-9, Day7-M5-12, and Day7-M4-1) were found to be non-027 *C. difficile*, as shown by >22,000 SNPs when compared to the genome of R20291. Day7-M3-1, Day7-M3-2, Day7-M5-9, Day7-M5-12 were discovered to be highly similar to ribotype 002 isolates, and Day7-M4-1 to M120 (ribotype 078). These were considered to be contamination, most likely occurring within the anaerobic cabinet at the time (subsequently containment protocols have been tightened considerably to control this highly transmissible organism). In the remaining 78 isolates, only 1 intergenic SNP (a G->C change at position 3876023) was found in Day7-M5-5, while all the other samples were identical. The fact that this SNP is present in one sample at Day 7 but absent from all samples at Day 90 suggests it was not fixed in the population.

4.4 Discussion

4.4.1 Strengths and limitations of spatial temporal genotype analysis

In this chapter, genetic diversity was analyzed for *C. difficile* ribotype 027 isolates from infected patients within a single hospital and its local capture areas. This analysis shows that the *C. difficile* 027 circulating in the sampling area have highly similar but in some cases distinguishable genotypes. Phylogenetic analysis can be used to divide these genotypes into three main groups, which differ in terms of their dominance over time and space. Interestingly, although not dominant, genotype clustering is more apparent during the latter part of the sampling period (October 2009 to May 2010), while a large number of genotypes co-exist earlier, particularly in the first half of 2009. This could be related to the fact that there were fewer reported CDI cases in late 2009 and 2010 (Table 4.1), which is concurrent with fewer local transmission events, therefore the same genotype is more likely to persist longer.

The co-existence of multiple genotypes and the overall lack of general clustering implies frequent transmission, underpinning the importance of applying frequent deep and effective disinfection regimes. However, this analysis did not identify obvious significant transmission events within the hospital. One reason could be the absence of carriers in the sampling. This collection only includes isolates from patients with clinical disease. Further, environmental samples were not included. Asymptomatic carriers of *C. difficile* can also excrete spores, which become potential sources of infection (Peach *et al.*, 1986). The percentage of asymptomatic carriers of *C. difficile* was found in an independent study to be 32% in patients with cystic fibrosis (Peach *et* *al.*, 1986). It is possible that data from such carriers will help construct transmission chains.

The hospital isolates analyzed in this chapter are highly similar, with many sharing identical genotypes, making them difficult to differentiate even by whole genome sequencing. Although the genotype groups A – C were supported by almost all alleles in the relevant strains, verifying the alleles using an independent genotyping platform would be useful.

4.4.2 Relapse, re-infection and multiple strain carriage

The analysis presented here suggests that the majority of recurrent CDI cases in this dataset are due to relapse rather than re-infection. However, this conclusion should at this stage be regarded as not being conclusive. Nine out of fourteen patients studied demonstrate *C. difficile* isolates of identical genotype for different infection episodes, which can be interpreted as evidence for relapse. However, it is also arguable that these patients could have acquired isolates of the same genotype from exogenous sources after the initial infections, since these genotypes are fairly prevalent during the study period. Patient H shows evidence for harbouring multiple strains, as the two genotypes differ by 4 SNPs, which is considerable divergence within this dataset. However, even here we cannot rule out the unusual accumulation of these SNPs by the same lineage within this patient. It is also intriguing that the genotype of Liv17/Liv17d (Figure 4.2) was unique to this patient. It is possible that the 4 SNPs occurred long before this study period (within Patient H or not), and both genotypes have been carried by Patient H since. The missing data in Liv8, Liv9, Liv14 and Liv15 makes it difficult to judge whether relapse or re-infection is true for Patient D, E and G, while the single SNP difference between Liv3 and Liv4 (isolates from two infection episodes of Patient B, collected 6 months apart) could be interpreted as either bacterial evolution within the host or re-infection by a new strain. The finding that the

Liv4 genotype is unique in our dataset implies that the former interpretation is more likely. All this analysis emphasises the importance of collecting larger comparative datasets.

The lack of obvious evidence for re-infection by a new strain in our dataset is in contrast with the early reports (Johnson *et al.*, 1989; ONeill *et al.*, 1991). However, as this study only focused on ribotype 027 isolates, the proportion of this type of re-infection might be expected to be lower. Multiple strain carriage was found in 1 out of 4 cases tested in this study, more common than suggested by O'Neill *et al.* (ONeill *et al.*, 1991). However, a recent report focused on ribotype 027 found more than one MLVA profile in 5 out of 39 faecal specimens, and suggested multiple strain carriage and rapid evolution as possible reasons (Tanner *et al.*, 2010). The two patients who exhibit *C. difficile* of different ribotype at two infection points strongly support exogenous re-infection, if not carriage of multiple strains. Sequencing multiple samples from patients will help us to understand this more clearly. It would also have been potentially more informative if the original sampling study was not limited to ribotype 027.

4.4.3 Insights from *C. difficile* colonization in mice

During a 90-day colonization period in mice, the *C. difficile* 027 strain BI-7 did not exhibit any detectable change in the core genome. One mutation arose but did not become a fixed variant. It is worth noting that the 3 mice showed different colonization levels of *C. difficile* and disease pathology, even though no genetic difference was discovered between the isolates they carried. This implies the colonization outcome can be more attributed to host factors, such as the immune system or the intestinal microbiota composition of the individual mouse, rather than the bacterial genotype.

Finally, the mutation rate of *C. difficile* ribotype 027 was estimated to be 1-2 mutations per genome per year in Chapter 3 and the current data are generally in agreement with this estimate.

Chapter 5

Final Discussion

In this thesis whole genome sequencing techniques were used to analyze genetic variation and the evolution of *C. difficile*, a recently emerged cause for antibiotic-associated diarrhoea. At the start of this project, only one completed *C. difficile* genome sequence was available (Sebaihia *et al.*, 2006). Through the analysis of this genome and sub-genomic comparisons with other *C. difficile*, it had been proposed that the genome of this species is highly dynamic; complementary analysis of *C. difficile* genetic variation had been carried out using MLST (Lemee *et al.*, 2004) and comparative genomic hybridization approaches (Stabler *et al.*, 2006). Through comparative analysis of multiple genome sequences between and within different ribotypes, further understanding was gained with respect to the genetic diversity of this species. These studies indicate that diversity is relatively large between ribotypes, but limited at least within ribotype 027, a recently emerged lineage associated with hospital outbreaks. The finding that multiple lineages are associated with virulence is consistent with previous results based on MLST data (Lemee *et al.*, 2004), suggesting genetic elements common to a number of *C. difficile* isolates underlie disease.

5.1 Significance of homologous recombination

In addition to supporting the previous suggestion that *C. difficile* has a highly dynamic genome (Sebaihia *et al.*, 2006; Stabler *et al.*, 2006), data outlined in Chapter 2 has also highlighted horizontal gene transfer and homologous recombination as two important mechanisms underlying this diversity. This is the first time homologous recombination involving large chromosomal region exchange was demonstrated in *C. difficile*. The impact of homologous recombination was previously considered to be low (Lemee *et al.*, 2004), but the findings in Chapter 2 and Chapter 3 suggest it is not a negligible influence, as 15% of the CF5 genome sequence can be attributed to imports, and homologous recombination blocks of >100 kb were found in multiple isolates belonging to ribotype 027. Identifying variants resulting from homologous recombination is an important consideration for future phylogenetic analysis of *C. difficile*, as these phenomena are sufficient to distort branch lengths in the phylogenetic tree.

5.2 Insights from the study of a global collection of BI/NAP1/027

The analysis in Chapter 3 presents the first detailed study of global transmission and whole genome evolution of a particular successful lineage of *C. difficile* – ribotype 027. The results show that resistance to fluoroquinolones has emerged twice independently, resulting in two resistant lineages, one of which has an apparent origin in the USA and which later spread to South Korea. Descendants of the other lineage include the majority of UK and all Australian isolates, although the origin of this lineage is unclear. In addition to showing rapid spread across continents, the analysis highlighted important genetic changes that are likely to underlie the success of modern day *C. difficile* BI/NAP1/027, including novel genomic islands and elements conferring antibiotic-resistance. The two genomic islands,Tn*6104* and Tn*6105*, found only in more recent BI/NAP1/027 isolates (Chapter 2 and Chapter 3) present interesting cases of genetic acquisition that can be pursued further through phenotypic analysis. The regulatory CDSs carried by each imply they have the potential to influence the transcriptome of modern day BI/NAP1/027 and therefore fitness or virulence. More insights could be gained by comparing gene expression between isolates with and without these genomic islands, perhaps by exploiting naturally existing BI/NAP1/027 isolates or genetic mutants obtained under laboratory conditions.

5.3 Selective pressure and gene candidates for functional study

The investigation of selective forces acting on the whole genome (Chapter 2) confirmed it is under purifying selection over the long time frame. Possibly more meaningful are the identification of CDSs under positive selection (Chapter 2) and CDSs harbouring homosplasic SNPs (Chapter 3). Although both studies yielded fairly small sets of CDSs, both contain a number of surface proteins and regulators, including membrane proteins, a putative exported protein, a putative signalling protein and a two-component regulator pair. The latter set also contains three CDSs that are known antibiotic drug targets. These surface proteins and regulators could potentially have significant functional impact on the organism, possibly through interacting with host immune systems or modifying gene transcription in bacteria. They represent key subjects for future functional study.

5.4 Insights from local hospital transmission study

Chapter 4 explored the use of whole genome sequencing in monitoring local transmission of *C. difficile* BI/NAP1/027. This is the first time high-throughput whole genome sequencing was used to analyze the within-hospital epidemiology of this organism. Although the study could have benefitted from a larger sample collection with isolates from more diverse sources, patterns of local persistence were observed. The findings provided evidence for relapses involving the same *C. difficile* as well as, re-infection with new strains (separate *C. difficile* lineages) and carriage of multiple strains, all potentially underlying causes for recurrent CDI. However, a clearer picture remains to be drawn with respect to the frequencies of each.

5.5 SNPs as genetic markers for genotyping

The analysis of BI/NAP1/027 *C. difficile* genetic diversity in Chapter 3 and Chapter 4 is by far the highest resolution sequence-based study on the organism. The study identified SNPs that discriminate between early and present day isolates, as well as between and within clades. These SNPs can be used as markers for genotyping projects with the aim of differentiating isolates within larger sample collections and interrogating their origins, a common goal of epidemiology studies. A proposed set of markers would include SNPs on major branches leading to each clade in the phylogenetic tree, and SNPs within each clade that define well-supported lineages, particularly the SNPs indicating different geographical origins. The small number of SNPs means the genetic marker sets can be maintained within a manageable size, and the study carried out relatively cheaply. A total of 48 SNPs have been chosen from this dataset and used to design genotyping assays to monitor BI/NAP1/027 at Royal Liverpool University Hospital (studies in progress).

The SNPs identified that differ between isolates belonging to different ribotypes (Chapter 2) may also potentially be used as a substitute for ribotyping. A number of SNPs unique to each ribotype have been selected for this purpose. It remains to be seen how accurate this approach is in differentiating between ribotypes.

In conclusion, the work outlined in this thesis begins the analysis of the *C. difficile* species at the whole genome level, facilitating comparative genomic analysis of potential practical benefit.

References

Achtman, M. (2008). Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. Annu Rev Microbiol 62, 53-70.

Achtman, M., Morelli, G., Zhu, P., Wirth, T., Diehl, I., Kusecek, B., Vogler, A.J., Wagner, D.M., Allender, C.J., Easterday, W.R., *et al.* (2004). Microevolution and history of the plague bacillus, *Yersinia pestis*. Proc Natl Acad Sci U S A 101, 17837-17842.

Achtman, M., and Wagner, M. (2008). Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol 6, 431-440.

Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A., and Carniel, E. (1999). *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. Proc Natl Acad Sci U S A 96, 14043-14048.

Akerlund, T., Persson, I., Unemo, M., Noren, T., Svenungsson, B., Wullt, M., and Burman, L.G. (2008). Increased sporulation rate of epidemic *Clostridium difficile* Type 027/NAP1. J Clin Microbiol 46, 1530-1533.

al Saif, N., and Brazier, J.S. (1996). The distribution of *Clostridium difficile* in the environment of South Wales. Journal of medical microbiology 45, 133-137.

Albrecht, M., Sharma, C.M., Reinhardt, R., Vogel, J., and Rudel, T. (2010). Deep sequencing-based discovery of the *Chlamydia trachomatis* transcriptome. Nucleic Acids Res 38, 868-877.

Alfa, M.J., Kabani, A., Lyerly, D., Moncrief, S., Neville, L.M., Al-Barrak, A., Harding, G.K., Dyck, B., Olekson, K., and Embil, J.M. (2000). Characterization of a toxin A-negative, toxin B-positive strain of *Clostridium difficile* responsible for a nosocomial outbreak of *Clostridium difficile*-associated diarrhea. J Clin Microbiol 38, 2706-2714.

Assefa, S., Keane, T.M., Otto, T.D., Newbold, C., and Berriman, M. (2009). ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics 25, 1968-1969.

Aury, J.M., Cruaud, C., Barbe, V., Rogier, O., Mangenot, S., Samson, G., Poulain, J., Anthouard, V., Scarpelli, C., Artiguenave, F., *et al.* (2008). High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. BMC Genomics 9, 603.

Bakken, J.S. (2009). Fecal bacteriotherapy for recurrent *Clostridium difficile* infection. Anaerobe 15, 285-289.

