Quantitative study of *Clostridium difficile* transmission using extensive epidemiological data and whole genome sequencing

David Eyre

Green Templeton College

Nuffield Department of Medicine

A thesis submitted for the degree of Doctor of Philosophy

Hilary 2013

Quantitative study of *Clostridium difficile* transmission using extensive epidemiological data and whole genome sequencing

David Eyre, Green Templeton College, a thesis submitted for the degree of Doctor of Philosophy, Hilary, 2013.

Clostridium difficile is a leading healthcare-associated infection, which causes diarrhoea, and is almost exclusively precipitated by antibiotic exposure. Traditionally *C. difficile* infection (CDI) has been considered predominantly transmitted within hospitals. However, endemic spread hampers identification of the source of infections, and therefore control and prevention of disease.

A cohort of consecutive hospital and community CDI cases in Oxfordshire from September 2007 to March 2011 was investigated. For each case hospital admission, ward movement and demographic data were available allowing contact events between cases to be reconstructed. Initially 944 cases to March 2010 underwent multilocus sequence typing (MLST), subdividing the endemic cases into 69 distinct lineages and demonstrating unexpectedly that ward-based contact with known symptomatic CDI cases only accounts for <25% of disease.

To better determine the extent of transmission arising from symptomatic patients, irrespective of the route transmission, isolates from 1223 cases to March 2011 underwent whole genome sequencing. Serially sampled patients with recurrent or on-going disease were used to estimate rates of *C. difficile* evolution and within-host diversity and to show 0-2 single nucleotide variants (SNVs) are expected between transmitted isolates obtained <124 days apart (95% prediction interval). Mixed infection with more than one strain was investigated, but probably plays only a minor role in onward transmission.

In the Oxfordshire CDI cohort, 333/957 (35%) CDI from April 2008 – March 2011 were within 2 SNVs of ≥1 previous case since September 2007 (consistent with transmission). 428/957 (45%) were >10SNVs from all previous cases: these distinct subtypes continued to be identified consistently throughout the study, suggesting cases arise from a considerable reservoir of *C. difficile*. Surprisingly, declines in the incidence of genetically-related CDI were similar to those in genetically distinct CDI suggesting interventions not just targeting symptomatic individuals, e.g. antimicrobial stewardship, have played a significant role in recent CDI declines.

Finally, the feasibility of studying asymptomatic inpatients as potential source of the unexplained transmission was investigated.

This thesis provides convincing evidence, in a setting with typical CDI incidence and infection control practice, that only the minority of CDI arises from other symptomatic cases. It demonstrates that much CDI arises from genetically diverse reservoirs, with each exposure resulting in relatively few secondary cases. Future control strategies therefore need to focus on identifying these reservoirs, one of which is plausibly asymptomatic inpatients, and also on interventions that prevent the transition from exposure and colonisation to disease, such as antimicrobial stewardship.

TABLE OF CONTENTS

1 Preface 7

1.1 Acknowledgements 7

1.2 Funding 9

1.3 Ethics Statements 10

1.4 Declaration and attributions 11

1.4.1 Chapter 3 attributions 11

1.4.2 Chapters 4 and 5 attributions 11

1.4.3 Chapter 6 attributions 13

1.5 Papers arising 14

1.5.1 Directly from this thesis 14

1.5.2 Other papers during the course of the thesis 15

1.6 Abbreviations 18

2 Literature review 21

2.1 Introduction 21

2.2 Microbiology 21

2.3 Pathophysiology 25

2.4 The scale of the problem 28

2.5 Risk factors 30

2.6 Transmission epidemiology 35

2.6.1 Evidence for the hospital as a vector 36

2.6.2 Infectious periods 40

2.6.3 Incubation periods 41

2.6.4 Transmission in long-term care facilities 42

2.6.5 Environmental sources of C. difficile 42

2.7 Prevention 44
2.8 Laboratory investigation 47
2.8.1 Diagnostic testing *47*
2.8.2 Molecular typing *49*
2.9 Management and treatment 52
2.10 Whole genome sequencing 55
2.10.1 Sequencing technology *55*
2.10.2 Data processing *60*
2.10.3 Clostridium difficile sequencing *62*
2.11 Thesis outline 63
3 Characterisation of ward-based *C. difficile* transmission with epidemiogloical and genotyping data 65
3.1 Introduction 65
3.2 Methods 66
3.2.1 Setting *66*
3.2.2 Samples and multilocus sequence typing *66*
3.2.3 Analysis *70*
3.2.4 Sensitivity analysis *75*
3.2.5 Controls *76*
3.2.6 Statistics and computation *78*
3.3 Results 78
3.3.1 Samples and STs *78*
3.3.2 Infectious/incubation periods *82*
3.3.3 Proportion of cases with a credible donor *87*
3.3.4 Ward contamination beyond discharge *89*
3.3.5 Sensitivity analyses *90*
3.3.6 Missing data simulation *93*

3.4 Discussion 94

4 Prerequisites for applying whole genome data 102

4.1 Introduction 102

4.2 Whole genome sequencing methods 102

4.2.1 Sample preparation and sequencing 102

4.2.2 Sequence data processing 103

4.3 Reproducibility 105

4.4 Within host diversity and evolution 106

4.4.1 Serial patient samples 107

4.4.2 Within sample diversity 118

4.4.3 Gut model 120

4.5 Population diversity 128

4.6 Mixed infection 131

4.6.1 Detecting mixed infection with genotyping 132

4.6.2 Estimating mixed infection with whole genome data 136

5 Characterising Oxfordshire *C. difficile* transmission using whole genome sequencing and epidemiological data 171

5.1 Introduction 171

5.2 Methods 173

5.2.1 Setting 173

5.2.2 Sequencing 174

5.2.3 Genetic analyses 174

5.2.4 Epidemiological analysis 180

5.3 Results 181

5.3.1 Genetic diversity within Oxfordshire cases 184

5.3.2 Epidemiological relationships between genetically-related cases 188

5.3.3 Epidemiologically unexplained genetically-related cases *191*

5.3.4 Comparison of discriminatory power of whole genome sequencing versus conventional genotyping results *193*

5.3.5 Population diversity *195*

5.3.6 Temporal trends in transmission *197*

5.4 Discussion 201

6 Asymptomatic Carriage Study 206

6.1 Introduction 206

6.2 Methods 207

6.2.1 Participants *208*

6.2.2 Samples *209*

6.2.3 Epidemiological data *210*

6.2.4 Analysis *211*

6.3 Results 212

6.3.1 Participants and samples *212*

6.3.2 Relationship between carriage and previous / subsequent disease *213*

6.3.3 Risk factors for carriage *215*

6.3.4 Genetic relationships between carriage and symptomatic isolates *219*

6.4 Discussion 223

7 Conclusions and future work 228

7.1 Understanding transmission 228

7.2 Infection control implications 231

7.3 Sequencing technology developments 234

7.4 Conclusions 236

8 References 237

1 PREFACE

1.1 ACKNOWLEDGEMENTS

The last 3 years have been a really enjoyable time. In part this has been because I have been able to work on unravelling a challenging and exciting problem, but it is also the result of working with some wonderful colleagues and supervisors, who have been great friends and encouragers.

I should start by thanking all those whose vision and hard work set up the resources that made this thesis possible. Dai Griffiths, Ali Vaughan and others have been collecting and saving *C. difficile* positive stool samples and isolates since I was a medical student! This provided a unique collection of all samples from a geographic region over more than 6 years. The epidemiological data used depended on linking numerous hospital databases, for which I am very grateful to David Wyllie, John Finney and rest of the Infections in Oxfordshire Research Database team, as well as to Sarah Walker for ensuring the excellent quality of the final data I used.

Particular thanks must go to my three supervisors, Tim Peto, Sarah Walker, and Rosalind Harding. Tim's overarching enthusiasm for understanding transmission has been memorable; it has extended to many late night emails, still thinking about *C. difficile* in the swimming pool on holiday, and numerous excited discussions about the latest graph! I am also really grateful to Sarah for all the time she has given me, and for endless comments and encouragement on emails, abstracts, posters, talks, manuscript drafts, and this thesis. I have really enjoyed our weekly meetings, and I am sure have been much more productive as a result, as well as

having someone else who is concerned about the welfare of our cat! I would like to thank Rosalind for getting me started thinking about phylogenetics, trees and sequence data. Ros' concerns about 'Ns' and missing sequence data got me thinking deeply about best to analyse and compare sequences. I enjoyed our recent discussion about taking on being 'Keeper of the Groves', which has inspired me to do something with my own garden now all the writing is finished!

Derrick Crook has been a great unofficial fourth supervisor and supporter. I would like to thank Daniel Wilson for several useful discussions on some of the more mathematical parts of this thesis. Mark Wilcox has provided some really helpful advice, as well as promoting our work. Dai and Ali in the lab have been amazing; from being super patient reminding me what a pipette is, to helping extract DNA for seemly infinite numbers of isolates. I have had a lot of fun sharing an office and learning together with some amazing friends, Nick, John, Tim, Nicole, Claire, and Bernadette.

And then lastly, I am very grateful to Katie, for letting me go to work and have fun for 3 years, for tirelessly going to work herself, and doing on calls so I didn't have to, and to our little baby waiting just long enough to be born so I could write this thesis.

