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.*

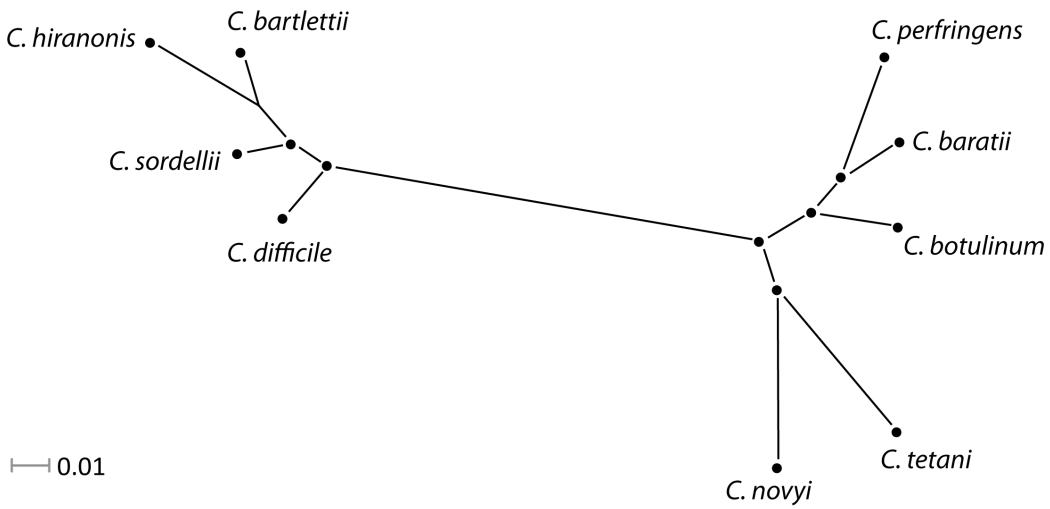

Figure 1.1: Phylogenetic tree of *C. difficile* and other *Clostridium* species. Created using PHYML (Guindon *et al.*, 2010) with 16S rRNA gene sequences retrieved from RDP Project (Cole *et al.*, 2009). Scale bar indicates substitutions per site.

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.

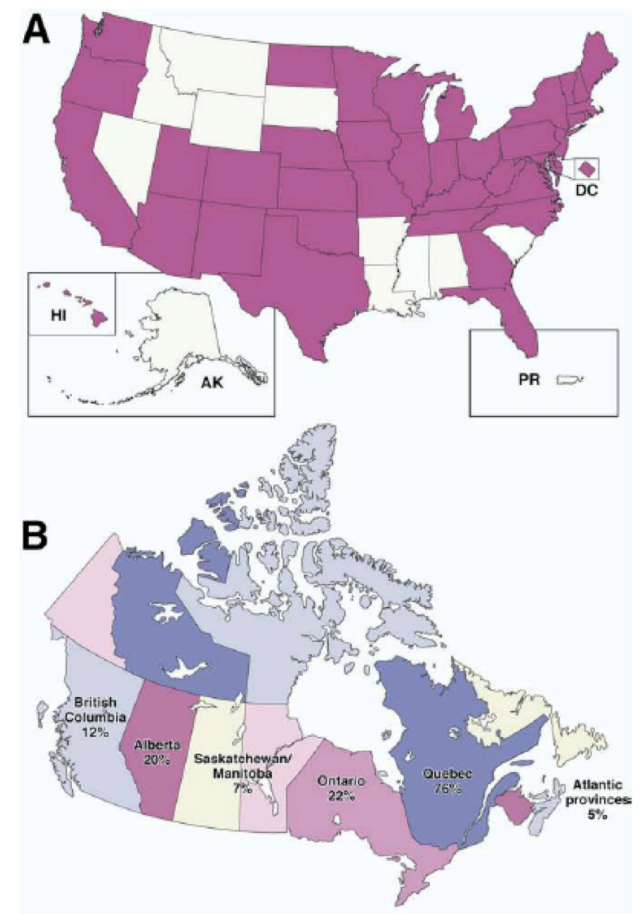

Figure 1.3: Distribution of BI/NAP1/027 *C. difficile* in the USA and Canada. Reproduced from (OConnor *et al.*, 2009). (*A*) States in the USA with >1 hospital that has reported CDI associated with BI/NAP1/027 as of October 2008 (shown in red). (*B*) Percentage of BI/NAP1/027 *C. difficile* isolates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