Barbut, F., and Petit, J.C. (2001). Epidemiology of *Clostridium difficile*-associated infections. Clin Microbiol Infect 7, 405-410.

Barrick, J.E., and Lenski, R.E. (2009). Genome-wide mutational diversity in an evolving population of *Escherichia coli*. Cold Spring Harb Symp Quant Biol 74, 119-129.

Barrick, J.E., Yu, D.S., Yoon, S.H., Jeong, H., Oh, T.K., Schneider, D., Lenski, R.E., and Kim, J.F. (2009). Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. Nature 461, 1243-1247.

Bartlett, J.G. (1994). *Clostridium difficile*: history of its role as an enteric pathogen and the current state of knowledge about the organism. Clin Infect Dis 18 Suppl 4, S265-272.

Bartlett, J.G. (2006). Narrative review: the new epidemic of *Clostridium difficile*-associated enteric disease. Ann Intern Med 145, 758-764.

Bartlett, J.G. (2008). Historical perspectives on studies of *Clostridium difficile* and *C. difficile* infection. Clin Infect Dis 46 Suppl 1, S4-11.

Bartlett, J.G. (2010). *Clostridium difficile*: progress and challenges. Ann N Y Acad Sci 1213, 62-69.

Bartlett, J.G., Chang, T.W., Gurwith, M., Gorbach, S.L., and Onderdonk, A.B. (1978). Antibiotic-associated pseudomembranous colitis due to toxin-producing clostridia. N Engl J Med 298, 531-534.

Bartlett, J.G., and Gerding, D.N. (2008). Clinical recognition and diagnosis of *Clostridium difficile* infection. Clin Infect Dis 46 Suppl 1, S12-18.

Bartlett, J.G., and Perl, T.M. (2005). The new *Clostridium difficile*--what does it mean? N Engl J Med 353, 2503-2505.

Bauer, M.P., Notermans, D.W., van Benthem, B.H., Brazier, J.S., Wilcox, M.H., Rupnik, M., Monnet, D.L., van Dissel, J.T., and Kuijper, E.J. (2011). *Clostridium difficile* infection in Europe: a hospital-based survey. Lancet 377, 63-73.

Belanger, S.D., Boissinot, M., Clairoux, N., Picard, F.J., and Bergeron, M.G. (2003). Rapid detection of *Clostridium difficile* in feces by real-time PCR. J Clin Microbiol 41, 730-734.

Bennett, S. (2004). Solexa ltd. Pharmacogenomics 5, 433-438.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., and Bignell, H.R. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53-59.

Bentley, S.D., and Parkhill, J. (2004). Comparative genomic structure of prokaryotes. Annu Rev Genet 38, 771-792.

Beres, S.B., Carroll, R.K., Shea, P.R., Sitkiewicz, I., Martinez-Gutierrez, J.C., Low, D.E., McGeer, A., Willey, B.M., Green, K., Tyrrell, G.J., *et al.* (2010). Molecular complexity of successive bacterial epidemics deconvoluted by comparative pathogenomics. Proc Natl Acad Sci U S A 107, 4371-4376.

Bernal, A., Ear, U., and Kyrpides, N. (2001). Genomes Online Database (GOLD): a monitor of genome projects world-wide. Nucleic Acids Res 29, 126-127.

Bielawski, J.P., and Yang, Z. (2003). Maximum likelihood methods for detecting adaptive evolution after gene duplication. J Struct Funct Genomics 3, 201-212.

Bignardi, G.E. (1998). Risk factors for *Clostridium difficile* infection. J Hosp Infect 40, 1-15.

Bouakaze, C., Keyser, C., de Martino, S.J., Sougakoff, W., Veziris, N., Dabernat, H., and Ludes, B. (2010). Identification and genotyping of *Mycobacterium tuberculosis* complex species by use of a SNaPshot Minisequencing-based assay. J Clin Microbiol 48, 1758-1766.

Brazier, J.S., Raybould, R., Patel, B., Duckworth, G., Pearson, A., Charlett, A., and Duerden, B.I. (2008). Distribution and antimicrobial susceptibility patterns of *Clostridium difficile* PCR ribotypes in English hospitals, 2007-08. Euro Surveill 13.

Brinkman, F.S., and Parkhill, J. (2008). Population genomics: modeling the new and a renaissance of the old. Curr Opin Microbiol 11, 439-441.

Brochet, M., Rusniok, C., Couve, E., Dramsi, S., Poyart, C., Trieu-Cuot, P., Kunst, F., and Glaser, P. (2008). Shaping a bacterial genome by large chromosomal replacements, the evolutionary history of *Streptococcus agalactiae*. Proc Natl Acad Sci U S A 105, 15961-15966.

Brouwer, M.S.M., Warburton, P.J., Roberts, A.P., Mullany, P., and Allan, E. (2011). Genetic Organisation, Mobility and Predicted Functions of Genes on Integrated, Mobile Genetic Elements in Sequenced Strains of *Clostridium difficile*. PLoS ONE 6, e23014.

Burdon, D.W., George, R.H., Mogg, G.A., Arabi, Y., Thompson, H., Johnson, M., Alexander-Williams, J., and Keighley, M.R. (1981). Faecal toxin and severity of antibiotic-associated pseudomembranous colitis. J Clin Pathol 34, 548-551.

Burns, D.A., Heap, J.T., and Minton, N.P. (2010). The diverse sporulation characteristics of *Clostridium difficile* clinical isolates are not associated with type. Anaerobe 16, 618-622.

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I.A., Belmonte, M.K., Lander, E.S., Nusbaum, C., and Jaffe, D.B. (2008). ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res 18, 810-820.

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*. Molecular Microbiology 40, 1187-1199.

Campbell, P.J., Stephens, P.J., Pleasance, E.D., O'Meara, S., Li, H., Santarius, T., Stebbings, L.A., Leroy, C., Edkins, S., Hardy, C., *et al.* (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 40, 722-729.

Carter, G.P., Awad, M.M., Kelly, M.L., Rood, J.I., and Lyras, D. (2011). TcdB or not TcdB: a tale of two *Clostridium difficile* toxins. Future Microbiol 6, 121-123.

Carter, G.P., Rood, J.I., and Lyras, D. (2010). The role of toxin A and toxin B in *Clostridium difficile*-associated disease: Past and present perspectives. Gut Microbes 1, 58-64.

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.

Caugant, D.A., Mocca, L.F., Frasch, C.E., Froholm, L.O., Zollinger, W.D., and Selander, R.K. (1987). Genetic structure of *Neisseria meningitidis* populations in relation to serogroup, serotype, and outer membrane protein pattern. J Bacteriol 169, 2781-2792.

Chain, P.S., Grafham, D.V., Fulton, R.S., Fitzgerald, M.G., Hostetler, J., Muzny, D., Ali, J., Birren, B., Bruce, D.C., Buhay, C., *et al.* (2009). Genomics. Genome project standards in a new era of sequencing. Science 326, 236-237.

Chaisson, M.J., and Pevzner, P.A. (2008). Short read fragment assembly of bacterial genomes. Genome Res 18, 324-330.

Cheknis, A.K., Sambol, S.P., Davidson, D.M., Nagaro, K.J., Mancini, M.C., Hidalgo-Arroyo, G.A., Brazier, J.S., Johnson, S., and Gerding, D.N. (2009). Distribution of *Clostridium* *difficile* strains from a North American, European and Australian trial of treatment for *C. difficile* infections: 2005-2007. Anaerobe 15, 230-233.

Chernak, E. (2005). Severe *Clostridium difficile*–associated disease in populations previously at low risk — four states. Morb Mortal Wkly Rep 54, 1201-1205.

Christensen, S.K., and Gerdes, K. (2003). RelE toxins from bacteria and Archaea cleave mRNAs on translating ribosomes, which are rescued by tmRNA. Mol Microbiol 48, 1389-1400.

Cohen, S.H., Tang, Y.J., and Silva, J., Jr. (2001). Molecular typing methods for the epidemiological identification of *Clostridium difficile* strains. Expert Rev Mol Diagn 1, 61-70.

Cole, J.R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R.J., Kulam-Syed-Mohideen, A.S., McGarrell, D.M., Marsh, T., Garrity, G.M., *et al.* (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37, D141-145.

Collins, M.D., Lawson, P.A., Willems, A., Cordoba, J.J., Fernandez-Garayzabal, J., Garcia, P., Cai, J., Hippe, H., and Farrow, J.A. (1994). The phylogeny of the genus *Clostridium*: proposal of five new genera and eleven new species combinations. Int J Syst Bacteriol 44, 812-826.

Croucher, N.J., Fookes, M.C., Perkins, T.T., Turner, D.J., Marguerat, S.B., Keane, T., Quail, M.A., He, M., Assefa, S., Bahler, J., *et al.* (2009). A simple method for directional transcriptome sequencing using Illumina technology. Nucleic Acids Res 37, e148.

Croucher, N.J., Harris, S.R., Fraser, C., Quail, M.A., Burton, J., van der Linden, M., McGee, L., von Gottberg, A., Song, J.H., Ko, K.S., *et al.* (2011). Rapid pneumococcal evolution in response to clinical interventions. Science 331, 430-434.

Dallal, R.M., Harbrecht, B.G., Boujoukas, A.J., Sirio, C.A., Farkas, L.M., Lee, K.K., and Simmons, R.L. (2002). Fulminant *Clostridium difficile*: an underappreciated and increasing cause of death and complications. Ann Surg 235, 363-372.

Darling, A.E., Mau, B., and Perna, N.T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS ONE 5, e11147.

Daubin, V., and Ochman, H. (2004). Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. Genome Res 14, 1036-1042.

Delcher, A.L., Harmon, D., Kasif, S., White, O., and Salzberg, S.L. (1999). Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27, 4636-4641.

Dethlefsen, L., McFall-Ngai, M., and Relman, D.A. (2007). An ecological and evolutionary perspective on human-microbe mutualism and disease. Nature 449, 811-818.

Didelot, X., Achtman, M., Parkhill, J., Thomson, N.R., and Falush, D. (2007). A bimodal pattern of relatedness between the *Salmonella* Paratyphi A and Typhi genomes: convergence or divergence by homologous recombination? Genome Res 17, 61-68.

Didelot, X., and Falush, D. (2007). Inference of bacterial microevolution using multilocus sequence data. Genetics 175, 1251-1266.

Didelot, X., and Maiden, M.C. (2010). Impact of recombination on bacterial evolution. Trends Microbiol 18, 315-322.

Ding, F., Tang, P., Hsu, M.H., Cui, P., Hu, S., Yu, J., and Chiu, C.H. (2009). Genome evolution driven by host adaptations results in a more virulent and antimicrobial-resistant *Streptococcus pneumoniae* serotype 14. BMC Genomics 10, 158.

Dingle, K.E., Griffiths, D., Didelot, X., Evans, J., Vaughan, A., Kachrimanidou, M., Stoesser, N., Jolley, K.A., Golubchik, T., Harding, R.M., *et al.* (2011). Clinical *Clostridium difficile*: Clonality and Pathogenicity Locus Diversity. PLoS ONE 6, e19993.

Dobrindt, U., Hochhut, B., Hentschel, U., and Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol 2, 414-424.

Dohm, J.C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 36, e105.

Donati, C., Hiller, N.L., Tettelin, H., Muzzi, A., Croucher, N.J., Angiuoli, S.V., Oggioni, M., Dunning Hotopp, J.C., Hu, F.Z., Riley, D.R., *et al.* (2010). Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. Genome Biol 11, R107.

Doolittle, W.F. (1999). Phylogenetic classification and the universal tree. Science 284, 2124-2129.

Doolittle, W.F., and Papke, R.T. (2006). Genomics and the bacterial species problem. Genome Biol 7, 116.

Dridi, L., Tankovic, J., Burghoffer, B., Barbut, F., and Petit, J.C. (2002). *gyrA* and *gyrB* mutations are implicated in cross-resistance to Ciprofloxacin and moxifloxacin in *Clostridium difficile*. Antimicrob Agents Chemother 46, 3418-3421.

Drudy, D., Fanning, S., and Kyne, L. (2007a). Toxin A-negative, toxin B-positive *Clostridium difficile*. Int J Infect Dis 11, 5-10.

Drudy, D., Harnedy, N., Fanning, S., Hannan, M., and Kyne, L. (2007b). Emergence and control of fluoroquinolone-resistant, toxin A-negative, toxin B-positive *Clostridium difficile*. Infect Control Hosp Epidemiol 28, 932-940.

Drummond, A.J., and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7, 214.

Drummond, A.J., Rambaut, A., Shapiro, B., and Pybus, O.G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 22, 1185-1192.

Duerden, B.I. (2010). Contribution of a government target to controlling *Clostridium difficile* in the NHS in England. Anaerobe.

Dupuy, B., and Sonenshein, A.L. (1998). Regulated transcription of *Clostridium difficile* toxin genes. Mol Microbiol 27, 107-120.

Endo, T., Ikeo, K., and Gojobori, T. (1996). Large-scale search for genes on which positive selection may operate. Mol Biol Evol 13, 685-690.

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8, 186-194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8, 175-185.

Fagan, R.P., Albesa-Jove, D., Qazi, O., Svergun, D.I., Brown, K.A., and Fairweather, N.F. (2009). Structural insights into the molecular organization of the S-layer from *Clostridium difficile*. Mol Microbiol 71, 1308-1322.

