Chapter 3

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

3.1 Introduction

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

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

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

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

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

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

3.2 Materials and Methods

3.2.1 Bacterial isolates

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

3.2.2 Sequencing, mapping, and SNP detection

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

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

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

3.2.3 Assessing accuracy of SNP detection

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

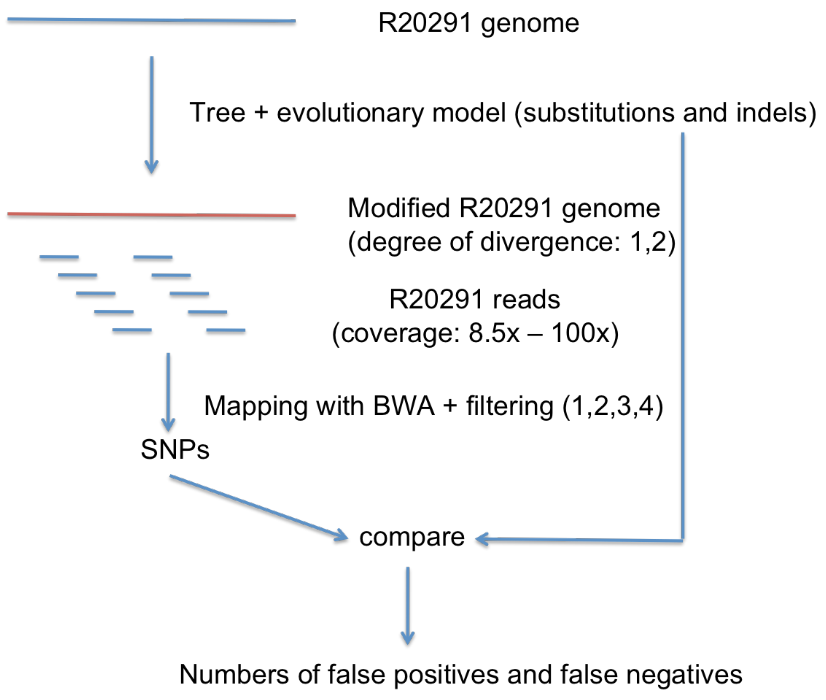

R20291 genome
Tree + evolutionary model (substitutions and indels)
Modified R20291 genome
(degree of divergence: 1,2)
R20291 reads
(coverage: 8.5x – 100x)
Mapping with BWA + filtering (1,2,3,4)
SNPs
compare
Numbers of false positives and false negatives

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

3.2.4 Phylogenetic analysis

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

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

3.2.5 Core and accessory genome

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

3.2.6 Identification of homoplasic characters and homologous recombination

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

3.2.7 Population history and mutation rate

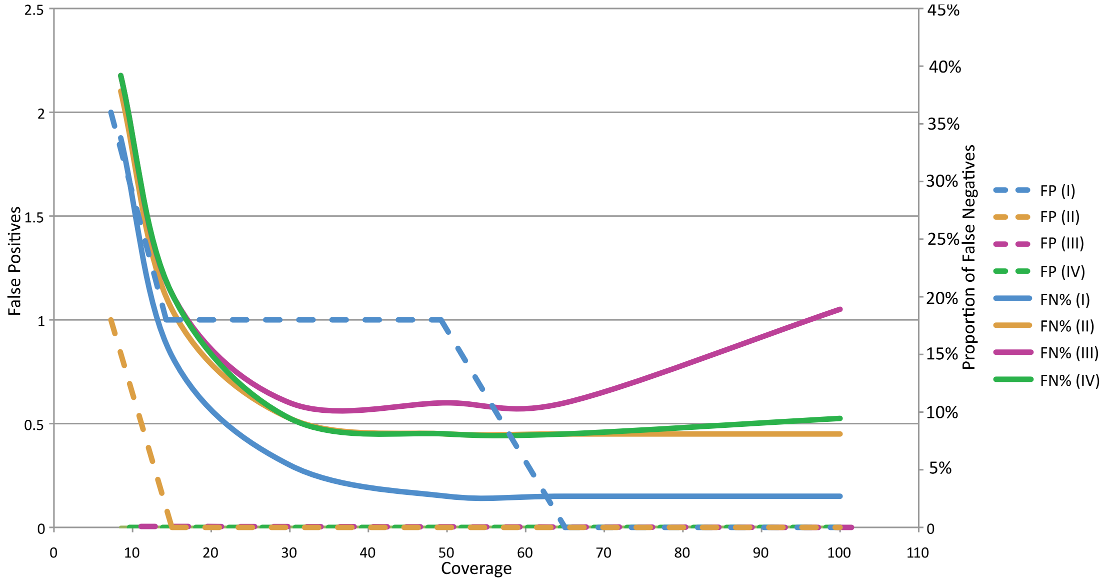
The apparent mutation rate was estimated using two methods: 1), A full maximum likelihood model assuming a rapid expansion which results in perfect star genealogies, implemented in an R script (Morelli *et al.*, 2010); and 2), the program BEAST (Drummond and Rambaut, 2007). BEAST analyses were carried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

3.3 Results

3.3.1 Assessment of SNP detection method

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

Scale 1 (74 SNPs)



Scale 2 (869 SNPs)

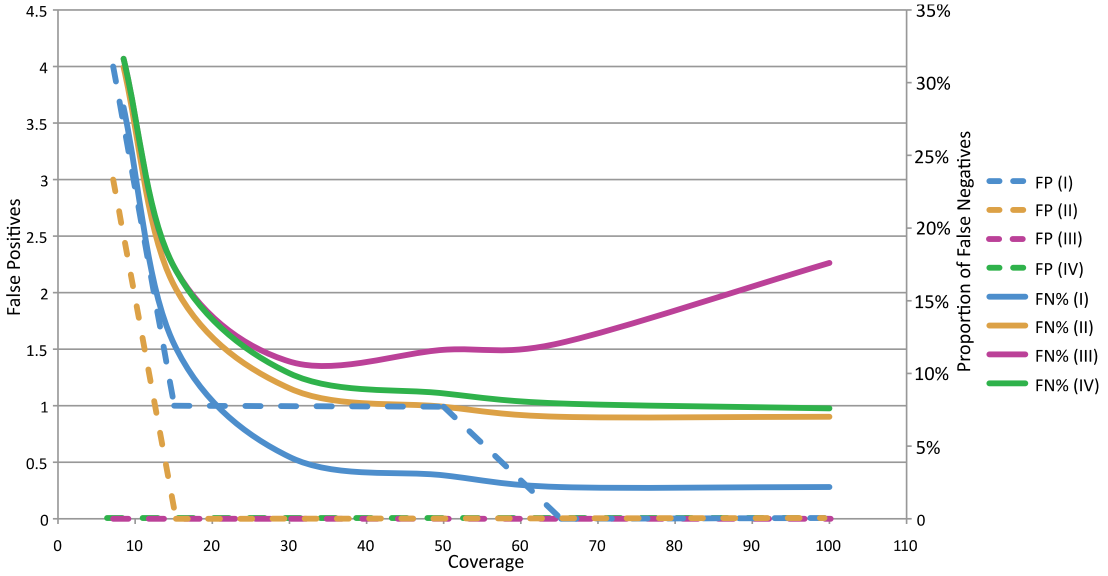


Figure 3.2: Numbers of false positive SNPs (left axis) and percentage of false negative SNPs (right axis) in relation to sequencing data coverage (x-axis) and different SNP filtering and validation measures (coloured lines). Dashed lines represent the absolute numbers of false positive SNPs; solid lines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