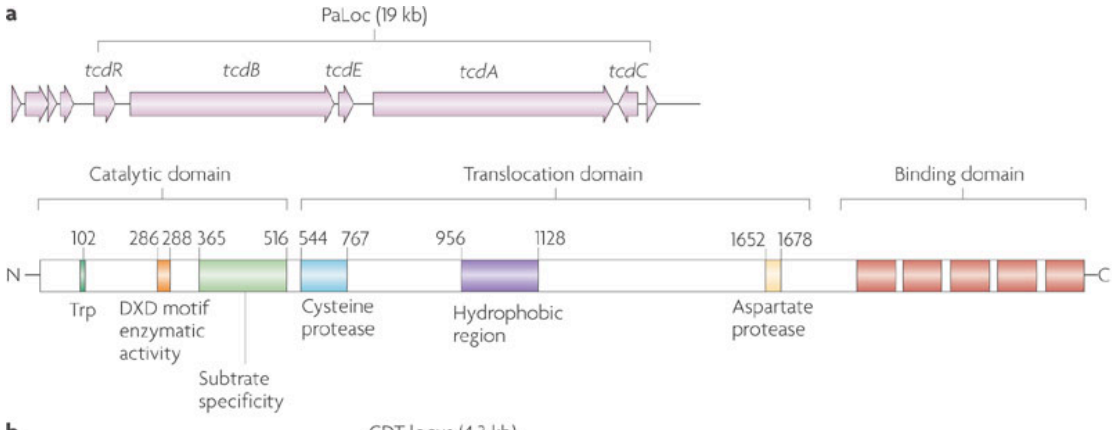

Figure 1.4: *C. difficile* pathogenicity locus (PaLoc) and domain organization of toxin B (below). Reproduced from (Rupnik *et al.*, 2009).

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).

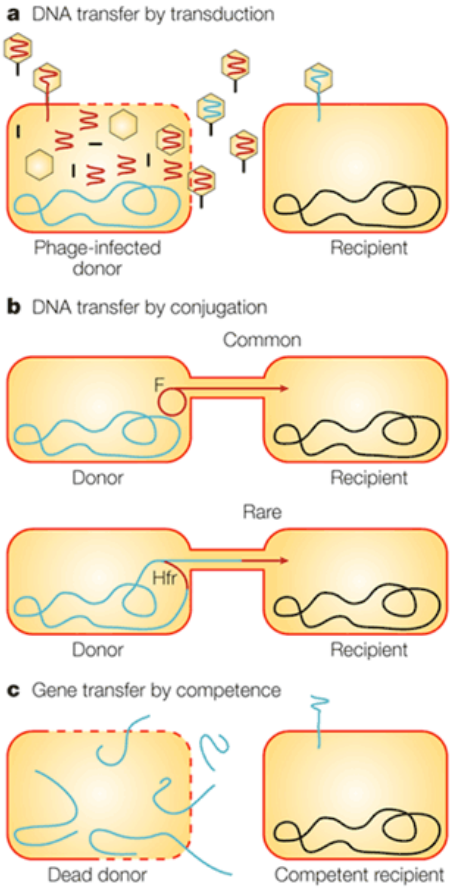

Figure 1.5: Mechanisms for DNA transfer in bacteria. Reproduced from (Redfield, 2001). a) Phage-mediated transduction. DNA from the host chromosome is accidentally packaged into the phage particle and transferred to the recipient cell. b) Conjugation by F plasmid or Hfr-mediated conjugation. c) After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