Falush, D. (2009). Toward the use of genomics to study microevolutionary change in bacteria. PLoS Genet 5, e1000627.

Falush, D., and Bowden, R. (2006). Genome-wide association mapping in bacteria? Trends Microbiol 14, 353-355.

Falush, D., Kraft, C., Taylor, N.S., Correa, P., Fox, J.G., Achtman, M., and Suerbaum, S. (2001). Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. Proc Natl Acad Sci U S A 98, 15056-15061.

Falush, D., Torpdahl, M., Didelot, X., Conrad, D.F., Wilson, D.J., and Achtman, M. (2006). Mismatch induced speciation in *Salmonella*: model and data. Philos Trans R Soc Lond B Biol Sci 361, 2045-2053.

Fawley, W.N., Underwood, S., Freeman, J., Baines, S.D., Saxton, K., Stephenson, K., Owens, R.C., Jr., and Wilcox, M.H. (2007). Efficacy of hospital cleaning agents and germicides against epidemic *Clostridium difficile* strains. Infect Control Hosp Epidemiol 28, 920-925.

Feil, E.J., Cooper, J.E., Grundmann, H., Robinson, D.A., Enright, M.C., Berendt, T., Peacock, S.J., Smith, J.M., Murphy, M., Spratt, B.G., *et al.* (2003). How clonal is *Staphylococcus aureus*? J Bacteriol 185, 3307-3316.

Feil, E.J., Enright, M.C., and Spratt, B.G. (2000). Estimating the relative contributions of mutation and recombination to clonal diversification: a comparison between *Neisseria meningitidis* and *Streptococcus pneumoniae*. Res Microbiol 151, 465-469.

Feil, E.J., Holmes, E.C., Bessen, D.E., Chan, M.S., Day, N.P., Enright, M.C., Goldstein, R., Hood, D.W., Kalia, A., Moore, C.E., *et al.* (2001). Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc Natl Acad Sci U S A 98, 182-187.

Feil, E.J., Li, B.C., Aanensen, D.M., Hanage, W.P., and Spratt, B.G. (2004). eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J Bacteriol 186, 1518-1530.

Feil, E.J., and Spratt, B.G. (2001). Recombination and the population structures of bacterial pathogens. Annu Rev Microbiol 55, 561-590.

Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L., *et al.* (2008). The Pfam protein families database. Nucleic Acids Res 36, D281-288.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269, 496-512.

Fletcher, W., and Yang, Z. (2009). INDELible: a flexible simulator of biological sequence evolution. Mol Biol Evol 26, 1879-1888.

Fraser, C., Alm, E.J., Polz, M.F., Spratt, B.G., and Hanage, W.P. (2009). The bacterial species challenge: making sense of genetic and ecological diversity. Science 323, 741-746.

Fraser, C., Hanage, W.P., and Spratt, B.G. (2007). Recombination and the nature of bacterial speciation. Science 315, 476-480.

Fraser-Liggett, C.M. (2005). Insights on biology and evolution from microbial genome sequencing. Genome Res 15, 1603-1610.

Gagneux, S., Long, C.D., Small, P.M., Van, T., Schoolnik, G.K., and Bohannan, B.J. (2006). The competitive cost of antibiotic resistance in *Mycobacterium tuberculosis*. Science 312, 1944-1946.

Gawronski, J.D., Wong, S.M., Giannoukos, G., Ward, D.V., and Akerley, B.J. (2009). Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for *Haemophilus* genes required in the lung. Proc Natl Acad Sci U S A 106, 16422-16427.

George, W.L., Rolfe, R.D., Harding, G.K., Klein, R., Putnam, C.W., and Finegold, S.M. (1982). *Clostridium difficile* and cytotoxin in feces of patients with antimicrobial agent-associated pseudomembranous colitis. Infection 10, 205-208.

Gerding, D.N., Muto, C.A., and Owens, R.C., Jr. (2008). Measures to control and prevent *Clostridium difficile* infection. Clin Infect Dis 46 Suppl 1, S43-49.

Gibbons, R.J., and Kapsimalis, B. (1967). Estimates of the overall rate of growth of the intestinal microflora of hamsters, guinea pigs, and mice. J Bacteriol 93, 510-512.

Gill, S.R., Pop, M., Deboy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel, B.S., Gordon, J.I., Relman, D.A., Fraser-Liggett, C.M., and Nelson, K.E. (2006). Metagenomic analysis of the human distal gut microbiome. Science 312, 1355-1359.

Gogarten, J.P., and Townsend, J.P. (2005). Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol 3, 679-687.

Goldberg, S.M., Johnson, J., Busam, D., Feldblyum, T., Ferriera, S., Friedman, R., Halpern, A., Khouri, H., Kravitz, S.A., Lauro, F.M., *et al.* (2006). A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proc Natl Acad Sci U S A 103, 11240-11245.

Goorhuis, A., Bakker, D., Corver, J., Debast, S.B., Harmanus, C., Notermans, D.W., Bergwerff, A.A., Dekker, F.W., and Kuijper, E.J. (2008). Emergence of *Clostridium difficile* infection due to a new hypervirulent strain, polymerase chain reaction ribotype 078. Clin Infect Dis 47, 1162-1170.

Gordon, D., Desmarais, C., and Green, P. (2001). Automated finishing with autofinish. Genome Res 11, 614-625.

Govoni, G., and Gros, P. (1998). Macrophage NRAMP1 and its role in resistance to microbial infections. Inflamm Res 47, 277-284.

Griffiths, D., Fawley, W., Kachrimanidou, M., Bowden, R., Crook, D.W., Fung, R., Golubchik, T., Harding, R.M., Jeffery, K.J., Jolley, K.A., *et al.* (2009). Multilocus Sequence Typing of *Clostridium difficile*. J Clin Microbiol.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59, 307-321.

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52, 696-704.

Gupta, R., Nagarajan, A., and Wajapeyee, N. (2010). Advances in genome-wide DNA methylation analysis. Biotechniques 49, iii-xi.

Gupta, S., and Maiden, M.C. (2001). Exploring the evolution of diversity in pathogen populations. Trends Microbiol 9, 181-185.

Guttman, D.S., and Dykhuizen, D.E. (1994). Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. Science 266, 1380-1383.

Hall, I.C., and O'Toole, E. (1935). Intestinal flora in new-born infants: with a description of a new pathogenic anaerobe, *Bacillus difficilis*. Archives of Pediatrics and Adolescent Medicine 49, 390.

Hanage, W.P., Fraser, C., and Spratt, B.G. (2005). Fuzzy species among recombinogenic bacteria. BMC Biol 3, 6.

Hanage, W.P., Fraser, C., and Spratt, B.G. (2006). Sequences, sequence clusters and bacterial species. Philos Trans R Soc Lond B Biol Sci 361, 1917-1927.

Hanage, W.P., Fraser, C., Tang, J., Connor, T.R., and Corander, J. (2009). Hyper-recombination, diversity, and antibiotic resistance in pneumococcus. Science 324, 1454-1457.

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev 68, 669-685.

Harismendy, O., Ng, P.C., Strausberg, R.L., Wang, X., Stockwell, T.B., Beeson, K.Y., Schork, N.J., Murray, S.S., Topol, E.J., Levy, S., *et al.* (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol 10, R32.

Harris, S.R., Feil, E.J., Holden, M.T., Quail, M.A., Nickerson, E.K., Chantratita, N., Gardete, S., Tavares, A., Day, N., Lindsay, J.A., *et al.* (2010). Evolution of MRSA during hospital transmission and intercontinental spread. Science 327, 469-474.

Hatt, J.K., and Youngman, P. (2000). Mutational analysis of conserved residues in the putative DNA-binding domain of the response regulator Spo0A of *Bacillus subtilis*. J Bacteriol 182, 6975-6982.

Hayes, F. (1998). A family of stability determinants in pathogenic bacteria. J Bacteriol 180, 6415-6418.

Hernandez, D., Francois, P., Farinelli, L., Osteras, M., and Schrenzel, J. (2008). De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res 18, 802-809.

Hiller, N.L., Janto, B., Hogg, J.S., Boissy, R., Yu, S., Powell, E., Keefe, R., Ehrlich, N.E., Shen, K., Hayes, J., *et al.* (2007). Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. J Bacteriol 189, 8186-8195.

Hogg, J.S., Hu, F.Z., Janto, B., Boissy, R., Hayes, J., Keefe, R., Post, J.C., and Ehrlich, G.D. (2007). Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. Genome Biol 8, R103.

Holmes, E.C., Urwin, R., and Maiden, M.C. (1999). The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis*. Mol Biol Evol 16, 741-749.

Holt, K.E., Baker, S., Dongol, S., Basnyat, B., Adhikari, N., Thorson, S., Pulickal, A.S., Song, Y., Parkhill, J., Farrar, J.J., *et al.* (2010). High-throughput bacterial SNP typing identifies distinct clusters of *Salmonella* Typhi causing typhoid in Nepalese children. BMC Infect Dis 10, 144.

Holt, K.E., Parkhill, J., Mazzoni, C.J., Roumagnac, P., Weill, F.X., Goodhead, I., Rance, R., Baker, S., Maskell, D.J., Wain, J., *et al.* (2008). High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. Nat Genet 40, 987-993.

Horner, D.S., Pavesi, G., Castrignano, T., De Meo, P.D., Liuni, S., Sammeth, M., Picardi, E., and Pesole, G. (2010). Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. Brief Bioinform 11, 181-197.

Huang, H., Fang, H., Weintraub, A., and Nord, C.E. (2009). Distinct ribotypes and rates of antimicrobial drug resistance in *Clostridium difficile* from Shanghai and Stockholm. Clin Microbiol Infect.

Hubert, B., Loo, V.G., Bourgault, A.M., Poirier, L., Dascal, A., Fortin, E., Dionne, M., and Lorange, M. (2007). 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 44, 238-244.

Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L., and Welch, D.M. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 8, R143.

Huson, D.H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23, 254-267.

Hutchison, C.A., 3rd (2007). DNA sequencing: bench to bedside and beyond. Nucleic Acids Res 35, 6227-6237.

Jain, R., Rivera, M.C., and Lake, J.A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A 96, 3801-3806.

Janvilisri, T., Scaria, J., Thompson, A.D., Nicholson, A., Limbago, B.M., Arroyo, L.G., Songer, J.G., Grohn, Y.T., and Chang, Y.F. (2009). Microarray identification of *Clostridium difficile* core components and divergent regions associated with host origin. J Bacteriol 191, 3881-3891.

Jasni, A.S., Mullany, P., Hussain, H., and Roberts, A.P. (2010). Demonstration of conjugative transposon (Tn*5397*)-mediated horizontal gene transfer between *Clostridium difficile* and *Enterococcus faecalis*. Antimicrob Agents Chemother 54, 4924-4926.

Jobstl, M., Heuberger, S., Indra, A., Nepf, R., Kofer, J., and Wagner, M. (2010). *Clostridium difficile* in raw products of animal origin. Int J Food Microbiol 138, 172-175.

Johnson, S. (2009). Recurrent *Clostridium difficile* infection: a review of risk factors, treatments, and outcomes. J Infect 58, 403-410.

Johnson, S., Adelmann, A., Clabots, C.R., Peterson, L.R., and Gerding, D.N. (1989). Recurrences of *Clostridium difficile* diarrhea not caused by the original infecting organism. J Infect Dis 159, 340-343.

Johnson, S., Kent, S.A., O'Leary, K.J., Merrigan, M.M., Sambol, S.P., Peterson, L.R., and Gerding, D.N. (2001). Fatal pseudomembranous colitis associated with a variant *Clostridium difficile* strain not detected by toxin A immunoassay. Ann Intern Med 135, 434-438.

Johnson, S., Samore, M.H., Farrow, K.A., Killgore, G.E., Tenover, F.C., Lyras, D., Rood, J.I., DeGirolami, P., Baltch, A.L., Rafferty, M.E., *et al.* (1999). Epidemics of diarrhea caused by a clindamycin-resistant strain of *Clostridium difficile* in four hospitals. N Engl J Med 341, 1645-1651.

Jolley, K.A., Feil, E.J., Chan, M.S., and Maiden, M.C. (2001). Sequence type analysis and recombinational tests (START). Bioinformatics 17, 1230-1231.

Jolley, K.A., Kalmusova, J., Feil, E.J., Gupta, S., Musilek, M., Kriz, P., and Maiden, M.C. (2000). Carried meningococci in the Czech Republic: a diverse recombining population. J Clin Microbiol 38, 4492-4498.

Joseph, R., Demeyer, D., Vanrenterghem, D., van den Berg, R., Kuijper, E., and Delmee, M. (2005). First isolation of *Clostridium difficile* PCR ribotype 027, toxinotype III in Belgium. Euro Surveill 10, E051020 051024.

Jukes, T.H., and Cantor, C.R. (1969). Evolution of protein molecules. Mammalian protein metabolism 3, 21ñ132.

Kalia, V.C., Mukherjee, T., Bhushan, A., Joshi, J., Shankar, P., and Huma, N. (2011). Analysis of the unexplored features of rrs (16S rDNA) of the Genus *Clostridium*. BMC Genomics 12, 18.

Keessen, E.C., Gaastra, W., and Lipman, L.J. (2011). *Clostridium difficile* infection in humans and animals, differences and similarities. Vet Microbiol.