3.3.2 Phylogenetic relationship

3.3.2.1 Maximum-likelihood phylogeny and phylogenetic networks

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

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

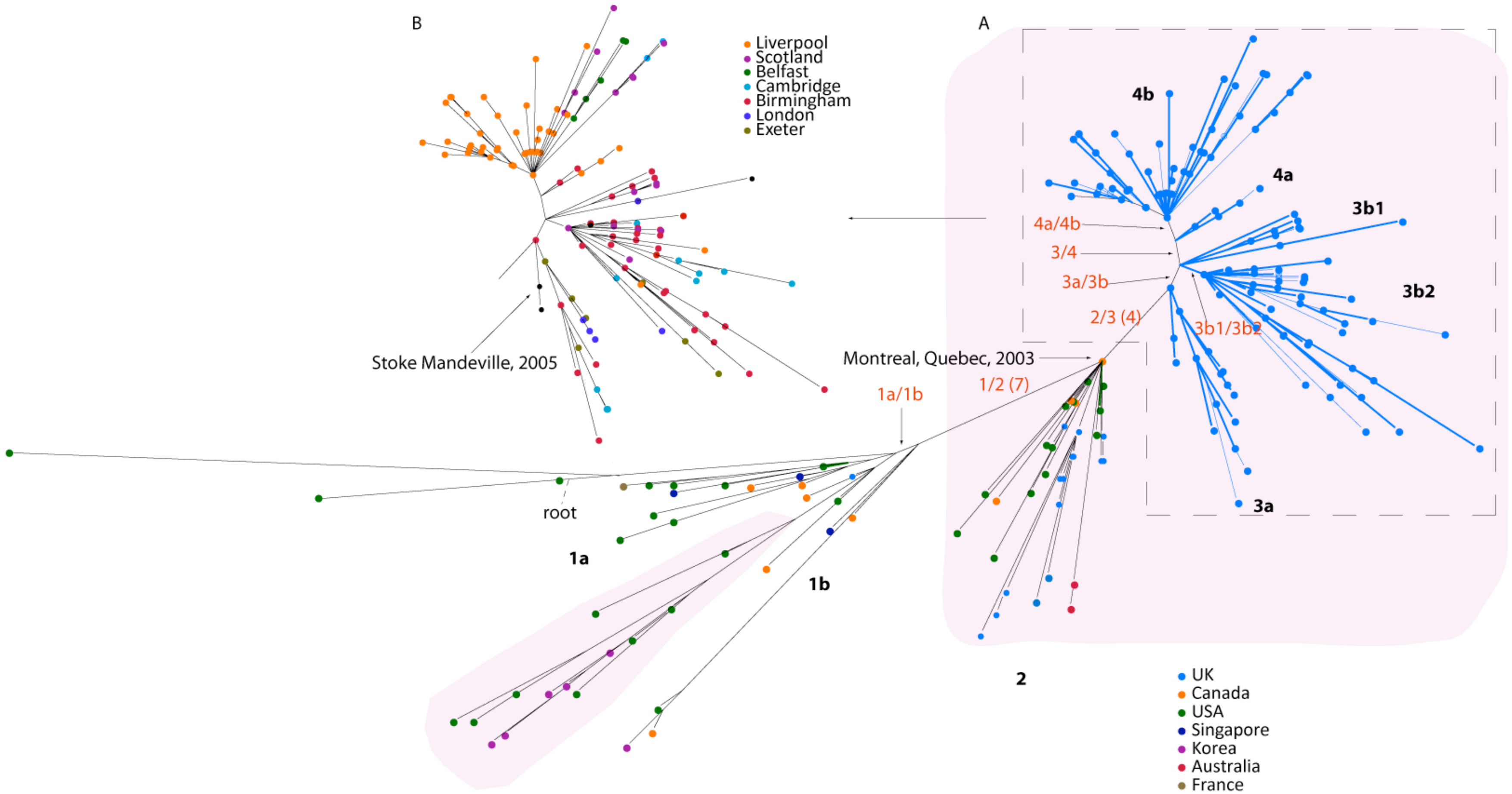


Figure 3.3: Global phylogeny of *C. difficile* BI/NAP1/027. A). Phylogenetic tree of a global collection of 339 isolates based on core genome SNPs. Nodes are coloured according to origin of isolate. Letters in bold denote the clade names (1a, 1b, 2, 3a, 3b1, 3b2, 4a and 4b). Orange text indicates SNPs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

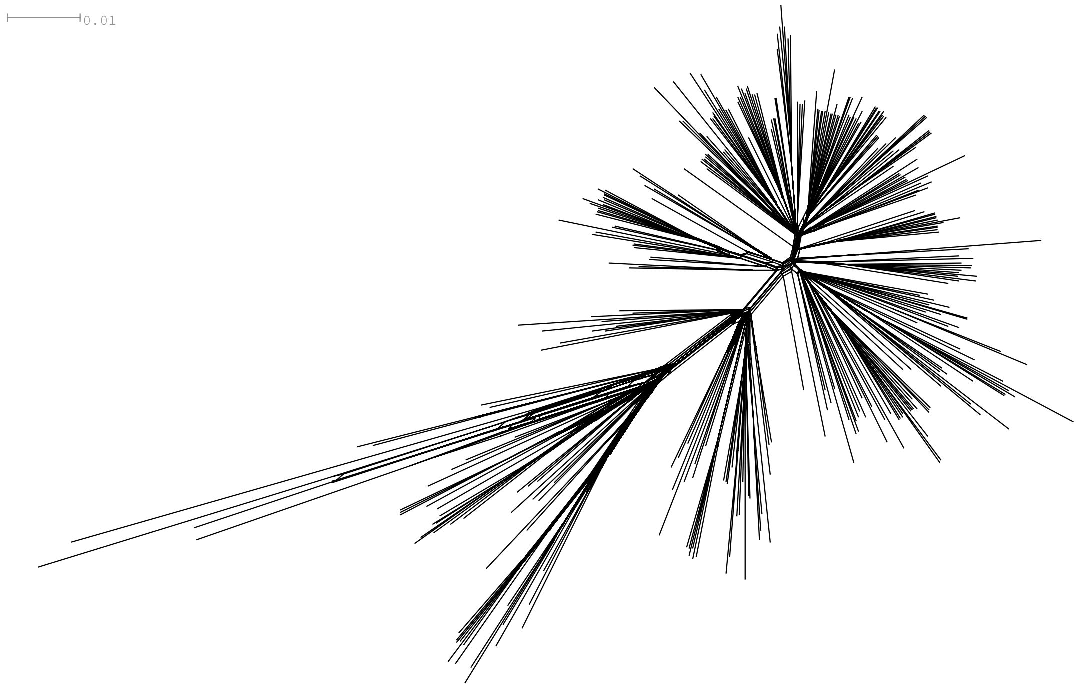


Figure 3.4. NeighbourNet network of the *C. difficile* BI/NAP1/027 lineage. Label indicates genetic distance per site.

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

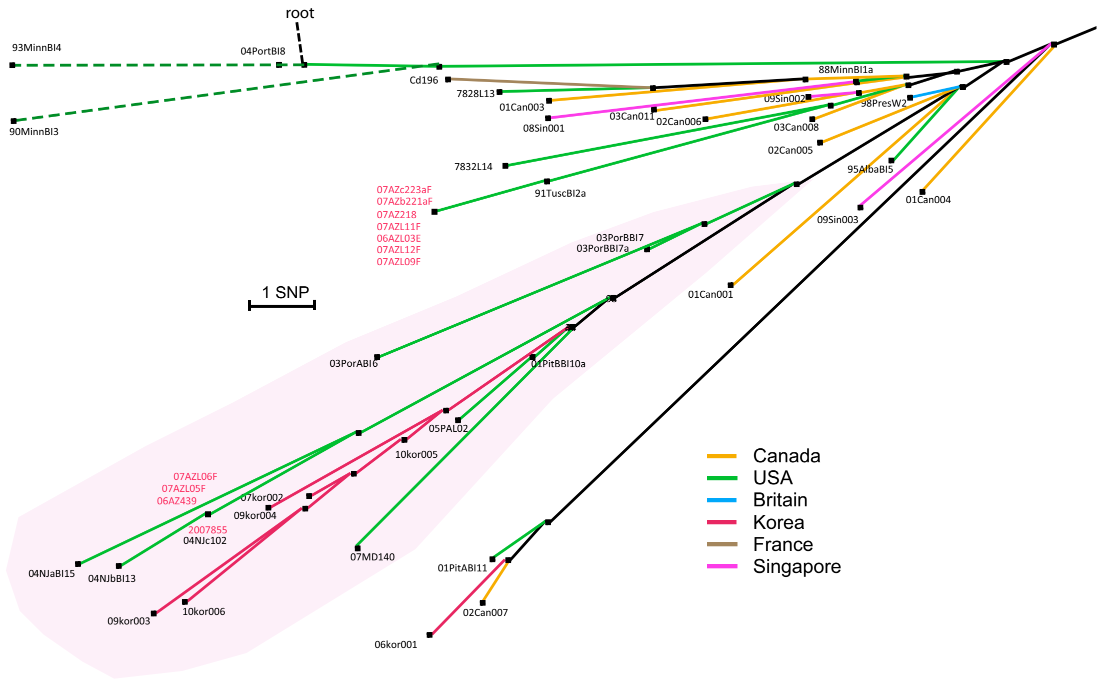


Figure 3.5: Expansion of clade 1 from Figure 3.3. Branches are coloured according to sampling locations of the isolates the branches directly lead to. Isolate names are shown in black (human isolates) or red (isolates from animals and food sources). The lineage harbouring the *gyrA* mutation is shown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