1.2 FUNDING

This thesis was supported by a National Institute for Health Research Doctoral Research Fellowship (NIHR-DRF-2010-03-40).

Additional funding support was provided by the National Institute for Health Research Oxford Biomedical Research Centre; the UKCRC Modernising Medical Microbiology Consortium, funded under the UKCRC Translational Infection Research Initiative supported by Medical Research Council, Biotechnology and Biological Sciences Research Council and the National Institute for Health Research on behalf of the Department of Health (Grant G0800778) and the Wellcome Trust (Grant 087646/Z/08/Z); and Wellcome Trust core funding (Grant 090532/Z/09/Z).

1.3 ETHICS STATEMENTS

Chapter 3: This was a pre-specified analysis within the Infections in Oxfordshire Research Database, an anonymized linked electronic database approved by the Oxford Research Ethics Committee (09/H0606/85) and the National Information Governance Board (5-07(a)/2009).

Chapter 4 (serially sampled patients, mixed infection study) & Chapter 5: These studies were approved by the Berkshire Research Ethics Committee (10/H0505/83) and the National Information Governance Board (8-05(e)/2010) without requiring individual patient consent, as part of collection of studies undertaken by the Modernising Medical Microbiology Consortium.

Chapter 6: This study, the "*Clostridium difficile* asymptomatic carriage and acquisition study," was approved by the Oxford C Research Ethics Committee (11/SC/0197).

1.4 Declaration and attributions

I, David Eyre, designed and conducted all the analyses presented in this thesis.

A detailed description of work undertaken and assistance obtained with processing of laboratory samples and specific methods is set out below.

1.4.1 Chapter 3 attributions

I designed and conducted the combined analysis of the genotyping and epidemiological data presented under the supervision of Dr Sarah Walker and Prof Tim Peto.

Laboratory multilocus sequencing typing of *C. difficile* clinical samples was predominantly undertaken prior to the start of this thesis by Dai Griffiths, Ali Vaughan, with assistance from Dr Kate Dingle, Dr Melina Kachrimanidou, Dr Nicole Stoessor, and Dr Bernadette Young. Remaining samples were sequenced typed by myself, Dai Griffiths, and Ali Vaughan. Linked anonymized epidemiological data was provided from the Infections in Oxford Research Database, developed by John Finney and Dr David Wyllie.

1.4.2 Chapters 4 and 5 attributions

I designed and conducted the analysis of the *C. difficile* whole genome data presented, including the assessment of sequencing reproducibility, within host variation and Oxfordshire transmission, under the supervision of Dr Sarah Walker, Dr Rosalind Harding and Prof Tim Peto.

DNA was extracted from cultured frozen Oxfordshire isolates by myself, Dai Griffiths and Ali Vaughan. DNA sequencing was undertaken by the Wellcome Trust

Centre for Human Genetics. Initial pre-processing of raw sequence data was undertaken using a data processing pipeline developed by the Department of Statistics, University of Oxford, in particular by Dr Camilla Ip, Dr Tanya Golubchik, Dr Elizabeth Batty, Dr Madeleine Cule and Dr Daniel Wilson. During this thesis I contributed to development of this pipeline, and the final version used includes refinements I made.

I designed and conducted the within-sample diversity experiments including selection of samples, culture, DNA preparation and analysis.

Supervision on models of within host evolution was provided by Dr Daniel Wilson and Dr Sarah Walker.

The *C. difficile* gut model was conceived by Prof Mark Wilcox and colleagues at the University of Leeds. DNA from three gut model experiments was provided by Prof Mark Wilcox, the timing and nature of the specific samples sequenced were determined by myself, Prof Mark Wilcox, and Dr Sarah Walker.

I developed the mixed infection estimator in section 4.6 with supervision provided by Dr Daniel Wilson and Dr Sarah Walker, and conducted the simulations and analyses presented.

The hidden Markov model described in chapter 5 was developed by Dr Daniel Wilson, based on concepts from ClonalFrame developed by Dr Xavier Didelot, and is used with their permission.

1.4.3 CHAPTER 6 ATTRIBUTIONS

I conceived and designed the study protocol, and obtained ethical approval for the study. I recruited all study participants, and collected questionnaire data. Stools samples were sent from study wards by ward nursing staff. I cultured the stool samples and extracted DNA. I analysed the data, supervised by Dr Sarah Walker and Prof Tim Peto.

1.5 PAPERS ARISING

1.5.1 DIRECTLY FROM THIS THESIS

Walker AS*, **Eyre DW***, **Wyllie DH**, **Dingle KE**, **Harding RM**, **O'Connor L**, **Griffiths D**, **Vaughan A**, **Finney J**, **Wilcox MH**, **Crook DW**, **Peto TEA**. 2012. Characterisation of Clostridium difficile hospital ward-based transmission using extensive epidemiological data and molecular typing. Plos Medicine **9**:e1001172:1–12. *Equal contributors.

Eyre DW, **Walker AS**, **Griffiths D**, **Wilcox MH**, **Wyllie DH**, **Dingle KE**, **Crook DW**, **Peto TEA**. 2012. Clostridium difficile Mixed Infection and Reinfection. J Clin Microbiol **50**:142–144.

Didelot X, **Eyre DW**, **Cule ML**, **Ip CL**, **Ansari MA**, **Griffiths D**, **Vaughan A**, **O'Conner L**, **Golubchik T**, **Batty E**, **Piazza P**, **Wilson DJ**, **Bowden R**, **Donnelly PJ**, **Dingle KE**, **Wilcox MH**, **Walker AS**, **Crook DW**, **Peto TE**, **Harding RM**. 2012. Microevolutionary analysis of Clostridium difficile genomes to investigate transmission. Genome Biol **13**:R118.

Eyre DW, Walker AS, Freeman J, Baines SD, Fawley WN, Chilton CH, Griffiths D, Vaughan A, Crook DW, Peto TEA, Wilcox MH. 2013. Short-term genome stability of serial *Clostridium difficile* ribotype 027 isolates in an experimental gut model and recurrent human disease. PLOS One **8**:e63540.

Eyre DW, Cule ML, Griffiths D, Crook DW, Peto TEA, Walker AS, Wilson DJ. 2013. Detection of mixed infection from bacterial whole genome sequence data

allows assessment of its role in *Clostridium difficile* transmission. PLOS Comput Biol **9**:e1003059.

Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A, O'Connor L, Ip CLC, Golubchik T, Batty EM, Finney JM, Wyllie DH, Didelot X, Piazza P, Bowden R, Dingle KE, Harding RH, Crook DW, Wilcox MH, Peto TEA, Walker AS. 2013. Whole genome sequencing reveals diverse sources for *C. difficile* infection. [In press].

1.5.2 OTHER PAPERS DURING THE COURSE OF THE THESIS

Eyre DW, **Golubchik T**, **Gordon NC**, **Bowden R**, **Piazza P**, **Batty EM**, **Ip CLC**, **Wilson DJ**, **Didelot X**, **O'Connor L**, **Lay R**, **Buck D**, **Kearns AM**, **Shaw A**, **Paul J**, **Wilcox MH**, **Donnelly PJ**, **Peto TEA**, **Walker AS**, **Crook DW**. 2012. A pilot study of rapid benchtop sequencing of Staphylococcus aureus and Clostridium difficile for outbreak detection and surveillance. BMJ Open **2**:e001124.

Eyre DW, **Walker AS**, **Wyllie D**, **Dingle KE**, **Griffiths D**, **Finney J**, **O'Connor L**, **Vaughan A**, **Crook DW**, **Wilcox MH**, **Peto TEA**, **for the Infections in Oxfordshire Research Database**. 2012. Predictors of First Recurrence of Clostridium difficile Infection: Implications for Initial Management. Clin Infect Dis **55**:S77–S87.

Dingle KE, **Didelot X**, **Ansari MA**, **Eyre DW**, **Vaughan A**, **Griffiths D**, **Ip CLC**, **Batty EM**, **Golubchik T**, **Bowden R**, **Jolley KA**, **Hood DW**, **Fawley WN**, **Walker AS**, **Peto TE**, **Wilcox MH**, **Crook DW**. 2013. Recombinational Switching of the Clostridium difficile S-Layer and a Novel Glycosylation Gene Cluster Revealed by Large-Scale Whole-Genome Sequencing. J Infect Dis **207**:675–686.

Walker TM, **Ip CL**, **Harrell RH**, **Evans JT**, **Kapatai G**, **Dedicoat MJ**, **Eyre DW**, **Wilson DJ**, **Hawkey PM**, **Crook DW**, **Parkhill J**, **Harris D**, **Walker AS**, **Bowden R**, **Monk P**, **Smith EG**, **Peto TE**. 2013. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis **13**:137–146.

Walker AS, **Eyre DW**, **Crook DW**, **Wilcox MH**, **Peto TEA**. 2013. Response to Walk et al "Clostridium difficile ribotype does not predict severe infection." Clin Infect Dis **56**:1845-1846.

Stoesser NE, **Martin J**, **Mawer D**, **Eyre DW**, **Walker AS**, **Peto TEA**, **Crook DW**, **Wilcox MH**. 2013. Risk Factors for Clostridium difficile Acquisition in Infants: Importance of Study Design. Clin Infect Dis **56**:1680-1681.