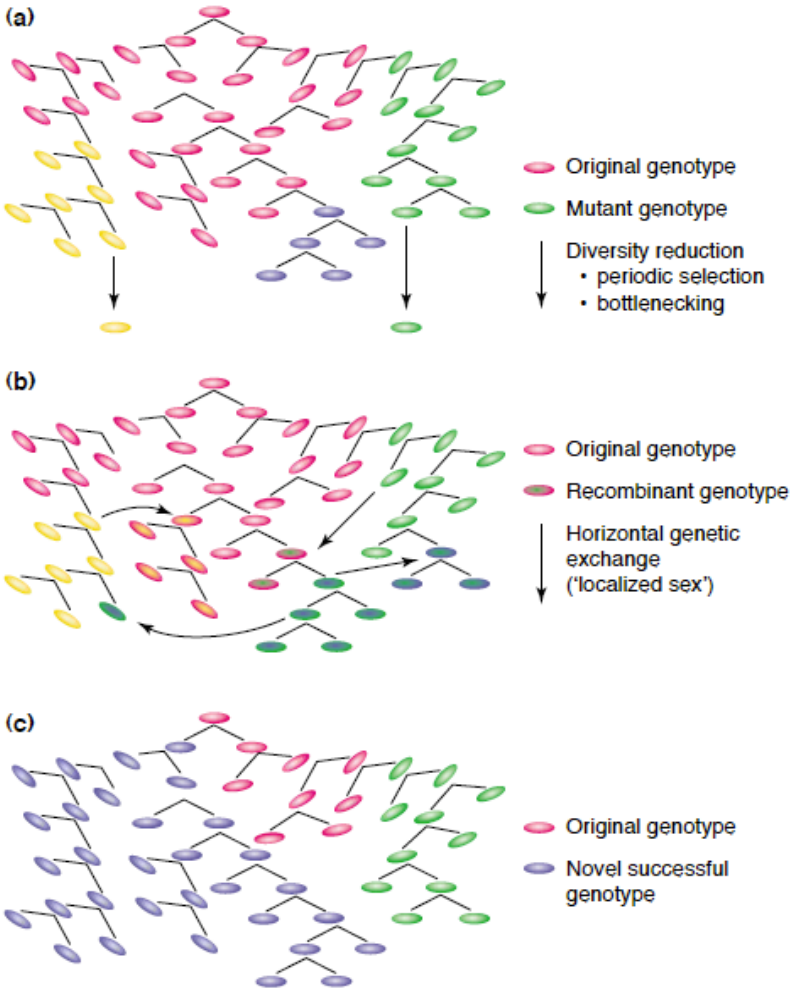

Figure 1.6: Models for bacterial population structure shaped by selection and demographic processes: (a) a clonal population with mutation and stochastic events causing diversity reduction; (b) a population with mutation, horizontal genetic exchange and recombination; and (c) dominance by a successful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

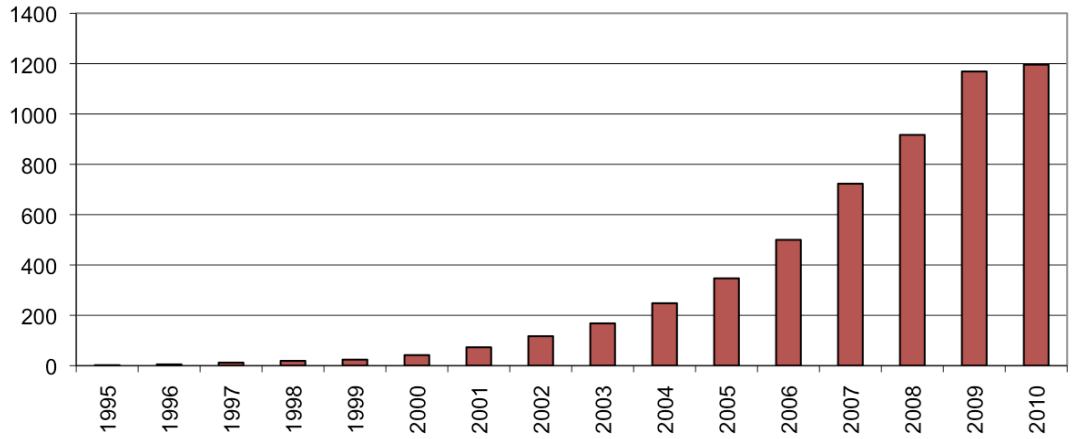

Figure 1.7: Number of sequenced complete genomes in each year (y-axis) from 1995 to 2010 (Genbank). Data sourced from Genomes Online Database.

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.