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

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

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

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

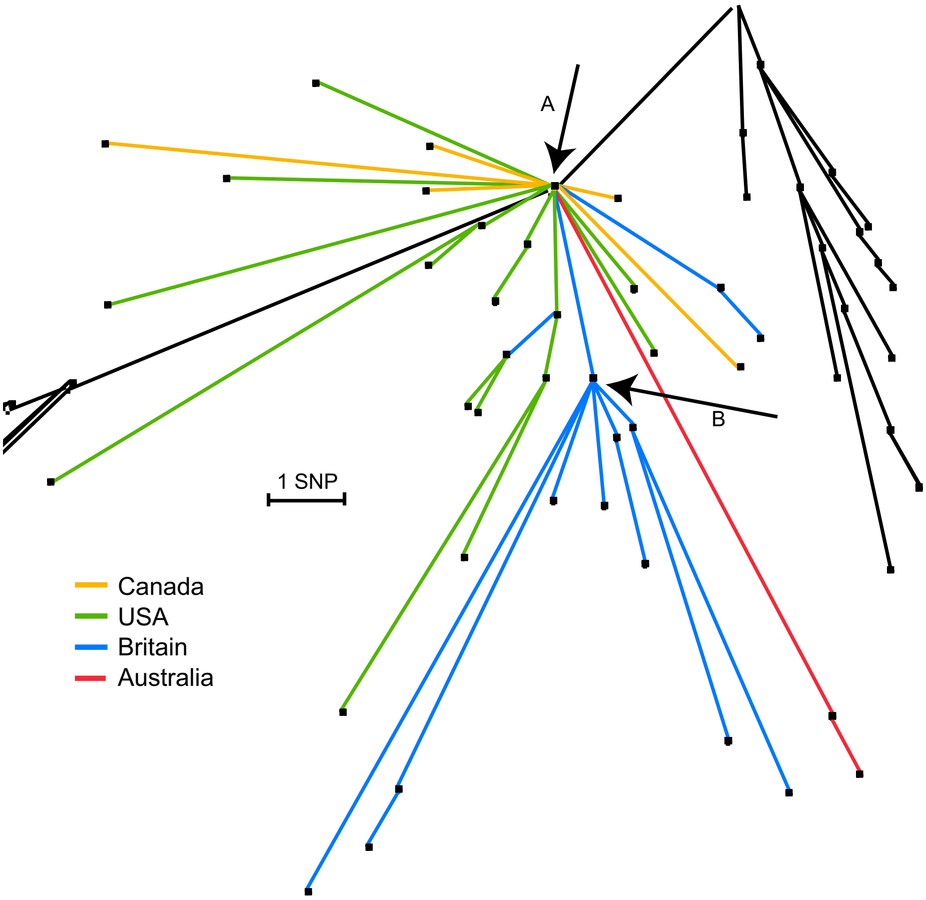
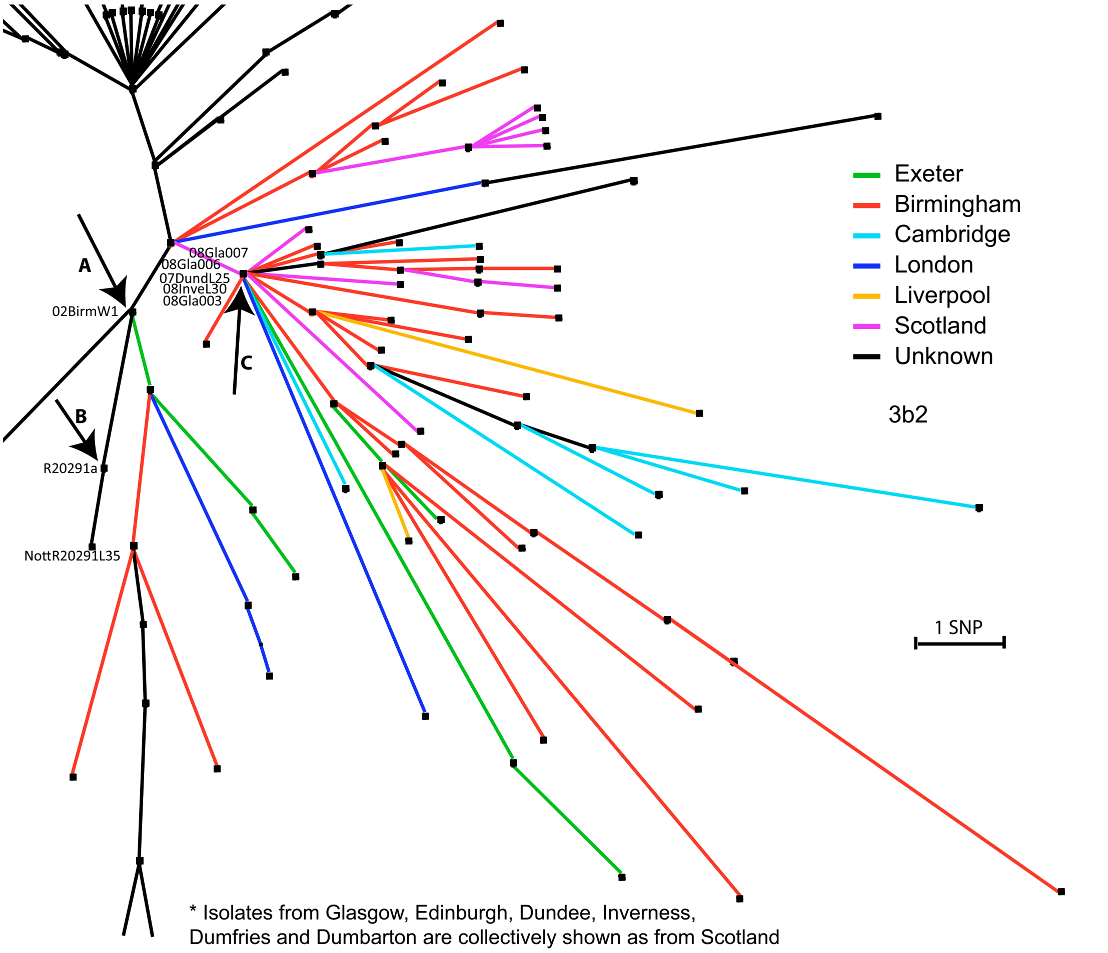


Figure 3.6: Expansion of clade 2 from Figure 3.3. Branches are coloured according to sampling locations of the strains the branches directly lead to. Arrows point to the isolate or lineage that are mentioned in the text.

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



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

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

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

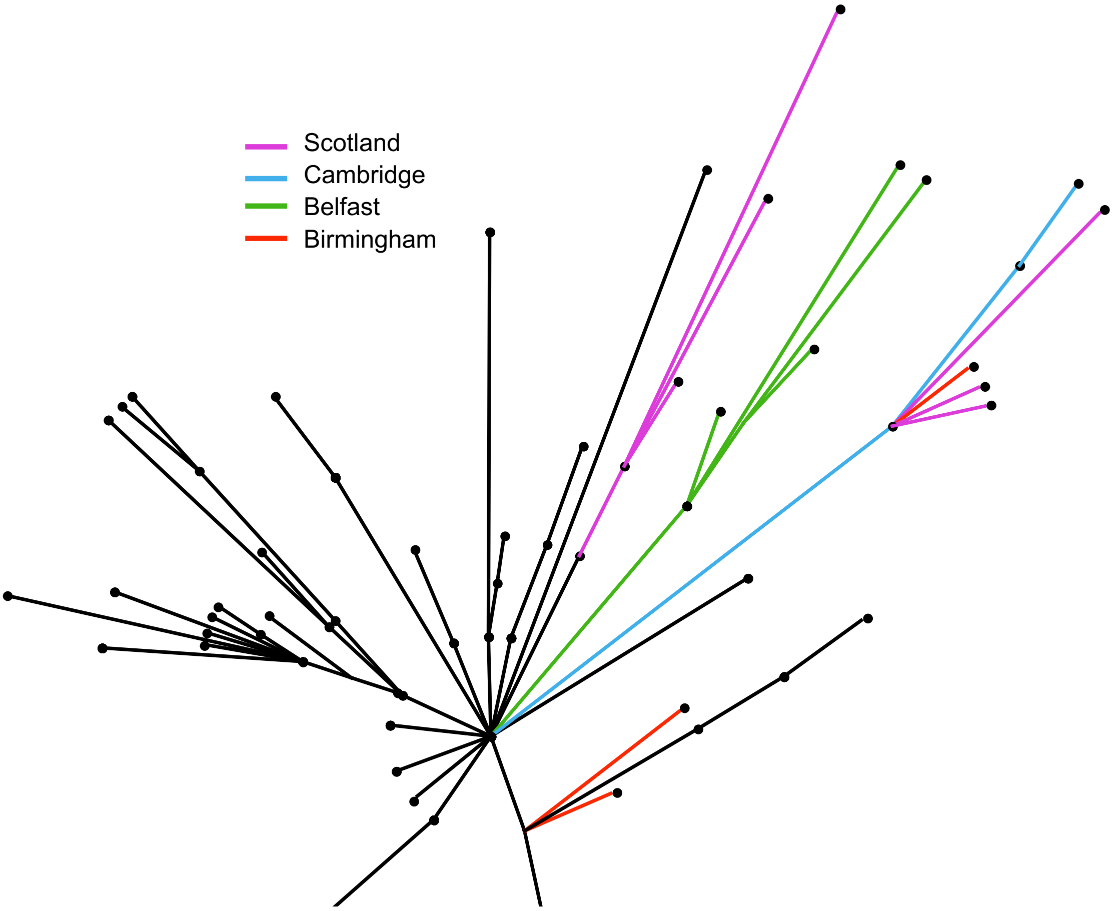


Figure 3.8: Expansion of clade 4 from Figure 3.3. Branches are coloured according to sampling locations of the strains the branches directly lead to, except isolates from Liverpool (shown in black).