Walker AS, **Eyre DW**, **Wyllie DH**, **Dingle KE**, **Griffiths D**, **Shine B**, **Oakley S**, **O'Connor L**, **Finney J**, **Vaughan A**, **Crook DW**, **Wilcox MH**, **Peto TEA**, **on behalf of the Infections in Oxfordshire Research Database (IORD)**. 2013. Relationship Between Bacterial Strain Type, Host Biomarkers and Mortality in Clostridium difficile Infection. Clin Infect Dis **56**:1589-1600.

Walker AS, Eyre DW, Wyllie DH, Dingle KE, Griffiths D, Shine B, Oakley S, O'Connor L, Finney JM, Vaughan A, Crook DW, Wilcox MH, Peto TEA and Infections in Oxfordshire Research Database. 2013. Relationship Between Bacterial Strain Type, Host Biomarkers, and Mortality in Clostridium difficile Infection. Clin Infect Dis 2013 **56**:1589–1600.

Stoesser N, Batty EM, Eyre DW, Morgan M, Wyllie DH, Del Ojo Elias C, Johnson JR, Walker AS, Peto TEA, Crook DW. 2013. Predicting antimicrobial susceptibilities for Escherichia coli and Klebsiella pneumoniae isolates using whole genomic sequence data. J Antimicrob Chemother [Online ahead of print].

Walker AS, **Eyre DW**, **Crook DW**, **Peto TEA, Wilcox MH**. 2013. Reply to Walk et al. Clin Infect Dis [Online ahead of print].

Eyre DW, Fawley WN, Best EL, Griffiths D, Stoesser NE, Crook DW, Peto TEA, Walker AS, Wilcox MH. 2013. Comparison of multilocus variable number tandem repeat analysis and whole genome sequencing for investigation of *Clostridium difficile* transmission. J Clin Microbiol [In Press].

1.6 ABBREVIATIONS

AAD, antibiotic-associated diarrhoea

AIC, Akaike information criterion

CD196, Clostridium difficile strain CD196

CD630, Clostridium difficile strain 630

CDI, Clostridium difficile infection

CI, Confidence Interval

Contigs, contiguous sequences

DI, Simpson Diversity Index

Gb, gigabase

GDH, glutamate dehydrogenase

HPA, Health Protection Agency

IORD, Infections in Oxfordshire Research Database

kb, kilobase

MIC, minimum inhibitory concentration

MIE, mixed infection estimator, an algorithm developed as part of this thesis

MLVA, multiple-locus variable-number tandem-repeat analysis

NAAT, nucleic acid amplification test

OUH, Oxford University Hospitals NHS Trust

PaLoc, pathogenicity locus

PCR, polymerase chain reaction

PFGE, pulsed-field gel electrophoresis

PMC, pseudomembranous colitis

r/m, the relative impact of recombination to mutation, i.e. the ratio of the number of substitutions introduced by recombination to the number of substitutions introduced by mutation

REA, restriction enzyme analysis

RMSE, root mean square error

RR, relative risk

SD, standard deviation

SNV, single nucleotide variant

ST, Sequence type

UK, United Kingdom

USA, United States of America

2 LITERATURE REVIEW

2.1 INTRODUCTION

Clostridium difficile is a leading healthcare-associated infection which is usually precipitated by antibiotic exposure. Antibiotic-associated diarrhoea and enterocolitis have been recognised throughout the antibiotic era,[1] with *C. difficile* established as the causative agent in the late 1970s.[2] More recently, the threat *C. difficile* poses to health increased significantly with the emergence in the early 2000s of a hypervirulent strain (ribotype-027, sequence type 1 (ST1)) associated with hospital outbreaks and high patient morbidity and mortality.[3] *C. difficile* has therefore become a major priority for patients, healthcare professionals and policy makers. Multiple interventions have been introduced aiming to prevent cases of *C. difficile* infection (CDI), including judicious use of provocative antibiotics, rapid isolation and treatment of presumptive cases, use of contact precautions and vigorous cleaning of the environment around cases.[4] These measures are believed to have reduced CDI incidence[5,6] but there is a lack of robust data identifying prevention of person-to-person transmission. Furthermore, despite these efforts CDI remains a considerable problem.

2.2 MICROBIOLOGY

C. difficile is an obligate anaerobic, Gram-positive, spore-forming bacillus[7] first isolated in 1935[8] as part of a systematic survey of the intestinal flora of new born children. This study was motivated in part by the recent availability of improved techniques for culturing anaerobic bacteria.[9] Louis Pasteur is credited with first discovering anaerobic bacteria in 1861 by culturing *Clostridium*

butyricum and also developed the terms aerobic and anaerobic for describing if microorganisms required oxygen for growth.[10] Classification of sporulating anaerobes based on a combination of morphology, in particular spore position, and biochemical properties, e.g. [11], allowed *C. difficile* to be identified as a distinct species. Successful culture required broth fermentation in specially sealed tubes to prevent gradual ingress of oxygen during its relatively slow growth.[8] This difficulty isolating the organism and its slow growth compared to other *Clostridium* species led to its name, *difficile*.[7,8] Although *C. difficile* was noted to produce an exotoxin highly pathogenic in guinea-pigs and rabbits when first discovered,[8] it was not until 1978 that *C. difficile* was established as a significant human pathogen as a result of investigations into the cause of pseudomembranous colitis (PMC).[2]

PMC is defined by inflammation of the colon involving formation of raised plaques which may join to form a confluent pseudomembrane. It was first described in 1893 by JM Finney, i.e. pre-antibiotic usage, following gastric surgery and post-operative alcohol enemas.[12] By the 1950s PMC was thought to be due to *Staphylococcus aureus*, but as 15-30% of healthy individuals are also colonised[12] and the original samples were obtained post-mortem[13] its role was questioned. Interestingly though, these investigations did lead to the discovery that vancomycin, used in this case to treat presumed *S. aureus*, provides successful treatment of PMC (in actual fact treating *C. difficile*).[14] An association between PMC onset and recent use of antibiotics was noted, in particular following clindamycin, with a 10% incidence of PMC in a prospective study of 200 patients treated with clindamycin for various reasons.[15] Other antibiotics were also

implicated including ampicillin, cephazolin, and cotrimoxazole.[13] It was unclear whether antibiotic-associated PMC was a result of antibiotic toxicity or hypersensitivity, or as other authors suggested due to changes in the intestinal flora.[16] Interest in a possible viral aetiology[17] led investigators to incubate faecal suspensions from patients with PMC in tissue cultures, and a toxic effect was noted with a characteristic cytopathic appearance. No virus particles were identified, but the toxin was noted to be present in high titres during acute illness, but not after convalescence or in patients with other causes of colitis. The toxin was characterised as heat labile and with a molecular weight of over 6000, however initial attempts to replicate its toxicity with supernatants from bacterial cultures were unsuccessful.[18]

Another line of evidence for an infectious cause for PMC came from a hamster model of disease.[19] Hamsters given clindamycin develop enterocolitis, and it was shown that transfer of caecal contents to another hamster (not given clindamycin) could transmit the disease. A *Clostridium* species was recovered from broth cultures of the caecal contents of diseased animals, which if administered also caused disease, and could be recovered from the 'secondary cases'. A cell-free supernatant of the culture also induced enterocolitis, but not when incubated with gas gangrene antitoxin.[20] An extension of this experiment provided evidence that a toxin producing *Clostridium* species is responsible for antibiotic-associated PMC in humans. Stools from patients with PMC injected intracecally into hamsters caused enterocolitis and cell cytotoxicity, both of which were neutralised by gas gangrene antitoxin. The toxicity was reproduced using cultures of *Clostridium* isolates from the stools. Morphological and biochemical characterisation of the

isolates confirmed *Clostridium difficile* as the causative organism.[2] Thus by using a combination of patients, healthy controls, and the hamster model Koch's postulates[21,22] were satisfied: *C. difficile* was found in those with PMC, but not in healthy controls; *C. difficile* was successfully isolated and grown in pure culture; and finally shown to induce disease in experimental animals.