Keim, P., Kalif, A., Schupp, J., Hill, K., Travis, S.E., Richmond, K., Adair, D.M., Hugh-Jones, M., Kuske, C.R., and Jackson, P. (1997). Molecular evolution and diversity in *Bacillus anthracis* as detected by amplified fragment length polymorphism markers. J Bacteriol 179, 818-824.

Keim, P., and Smith, K.L. (2002). *Bacillus anthracis* evolution and epidemiology. Curr Top Microbiol Immunol 271, 21-32.

Killgore, G., Thompson, A., Johnson, S., Brazier, J., Kuijper, E., Pepin, J., Frost, E.H., Savelkoul, P., Nicholson, B., van den Berg, R.J., *et al.* (2008). Comparison of seven techniques for typing international epidemic strains of *Clostridium difficile*: restriction endonuclease analysis, pulsed-field gel electrophoresis, PCR-ribotyping, multilocus sequence typing, multilocus variable-number tandem-repeat analysis, amplified fragment length polymorphism, and surface layer protein A gene sequence typing. J Clin Microbiol 46, 431-437.

Kim, H., Lee, Y., Moon, H.W., Lim, C.S., Lee, K., and Chong, Y. (2011). Emergence of *Clostridium difficile* Ribotype 027 in Korea. Korean J Lab Med 31, 191-196.

Kim, H., Riley, T.V., Kim, M., Kim, C.K., Yong, D., Lee, K., Chong, Y., and Park, J.W. (2008). Increasing prevalence of toxin A-negative, toxin B-positive isolates of *Clostridium difficile* in Korea: impact on laboratory diagnosis. J Clin Microbiol 46, 1116-1117.

Kimura, M. (1983). The Neutral Theory of Molecular Evolution. Cambridge University Press.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., *et al.* (2007). Paired-end mapping reveals extensive structural variation in the human genome. Science 318, 420-426.

Koski, L.B., Morton, R.A., and Golding, G.B. (2001). Codon bias and base composition are poor indicators of horizontally transferred genes. Mol Biol Evol 18, 404-412.

Kryazhimskiy, S., and Plotkin, J.B. (2008). The population genetics of dN/dS. PLoS Genet 4, e1000304.

Kuehne, S.A., Cartman, S.T., Heap, J.T., Kelly, M.L., Cockayne, A., and Minton, N.P. (2010). The role of toxin A and toxin B in *Clostridium difficile* infection. Nature 467, 711-713.

Kuijper, E.J., Barbut, F., Brazier, J.S., Kleinkauf, N., Eckmanns, T., Lambert, M.L., Drudy, D., Fitzpatrick, F., Wiuff, C., Brown, D.J., *et al.* (2008). Update of *Clostridium difficile* infection due to PCR ribotype 027 in Europe, 2008. Euro Surveill 13.

Kuijper, E.J., van Dissel, J.T., and Wilcox, M.H. (2007). *Clostridium difficile*: changing epidemiology and new treatment options. Curr Opin Infect Dis 20, 376-383.

Kurtz, S., Choudhuri, J.V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29, 4633-4642.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004). Versatile and open software for comparing large genomes. Genome Biol 5, R12.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25.

Langridge, G.C., Phan, M.D., Turner, D.J., Perkins, T.T., Parts, L., Haase, J., Charles, I., Maskell, D.J., Peters, S.E., Dougan, G., *et al.* (2009). Simultaneous assay of every *Salmonella* Typhi gene using one million transposon mutants. Genome Res 19, 2308-2316.

Laughon, B.E., Viscidi, R.P., Gdovin, S.L., Yolken, R.H., and Bartlett, J.G. (1984). Enzyme immunoassays for detection of *Clostridium difficile* toxins A and B in fecal specimens. J Infect Dis 149, 781-788.

Lawley, T.D., Clare, S., Walker, A.W., Goulding, D., Stabler, R.A., Croucher, N., Mastroeni, P., Scott, P., Raisen, C., Mottram, L., *et al.* (2009). Antibiotic treatment of *Clostridium difficile* carrier mice triggers a supershedder state, spore-mediated transmission, and severe disease in immunocompromised hosts. Infect Immun 77, 3661-3669.

Lawrence, J.G. (2002). Gene transfer in bacteria: speciation without species? Theor Popul Biol 61, 449-460.

Lawrence, J.G. (2006). Evolution of microbial pathogens, Chapter 2 (Wiley-Blackwell).

Lefebure, T., and Stanhope, M.J. (2007). Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. Genome Biol 8, R71.

Lemee, L., Dhalluin, A., Pestel-Caron, M., Lemeland, J.F., and Pons, J.L. (2004). Multilocus sequence typing analysis of human and animal *Clostridium difficile* isolates of various toxigenic types. J Clin Microbiol 42, 2609-2617.

Lemey, P., Rambaut, A., Drummond, A.J., and Suchard, M.A. (2009). Bayesian phylogeography finds its roots. PLoS Comput Biol 5, e1000520.

Levin, B.R. (1981). Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations. Genetics 99, 1-23.

Lewis, T., Loman, N.J., Bingle, L., Jumaa, P., Weinstock, G.M., Mortiboy, D., and Pallen, M.J. (2010). High-throughput whole-genome sequencing to dissect the epidemiology of *Acinetobacter baumannii* isolates from a hospital outbreak. J Hosp Infect 75, 37-41.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18, 1851-1858.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., *et al.* (2010). De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20, 265-272.

Linder, J.A., Huang, E.S., Steinman, M.A., Gonzales, R., and Stafford, R.S. (2005). Fluoroquinolone prescribing in the United States: 1995 to 2002. Am J Med 118, 259-268.

Lloyd, A.L., Rasko, D.A., and Mobley, H.L. (2007). Defining genomic islands and uropathogen-specific genes in uropathogenic *Escherichia coli*. J Bacteriol 189, 3532-3546.

Loo, V.G., Poirier, L., Miller, M.A., Oughton, M., Libman, M.D., Michaud, S., Bourgault, A.M., Nguyen, T., Frenette, C., Kelly, M., *et al.* (2005). A predominantly clonal multi-institutional outbreak of *Clostridium difficile*-associated diarrhea with high morbidity and mortality. N Engl J Med 353, 2442-2449.

Louie, T.J., Miller, M.A., Mullane, K.M., Weiss, K., Lentnek, A., Golan, Y., Gorbach, S., Sears, P., and Shue, Y.K. (2011). Fidaxomicin versus vancomycin for *Clostridium difficile* infection. N Engl J Med 364, 422-431.

Lovett, S.T., Hurley, R.L., Sutera, V.A., Jr., Aubuchon, R.H., and Lebedeva, M.A. (2002). Crossing over between regions of limited homology in *Escherichia coli*. RecA-dependent and RecA-independent pathways. Genetics 160, 851-859.

Lyerly, D.M., Saum, K.E., MacDonald, D.K., and Wilkins, T.D. (1985). Effects of *Clostridium difficile* toxins given intragastrically to animals. Infect Immun 47, 349-352.

Lyras, D., O'Connor, J.R., Howarth, P.M., Sambol, S.P., Carter, G.P., Phumoonna, T., Poon, R., Adams, V., Vedantam, G., Johnson, S., *et al.* (2009). Toxin B is essential for virulence of *Clostridium difficile*. Nature 458, 1176-1179.

MacCannell, D.R., Louie, T.J., Gregson, D.B., Laverdiere, M., Labbe, A.C., Laing, F., and Henwick, S. (2006). Molecular analysis of *Clostridium difficile* PCR ribotype 027 isolates from Eastern and Western Canada. J Clin Microbiol 44, 2147-2152.

MacLean, D., Jones, J.D., and Studholme, D.J. (2009). Application of 'next-generation' sequencing technologies to microbial genetics. Nat Rev Microbiol 7, 287-296.

Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., *et al.* (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A 95, 3140-3145.

Maiden, M.C., and Urwin, R. (2006). Evolution of microbial pathogens, Chapter 3 (Wiley-Blackwell).

Majewski, J., Zawadzki, P., Pickerill, P., Cohan, F.M., and Dowson, C.G. (2000). Barriers to genetic exchange between bacterial species: *Streptococcus pneumoniae* transformation. J Bacteriol 182, 1016-1023.

Mardis, E.R. (2011). A decade's perspective on DNA sequencing technology. Nature 470, 198-203.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376-380.

Matamouros, S., England, P., and Dupuy, B. (2007). *Clostridium difficile* toxin expression is inhibited by the novel regulator TcdC. Mol Microbiol 64, 1274-1288.

Maurelli, A.T. (2007). Black holes, antivirulence genes, and gene inactivation in the evolution of bacterial pathogens. FEMS Microbiol Lett 267, 1-8.

McDonald, L.C., Killgore, G.E., Thompson, A., Owens, R.C., Jr., Kazakova, S.V., Sambol, S.P., Johnson, S., and Gerding, D.N. (2005). An epidemic, toxin gene-variant strain of *Clostridium difficile*. N Engl J Med 353, 2433-2441.

Medini, D., Donati, C., Tettelin, H., Masignani, V., and Rappuoli, R. (2005). The microbial pan-genome. Curr Opin Genet Dev 15, 589-594.

Medini, D., Serruto, D., Parkhill, J., Relman, D.A., Donati, C., Moxon, R., Falkow, S., and Rappuoli, R. (2008). Microbiology in the post-genomic era. Nat Rev Microbiol 6, 419-430.

Merrigan, M., Venugopal, A., Mallozzi, M., Roxas, B., Viswanathan, V.K., Johnson, S., Gerding, D.N., and Vedantam, G. (2010). Human hypervirulent *Clostridium difficile* strains exhibit increased sporulation as well as robust toxin production. J Bacteriol 192, 4904-4911.

Metzker, M.L. (2010). Sequencing technologies - the next generation. Nat Rev Genet 11, 31-46.

Milkman, R. (1997). Recombination and population structure in *Escherichia coli*. Genetics 146, 745-750.

Miller, J.R., Koren, S., and Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. Genomics 95, 315-327.

Mogg, G.A., Burdon, D.W., and Keighley, M. (1979). Oral metronidazole in *Clostridium difficile* colitis. Br Med J 2, 335.

Moir, A. (2006). How do spores germinate? J Appl Microbiol 101, 526-530.

Monot, M., Honore, N., Garnier, T., Zidane, N., Sherafi, D., Paniz-Mondolfi, A., Matsuoka, M., Taylor, G.M., Donoghue, H.D., Bouwman, A., *et al.* (2009). Comparative genomic and phylogeographic analysis of *Mycobacterium leprae*. Nat Genet 41, 1282-1289.

Morelli, G., Song, Y., Mazzoni, C.J., Eppinger, M., Roumagnac, P., Wagner, D.M., Feldkamp, M., Kusecek, B., Vogler, A.J., Li, Y., *et al.* (2010). *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. Nat Genet.

Mullany, P., Wilks, M., Lamb, I., Clayton, C., Wren, B., and Tabaqchali, S. (1990). Genetic analysis of a tetracycline resistance element from *Clostridium difficile* and its conjugal transfer to and from *Bacillus subtilis*. J Gen Microbiol 136, 1343-1349.

Muto, C.A., Pokrywka, M., Shutt, K., Mendelsohn, A.B., Nouri, K., Posey, K., Roberts, T., Croyle, K., Krystofiak, S., Patel-Brown, S., *et al.* (2005). A large outbreak of *Clostridium difficile*-associated disease with an unexpected proportion of deaths and colectomies at a teaching hospital following increased fluoroquinolone use. Infect Control Hosp Epidemiol 26, 273-280.

Nagarajan, N., and Pop, M. (2010). Sequencing and genome assembly using next-generation technologies. Methods Mol Biol 673, 1-17.

Nakamura, Y., Itoh, T., Matsuda, H., and Gojobori, T. (2004). Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat Genet 36, 760-766.

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3, 418-426.

Ning, Z., Cox, A.J., and Mullikin, J.C. (2001). SSAHA: a fast search method for large DNA databases. Genome Res 11, 1725-1729.

Noren, T., Akerlund, T., Wullt, M., Burman, L.G., and Unemo, M. (2007). Mutations in fusA associated with posttherapy fusidic acid resistance in *Clostridium difficile*. Antimicrob Agents Chemother 51, 1840-1843.

Nyc, O., Pituch, H., Matejkova, J., Obuch-Woszczatynski, P., and Kuijper, E.J. (2011). *Clostridium difficile* PCR ribotype 176 in the Czech Republic and Poland. Lancet 377, 1407.

O'Connor, J.R., Johnson, S., and Gerding, D.N. (2009). *Clostridium difficile* infection caused by the epidemic BI/NAP1/027 strain. Gastroenterology 136, 1913-1924.

Ochman, H., Elwyn, S., and Moran, N.A. (1999). Calibrating bacterial evolution. Proc Natl Acad Sci U S A 96, 12638-12643.

OConnor, J.R., Galang, M.A., Sambol, S.P., Hecht, D.W., Vedantam, G., Gerding, D.N., and Johnson, S. (2008). Rifampin and rifaximin resistance in clinical isolates of *Clostridium difficile*. Antimicrob Agents Chemother 52, 2813-2817.

OConnor, J.R., Johnson, S., and Gerding, D.N. (2009). *Clostridium difficile* infection caused by the epidemic BI/NAP1/027 strain. Gastroenterology 136, 1913-1924.

Onderdonk, A.B., Cisneros, R.L., and Bartlett, J.G. (1980). *Clostridium difficile* in gnotobiotic mice. Infect Immun 28, 277-282.

