Chapter 5

Final Discussion

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

5.1 Significance of homologous recombination

In addition to supporting the previous suggestion that *C. difficile* has a highly dynamic genome (Sebaihia *et al.*, 2006; Stabler *et al.*, 2006), data outlined in Chapter 2 has also highlighted horizontal gene transfer and homologous recombination as two important mechanisms underlying this diversity. This is

the first time homologous recombination involving large chromosomal region exchange was demonstrated in *C. difficile*. The impact of homologous recombination was previously considered to be low (Lemee *et al.*, 2004), but the findings in Chapter 2 and Chapter 3 suggest it is not a negligible influence, as 15% of the CF5 genome sequence can be attributed to imports, and homologous recombination blocks of >100 kb were found in multiple isolates belonging to ribotype 027. Identifying variants resulting from homologous recombination is an important consideration for future phylogenetic analysis of *C. difficile*, as these phenomena are sufficient to distort branch lengths in the phylogenetic tree.

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

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

these genomic islands, perhaps by exploiting naturally existing BI/NAP1/027 isolates or genetic mutants obtained under laboratory conditions.

5.3 Selective pressure and gene candidates for functional study

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

5.4 Insights from local hospital transmission study

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

5.5 SNPs as genetic markers for genotyping

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

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

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