The same study identifying *C. difficile* as the cause of PMC, also detected *C. difficile* in the stool of 1/54 patients with antibiotic-associated diarrhoea, without PMC.[2] Further work confirmed *C. difficile* causes a spectrum of disease, including 192/710 (27%) of antibiotic-induced diarrhoea episodes in patients without evidence of PMC.[23] However many cases of antibiotic-associated diarrhoea (AAD) do not have a known infectious cause.[24,25] AAD first became a clinical concern in the 1950s with the introduction of broad spectrum antibiotics (tetracycline and chloramphenicol) and rates range from 5% to 39% depending on the antibiotic used.[25] Disruption of the normal bowel flora is a likely cause of the ≥60-70% of AAD that is non-infective; a meta-analysis of concomitant use of probiotics showed some benefit. The overall relative risk of ADD with probiotics was 0.58 (95% confidence interval, CI, 0.50 – 0.68), giving a number needed to treat of 13 (95%CI 10 – 19). However significant heterogeneity exists between trials reflecting the different populations, antibiotics and probiotics studied.[26]

C. difficile is found in two important forms. During logarithmic growth the vegetative form of *C. difficile* dominates.[7] *C. difficile* also forms highly resistant spores, resistant to heat, radiation, chemical treatment and antibiotics, and that germinate in response to cholate derivatives, found in bile, and the amino acid glycine.[7,27] Successful purification of spores, has allowed electron-microscopy

studies of their structure and characterisation of their infectivity; as few as 7 spores per cm^2 is sufficient to infect 50% of mice.[28]

2.3 PATHOPHYSIOLOGY

C. difficile is acquired by faeco-oral transmission.[29] Spores are thought to be the principle ingested source of infection, given their resistance to gastric acid.[30] However, vegetative forms are also likely to be infectious as suppression of gastric acid to which the vegetative forms are sensitive[31] is associated with increased risk of CDI.[32,33] Stools from patients contain 10-fold more vegetative forms than spores. Although the vegetative forms die rapidly in dry conditions, on moist surfaces they remain viable for up to 6 hours. Survival of vegetative forms in gastric acid contents is marginal, but experimental elevation of pH above 5 resulted in survival comparable to a phosphate-buffered saline control. Vegetative forms were also able to survive inoculation into gastric contents with a pH above 5 obtained from patients on proton pump inhibitors.[31]

Germinating spores develop filamentous projections which facilitate attachment to the colonic mucosa.[34] *C. difficile* may then cause a range of disease from asymptomatic colonisation, self-limiting mild and watery diarrhoea, to pseudomembranous colitis, fulminant colitis and death.[35] Complications of CDI include multiple recurrences, intestinal perforation, toxic megacolon, sepsis, requirement for colectomy and death.[36]

Asymptomatic colonisation is relatively common demonstrating acquisition alone is insufficient to cause disease. Up to 3% of healthy adults may carry *C. difficile* asymptomatically.[37] Colonisation in children is more common, pooled data from

30 studies and 5887 children show ~35% of children in the first year of life are colonised with rates falling to ~15% by 1-8 years and then falling to adult levels of ~3% thereafter.[38] In hospitalised adults, asymptomatic carriage rates reported in a large Canadian study were 184/4143 (4.4%) at admission and then a further 123 (3.0%) developed asymptomatic carriage during their hospital stay. However the wards included were selected to be a mix of previously high and low CDI incidence wards, and therefore despite the considerable size of the study generalising the findings is difficult.[39] Five earlier studies, with a total of 1755 patients across a range of specialties, report rates of asymptomatic carriage rates at admission of 6–11%, and acquisition rates of 4-21%, with more than 63% of these patients remaining asymptomatic.[40]

Perturbation of normal gut flora, in the majority of cases by antibiotics, allows *C. difficile* to proliferate and cause toxin mediated disease.[7] The relative composition of the gut microbiome has been studied most recently by sequencing of 16S ribosomal RNA genes. In elderly subjects asymptomatically colonised with *C. difficile* the composition of the microbiome was similar to culture-negative patients at the phylum and family level taxonomic level. However in patients with CDI marked changes in the microbiome were seen, with a significant reduction in the diversity present and a change in the dominant organisms.[41] Similar techniques have been used to show profound changes occur in gut flora following antibiotic exposure. The specific changes seen vary by antibiotic, but suppression of the normal beneficial members of the anaerobic microbiome provides an opportunity for *C. difficile* to flourish.[42,43]

Toxins A and B are the two principal *C. difficile* virulence determinants and are part of a conserved 19.6kb pathogenicity locus (PaLoc).[44] The PaLoc contains five genes, *tcdA* and *tcdB* encoding the 2 toxin genes and 3 small open reading frames *tcdR* (a positive regulator of toxin production), *tcdE* (encoding a holin which enables release of toxin from the cell) and *tcdC* (hypothesised to be negative regulator, but now disputed, see below).[7] Toxins A and B enter host cells via receptor-mediated endocytosis, and following entry into the host cytosol both affect the same target, irreversibly glycosylating and inactivating Ras family small GTPases, including Rho-GTPases. This leads to disruption of control of the actin cytoskeleton, membrane blebbing and eventual apoptosis of the cell.[45]

The presence of either toxin A or B is sufficient to cause cytotoxic activity *in vitro* and disease in an animal model. Kuehne *et al* created isogenic mutants of *C. difficile* producing either toxin A or toxin B which both caused fulminant disease in a hamster model and cytotoxicity in 2 cell lines. Toxin A and B specific neutralising antibodies prevented cell cytotoxicity as expected. A double knock-out strain without toxin A or B caused neither cell toxicity nor any symptoms in colonised hamsters.[46] Other researchers have found conflicting results, with only toxin B and not toxin A required for disease in another isogenic mutant study [47] and earlier work suggesting the opposite.[45] The disease models used and the remainder of the genome in the strains studied may explain the variation in results seen.[46]

Most strains causing disease in humans carry both toxins, however several outbreaks have been reported of A-B+ strains,[48,49] which are associated with the typical range of disease from asymptomatic colonisation to severe colitis.[50]

The hypervirulent ribotype-027/ST1 strain also carries an additional binary toxin,[51] an actin ADP-ribosyltransferase which also disrupts the cell actin cytoskeleton.[52] This strain also contains a truncation in the *tcdC* gene originally hypothesised to be a negative regulator with the truncation resulting in increased toxin production,[53] however this has been recently questioned.[54]

2.4 THE SCALE OF THE PROBLEM

The incidence of CDI annually in England has fallen consistently since 2007, from 55498 cases in 2007/8 (when enhanced surveillance of all cases in those above 2 years was introduced), a rate of 13875 per quarter to 3708 reports between January and March 2012 (**Figure 1**). However it still remains a considerable problem, with ~19000 cases in England in 2011.[55]

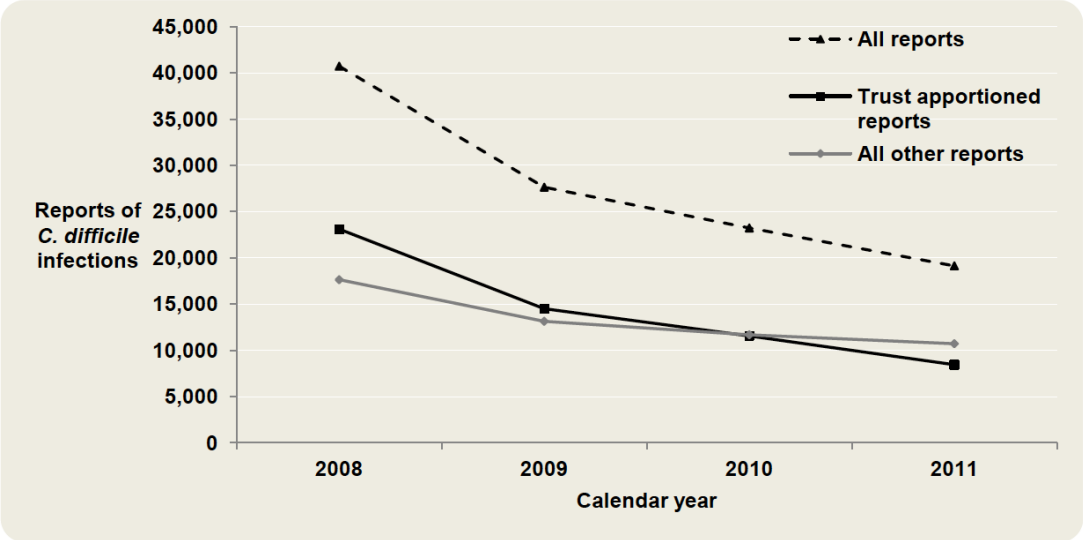


Figure 1. Clostridium difficile infection reports, England, 2008-2011. Source: Health Protection Agency.[55]

The number of cases apportioned to acute hospital Trusts (defined by the UK Health Protection Agency as a case diagnosed while an inpatient and >72 hours

after admission) has fallen more rapidly such that non-Trust apportioned cases now account for the majority (**Figure 1**). Trust apportioned incidence has fallen from 8.8 cases/10000 bed-days at the start of 2008[56] to 1.8 cases/10000 bed-days in the first quarter of 2012.[55] This contrasts with inpatient CDI incidence in North America between 1997 and 2005 of 3.8 - 9.5/10000 bed-days[29] where incidence has continued to rise over the last 5 years.[57] The incidence across Europe in 2008 was ~5.5 per 10000 bed-days.[58]

Historically reported rates of attributable mortality in *C. difficile* were relatively low ~2%,[29] however the emergence of the ribotype-027/ST1 strain in the early 2000s was associated with 30-day attributable mortality of 6.9% and an overall 30-day mortality of 25%.[3] All cause 28-day mortality in a UK setting following ribotype-027/ST1 infection was 23% compared to 11% following ribotype-106/ST42.[58] In the Netherlands attributable mortality following ribotype-027 infection was 6.3% compared to 1.2% with other ribotypes.[59] Over a similar time period as the emergence of ribotype-027 in England and Wales, between 2004 and 2007, there was a 4-fold rise in CDI on death certificates.[60] Since 2007 rates have fallen back again by more than 70%.[61] CDI was involved in 1.1% of all deaths in England and Wales between 2006 and 2010 and just less than 2% of all hospital deaths.[61]