ONeill, G., Ogunsola, F., Brazier, J., and Duerden, B. (1996). Modification of a PCR Ribotyping Method for Application as a Routine Typing Scheme for *Clostridium difficile*. Anaerobe 2, 205-209.

ONeill, G.L., Beaman, M.H., and Riley, T.V. (1991). Relapse versus reinfection with *Clostridium difficile*. Epidemiol Infect 107, 627-635.

Pallen, M.J., Loman, N.J., and Penn, C.W. (2010). High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. Curr Opin Microbiol 13, 625-631.

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.

Parkhill, J. (2008). Time to remove the model organism blinkers. Trends Microbiol 16, 510-511.

Peach, S.L., Borriello, S.P., Gaya, H., Barclay, F.E., and Welch, A.R. (1986). Asymptomatic carriage of *Clostridium difficile* in patients with cystic fibrosis. J Clin Pathol 39, 1013-1018.

Pearson, T., Busch, J.D., Ravel, J., Read, T.D., Rhoton, S.D., U'Ren, J.M., Simonson, T.S., Kachur, S.M., Leadem, R.R., Cardon, M.L., *et al.* (2004). Phylogenetic discovery bias in *Bacillus anthracis* using single-nucleotide polymorphisms from whole-genome sequencing. Proc Natl Acad Sci U S A 101, 13536-13541.

Pearson, W.R., and Lipman, D.J. (1988). Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 85, 2444-2448.

Pepin, J., Valiquette, L., Alary, M.E., Villemure, P., Pelletier, A., Forget, K., Pepin, K., and Chouinard, D. (2004). *Clostridium difficile*-associated diarrhea in a region of Quebec from 1991 to 2003: a changing pattern of disease severity. CMAJ 171, 466-472.

Perez-Losada, M., Browne, E.B., Madsen, A., Wirth, T., Viscidi, R.P., and Crandall, K.A. (2006). Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. Infect Genet Evol 6, 97-112.

Perkins, T.T., Kingsley, R.A., Fookes, M.C., Gardner, P.P., James, K.D., Yu, L., Assefa, S.A., He, M., Croucher, N.J., Pickard, D.J., *et al.* (2009). A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella* typhi. PLoS Genet 5, e1000569.

Perna, N.T., Plunkett, G., 3rd, Burland, V., Mau, B., Glasner, J.D., Rose, D.J., Mayhew, G.F., Evans, P.S., Gregor, J., Kirkpatrick, H.A., *et al.* (2001). Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 409, 529-533.

Peterson, L.R., Manson, R.U., Paule, S.M., Hacek, D.M., Robicsek, A., Thomson, R.B., Jr., and Kaul, K.L. (2007). Detection of toxigenic *Clostridium difficile* in stool samples by real-time polymerase chain reaction for the diagnosis of *C. difficile*-associated diarrhea. Clin Infect Dis 45, 1152-1160.

Pinner, E., Gruenheid, S., Raymond, M., and Gros, P. (1997). Functional complementation of the yeast divalent cation transporter family SMF by NRAMP2, a member of the mammalian natural resistance-associated macrophage protein family. J Biol Chem 272, 28933-28938.

Pop, M., and Salzberg, S.L. (2008). Bioinformatics challenges of new sequencing technology. Trends Genet 24, 142-149.

Popoff, M.R., Rubin, E.J., Gill, D.M., and Boquet, P. (1988). Actin-specific ADP-ribosyltransferase produced by a *Clostridium difficile* strain. Infect Immun 56, 2299-2306.

Posada, D. (2008). jModelTest: phylogenetic model averaging. Molecular Biology and Evolution 25, 1253.

Quail, M.A., Kozarewa, I., Smith, F., Scally, A., Stephens, P.J., Durbin, R., Swerdlow, H., and Turner, D.J. (2008). A large genome center's improvements to the Illumina sequencing system. Nat Methods 5, 1005-1010.

Rathnayake, I.U., Hargreaves, M., and Huygens, F. (2011). Genotyping of *Enterococcus faecalis* and *Enterococcus faecium* isolates by use of a set of eight single nucleotide polymorphisms. J Clin Microbiol 49, 367-372.

Redelings, M.D., Sorvillo, F., and Mascola, L. (2007). Increase in *Clostridium difficile*-related mortality rates, United States, 1999-2004. Emerg Infect Dis 13, 1417-1419.

Redfield, R.J. (2001). Do bacteria have sex? Nat Rev Genet 2, 634-639.

Reid, S.D., Herbelin, C.J., Bumbaugh, A.C., Selander, R.K., and Whittam, T.S. (2000). Parallel evolution of virulence in pathogenic *Escherichia coli*. Nature 406, 64-67.

Richards, M., Knox, J., Elliott, B., Mackin, K., Lyras, D., Waring, L.J., and Riley, T.V. (2011). Severe infection with *Clostridium difficile* PCR ribotype 027 acquired in Melbourne, Australia. Med J Aust 194, 369-371.

Riggs, M.M., Sethi, A.K., Zabarsky, T.F., Eckstein, E.C., Jump, R.L., and Donskey, C.J. (2007). Asymptomatic carriers are a potential source for transmission of epidemic and nonepidemic *Clostridium difficile* strains among long-term care facility residents. Clin Infect Dis 45, 992-998.

Roach, J.C., Boysen, C., Wang, K., and Hood, L. (1995). Pairwise end sequencing: a unified approach to genomic mapping and sequencing. Genomics 26, 345-353.

Roberts, A.P., and Mullany, P. (2009). A modular master on the move: the Tn*916* family of mobile genetic elements. Trends Microbiol 17, 251-258.

Roberts, A.P., and Mullany, P. (2011). Tn*916*-like genetic elements: a diverse group of modular mobile elements conferring antibiotic resistance. FEMS Microbiol Rev.

Roberts, M.C., McFarland, L.V., Mullany, P., and Mulligan, M.E. (1994). Characterization of the genetic basis of antibiotic resistance in *Clostridium difficile*. J Antimicrob Chemother 33, 419-429.

Robinson, D.A., and Enright, M.C. (2004). Evolution of *Staphylococcus aureus* by large chromosomal replacements. J Bacteriol 186, 1060-1064.

Rocha, E.P., Smith, J.M., Hurst, L.D., Holden, M.T., Cooper, J.E., Smith, N.H., and Feil, E.J. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol 239, 226-235.

Roumagnac, P., Weill, F.X., Dolecek, C., Baker, S., Brisse, S., Chinh, N.T., Le, T.A., Acosta, C.J., Farrar, J., Dougan, G., *et al.* (2006). Evolutionary history of *Salmonella* typhi. Science 314, 1301-1304.

Rouphael, N.G., O'Donnell, J.A., Bhatnagar, J., Lewis, F., Polgreen, P.M., Beekmann, S., Guarner, J., Killgore, G.E., Coffman, B., Campbell, J., *et al.* (2008). *Clostridium difficile*-

associated diarrhea: an emerging threat to pregnant women. Am J Obstet Gynecol 198, 635 e631-636.

Rubin, E.M., Lucas, S., Richardson, P., Rokhsar, D., and Pennacchio, L. (2004). Finishing the euchromatic sequence of the human genome. Nature 431.

Rupnik, M., Avesani, V., Janc, M., von Eichel-Streiber, C., and Delmee, M. (1998). A novel toxinotyping scheme and correlation of toxinotypes with serogroups of *Clostridium difficile* isolates. J Clin Microbiol 36, 2240-2247.

Rupnik, M., Wilcox, M.H., and Gerding, D.N. (2009). *Clostridium difficile* infection: new developments in epidemiology and pathogenesis. Nat Rev Microbiol 7, 526-536.

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.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M., and Smith, M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265, 687-695.

Sanger, F., and Coulson, A.R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94, 441-448.

Savage, A.M., and Alford, R.H. (1983). Nosocomial spread of *Clostridium difficile*. Infect Control 4, 31-33.

Saxton, K., Baines, S.D., Freeman, J., O'Connor, R., and Wilcox, M.H. (2009). Effects of exposure of *Clostridium difficile* PCR ribotypes 027 and 001 to fluoroquinolones in a human gut model. Antimicrob Agents Chemother 53, 412-420.

Scaria, J., Ponnala, L., Janvilisri, T., Yan, W., Mueller, L.A., and Chang, Y.F. (2010). Analysis of ultra low genome conservation in *Clostridium difficile*. PLoS ONE 5, e15147.

Schmidt, M.L., and Gilligan, P.H. (2009). *Clostridium difficile* testing algorithms: what is practical and feasible? Anaerobe 15, 270-273.

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.

Sekulovic, O., Meessen-Pinard, M., and Fortier, L.C. (2011). Prophage-stimulated toxin production in *Clostridium difficile* NAP1/027 lysogens. Journal of Bacteriology 193, 2726.

Selander, R.K., and Levin, B.R. (1980). Genetic diversity and structure in *Escherichia coli* populations. Science 210, 545-547.

Setlow, P. (2003). Spore germination. Curr Opin Microbiol 6, 550-556.

Sharma, C.M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., Chabas, S., Reiche, K., Hackermuller, J., Reinhardt, R., *et al.* (2010). The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature 464, 250-255.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nat Biotechnol 26, 1135-1145.

Sheppard, S.K., McCarthy, N.D., Falush, D., and Maiden, M.C. (2008). Convergence of *Campylobacter* species: implications for bacterial evolution. Science 320, 237-239.

Sheridan, P.P., Freeman, K.H., and Brenchley, J.E. (2003). Estimated minimal divergence times of the major bacterial and archaeal phyla. Geomicrobiology Journal 20, 1-14.

Silva, J., Jr., Batts, D.H., Fekety, R., Plouffe, J.F., Rifkin, G.D., and Baird, I. (1981). Treatment of *Clostridium difficile* colitis and diarrhea with vancomycin. Am J Med 71, 815-822.

Skippington, E., and Ragan, M.A. (2011). Lateral genetic transfer and the construction of genetic exchange communities. FEMS Microbiol Rev.

Smith, J.M., Dowson, C.G., and Spratt, B.G. (1991). Localized sex in bacteria. Nature 349, 29-31.

Smith, J.M., Smith, N.H., O'Rourke, M., and Spratt, B.G. (1993). How clonal are bacteria? Proc Natl Acad Sci U S A 90, 4384-4388.

Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B., and Hood, L.E. (1986). Fluorescence detection in automated DNA sequence analysis. Nature 321, 674-679.

Smith, N.H., Gordon, S.V., de la Rua-Domenech, R., Clifton-Hadley, R.S., and Hewinson, R.G. (2006). Bottlenecks and broomsticks: the molecular evolution of *Mycobacterium bovis*. Nat Rev Microbiol 4, 670-681.

Songer, J.G., Jones, R., Anderson, M.A., Barbara, A.J., Post, K.W., and Trinh, H.T. (2007). Prevention of porcine *Clostridium difficile*-associated disease by competitive exclusion with nontoxigenic organisms. Vet Microbiol 124, 358-361.

Spigaglia, P., Barbanti, F., Louie, T., Barbut, F., and Mastrantonio, P. (2009). Molecular analysis of the *gyrA* and *gyrB* quinolone resistance-determining regions of fluoroquinolone-resistant *Clostridium difficile* mutants selected in vitro. Antimicrob Agents Chemother 53, 2463-2468.

Spigaglia, P., Barbanti, F., Mastrantonio, P., Brazier, J.S., Barbut, F., Delmee, M., Kuijper, E., and Poxton, I.R. (2008). Fluoroquinolone resistance in *Clostridium difficile* isolates from a prospective study of *C. difficile* infections in Europe. Journal of medical microbiology 57, 784-789.

Spigaglia, P., Carattoli, A., Barbanti, F., and Mastrantonio, P. (2010). Detection of *gyrA* and *gyrB* mutations in *Clostridium difficile* isolates by real-time PCR. Mol Cell Probes 24, 61-67.

Spigaglia, P., Carucci, V., Barbanti, F., and Mastrantonio, P. (2005). ErmB determinants and Tn*916*-Like elements in clinical isolates of *Clostridium difficile*. Antimicrob Agents Chemother 49, 2550-2553.

Spigaglia, P., and Mastrantonio, P. (2002). Molecular analysis of the pathogenicity locus and polymorphism in the putative negative regulator of toxin production (TcdC) among *Clostridium difficile* clinical isolates. J Clin Microbiol 40, 3470-3475.

Spratt, B.G. (2004). Exploring the concept of clonality in bacteria. Methods Mol Biol 266, 323-352.

Spratt, B.G., Hanage, W.P., and Feil, E.J. (2001). The relative contributions of recombination and point mutation to the diversification of bacterial clones. Curr Opin Microbiol 4, 602-606.

Spratt, B.G., and Maiden, M.C. (1999). Bacterial population genetics, evolution and epidemiology. Philos Trans R Soc Lond B Biol Sci 354, 701-710.

Sreevatsan, S., Pan, X., Stockbauer, K.E., Connell, N.D., Kreiswirth, B.N., Whittam, T.S., and Musser, J.M. (1997). Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. Proc Natl Acad Sci U S A 94, 9869-9874.

Srivatsan, A., Han, Y., Peng, J., Tehranchi, A.K., Gibbs, R., Wang, J.D., and Chen, R. (2008). High-precision, whole-genome sequencing of laboratory strains facilitates genetic studies. PLoS Genet 4, e1000139.