3.3.2.2 Phylogenetic relationships inferred with Bayesian analysis

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

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

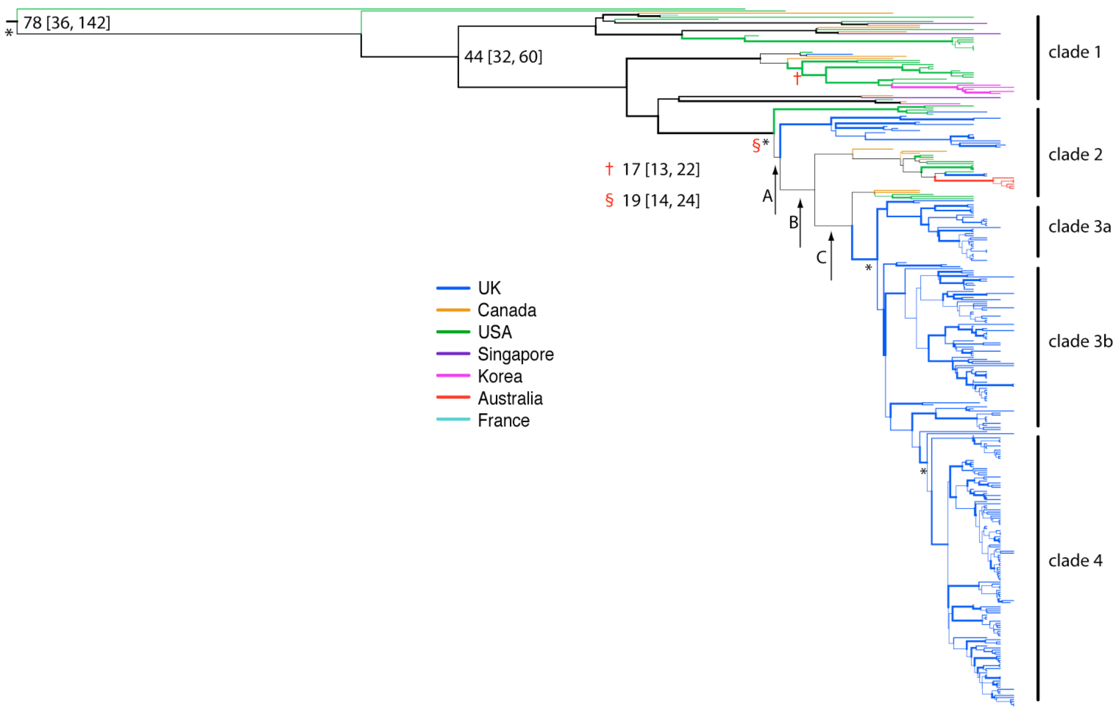


Figure 3.9: A summarized tree based on sampled trees from a BEAST run with a relaxed log exponential clock model. Labels on the right correspond to clade names in Figure 3.3. Branches are coloured according to origins of isolates. Line widths of internal branches indicate how well-supported the branches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

3.3.3 Discriminatory SNPs and potential functional consequences

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

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

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

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

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

3.3.4 Signatures of recombination

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

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

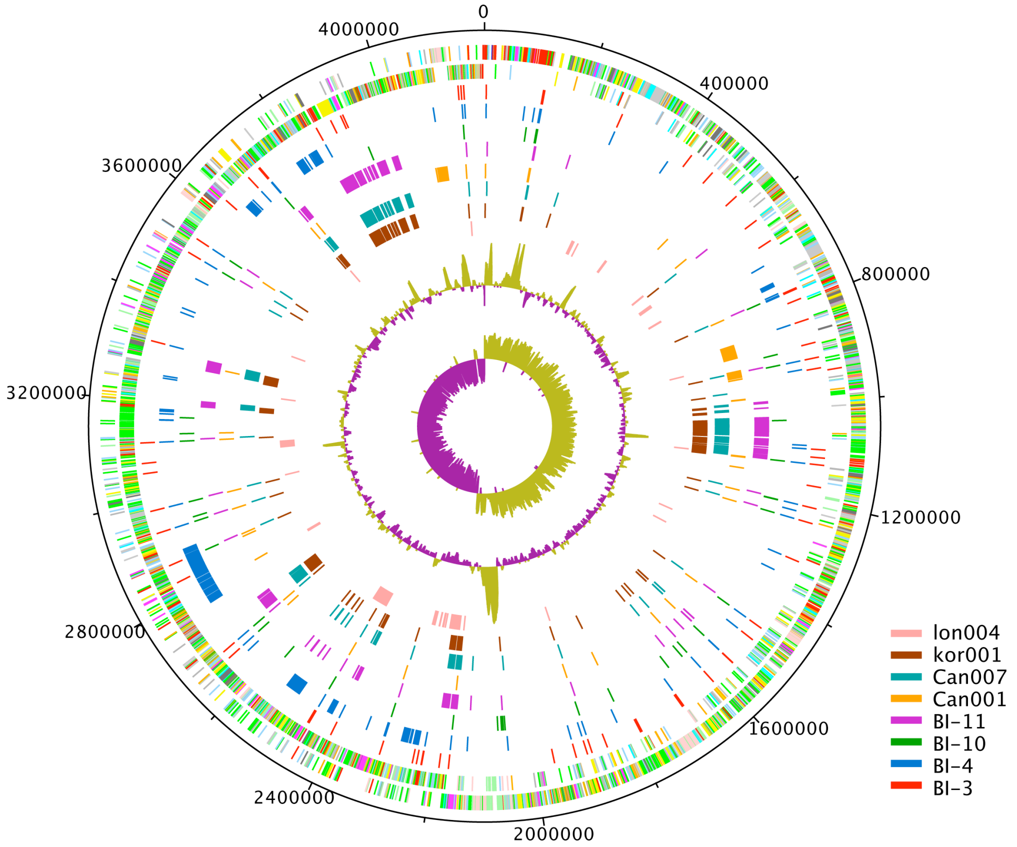


Figure 3.10: SNPs between R20291 and eight *C. difficile* isolates showing homologous recombination blocks. Outer circle: CDSs of *C. difficile* R20291 genome, shown on a pair of concentric rings representing both coding strands; two inner circles: G+C% content plot and GC deviation plot (>0% olive,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

3.3.5 Homoplasic SNPs and convergent evolution

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

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

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

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

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

Table 3.2. Homoplasic SNPs and associated gene products.