Estimated annual costs of CDI in the USA range from $0.4 – 3 billion, although the authors of these papers acknowledge these estimates are conservative, and mostly only account for hospital related costs.[36] A recent simulation study estimated the annual costs of *C. difficile* in the USA as ≥$1043 million to hospital institutions and healthcare funders, and ≥$796 million to society more widely.[62] Costs per

CDI are estimated in the region of several thousand pounds, with recurrent disease costing considerably more. Attributable additional length of stay from CDI across a number of studies ranges from 3 – 21 days.[63] CDI is also associated with increased risk of re-admission over the next 180 days and discharge to a long-term care facility or other hospital, rather than home.[64]

2.5 RISK FACTORS

Prior antibiotic exposure is the principle risk factor for CDI, with broad-spectrum cephalosporins, clindamycin and fluroquinolones most frequently implicated.[58,65] The pooled odds ratios associated with different antibiotics in a meta-analysis of 49 studies are shown in **Figure 2**.[66] However, there is large variation in the risks reported for different antibiotics, with conflicting data from different studies, in part due the relatively small sample sizes in many, varying experimental design including varying control selection in case-control studies, and the different populations studied.[65] Importantly the relative risk of any given antibiotic is likely to depend on the resistance present in locally circulating strains of *C. difficile*. Over time a pattern of detection of *C. difficile* resistance to the dominant antibiotic, followed by identification of the antibiotic as a risk factor is seen, first with clindamycin and then cephalosporins, and more recently fluroquinolones.[65,67]

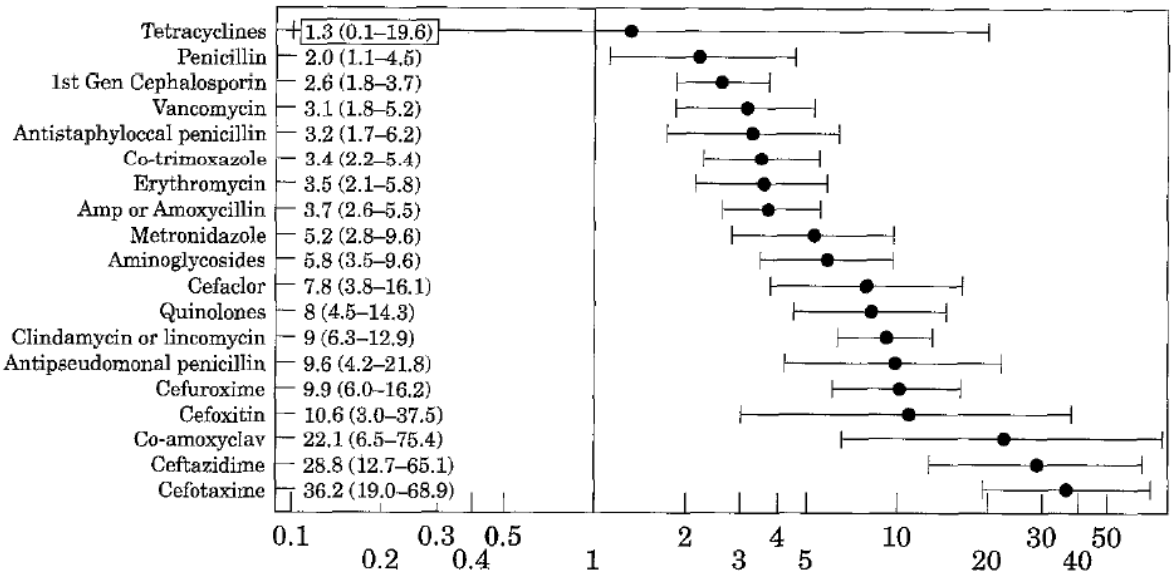


Figure 2. Antibiotic risk factors for CDI. Pooled odds ratio for CDI with 95% CI for each antibiotic exposure. Source:[66]

The combination of multiple antibiotics increases risk. In a case-control study patients with CDI were more likely than controls to have received multiple antibiotics within 14 days of the onset of CDI symptoms (80/109 versus 56/108, $p<0.002$).[68]. A cohort study of patients admitted to community hospital identified the number of antibiotics taken as a risk factor for CDI (RR, 1.49; 95% CI, 1.23–1.81). Increasing duration of therapy is also associated with increased risk. Data from a Canadian cohort study are shown in **Table 1**, showing increased CDI hazard ratios with duration of therapy across a range of antibiotic classes, after controlling for age, length of hospital stay, comorbidities and previous CDI, albeit with relatively broad confidence intervals.[67]

Antibiotic class	Adjusted hazards ratio (95% CI),[a] by duration of therapy		
	1–3 days	4–6 days	⩾7 days
Fluoroquinolones	2.42 (1.62–3.62)	2.99 (2.06–4.35)	4.33 (3.21–5.84)
First-generation cephalosporins	1.07 (0.66–1.75)	2.61 (1.28–5.31)	3.14 (1.98–4.98)
Cefuroxime and oral second-generation cephalosporins	1.20 (0.73–1.98)	1.80 (1.17–2.76)	1.80 (1.20–2.69)
Cefoxitin	3.41 (2.07–5.60)	2.58 (0.36–18.63)	2.14 (0.29–15.54)
Third-generation cephalosporins	1.41 (0.94–2.10)	1.53 (0.93–2.53)	1.75 (1.08–2.83)
Macrolides	1.38 (0.80–2.40)	1.62 (0.88–2.97)	2.09 (1.12–3.90)
Clindamycin	1.15 (0.47–2.83)	2.35 (0.86–6.43)	2.38 (1.15–4.93)
Intravenous β-lactam/β-lactamase inhibitors	1.75 (0.96–3.18)	1.98 (1.13–3.50)	1.82 (1.15–2.88)

Table 1. Impact of duration of antibiotic therapy on CDI. Source:[67]

The exact temporal relationship between antibiotic exposure and *C. difficile* acquisition is not entirely clear. The duration of increased susceptibility to CDI following antibiotics was investigated in a recent case-control study. Antibiotic exposure in the last 90 days increased the odds of CDI, with exposure 30-90 days before the onset of diarrhoea associated with lesser odds than exposure in the month before diarrhoea (**Figure 3**). A clear dose relationship was seen in this data too, as measured by defined daily doses of antibiotics received, with an approximate 3-fold increase in the odds of CDI with ≥14 days of antibiotics compared to under 3 days of antibiotics.[69]

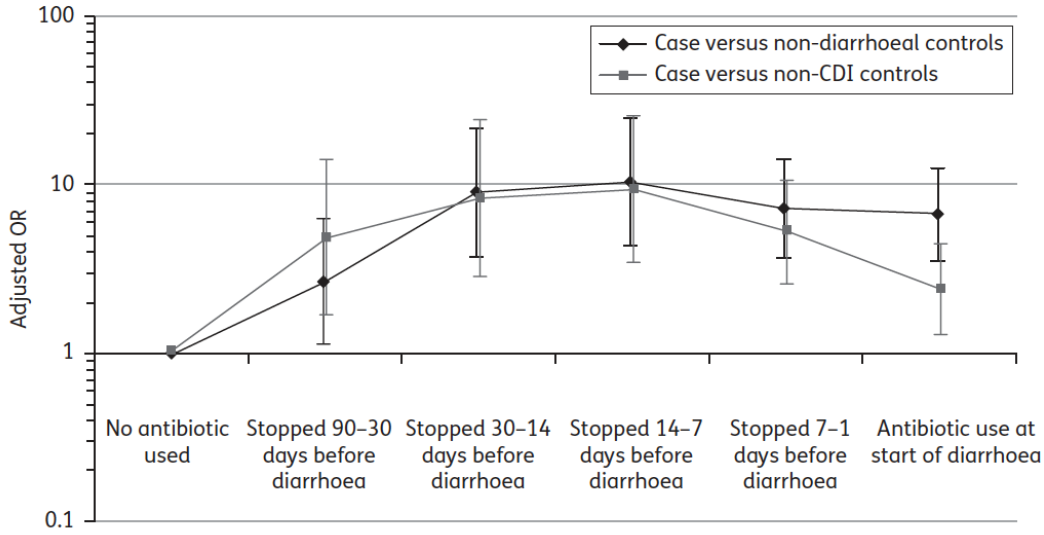


Figure 3. Relationship between timing of antibiotic exposure and the risk of CDI. Source:[69]

Data from prospective cohort studies[39,70,71] and case control studies[72] demonstrates increasing age is a strong risk factor, for example in one study the adjusted relative risk of CDI in patients aged 61-75 years was 9.6 compared to those aged 18-40 years,[70] and in another study, each year of increasing age was associated with increased odds of 1.02 (95%CI 1.00 – 1.04) of CDI.[39] In a survey of US short-stay hospitals between 1996 and 2003 CDI rates were 7-fold higher in patients >65 years compared with the group with the next highest rate, those 45-65 years.[73] Increased severity of underlying illness is also associated with increased risk.[70,71]

Gastric acid suppressive medication is a probable, but debated, risk factor. In a recent and comprehensive meta-analysis of 47 studies the pooled odds ratio of CDI in patients taking proton pump inhibitors was 1.65 (95%CI 1.47 – 1.85). However evidence of publication bias was apparent on a funnel plot, and using a regression

based method to adjust for this the pooled odds ratio was 1.51 (95%CI 1.26 – 1.83). In a meta-regression lower effect estimates were seen in studies that asked patients to recall proton pump inhibitor use, compared with those that used medical records, and in studies using unadjusted effect estimates compared with those using adjusted effect estimates. In the general population the number needed to harm based on these estimates and rates of CDI may be as large as 3925, but in hospitalised patients receiving antibiotics (considering their risk of CDI up to 14 days post discharge) the number needed to harm falls to 50.[32] A previous meta-analysis demonstrates proton pump inhibitors are associated with greater risk than H_2 receptor antagonists.[33] Further support for proton pump inhibitors as a risk factor comes from data showing a dose response relationship with increased doses and greater treatment durations associated with increased risk, and from a study showing four-fold increased CDI recurrence risk while taking proton pump inhibitors.[74]

Hospital exposure is an established risk factor, even after accounting for increased antibiotic exposure in hospital[75], in part mediated by shared time and space in hospital with CDI cases.[76] This is discussed in the section that follows on transmission epidemiology.