Stabler, R.A., Gerding, D.N., Songer, J.G., Drudy, D., Brazier, J.S., Trinh, H.T., Witney, A.A., Hinds, J., and Wren, B.W. (2006). Comparative phylogenomics of *Clostridium difficile* reveals clade specificity and microevolution of hypervirulent strains. J Bacteriol 188, 7297-7305.

Stabler, R.A., He, M., Dawson, L., Martin, M., Valiente, E., Corton, C., Lawley, T.D., Sebaihia, M., Quail, M.A., Rose, G., *et al.* (2009). Comparative genome and phenotypic analysis of *Clostridium difficile* 027 strains provides insight into the evolution of a hypervirulent bacterium. Genome Biol 10, R102.

Stackebrandt, E. Taxonomic parameters revisited: tarnished gold standards. Microbiology Today 33, 153-155.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688-2690.

Stamper, P.D., Babiker, W., Alcabasa, R., Aird, D., Wehrlin, J., Ikpeama, I., Gluck, L., and Carroll, K.C. (2009). Evaluation of a new commercial TaqMan PCR assay for direct detection of the *Clostridium difficile* toxin B gene in clinical stool specimens. J Clin Microbiol 47, 3846-3850.

Stubbs, S.L., Brazier, J.S., O'Neill, G.L., and Duerden, B.I. (1999). PCR targeted to the 16S-23S rRNA gene intergenic spacer region of *Clostridium difficile* and construction of a library consisting of 116 different PCR ribotypes. J Clin Microbiol 37, 461-463.

Studholme, D.J., Ibanez, S.G., MacLean, D., Dangl, J.L., Chang, J.H., and Rathjen, J.P. (2009). A draft genome sequence and functional screen reveals the repertoire of type III secreted proteins of *Pseudomonas syringae* pathovar tabaci 11528. BMC Genomics 10, 395.

Suchard, M.A., Weiss, R.E., and Sinsheimer, J.S. (2001). Bayesian selection of continuous-time Markov chain evolutionary models. Mol Biol Evol 18, 1001-1013.

Tanner, H.E., Hardy, K.J., and Hawkey, P.M. (2010). Coexistence of multiple multilocus variable-number tandem-repeat analysis subtypes of *Clostridium difficile* PCR ribotype 027 strains within fecal specimens. J Clin Microbiol 48, 985-987.

Taylor, N.S., Thorne, G.M., and Bartlett, J.G. (1981). Comparison of two toxins produced by *Clostridium difficile*. Infect Immun 34, 1036-1043.

Teasley, D.G., Gerding, D.N., Olson, M.M., Peterson, L.R., Gebhard, R.L., Schwartz, M.J., and Lee, J.T., Jr. (1983). Prospective randomised trial of metronidazole versus vancomycin for *Clostridium difficile*-associated diarrhoea and colitis. Lancet 2, 1043-1046.

Tedesco, F.J., Barton, R.W., and Alpers, D.H. (1974). Clindamycin-associated colitis. A prospective study. Ann Intern Med 81, 429-433.

Tettelin, H., and Feldblyum, T. (2009). Bacterial genome sequencing. Methods Mol Biol 551, 231-247.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., *et al.* (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus* agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A 102, 13950-13955.

Thomas, C.M., and Nielsen, K.M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol 3, 711-721.

Turcatti, G., Romieu, A., Fedurco, M., and Tairi, A.P. (2008). A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. Nucleic Acids Res 36, e25.

Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R., and Gordon, J.I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444, 1027-1031.

Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin, E.M., Rokhsar, D.S., and Banfield, J.F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428, 37-43.

Underwood, S., Guan, S., Vijayasubhash, V., Baines, S.D., Graham, L., Lewis, R.J., Wilcox, M.H., and Stephenson, K. (2009). Characterization of the sporulation initiation pathway of *Clostridium difficile* and its role in toxin production. J Bacteriol 191, 7296-7305.

van den Berg, R.J., Schaap, I., Templeton, K.E., Klaassen, C.H., and Kuijper, E.J. (2007). Typing and subtyping of *Clostridium difficile* isolates by using multiple-locus variable-number tandem-repeat analysis. J Clin Microbiol 45, 1024-1028.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. Science 304, 66-74.

Vernikos, G.S., and Parkhill, J. (2006). Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. Bioinformatics 22, 2196-2203.

Vogler, A.J., Busch, J.D., Percy-Fine, S., Tipton-Hunton, C., Smith, K.L., and Keim, P. (2002). Molecular analysis of rifampin resistance in *Bacillus anthracis* and *Bacillus cereus*. Antimicrob Agents Chemother 46, 511-513.

Von Abercron, S.M., Karlsson, F., Wigh, G.T., Wierup, M., and Krovacek, K. (2009). Low occurrence of *Clostridium difficile* in retail ground meat in Sweden. J Food Prot 72, 1732-1734.

Vonberg, R.P., Kuijper, E.J., Wilcox, M.H., Barbut, F., Tull, P., Gastmeier, P., van den Broek, P.J., Colville, A., Coignard, B., Daha, T., *et al.* (2008). Infection control measures to limit the spread of *Clostridium difficile*. Clin Microbiol Infect 14 Suppl 5, 2-20.

Vos, M. (2009). Why do bacteria engage in homologous recombination? Trends Microbiol 17, 226-232.

Vos, M., and Didelot, X. (2009). A comparison of homologous recombination rates in bacteria and archaea. ISME J 3, 199-208.

Voth, D.E., and Ballard, J.D. (2005). *Clostridium difficile* toxins: mechanism of action and role in disease. Clin Microbiol Rev 18, 247-263.

Walters, B.A., Roberts, R., Stafford, R., and Seneviratne, E. (1983). Relapse of antibiotic associated colitis: endogenous persistence of *Clostridium difficile* during vancomycin therapy. Gut 24, 206-212.

Wang, H., Roberts, A.P., Lyras, D., Rood, J.I., Wilks, M., and Mullany, P. (2000). Characterization of the ends and target sites of the novel conjugative transposon Tn*5397* from *Clostridium difficile*: excision and circularization is mediated by the large resolvase, TndX. J Bacteriol 182, 3775-3783.

Wang, J., Karnati, P.K., Takacs, C.M., Kowalski, J.C., and Derbyshire, K.M. (2005). Chromosomal DNA transfer in *Mycobacterium smegmatis* is mechanistically different from classical Hfr chromosomal DNA transfer. Mol Microbiol 58, 280-288.

Ward, T.J., Ducey, T.F., Usgaard, T., Dunn, K.A., and Bielawski, J.P. (2008). Multilocus genotyping assays for single nucleotide polymorphism-based subtyping of *Listeria monocytogenes* isolates. Appl Environ Microbiol 74, 7629-7642.

Warny, M., Pepin, J., Fang, A., Killgore, G., Thompson, A., Brazier, J., Frost, E., and McDonald, L.C. (2005). Toxin production by an emerging strain of *Clostridium difficile* associated with outbreaks of severe disease in North America and Europe. Lancet 366, 1079-1084.

Welch, R.A., Burland, V., Plunkett, G., 3rd, Redford, P., Roesch, P., Rasko, D., Buckles, E.L., Liou, S.R., Boutin, A., Hackett, J., *et al.* (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Natl Acad Sci U S A 99, 17020-17024.

Wilcox, M.H., Planche, T., Fang, F.C., and Gilligan, P. (2010). What is the current role of algorithmic approaches for diagnosis of *Clostridium difficile* infection? J Clin Microbiol 48, 4347-4353.

Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L.H., Karch, H., Reeves, P.R., Maiden, M.C., Ochman, H., *et al.* (2006). Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol Microbiol 60, 1136-1151.

Wozniak, R.A., and Waldor, M.K. (2010). Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. Nat Rev Microbiol 8, 552-563.

Wren, B.W., and Tabaqchali, S. (1987). Restriction endonuclease DNA analysis of *Clostridium difficile*. J Clin Microbiol 25, 2402-2404.

Wust, J., Sullivan, N.M., Hardegger, U., and Wilkins, T.D. (1982). Investigation of an outbreak of antibiotic-associated colitis by various typing methods. J Clin Microbiol 16, 1096-1101.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24, 1586-1591.

Yang, Z., and Bielawski, J.P. (2000). Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15, 496-503.

Yoon, S.S., and Brandt, L.J. (2010). Treatment of refractory/recurrent *C. difficile*-associated disease by donated stool transplanted via colonoscopy: a case series of 12 patients. J Clin Gastroenterol 44, 562-566.

Zar, F.A., Bakkanagari, S.R., Moorthi, K.M., and Davis, M.B. (2007). A comparison of vancomycin and metronidazole for the treatment of *Clostridium difficile*-associated diarrhea, stratified by disease severity. Clin Infect Dis 45, 302-307.

Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18, 821-829.

Appendix A

C. difficile BI/NAP1/027 global collection used in Chapter 3

The isolates were contributed by Munir Pirmohamed, Fabio Miyajima and Paul Roberts (University of Liverpool, UK), Brendan Wren (London School of Hygiene and Tropical Medicine, UK), Sharon Peacock and Fiona Cooke (Addenbrooke's Hospital, Cambridge, UK), Martin Curran and Niddy Walpole (Cambridge, UK), Derek Brown, John Coia and Gill Douce (Scottish Microbiology Reference Lab), Derek Fairley (Belfast Health and Social Care Trust, NI), Dale Gerding (Hines VA Hospital, USA), Glen Songer (University of Arizona, USA), Jon Brazier (Anaerobic Reference Laboratory in Cardiff, Wales), Tom Riley (University of Western Australia), Tse Hsien Koh (Singapore General Hospital), Stephen Michell and Emma Butt (University of Exeter, UK), Peter Hawkey, Katie Hardy and Sue Manzoor (West Midlands Public Health Laboratory, Birmingham, UK), Tom Louie and Kristine Cannon (University of Calgary, Canada), Hee-Jung Kim (Yonsei University, Korea), Kathy Bamford and Stephanie d' Arc (Imperial College of London, UK), Mark Wilcox and Warren Fawley (University of Leeds, UK), Ed Kuijper and Marjolein Hensgens (University of Leiden, The Netherlands).

Isolate Name	Alternative Isolate Name	Year	Location	Source	Coverage
Aus001	10Aus001	2010	Melbourne, Australia	Human	128
Aus002	10Aus002	2010	Melbourne, Australia	Human	123
Aus003	10Aus003	2010	Melbourne, Australia	Human	114
Aus004	10Aus004	2010	Melbourne, Australia	Human	144
Aus005	10Aus005	2010	Melbourne, Australia	Human	141
Aus006	10Aus006	2010	Melbourne, Australia	Human	129
Bel023	09Bel023	2009	Belfast, UK	Human	104
Bel024	09Bel024	2009	Belfast, UK	Human	93
Bel025			Belfast, UK	Human	59
Bel026	09Bel026	2009	Belfast, UK	Human	82
Bel027	09Bel027	2009	Belfast, UK	Human	95
Bel028	09Bel028	2009	Belfast, UK	Human	146
Bel029	09Bel029	2009	Belfast, UK	Human	110
Bel030	09Bel030	2009	Belfast, UK	Human	63
Bel031	09Bel031	2009	Belfast, UK	Human	99
Bel032	09Bel032	2009	Belfast, UK	Human	91
Bel033	10Bel033	2010	Belfast, UK	Human	121
Bir001	07Bir001H1	2007	Birmingham, UK	Human	114
Bir002	10Bir002H1	2010	Birmingham, UK	Human	122
Bir003	08Bir003H2	2008	Birmingham, UK	Human	139
Bir004	10Bir004H2	2010	Birmingham, UK	Human	115
Bir005	08Bir005H3	2008	Birmingham, UK	Human	123
Bir006	08Bir006H3	2008	Birmingham, UK	Human	119
Bir007	07Bir007H4	2007	Birmingham, UK	Human	132
Bir008	10Bir008H4	2010	Birmingham, UK	Human	123
Bir009	07Bir009H5	2007	Birmingham, UK	Human	112
Bir010	09Bir010H5	2009	Birmingham, UK	Human	97
Bir011	08Bir011H6	2008	Birmingham, UK	Human	99
Bir012	10Bir012H6	2010	Birmingham, UK	Human	103

Bir013	07Bir013H7	2007	Birmingham, UK	Human	105
Bir014	08Bir014H7	2008	Birmingham, UK	Human	128
Bir015	07Bir015H8	2007	Birmingham, UK	Human	105
Bir016	10Bir016H8	2010	Birmingham, UK	Human	102
Bir017	07Bir017H9	2007	Birmingham, UK	Human	101
Bir018	07Bir018H9	2007	Birmingham, UK	Human	130
Bir019	07Bir019H10	2007	Birmingham, UK	Human	100
Bir020	10Bir020H10	2010	Birmingham, UK	Human	86
Bir021	07Bir021H11	2007	Birmingham, UK	Human	86
Bir022	10Bir022H11	2010	Birmingham, UK	Human	104
Bir023	07Bir023H12	2007	Birmingham, UK	Human	141
Bir024	09Bir024H12	2009	Birmingham, UK	Human	135
Bir025	07Bir025H13	2007	Birmingham, UK	Human	177
Bir026	10Bir026H13	2010	Birmingham, UK	Human	136
Bir027	07Bir027H13	2007	Birmingham, UK	Human	128
Bir028	09Bir028H14	2009	Birmingham, UK	Human	134
Bir029	07Bir029H15	2007	Birmingham, UK	Human	145
Bir030	08Bir030H15	2008	Birmingham, UK	Human	139
Bir031	08Bir031H16	2008	Birmingham, UK	Human	124
Bir032	08Bir032H16	2008	Birmingham, UK	Human	111
Bir033	08Bir033H16	2008	Birmingham, UK	Human	118
Bir034	08Bir034H16	2008	Birmingham, UK	Human	65
Bir035	08Bir035H16	2008	Birmingham, UK	Human	91
Bir036	08Bir036H16	2008	Birmingham, UK	Human	136
Bir037	08Bir037H16	2008	Birmingham, UK	Human	104
Bir038	08Bir038H16	2008	Birmingham, UK	Human	91
Bir039	08Bir039H16	2008	Birmingham, UK	Human	115
Wal001	02BirmW1	2002	Birmingham, UK	Human	125
Cam001	08Cam001	2008	Cambridge, UK	Human	107
Cam002	08Cam002	2008	Cambridge, UK	Human	97
Cam003	08Cam003	2008	Cambridge, UK	Human	135
Cam004	08Cam004	2008	Cambridge, UK	Human	77