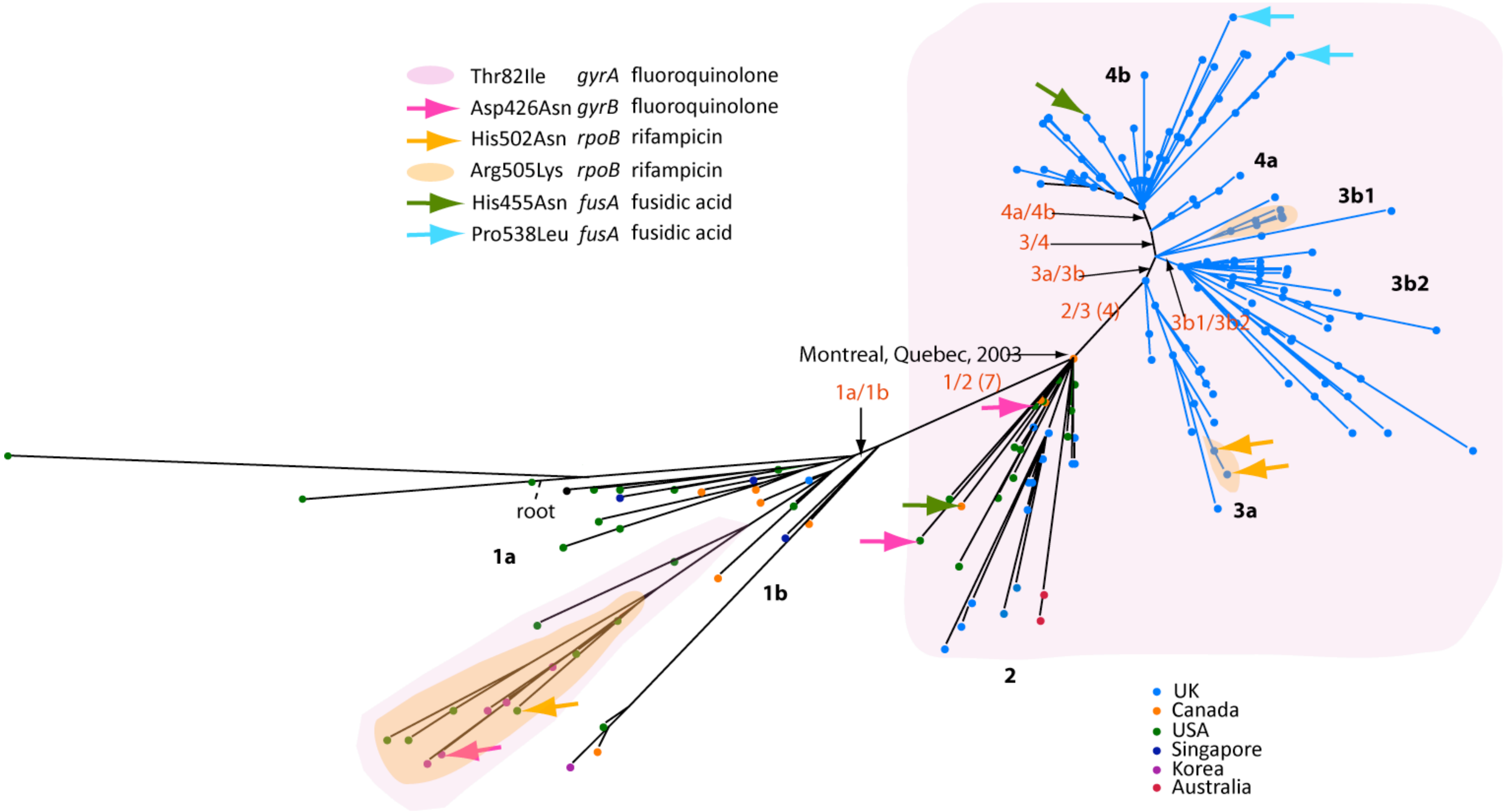

Thr82Ile gyrA fluoroquinolone
Asp426Asn gyrB fluoroquinolone
His502Asn rpoB rifampicin
Arg505Lys rpoB rifampicin
His455Asn fusA fusidic acid
Pro538Leu fusA fusidic acid
4b
4a
3b1
3b2
4a/4b
3/4
3a/3b
2/3 (4)
3b1/3b2
Montreal, Quebec, 2003
1/2 (7)
1a/1b
root
1a
1b
3a
2
UK
Canada
USA
Singapore
Korea
Australia

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

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

3.3.6 Estimating mutation rate

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

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

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

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

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

3.3.7 Population history inference

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

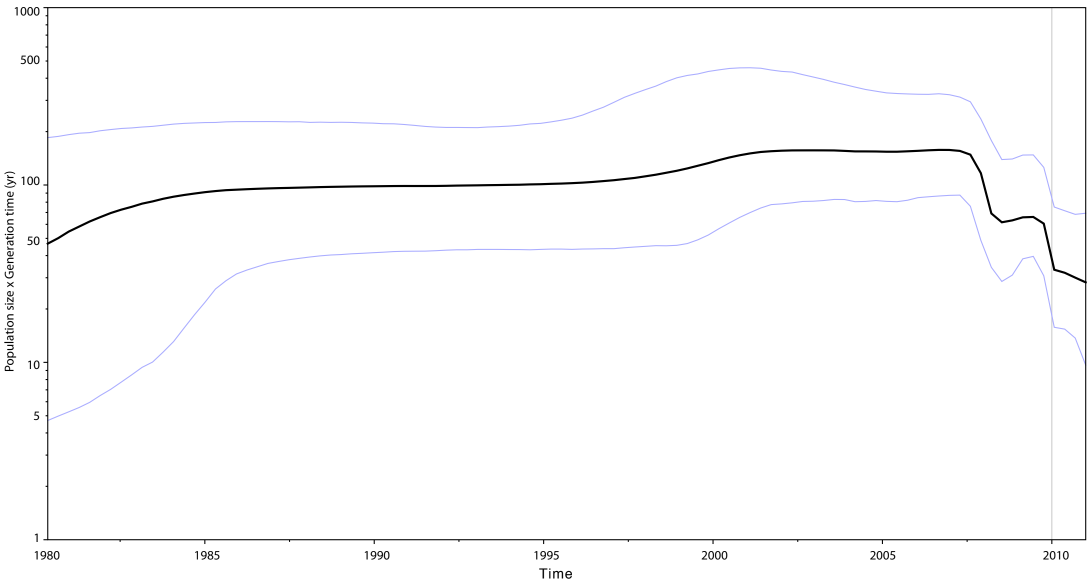


Figure 3.12: Bayesian skyline plot indicating changes in population size of this BI/NAP1/027 sample collection. Time (in years) is given on the x-axis; y-axis shows the product of population size and generation time (in years) in log scale. The black line represents median estimate; purple lines denote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

3.3.8 Horizontal gene transfer in BI/NAP1/027 lineage

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

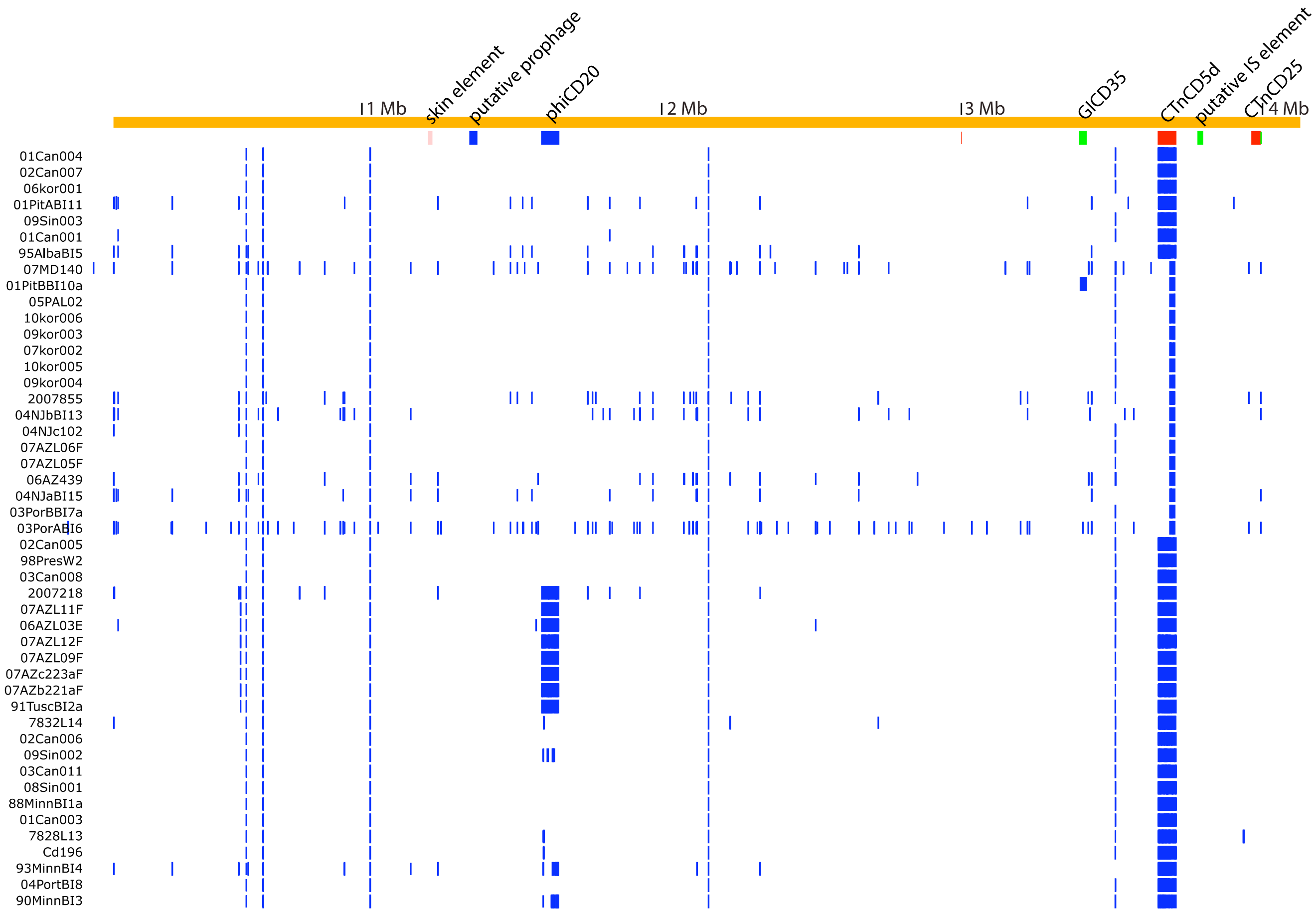


Figure 3.13: Genomic regions that are present in the reference (M7404) but absent (shown by blue boxes) from isolates in clade 1. The genome of M7404 (>4 Mb in size) is represented by an orange bar. The genomic islands carried by M7404 are depicted by coloured boxes beneath it. Names of isolates in clade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