Asymptomatic carriage of *C. difficile* appears to be protective against subsequent disease, at least in the short term. In 4 longitudinal studies including 810 patients admitted to hospital, 22/618 non-colonised patients developed CDI, compared to 2/192 patients colonised at admission (pooled risk difference -2.3% (95% CI -4.3,-0.3)). The mean (sd) follow up was 1.7 (1.3) weeks.[77] In a subset of the 810 patients, 282 long stay inpatients studied for 9 weeks, 0/51 asymptomatic carriers

developed CDI, whereas 7/229 non-colonised patients developed CDI during follow-up.[78] IgG antibodies to *C. difficile* toxin A may play in important protective role. Serial anti toxin A and toxin B antibody levels and stool cultures were obtained from patients admitted to hospital. Baseline antibody levels were similar in patients who acquired *C. difficile* and those who did not. However, amongst those who acquired *C. difficile*, patients who remained asymptomatic developed significantly greater rises in IgG antibodies to toxin A, compared with those who developed diarrhoea.[79]

2.6 TRANSMISSION EPIDEMIOLOGY

Traditionally insights into the transmission epidemiology of infectious disease arise from point source outbreaks (e.g. Norovirus[80], salmonella[81], HIV[82]) or challenge studies, (e.g. Typhoid[83], Syphilis[84], Malaria[85]). Here both the timing and source of exposure to the infectious agent are likely to be known allowing estimates of important parameters, such as incubation periods, infectious doses and attack rates. In contrast, in endemic disease, patients may be exposed to multiple potential sources of infection at different times obscuring the true source of infection and hindering assessment of epidemiological parameters. In the case of endemic *C. difficile* this assessment is further complicated by the pathogenesis of the disease where symptomatic disease is usually the consequence both of acquisition of *C. difficile* and exposure, a variable amount of time later, or previously, to antibiotics. Recent guidelines highlight the lack for evidence for infectious and incubation periods in *C. difficile*[29], which also impairs understanding of the extent of in-hospital transmission. Understanding these attributes, and the relative importance of potential sources of infection, are

important to delivering rational infection control policy and to assessing the infection control performance of institutions.

2.6.1 EVIDENCE FOR THE HOSPITAL AS A VECTOR

Several convincing lines of evidence point to the hospital as having an important role in the acquisition of *C. difficile*. Most compellingly, serial sampling of hospital inpatients demonstrates *C. difficile* colonisation and infection increases during an inpatient admission. McFarland *et al*, writing in 1989, serially sampled adult general medical patients admitted to hospital; 29/428 (7%) had positive *C. difficile* stool cultures at admission, of the remaining 399, 83 (21%) became culture positive during the course of their hospital stay, 25% by the end of week 1 of their admission, 60% by the end of week 2. 31 (37%) went on to develop diarrhoea. Exposure to a roommate culture-positive for *C. difficile* was associated with 1.73 (95% CI 1.15-2.55) relative risk of acquisition; matching immunoblot types were observed in 20/23 suspected transmission events. The hands of healthcare workers and the hospital environment around cases and asymptomatic patients were also often contaminated with *C. difficile*.[86] Similarly, Clabots *et al* serially sampled 634 patients admitted to a medical-surgical ward, and identified 65 culture-positive patients at admission, 54 inpatient acquisitions, and that acquisition risk increased linearly with length of stay. Prior hospitalisation was associated with *C. difficile*-positive stool cultures at admission in this study[87] and another.[88] More recently, of over 4000 patients followed during hospital admission, 4.4% were carriers at admission, 2.8% developed CDI and 3.0% *C. difficile* colonisation during admission. Risk of acquisition increased with length of stay as shown in **Figure 4**.[39]

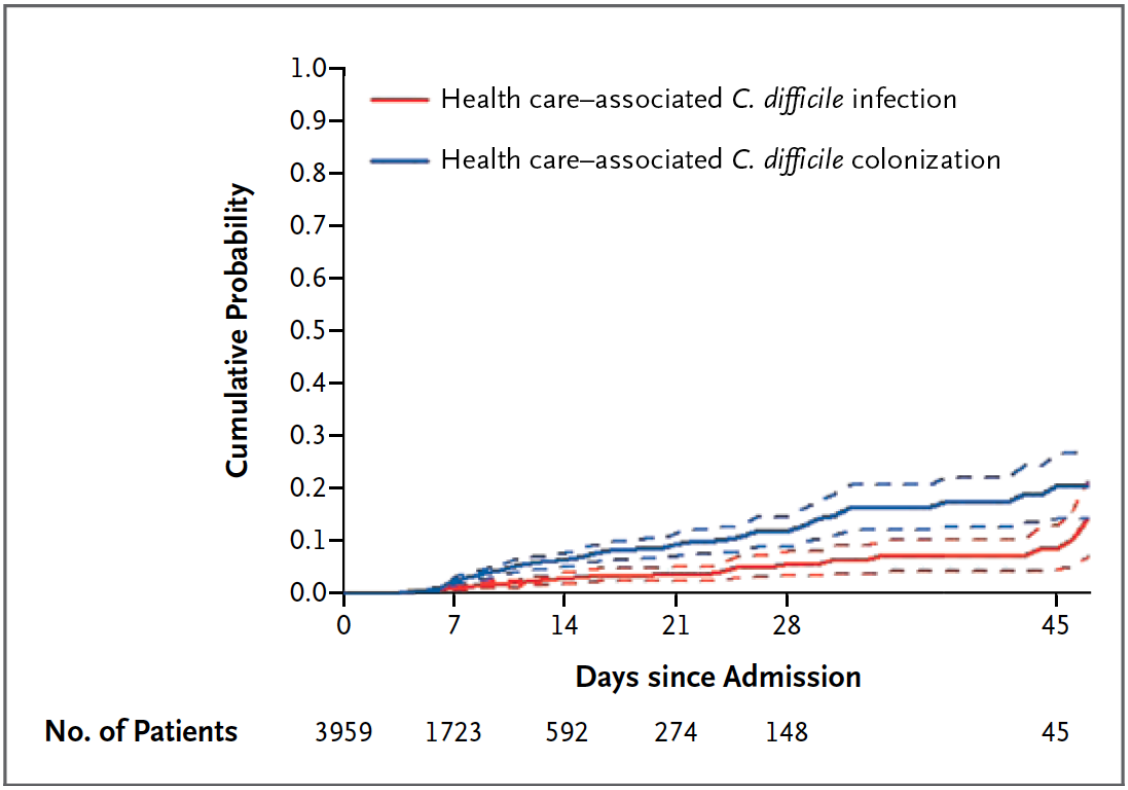


Figure 4. Times to health-care associated *C. difficile* infection and colonisation during hospitalization. The analysis presented excludes 184 with *C. difficile* colonisaiton on admission. The dashed lines indicate 95% confidence intervals. Source: [39]

Supporting the importance of CDI cases to onward transmission, exposure to symptomatic hospital inpatients is associated with increased risk of CDI. Mean exposure to symptomatic patients (mean CDI pressure, ∑case-days of exposure to symptomatic patients / length of stay at risk) was associated with CDI in retrospective cohort study including 382 CDIs. Mean CDI pressures of 0.3-1.4 were associated with 3.9 fold odds of CDI compared to pressures of <0.3.[76] Physical proximity to cases was reported to increase CDI risk (relative risk 1.86, 95% CI 1.06-3.28) in a community hospital cohort.[89]

Data from Clabots *et al*'s study,[87] discussed above, suggest asymptomatic patients are another important source of infection by introducing *C. difficile* to

hospitals and subsequently propagating it further. Samples underwent restriction endonuclease typing (REA). 19 REA types were found where both admissions with, and acquisitions of, the same REA type were observed. 16/19 followed a sequence of introduction, followed by secondary cases, with 15/16 introduced by an asymptomatic patient (2 from another ward in same institution, 2 from another hospital, 12 from home (7 with inpatient exposure in the last 30 days).[87] In contrast, in another study, inpatient contacts of index cases were cultured within 48 hours of the onset of symptoms in the index case. 12/99 contacts developed CDI, and 19 were asymptomatically colonised. *C. difficile* isolates were typed by pulsed-field gel electrophoresis (PFGE). 5/12 symptomatic contacts had matching PFGE types with the index case compared to 1/19 asymptomatic contacts, suggesting asymptomatic contacts played only a minor role in transmission.[90] Hence, while onward transmission from asymptomatically colonised patients appears possible, there is not clear data on its relative importance as a source of CDI.