Cam005	08Cam005	2008	Cambridge, UK	Human	97
Cam006	08Cam006	2008	Cambridge, UK	Human	103
Cam007	08Cam007	2008	Cambridge, UK	Human	75
Cam008	08Cam008	2008	Cambridge, UK	Human	102
Cam009	08Cam009	2008	Cambridge, UK	Human	74
Cam010	08Cam010	2008	Cambridge, UK	Human	92
Cam011	08Cam011	2008	Cambridge, UK	Human	90
Cam012	08Cam012	2008	Cambridge, UK	Human	77
Cam013	08Cam013	2008	Cambridge, UK	Human	149
Cam014	08Cam014	2008	Cambridge, UK	Human	110
Cam015	08Cam015	2008	Cambridge, UK	Human	84
Cam016	08Cam016	2008	Cambridge, UK	Human	162
Cam017	07Cam017	2007	Cambridge, UK	Human	85
Cam018a	07Cam018a	2007	Cambridge, UK	Human	109
Cam019	07Cam019	2007	Cambridge, UK	Human	84
Cam020	07Cam020	2007	Cambridge, UK	Human	140
Cam021	07Cam021	2007	Cambridge, UK	Human	100
Cam022	07Cam022	2007	Cambridge, UK	Human	162
Cam023	07Cam023	2007	Cambridge, UK	Human	102
Cam024	07Cam024	2007	Cambridge, UK	Human	107
Cam025	07Cam025	2007	Cambridge, UK	Human	118
Cam026	07Cam026	2007	Cambridge, UK	Human	90
Cam027	07Cam027	2007	Cambridge, UK	Human	44
Cam028	07Cam028	2007	Cambridge, UK	Human	172
Cam029	07Cam029	2007	Cambridge, UK	Human	84
Cam030	07Cam030	2007	Cambridge, UK	Human	92
Cam031	07Cam031	2007	Cambridge, UK	Human	340
Cam032	07Cam032	2007	Cambridge, UK	Human	99
Cam033	07Cam033	2007	Cambridge, UK	Human	77
Cam034	07Cam034	2007	Cambridge, UK	Human	42
Cam035	07Cam035	2007	Cambridge, UK	Human	54
Cam036	09Cam036	2009	Cambridge, UK	Human	161

Cam037	10Cam037	2010	Cambridge, UK	Human	109
Cam038	10Cam038	2010	Cambridge, UK	Human	66
Cam039	10Cam039	2010	Cambridge, UK	Human	86
Can001	01Can001	2001	Calgary, Canada	Human	118
Can002	01Can002	2001	Calgary, Canada	Human	97
Can003	01Can003	2001	Calgary, Canada	Human	95
Can004	01Can004	2001	Calgary, Canada	Human	127
Can005	02Can005	2002	Calgary, Canada	Human	97
Can006	02Can006	2002	Calgary, Canada	Human	88
Can007	02Can007	2002	Calgary, Canada	Human	186
Can008	03Can008	2003	Calgary, Canada	Human	104
Can009	03Can009	2003	Montreal, Canada	Human	102
Can010	03Can010	2003	Montreal, Canada	Human	107
Can011	03Can011	2003	Calgary, Canada	Human	101
Can012	03Can012	2003	Montreal, Canada	Human	105
M7404		2005	Montreal, Canada	Human	100
Exe001	07Exe001	2007	Exeter, UK	Human	141
Exe002	07Exe002	2007	Exeter, UK	Human	155
Exe003	08Exe003	2008	Exeter, UK	Human	147
Exe004	07Exe004	2007	Exeter, UK	Human	121
Exe005	07Exe005	2007	Exeter, UK	Human	113
Exe006	07Exe006	2007	Exeter, UK	Human	105
Exe009	07Exe009	2007	Exeter, UK	Human	153
Exe010	07Exe010	2007	Exeter, UK	Human	135
Exe011	07Exe011	2007	Exeter, UK	Human	117
Exe012	07Exe012	2007	Exeter, UK	Human	105
Exe013	07Exe013	2007	Exeter, UK	Human	107
Exe014	08Exe014	2008	Exeter, UK	Human	139
Exe015	07Exe015	2007	Exeter, UK	Human	134
Cd196		1985	Paris, France	Human	121
Gla001	07Gla001	2007	Glasgow, UK	Human	193
Gla002	08Gla002	2008	Glasgow, UK	Human	209

Gla003	08Gla003	2008	Glasgow, UK	Human	202
Gla004	08Gla004	2008	Glasgow, UK	Human	201
Gla005	08Gla005	2008	Glasgow, UK	Human	187
Gla006	08Gla006	2008	Glasgow, UK	Human	202
Gla007	08Gla007	2008	Glasgow, UK	Human	208
Gla008	08Gla008	2008	Glasgow, UK	Human	155
Gla009	08Gla009	2008	Glasgow, UK	Human	181
Gla010	08Gla010	2008	Glasgow, UK	Human	255
Gla012	09Gla012	2009	Glasgow, UK	Human	217
Gla013	09Gla013	2009	Glasgow, UK	Human	319
Gla014	09Gla014	2009	Glasgow, UK	Human	154
Gla015	09Gla015	2009	Glasgow, UK	Human	150
Gla016	09Gla016	2009	Glasgow, UK	Human	147
Gla017	09Gla017	2009	Glasgow, UK	Human	49
Gla018	09Gla018	2009	Glasgow, UK	Human	155
Gla019	09Gla019	2009	Glasgow, UK	Human	237
Gla020	09Gla020	2009	Glasgow, UK	Human	183
Gla021	09Gla021	2009	Glasgow, UK	Human	254
Gla022	09Gla022	2009	Glasgow, UK	Human	161
LSTM031	08GlasL31	2008	Glasgow, UK	Human	110
kor001	06kor001	2006	Seoul, Korea	Human	117
kor002	07kor002	2007	Seoul, Korea	Human	103
kor003	09kor003	2009	Seoul, Korea	Human	120
kor004	09kor004	2009	Seoul, Korea	Human	148
kor005	10kor005	2010	Seoul, Korea	Human	148
kor006	10kor006	2010	Seoul, Korea	Human	156
Liv005	09Liv005	2009	Liverpool, UK	Human	74
Liv008	09Liv008	2009	Liverpool, UK	Human	115
Liv009	09Liv009	2009	Liverpool, UK	Human	107
Liv020	09Liv020	2009	Liverpool, UK	Human	103
Liv021	09Liv021	2009	Liverpool, UK	Human	95
Liv051	08Liv051	2008	Liverpool, UK	Human	57

Liv052	08Liv052	2008	Liverpool, UK	Human	48
Liv053	08Liv053	2008	Liverpool, UK	Human	82
Liv054	08Liv054	2008	Liverpool, UK	Human	97
Liv055	08Liv055	2008	Liverpool, UK	Human	123
Liv056	08Liv056	2008	Liverpool, UK	Human	78
Liv057	08Liv057	2008	Liverpool, UK	Human	81
Liv058	08Liv058	2008	Liverpool, UK	Human	57
Liv059	08Liv059	2008	Liverpool, UK	Human	52
Liv060	08Liv060	2008	Liverpool, UK	Human	109
Liv061	08Liv061	2008	Liverpool, UK	Human	75
Liv062	08Liv062	2008	Liverpool, UK	Human	99
Liv063	09Liv063	2009	Liverpool, UK	Human	123
Liv064	09Liv064	2009	Liverpool, UK	Human	70
Liv065	09Liv065	2009	Liverpool, UK	Human	91
Liv066	09Liv066	2009	Liverpool, UK	Human	119
Liv067	09Liv067	2009	Liverpool, UK	Human	107
Liv068	09Liv068	2009	Liverpool, UK	Human	135
Liv069	09Liv069	2009	Liverpool, UK	Human	126
Liv070	09Liv070	2009	Liverpool, UK	Human	110
Liv071	09Liv071	2009	Liverpool, UK	Human	117
Liv072	09Liv072	2009	Liverpool, UK	Human	122
Liv073	09Liv073	2009	Liverpool, UK	Human	68
Liv074	09Liv074	2009	Liverpool, UK	Human	117
Liv075	09Liv075	2009	Liverpool, UK	Human	142
Liv076	09Liv076	2009	Liverpool, UK	Human	294
Liv077	09Liv077	2009	Liverpool, UK	Human	160
Liv078	09Liv078	2009	Liverpool, UK	Human	77
Liv079	09Liv079	2009	Liverpool, UK	Human	94
Liv081	08Liv081	2008	Liverpool, UK	Human	76
Liv082	08Liv082	2008	Liverpool, UK	Human	116
Liv083	08Liv083	2008	Liverpool, UK	Human	108
Liv084	08Liv084	2008	Liverpool, UK	Human	149

Liv085	08Liv085	2008	Liverpool, UK	Human	81
Liv086	09Liv086	2009	Liverpool, UK	Human	98
Liv087	09Liv087	2009	Liverpool, UK	Human	122
Liv088	09Liv088	2009	Liverpool, UK	Human	122
Liv089	09Liv089	2009	Liverpool, UK	Human	104
Liv090	09Liv090	2009	Liverpool, UK	Human	108
Liv091	09Liv091	2009	Liverpool, UK	Human	111
Liv092	09Liv092	2009	Liverpool, UK	Human	87
Liv093	09Liv093	2009	Liverpool, UK	Human	77
Liv094	09Liv094	2009	Liverpool, UK	Human	113
Liv095	09Liv095	2009	Liverpool, UK	Human	213
Liv096	09Liv096	2009	Liverpool, UK	Human	98
Liv097	09Liv097	2009	Liverpool, UK	Human	86
Liv098	09Liv098	2009	Liverpool, UK	Human	109
Liv1	08Liv1	2008	Liverpool, UK	Human	46
Liv100	10Liv100	2010	Liverpool, UK	Human	80
Liv101	10Liv101	2010	Liverpool, UK	Human	73
Liv102	10Liv102	2010	Liverpool, UK	Human	105
Liv103	10Liv103	2010	Liverpool, UK	Human	74
Liv104	10Liv104	2010	Liverpool, UK	Human	66
Liv10a	09Liv10a	2009	Liverpool, UK	Human	163
Liv11	08Liv11	2008	Liverpool, UK	Human	61
Liv12	09Liv12	2009	Liverpool, UK	Human	63
Liv127	10Liv127	2010	Liverpool, UK	Human	105
Liv13	09Liv13	2009	Liverpool, UK	Human	53
Liv131	10Liv131	2010	Liverpool, UK	Human	99
Liv132	10Liv132	2010	Liverpool, UK	Human	102
Liv136	10Liv136	2010	Liverpool, UK	Human	82
Liv14	09Liv14	2009	Liverpool, UK	Human	55
Liv145	09Liv145	2009	Liverpool, UK	Human	403
Liv146	09Liv146	2009	Liverpool, UK	Human	109
Liv147	09Liv147	2009	Liverpool, UK	Human	122

Liv148	09Liv148	2009	Liverpool, UK	Human	98
Liv149	09Liv149	2009	Liverpool, UK	Human	85
Liv15	09Liv15	2009	Liverpool, UK	Human	44
Liv16	09Liv16	2009	Liverpool, UK	Human	49
Liv17	09Liv17	2009	Liverpool, UK	Human	42
Liv170	09Liv170	2009	Liverpool, UK	Human	110
Liv171	09Liv171	2009	Liverpool, UK	Human	105
Liv172	09Liv172	2009	Liverpool, UK	Human	225
Liv173	09Liv173	2009	Liverpool, UK	Human	114
Liv174	09Liv174	2009	Liverpool, UK	Human	110
Liv175	09Liv175	2009	Liverpool, UK	Human	111
Liv176	09Liv176	2009	Liverpool, UK	Human	113
Liv177	09Liv177	2009	Liverpool, UK	Human	115
Liv178	09Liv178	2009	Liverpool, UK	Human	112
Liv179	09Liv179	2009	Liverpool, UK	Human	119
Liv18	08Liv18	2008	Liverpool, UK	Human	55
Liv180	09Liv180	2009	Liverpool, UK	Human	110
Liv181	09Liv181	2009	Liverpool, UK	Human	109
Liv182	09Liv182	2009	Liverpool, UK	Human	124
Liv183	09Liv183	2009	Liverpool, UK	Human	128
Liv184	09Liv184	2009	Liverpool, UK	Human	98
Liv185	09Liv185	2009	Liverpool, UK	Human	125
Liv186	09Liv186	2009	Liverpool, UK	Human	134
Liv187	09Liv187	2009	Liverpool, UK	Human	116
Liv188	07Liv188	2007	Liverpool, UK	Human	112
Liv189	07Liv189	2007	Liverpool, UK	Human	142
Liv19	09Liv19	2009	Liverpool, UK	Human	46
Liv190	07Liv190	2007	Liverpool, UK	Human	120
Liv2	09Liv2	2009	Liverpool, UK	Human	44
Liv20	09Liv20	2009	Liverpool, UK	Human	56
Liv21	09Liv21	2009	Liverpool, UK	Human	44
Liv3	09Liv3	2009	Liverpool, UK	Human	47