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

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

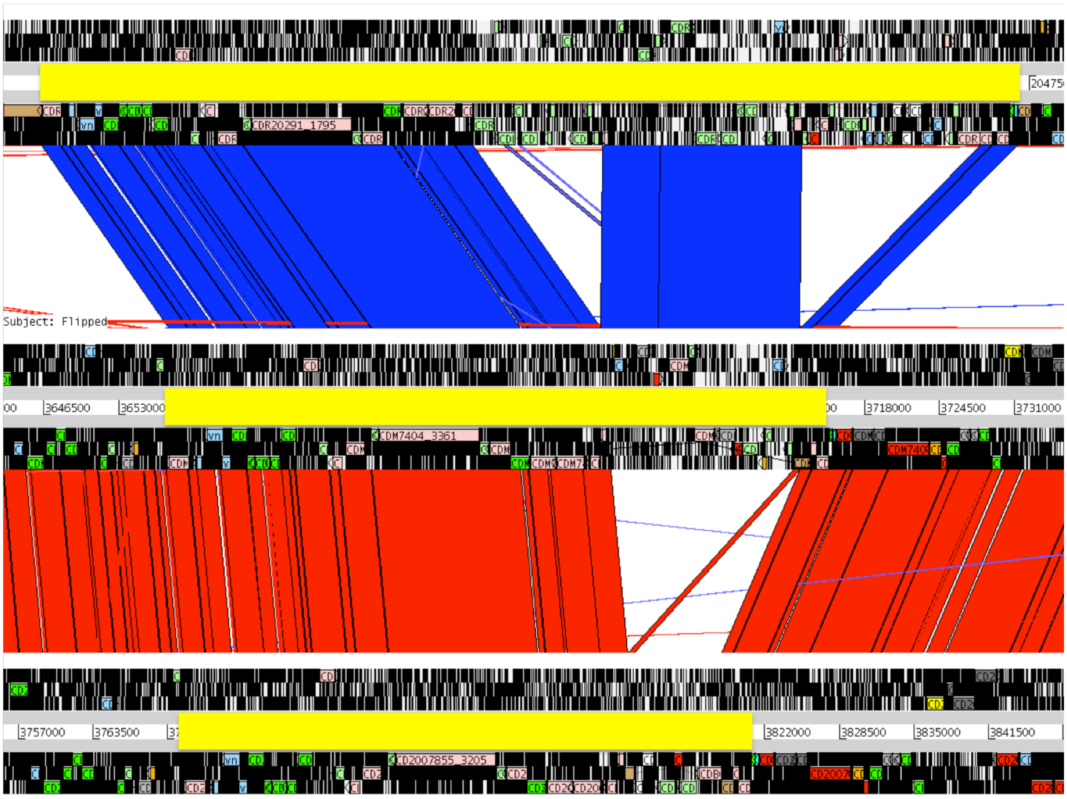


Figure 3.14: Comparison of three versions of CTn*CD5* (indicated by yellow boxes) in strains R20291 (top), M7404 (middle), and 2007855 (bottom).

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

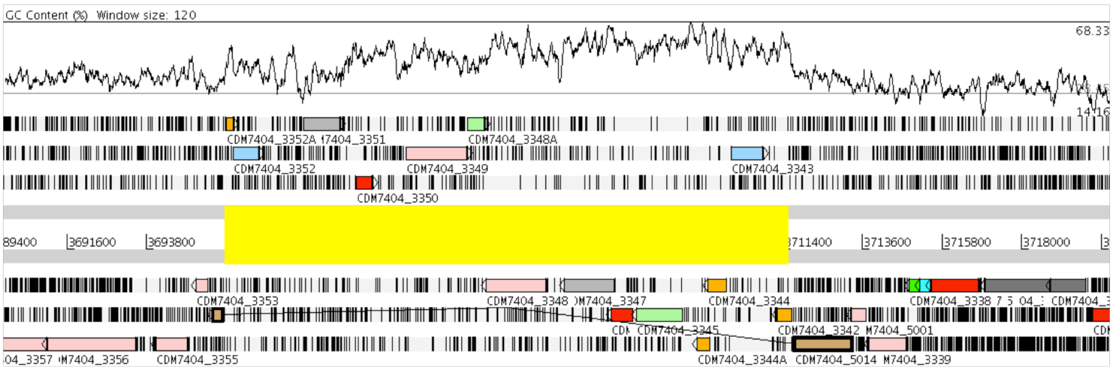


Figure 3.15: Genes gained along the branch connecting clades 1 and 2, shown in the genome of M7404. The graph on top depicts GC content. The extent of the insertion Tn*6105* is indicated by a yellow box.

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

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

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

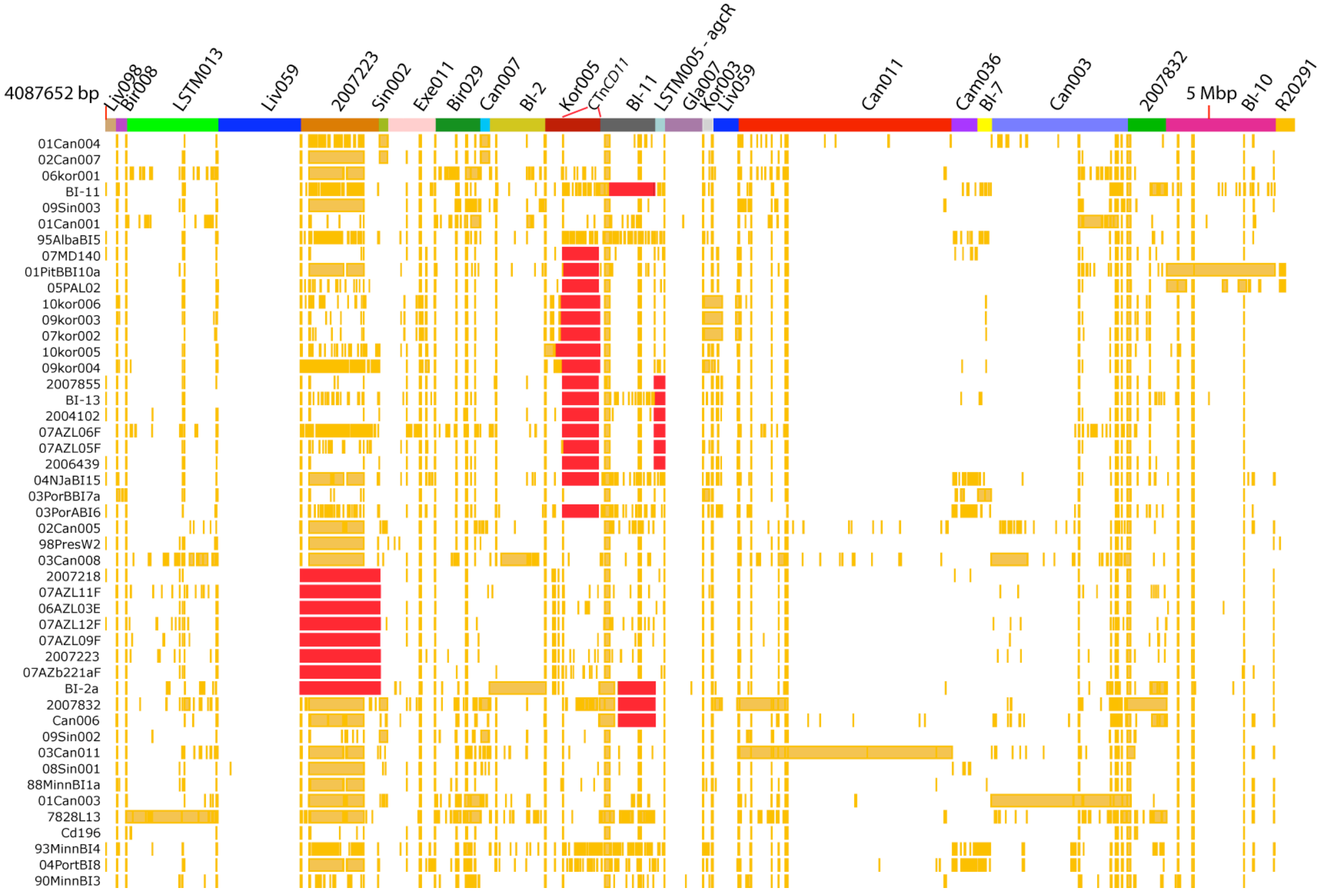


Figure 3.16: Accessory genome components of isolates in clade 1. The coloured boxes on top depict groups of contigs or genomic island sequences from different isolates. Sequence from the same isolate is shown in the same colour, with the isolate name labelled above. Yellow and red boxes depict same genome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