The role of hospitals in acquisition of *C. difficile* is also supported by the contrasting rates of CDI observed in healthcare facilities with those in the community. Clearly such comparisons are confounded by the different populations in each location, their demographics, underlying illnesses and importantly exposure to provocative antibiotics, however marked differences in incidence are seen. In endemic healthcare settings ~5-10 CDI cases per 10000 patient-days are seen compared with rates in the community in patients without recent healthcare contact of 8-25 cases per 100000 person-years.[91] In a large defined population

in Sweden, CDI rates in hospital inpatients were 1300-fold greater than the community, and antibiotic use only 37-fold higher in hospitals.[92]

Environmental contamination with *C. difficile* in hospitals is also widely reported. Surfaces and floors around infected patients have been found to be significantly contaminated with *C. difficile*, compared to other areas of the hospital.[93] Contamination is greater around symptomatic patients compared to patients colonised asymptomatically, for example 20% versus 7% of surfaces tested (compared to 3% in the room of a patient with *C. difficile* negative stool cultures).[93] In another study 7/88 (8%) of rooms occupied by culture negative patients were contaminated, compared to 11/38 (29%) rooms of asymptomatic culture-positive patients and 44/90 (49%) of the rooms of symptomatic CDI patients.[86] Deliberate contamination of the environment persisted for up to 5 months without cleaning activity, although numbers of recoverable organisms dropped by 2 orders of magnitude 24-48 hours after inoculation,[93] while surface disinfection reduced contamination 4-fold.[94] In another study contamination of rooms of infected patients was found up to 40 days after discharge.[35] In an outbreak setting, contamination of the environment has been found even in the rooms of uninfected patients, 16% in one series[95] and 33% in another small study.[96] *C.* difficile of the same genotype as symptomatic cases has been found on the hands[86,90] and uniforms[97] of clinical staff. Portable equipment and areas not directly used for patient care, for example physician and nurse work areas, also become contaminated.[95] In addition, the air around infected patients may become contaminated, in 12% of cases on limited sampling and in 7/10 cases in more intensive sampling.[98]

Although such data point to the hospital as in important vector for transmission, the relative importance of person-to-person spread versus environmental contamination, and of symptomatic patients versus asymptomatic carriers, is unclear. Despite this, symptomatic patients and their immediate environment have become regarded as the principle source of transmission[99] and therefore the main target of prevention efforts. However this has not always proved effective, in one study no change in incidence was seen after strengthening infection control procedures, but a marked reduction was achieved by targeting antimicrobial stewardship rather than just disease acquisition.[6]

2.6.2 INFECTIOUS PERIODS

Relatively little is known about the duration of the infectious period in CDI, for example how infectiousness relates to the onset and resolution of diarrhoea. Current guidance suggests that patients remain isolated until 48 hours after resolution of diarrhoea, [4,100] based on consensus and use of this threshold in specific outbreaks.[101] However, skin contamination at multiple sites (groin, chest, abdomen, forearms, hands) has been found to persist longer, being present in ~50% of cases at 7-10 days after resolution of symptoms.[102] Skin and environmental contamination appears to be supressed during treatment, but in another study was apparent in ~50% of patients from 1-4 weeks from the end of treatment and in ~20% up to 6 weeks from the end of treatment (**Figure 5**).[103] In 46 patients with resolved CDI, stool, skin and environmental contamination was observed in 50%, 29%, and 5% <1 month post resolution of symptoms, versus 46%, 18%, 5% 1-6 months post resolution of symptoms respectively. In 12 patients 6-24 months post CDI no *C. difficile* was recovered.[104]

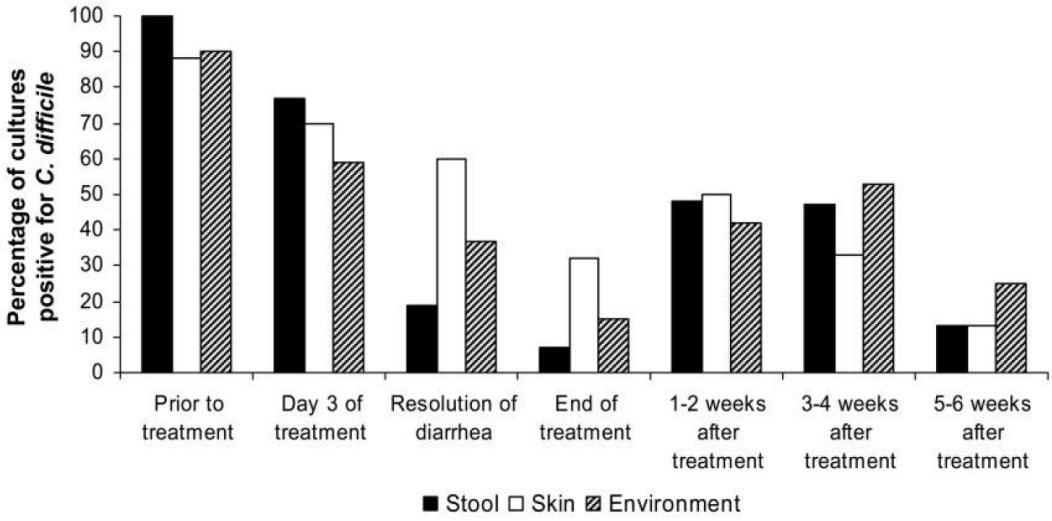


Figure 5. Percentage of stool, skin (chest and abdomen), and environmental (bed rail, bedside table, call button, toilet seat) cultures positive for *C. difficile* among 52 patients with CDI. The limit of detection for stool specimens was ~2 $\log_{10}$ colony forming units/gram. The numbers of patients who had samples cultured at each time point were 52 before treatment, 48 on day 3 of treatment, 43 after resolution of diarrhoea, 28 at the end of treatment, 22 at 1-2 weeks after treatment, 15 at 3-4 weeks after treatment and 8 at 5-6 weeks after treatment. Source: [103]

2.6.3 INCUBATION PERIODS

Conflicting data on incubation periods (time from acquisition of the organism to development of disease) comes from serial sampling of hospital inpatients and studies examining the duration of increased risk of CDI following hospital exposure. In 2 serial sampling studies the median time from first *C.* difficile-positive stool culture to onset of diarrhoea in patients previously culture-negative was 3 days (range -1 to 24 days),[88] and 2 days.[86] However, in a study of outpatients at a cancer centre, hospital exposure within the last 60 days was a risk factor for subsequent CDI, even after controlling for antibiotic exposure.[105] The number of hospital admissions in the last year was identified as a risk factor in another multivariate analysis.[106] Current consensus guidelines define disease

occurring after 48 hours of healthcare facility admission and up to 4 weeks post discharge as healthcare facility associated, between 4 – 12 weeks post discharge as indeterminate and >12 weeks post discharge as community associated.[29]

2.6.4 TRANSMISSION IN LONG-TERM CARE FACILITIES

High rates of *C. difficile* colonisation in long-term care facilities have been reported, likely as a consequence of the elderly population, a closed environment and significant antibiotic exposure. In the context of a CDI epidemic in local acute and long term care facilities, asymptomatic carriage rates in long term care residents were 35/68 (51%). Compared with non-colonised patients, asymptomatic residents had higher rates of skin and environmental contamination. *C. difficile* was recovered from investigators' hands after contact with asymptomatic residents.[107] In the absence of an outbreak, reported colonisation rates range from 4% to 20%, with a significant proportion of residents colonised on admission and a further 10-20% acquiring colonisation during their stay.[108,109] Most patients are colonised for less than 2 months; in 94 serially sampled patients in a long-term care facility, 78 without symptoms, 70% were colonised for <2 months, 7% for 3-6 months and 12% for more than 6 months.[110] Outbreaks of symptomatic disease in long term care facilities are also reported. [111,112]

2.6.5 ENVIRONMENTAL SOURCES OF *C. DIFFICILE*

Food, animal and environmental contamination with *C. difficile* have been repeatedly reported, but the extent to which this contributes to human disease is unclear.[58,113] Although overlap in the genotypes recovered and those causing disease is seen, direct evidence of transmission from the wider environment back to humans is sparse.