Liv38	09Liv38	2009	Liverpool, UK	Human	64
Liv4	09Liv4	2009	Liverpool, UK	Human	45
Liv40	09Liv40	2009	Liverpool, UK	Human	72
Liv41	09Liv41	2009	Liverpool, UK	Human	63
Liv48	09Liv48	2009	Liverpool, UK	Human	63
Liv49	09Liv49	2009	Liverpool, UK	Human	73
Liv5	09Liv5	2009	Liverpool, UK	Human	33
Liv6a	09Liv6a	2009	Liverpool, UK	Human	110
Liv7	09Liv7	2009	Liverpool, UK	Human	45
CD679		2009	London, UK	Human	110
ham001	02ham001	2002	London, UK	Human	121
ham002	07ham002	2007	London, UK	Human	133
ham003a	09ham003a	2009	London, UK	Human	127
Ham004	07Ham004	2007	London, UK	Human	129
Ham005	03Ham005	2003	London, UK	Human	128
ham006	07ham006	2007	London, UK	Human	83
ham007	03ham007	2003	London, UK	Human	182
ham009	09ham009	2009	London, UK	Human	123
ham010a	08ham010a	2008	London, UK	Human	97
ham011	04ham011	2004	London, UK	Human	106
lon001			London, UK	Human	92
lon004			London, UK	Human	86
lon005			London, UK	Human	95
lon006			London, UK	Human	126
LSTM35	NottR20291L35		Nottingham, UK	Human	97
LSTM36	MaidL36	2006	Maidstone, UK	Human	80
Wal002	98PresW2	1998	Preston, UK	Human	108
LSTM025	07DundL25	2007	Dundee, UK	Human	121
LSTM026	08DumfL26	2008	Dumfries, UK	Human	163
LSTM027	08AyrsL27	2008	Ayrshire, UK	Human	131
LSTM028	08EdinL28	2008	Edinburgh, UK	Human	158
LSTM029	08DumbL29	2008	Dumbarton, UK	Human	166

LSTM030	08InveL30	2008	Inverness, UK	Human	106
Sin001	08Sin001	2008	Singapore	Human	86
Sin002	09Sin002	2009	Singapore	Human	84
Sin003	09Sin003	2009	Singapore	Human	109
R20291a		2005	Stoke Mandeville, UK	Human	128
2004013	04MEa013	2004	Maine, USA	Human	37
2004102	04NJc102	2004	New Jersey, USA	Human	50
2004118	04MEb118	2004	Maine, USA	Human	30
2004163	04PA163	2004	Pennsylvania, USA	Human	32
2006439	06AZ439	2006	Arizona, USA	Food	38
2007140	07MD140	2007	Maryland, USA	Human	29
2007218		2007	Arizona, USA	Food	29
2007221a	07AZb221aF	2007	Arizona, USA	Food	128
2007223a	07AZc223aF	2007	Arizona, USA	Food	129
2007825		2007	USA	Human	27
2007833	07AZa833	2007	Arizona, USA	Human	9
2007837		2007	USA	Human	53
2007850		2007	USA	Human	57
2007855		2007	USA	Bovine	40
BI-1_L21	88MinnBI-1a	1988	Minneapolis, MN, USA	Human	143
BI-10a	01PitBBI-10a	2001	Pittsburgh, PA, USA	Human	119
BI-11	01PitABI-11	2001	Pittsburgh, PA, USA	Human	51
BI-12_L23	04CampBI-12a	2004	Camp Hill, PA, USA	Human	138
BI-13	04NJbBI-13	2004	New Jersey, USA	Human	62
BI-15	04NJaBI-15	2004	New Jersey, USA	Human	90
BI-17_L24	04MontBI-17a	2004	Montreal, Canada	Human	124
BI-2a	91TuscBI-2a	1991	Tuscon, AZ, USA	Human	100
BI-3a	90MinnBI-3a	1990	Minneapolis, MN, USA	Human	132
BI-4	93MinnBI-4	1993	Minneapolis, MN, USA	Human	105
BI-5	95AlbaBI-5	1995	Albany, NY, USA	Human	74
BI-6	03PorABI-6	2003	Portland, OR, USA	Human	38
BI-6p	04AtlaBI-6p	2004	Atlanta, GA, USA	Human	60

BI-7_L22	03PorBBI-7a	2003	Portland, OR, USA	Human	88
BI-8	04PortBI-8	2004	Portland, ME, USA	Human	78
BI-8_L32	08Ayr195	2008	Ayrshire, UK	Human	101
LSTM001	04MEL01	2004	Maine, USA	Human	153
LSTM002	05PAL02	2005	Pennsylvania, USA	Human	135
LSTM003	06AZL03E	2006	Arizona, USA	Equine	133
LSTM004	07CTNL04	2007	FOODNET-CT, USA	Human	132
LSTM005	07AZL05F	2007	Arizona, USA	Food	163
LSTM006	07AZL06F	2007	Arizona, USA	Food	120
LSTM007	07TNNL07	2007	FOODNET-TN, USA	Human	119
LSTM009	07AZL09F	2007	Arizona, USA	Food	137
LSTM011	07AZL11F	2007	Arizona, USA	Food	164
LSTM012	07AZL12F	2007	Arizona, USA	Food	131
LSTM013	7828L13	2007	USA	Human	168
LSTM014	7832L14	2007	USA	Human	133
LSTM016	7839L16	2007	USA	Human	140
LSTM019	7850L19H	2007	USA	Household	123
LSTM020	7851L20H	2007	USA	Household	153
LSTM33				Human	105
LSTM34				Human	107
LSTM37				Human	90
LSTM38				Human	95

Appendix B

C. difficile BI/NAP1/027 local hospital collection used in Chapter 4

CDU – Clinical Decision Unit (A&E)

HEC – Heart Emergency Centre (A&E)

AMAU – Acute Medical Admission Unit (A&E)

POCU – Post-Operative Critical Unit

Aintree – Aintree University Hospital

BGH – Broad Green Hospital

Wirral – Wirral University Teaching Hospital

Name	Date (dd/mm/yyyy)	Ward	StudyNo	Alternative Name
Liv1	05/12/2008	CDU	CDP080028	
Liv2	02/03/2009	3X	CDP080028	
Liv3	12/01/2009	5X	CDP080034	
Liv4	09/07/2009	GP	CDP080034	
Liv5	12/03/2009	6A	CDP080046	
Liv6	31/05/2009	9B	CDP080046	
Liv7	31/03/2009	11BGH	CDP080047	
Liv8	10/07/2009	HEC	CDP080047	
Liv9	03/01/2009	AMAU	CDP080033	
Liv10	19/02/2009	3X	CDP080033	
Liv11	14/08/2008	7Y	CDN080506	
Liv12	13/02/2009	3X	CDN080506	
Liv13	29/03/2009	7Y	CDN080506	
Liv14	23/02/2009	POCU	CDP080049	
Liv15	09/07/2009	2B	CDP080049	
Liv16	27/02/2009	GP	CDP080048	
Liv17	08/04/2009	6A	CDP080048	
Liv18	09/12/2008	5A	CDP080029	
Liv19	05/05/2009	AMAU	CDP080029	
Liv20	25/04/2009	3X	CDP080051	
Liv21	01/06/2009	7Y	CDP080051	
Liv38	07/10/2009	AMAU	CDN080567	
Liv40	19/11/2009	5A	CDA081055	
Liv41	29/12/2009	3A	CDA081055	
Liv48	16/10/2009	8Y	CDN080568	
Liv49	16/11/2009	3X	CDN080568	

Liv051	22/07/2008	2A	CDP080004
Liv052	18/11/2008	5X	CDP080005
Liv053	23/08/2008	6A	CDP080010
Liv054	06/11/2008	7Y	CDP080018
Liv055	19/11/2008	3X	CDP080020
Liv056	19/11/2008	5A	CDP080021
Liv057	23/11/2008	3X	CDP080022
Liv058	22/11/2008	6B	CDP080024
Liv059	21/11/2008	AMAU	CDP080025
Liv060	28/11/2008	7B	CDP080026
Liv061	04/12/2008	9X	CDP0800027
Liv062	31/12/2008	5Y	CDP080032
Liv063	21/01/2009	5A	CDP080036
Liv064	01/02/2009	5B	CDP080039
Liv065	03/02/2009	5B	CDP080040
Liv066	03/02/2009	5B	CDP080041
Liv067	26/02/2009	8B	CDP080043
Liv068	03/03/2009	AMAU	CDP080044
Liv069	20/04/2009	2B	CDP080052
Liv070	30/04/2009	8X	CDP080053
Liv071	08/05/2009	3X	CDP080054
Liv072	12/05/2009	8X	CDP080056
Liv073	21/07/2009	5BGH	CDP080062
Liv074	25/08/2009	AMAU	CDP080065
Liv075	25/08/2009	AMAU	CDP080068
Liv076	21/09/2009	8X	CDP080070
Liv077	28/10/2009	2B	CDP080073
Liv078	17/11/2009	3Y	CDP080075
Liv079	03/12/2009	3A	CDP080077
Liv081	28/07/2008	2B	CDN080505
Liv082	13/08/2008	3C	CDN080508
Liv083	03/10/2008	6A	CDN080515
Liv084	04/11/2008	3B	CDN080518
Liv085	10/12/2008	8Y/9X	CDN080521
Liv086	08/01/2009	5B	CDN080523
Liv087	28/01/2009	5Y	CDN0805027
Liv088	01/03/2009	5B	CDN080531
Liv089	06/04/2009	5A	CDN080533
Liv090	27/04/2009	7Y	CDN080538
Liv091	30/04/2009	8X	CDN080539
Liv092	22/05/2009	5A	CDN080542
Liv093	08/06/2009	8X	CDN080546
Liv094	21/07/2009	2X	CDN080550
Liv095	07/08/2009	9X	CDN080552
Liv096	10/08/2009	9B	CDN080553
Liv097	10/10/2009	5X	CDN080566
Liv098	28/10/2009	6A	CDN080569
Liv100	10/02/2010	11BGH	CDP080079
Liv101	25/02/2010	11BGH	CDP080081
Liv102	28/03/2010	AMAU	CDP080082
Liv103	26/02/2010	11BGH	CDN080578
Liv104	15/03/2010	11BGH	CDN080580
Liv127	20/05/2010	3X	CDN080580
Liv131	16/05/2010	Aintree	CDP08UHA0009
Liv132	21/02/2010	Aintree	CDN08UHA0501

Liv136	16/05/2010	Wirral	CDN08WTH0507	
Liv145	08/03/2009	3X	CDP080034	
Liv146	29/05/2009	GP	CDP080034	
Liv147	15/04/2009	5A	CDP080049	
Liv148	11/09/2009	GP	CDP080048	
Liv149	23/04/2009	7B	CDP080051	
Liv170	26/07/2009	2A	2A1	
Liv171	01/04/2009	2B	2B1a	
Liv172	26/05/2009	2A	2B1b	
Liv173	25/04/2009	2B	2B2a	
Liv174	29/06/2009	2A	2B2b	
Liv175	04/05/2009	2Y	2Y1	
Liv176	13/07/2009	2Y	2Y2	
Liv177	25/07/2009	2Y	2Y3	
Liv178	21/02/2009	5A	5A1	
Liv179	08/02/2009	5B	5B1	
Liv180	03/08/2009	5Y	5Y1	
Liv181	11/09/2009	5Y	5Y2	
Liv182	14/02/2009	7Y	7Y1	
Liv183	15/02/2009	7Y	7Y2	
Liv184	22/04/2009	7Y	7Y3	
Liv185	04/05/2009	7Y	7Y4	
Liv186	13/05/2009	7Y	7Y5	
Liv187	11/05/2009	ACRU (8X)	8X1	
Liv188	15/05/2007	5A	2746	
Liv189	02/07/2007	2B	3584	
Liv190	03/07/2007	7 BGH	3615	
Liv150	12/03/2009			Liv5a
Liv151	12/03/2009			Liv5b
Liv152	12/03/2009			Liv5c
Liv153	12/03/2009			Liv5d
Liv154	12/03/2009			Liv5e
Liv155	14/08/2008			Liv11a
Liv156	14/08/2008			Liv11b
Liv157	14/08/2008			Liv11c
Liv158	14/08/2008			Liv11d
Liv159	14/08/2008			Liv11e
Liv160	08/04/2009			Liv17a
Liv161	08/04/2009			Liv17b
Liv162	08/04/2009			Liv17c
Liv163	08/04/2009			Liv17d
Liv164	08/04/2009			Liv17e
Liv165	09/12/2008			Liv18a
Liv166	09/12/2008			Liv18b
Liv167	09/12/2008			Liv18c
Liv168	09/12/2008			Liv18d
Liv169	09/12/2008			Liv18e