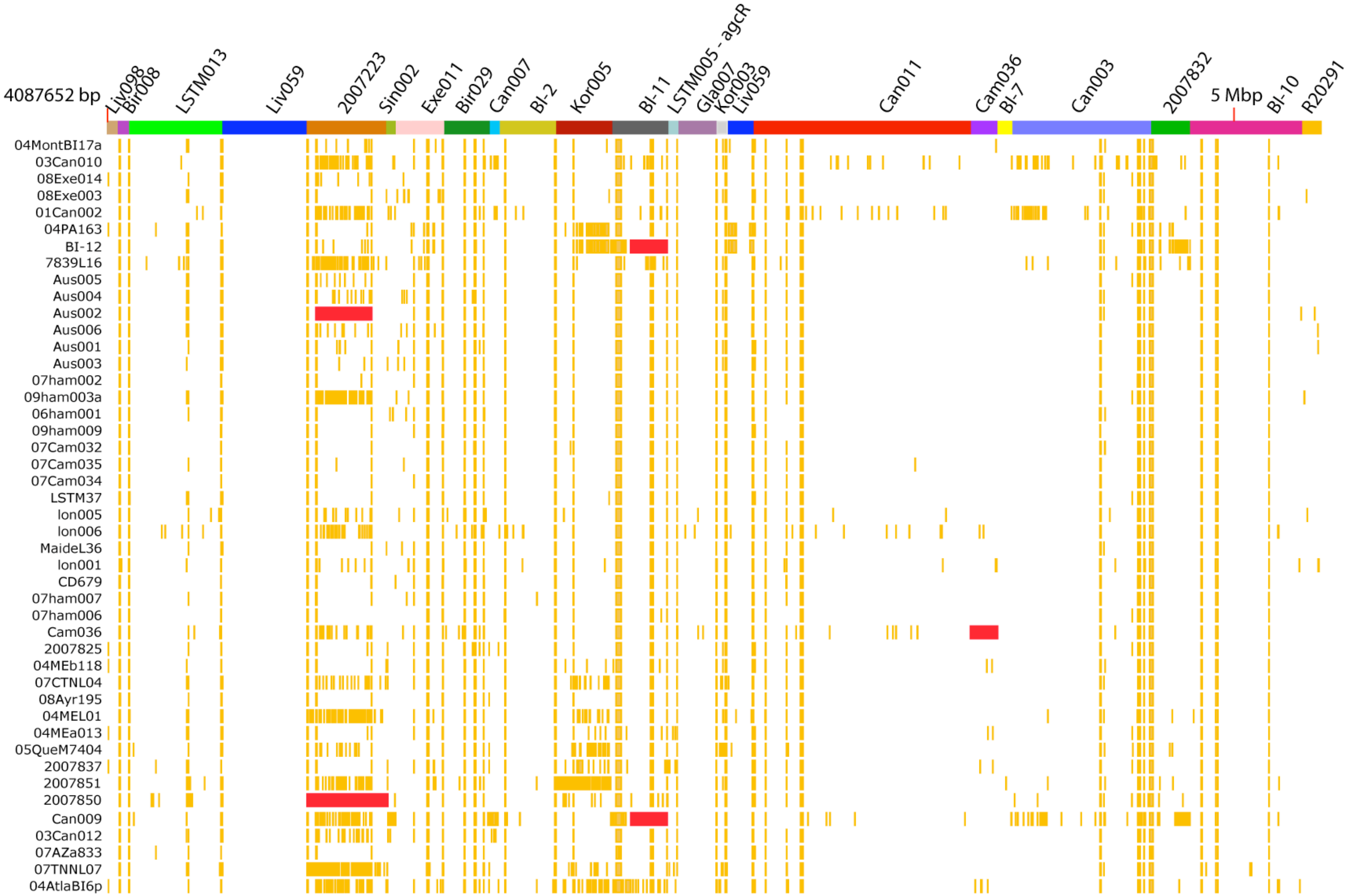


Figure 3.17: Accessory genome components of isolates in clade 2. Other descriptions for this figure are the same as Figure 3.16.

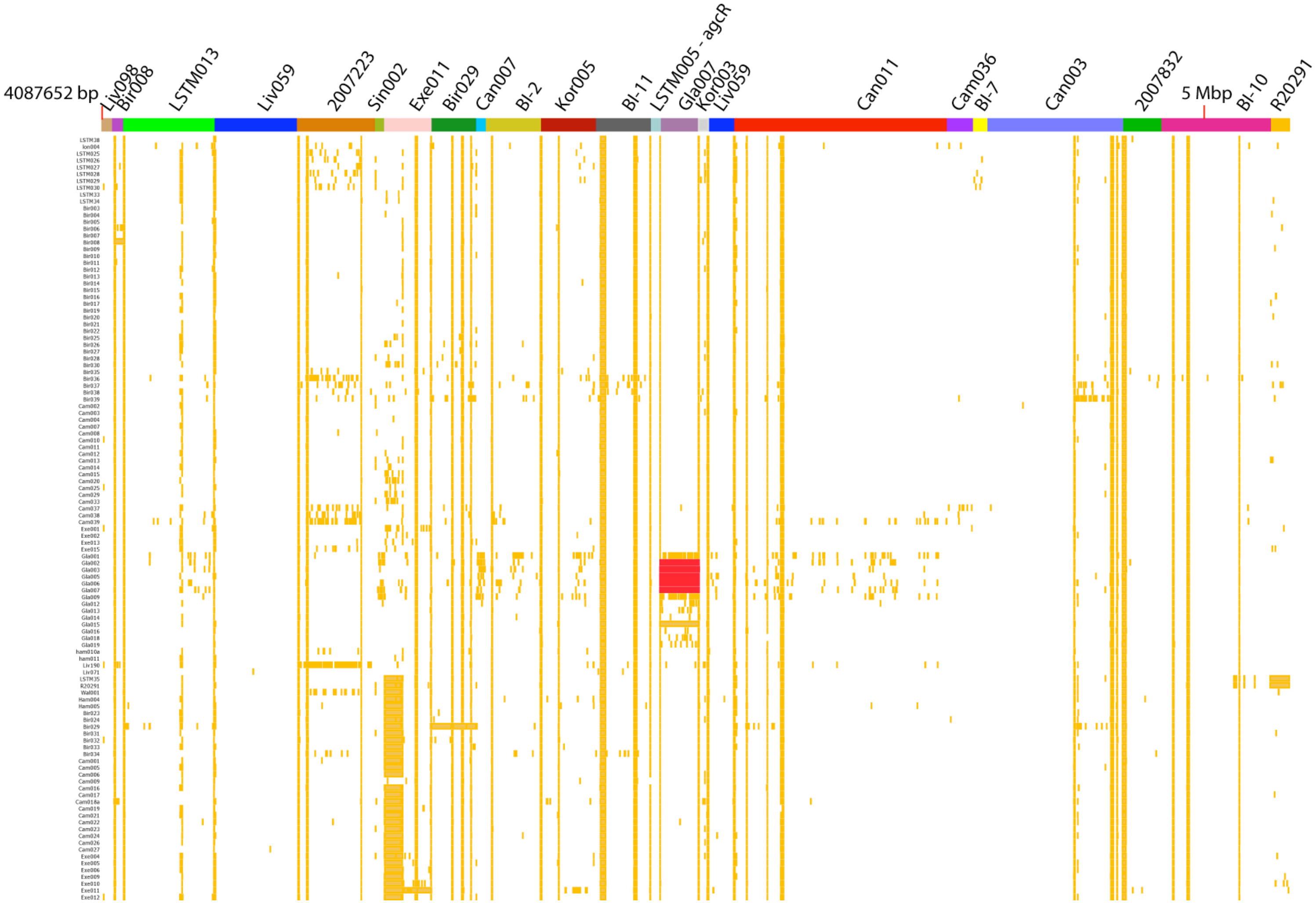


Figure 3.18: Accessory genome components of isolates in clade 3. Other descriptions for this figure are the same as Figure 3.16.