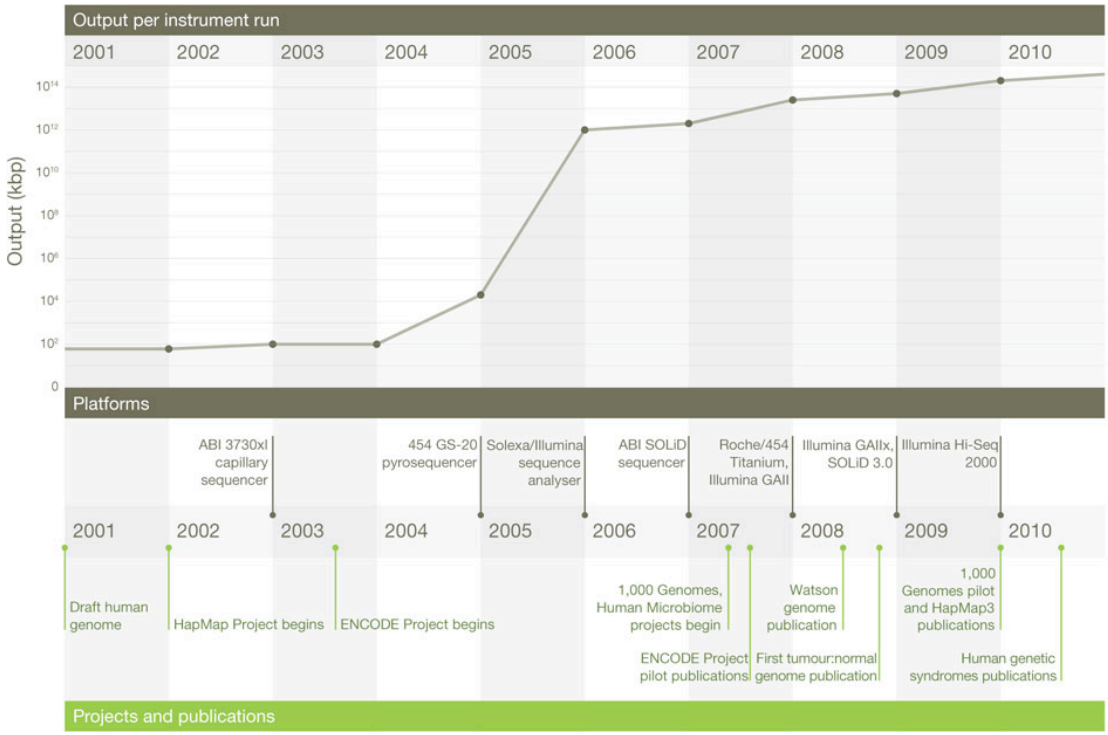

Figure 1.8: Increase in sequencing capacity during the first decade in 21st century. Reproduced from (Mardis, 2011). Top: amounts of data per run per instrument; middle: sequencing platforms; bottom: major genome projects.

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 Illumina HiSeq 2000	95 Gb/run (9 d) 600 Gb/run (11 d)	150 100	$2.0/Mb $0.1/Mb	99.8%

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).

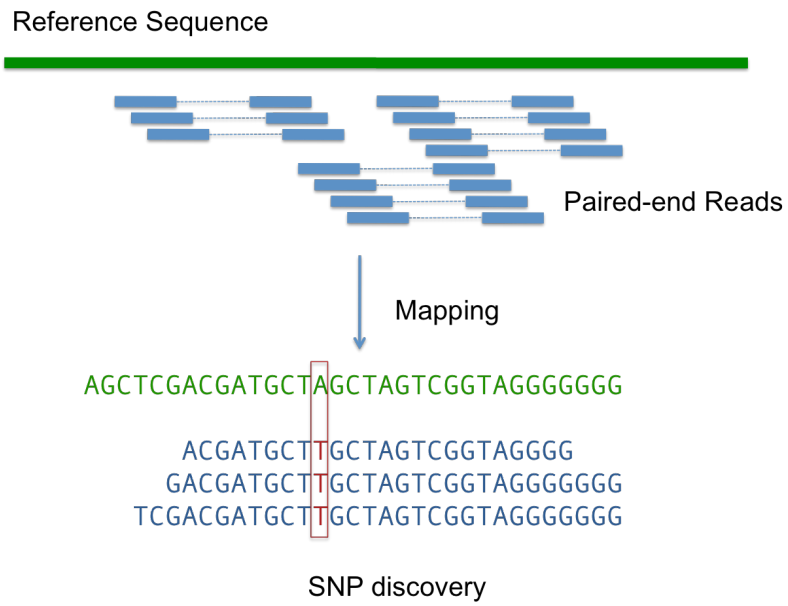

Figure 1.9: Mapping paired-end reads to a reference sequence and identify SNPs. Reproduced from (Nagarajan and Pop, 2010).

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.