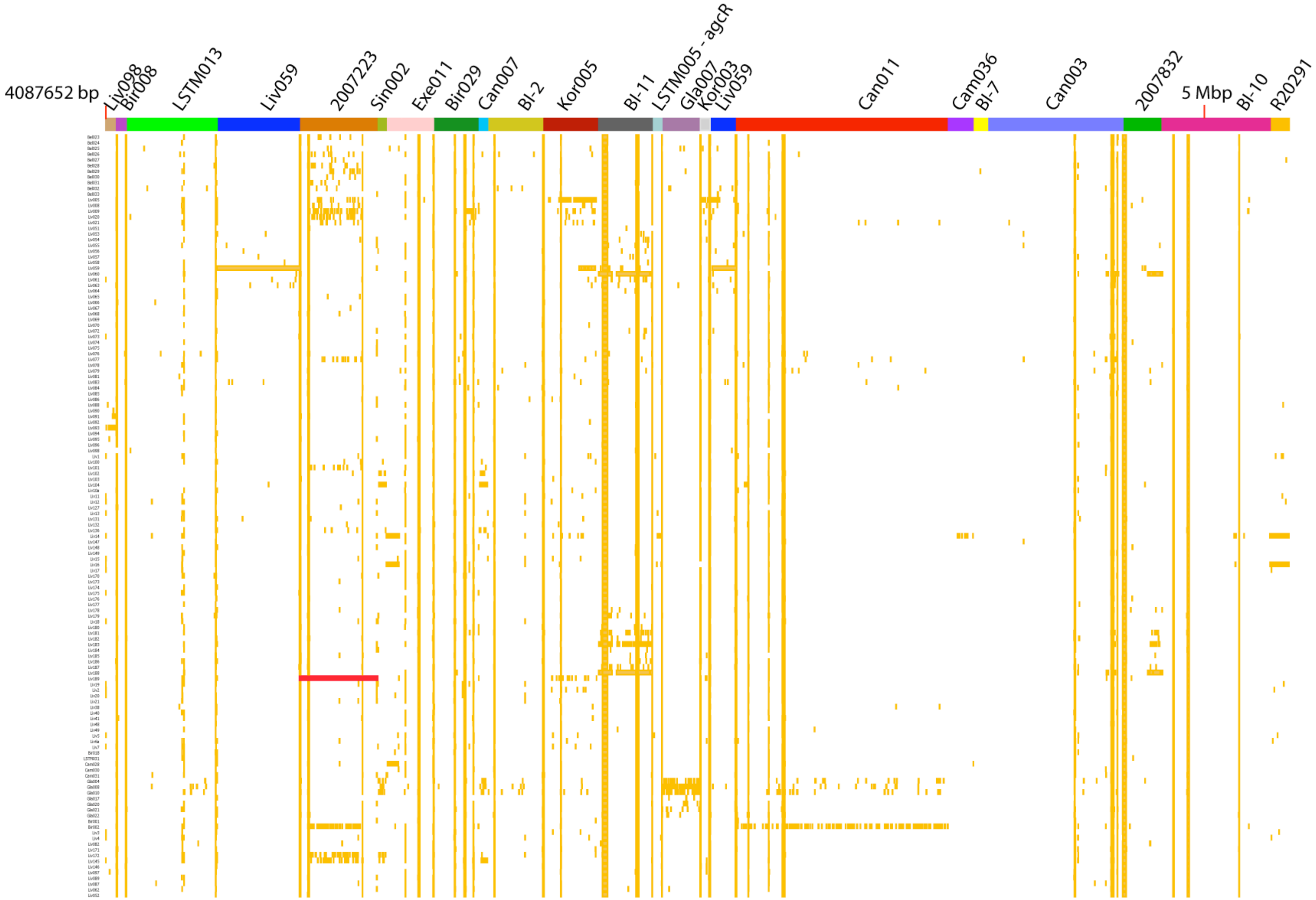


Figure 3.19: Accessory genome components of isolates in clade 4. Other descriptions for this figure are the same as Figure 3.16.

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

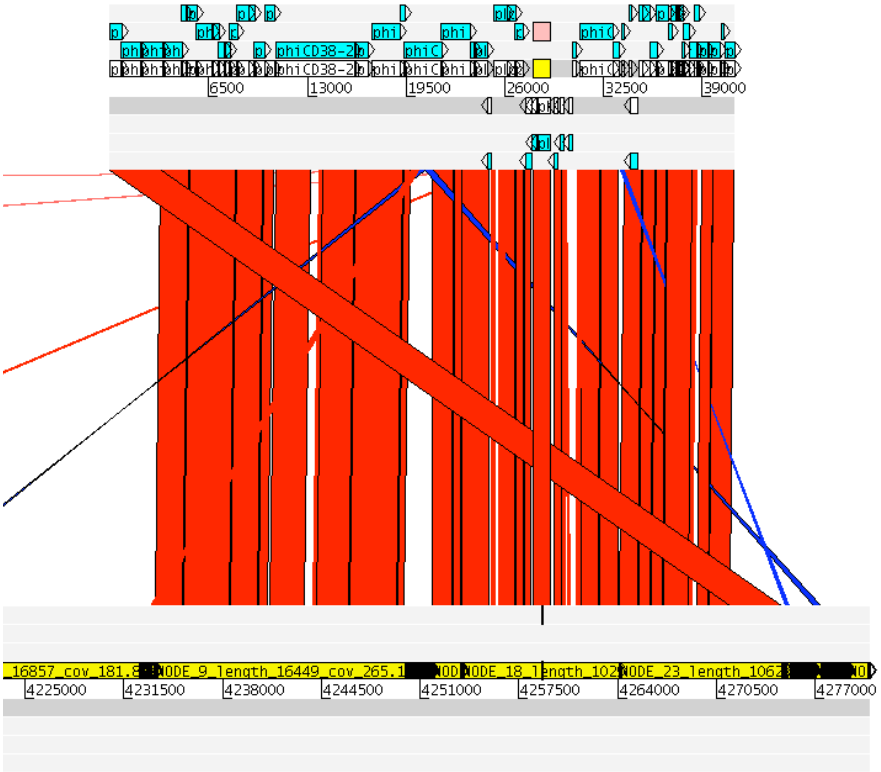


Figure 3.20: Comparison between phi*CD38-2* and unique contigs from isolate BI-2a.

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

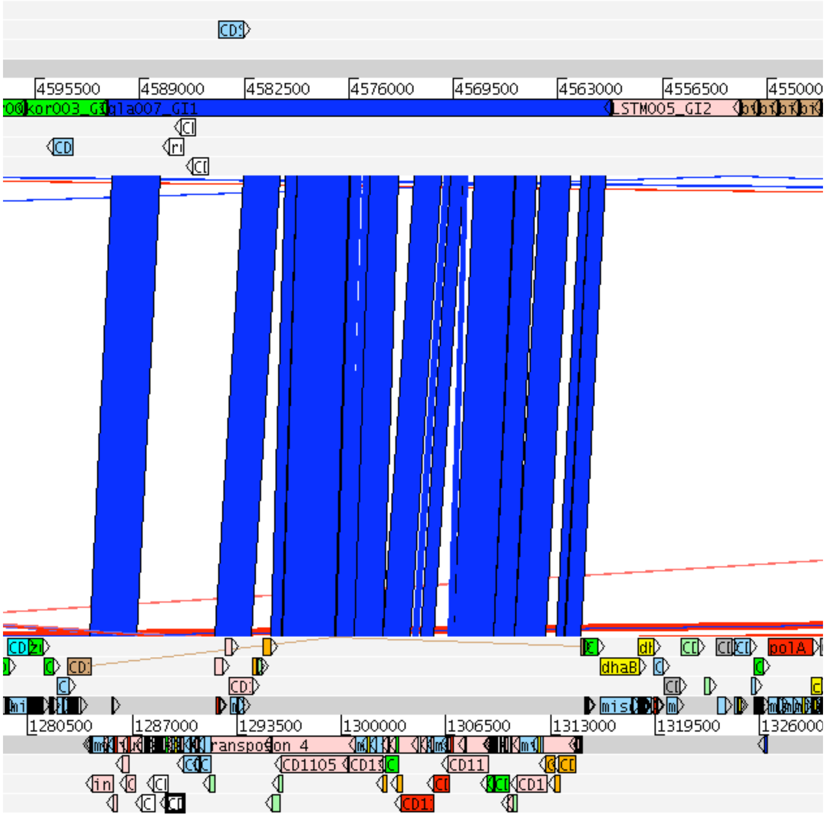


Figure 3.21: Comparison between a contig in Gla007 and CTn*4* in 630.

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

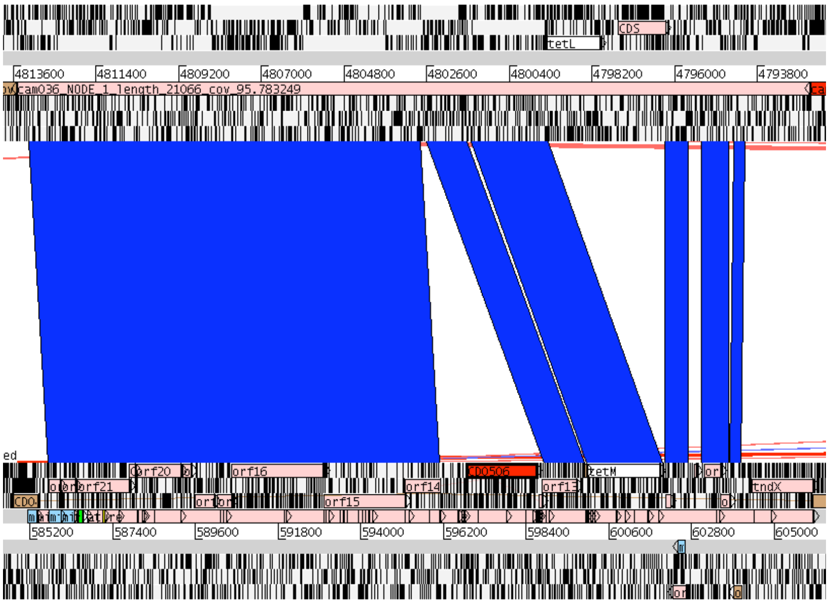


Figure 3.22: Comparison between a contig unique to Cam036 and CTn*3* in 630.

3.3.9 PaLoc region variation

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

3.4 Discussion

3.4.1 Two lineages associated with fluoroquinolone resistance

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

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

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

3.4.2 Other insights drawn from the phylogeny

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

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

3.4.3 Agreement with earlier analyses based on a smaller sample set

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

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

3.4.4 Antibiotic resistance

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

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

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

3.4.6 Potential mechanisms for large chromosomal region replacement

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

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