The possibility of food-borne transmission of *C. difficile* is supported by several studies showing contamination of processed meat products in Europe (0-3% of meat tested) and North America (~2-20%, with one study reporting rates of up to 42%).[113] Although only low levels of spores have been found on quantitative culture, for example, 10-240 spores per gram[114], these amounts are highly infectious in animal models of disease.[28] Toxigenic ribotypes have been recovered in both Europe and North America, and although ribotypes 027 and 078 have not been found in meat products in Europe, these are the main types found in food in North America.[113] *C. difficile* has also been found in 2.4% of raw vegetables,[115] 8% of ready-to-eat salads in the UK,[116] and 4.5% of retail vegetables in Canada.[117]

C. difficile has been isolated from a wide variety of mammals including cows, horses, pigs, elephants, bears, hamsters and non-human primates, as well as from poultry.[113] Household pets are a possible reservoir of disease, with *C. difficile* grown from the stools of ~40% of cats and dogs, and widespread contamination of veterinary clinics observed.[118] Although in this study no common REA types were found between cat/dog and local human isolates,[119] more recent data have identified ribotypes causing human disease in dogs,[120] however environmental isolates recovered from the dogs' households were of differing ribotypes. Other studies have also demonstrated overlap between ribotypes found in horses, dogs and humans.[113] Most investigative effort has focused on *C. difficile* in animals in the human food chain, in particular poultry, cows and pigs.[113] The dominant ribotype found in these animals is 078, which is particularly associated with pigs. Comparison of strains causing human disease in the Netherlands with isolates

from production animals found several overlapping ribotypes, and in the case of ribotype-078 strains that were genetically related by multiple-locus variable-number tandem-repeat analysis (MLVA) in pigs and humans.[121] Ribotype-078 is an increasing cause of *C. difficile* infection.[112]

C. difficile is also widely found in the environment. In a comprehensive survey in South Wales involving over 2500 samples, *C. difficile* was recovered from around half of rivers, lakes, and seawater samples taken from beaches. 21% of soil samples, 4/8 swimming pool samples and 1/18 tap water samples were also positive.[115] Another study found *C. difficile* in 68% of 69 river water samples taken in Slovenia.[122] *C. difficile* was found in all 18 samples analysed from 9 wastewater plants in Switzerland, 9 samples from the inflow and 9 from the treated outflow.[123] As a likely consequence of this widespread contamination of water, *C. difficile* has also been isolated from mussels, clams and zooplankton in the Gulf of Naples.[124]

2.7 PREVENTION

Significant effort has been expended to reduce CDI incidence. Multi-faceted prevention strategies to reduce *C. difficile* acquisition involve early diagnosis and swift isolation of cases, contact precautions including the use of gowns, gloves and fastidious hand washing, and environmental cleaning around cases.[4] The rationale for these strategies stems from the observations outlined above in the transmission epidemiology section, for example demonstrating skin,[102,103] environmental,[93] and healthcare worker[86,102] contamination around cases, combined with experience of dealing with other infectious pathogens in

hospitals.[4,125] There are relatively few specific studies assessing these interventions in *C. difficile*. Some examples include implementation of glove use, which resulted in a fall in CDI incidence on 2 wards, from 7.7 to 1.5 cases/1000 discharges, while on 2 control wards in the same hospital without the intervention no change in CDI incidence was seen.[126] In another example, use of hypochlorite rather than neutral cleaning products resulted in a fall in CDI incidence in 1 of 2 wards in a cross-over study.[127] Replacement of digital rectal thermometers with disposable thermometers reduced CDI incidence from 27 to 18 per 10000 patient-days in an acute hospital and 4 to 1 per 10000 patient-days in a skilled nursing facility. [128]

Successful control of outbreaks after introduction of packages of measures (or "bundles") has been reported, but only provides weak evidence for each individual intervention. For example, a large outbreak in Manchester, UK in 1992 was controlled after introduction of intensive staff education, monitoring, contact precautions, cohorting of patients, rigorous cleaning and restriction of antibiotic use.[129]

These interventions are complimented by antimicrobial stewardship programmes. The combination of restricting the use of high-risk antibiotics alongside introducing less provocative alternatives has been shown to reduce CDI incidence. In a 5 year study in 700 bed US hospital, restriction of clindamycin use reduced CDI incidence by over two-thirds.[130] In a study in a large Elderly Medicine unit in Leeds, UK, availability of a less provocative antibiotic, piperacillin-tazobactam, alone was insufficient to reduce CDI incidence, but active restriction of cephalosporin prescribing reduced CDI incidence by 52%. Following the

restriction, during a failure of piperacillin-tazobactam supply, rates of cephalosporin use rose five fold and CDI rates by 232%.[131] Restriction of third-generation cephalosporin use in a large tertiary hospital in Western Australia resulted in a fall in CDI incidence from 6.1 to 2.8 cases/10000 bed days.[132] In an interrupted time series analysis in Canada strengthening of the infection control policy had no impact of CDI incidence, however introduction of an antimicrobial stewardship programme to reduce cephalosporin, clindamycin, ciprofloxacin and macrolide use, resulted in a parallel fall in use of these antibiotics and in CDI incidence by 60% (**Figure 6**).[6] The authors suggest the lack of efficacy of the infection control interventions may have been the result of widespread environmental contamination given epidemic levels of CDI cases (~25-30 cases/10000 bed days).

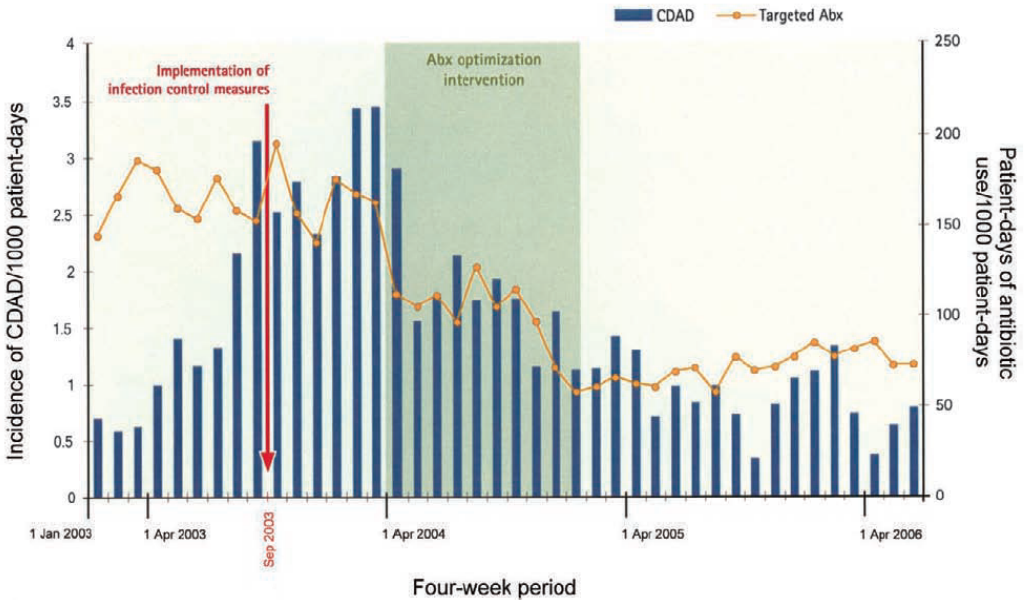


Figure 6. Targeted antibiotic consumption and nosocomial *C. difficile* infection incidence per 1000 patient-days of hospitalization. Source: [6]

Education and feedback to clinicians is important to the success of prevention programmes. In a UK hospital restriction of cephalosporin use and feedback on CDI rates and antibiotic usage resulted in a fall in CDI rates from 3.83 to 0.91 cases per 100 admissions and cephalosporin use fell by two-thirds. The lead clinician responsible for feedback was then absent and CDI rates rose back to 2.66 per 100 admissions, and cephalosporin use climbed. Re-introduction of feedback resulted in both falling again.[133]

2.8 LABORATORY INVESTIGATION

The diagnosis of *C. difficile* infection depends both on an appropriate clinical syndrome and detection of the organism. For example, Infectious Diseases Society of America and Society for Healthcare Epidemiology of America diagnostic criteria require diarrhoea, defined as the passage of ≥3 unformed stools within a 24 period, and stool test result positive for toxigenic *C. difficile* or its toxins or pseudomembranous colitis found at endoscopy or on histopathology.[29]

2.8.1 DIAGNOSTIC TESTING

C. difficile can either be detected directly through measures such as culture, molecular testing for specific antigens, for example glutamate dehydrogenase (GDH), or nucleic acid amplification tests, NAAT, for specific *C. difficile* genes; or it can be detected indirectly by detection of its toxins, for example enzyme immunoassays (EIA) and cell cytotoxicity assays.[134] Although cytotoxic culture and cell cytotoxicity assays are regarded as gold standards, the rapid turn around afforded by EIA-toxin tests led to their widespread adoption.[29] However the

limited sensitivity and specificity of EIA-testing[135] has prompted a search for improved testing methodologies.[136]

The most definitive study to date involves 12441 samples from 4 UK hospitals tested in parallel with several different methods, including cell cytotoxicity and cytotoxic culture as a gold standards, GDH EIA, 2 toxin EIA kits, and a toxin PCR (i.e. NAAT).[137] Importantly of the two gold standard methods, only cell cytotoxicity was associated with clinical outcome (mortality), suggesting that actual detection of toxin, rather than *C. difficile* with the capacity to make toxin, is most relevant when considering which patients have *C. difficile* disease (as opposed to carriage). This is important when considering the use of NAAT tests that detect toxin genes, rather than the presence of sufficient toxin to cause disease.[136] The study highlights the limited sensitivity of the two toxin EIA kits tested, 83.2% (EIA2 kit) and 67.0% verses cell cytotoxicity. In contrast GDH testing was 96.4% sensitive compared to cell cytotoxicity. To improve the specificity of testing, a two-step strategy is suggested with a sensitive screen, followed by a confirmatory toxin test. A proposed combination of GDH and EIA2 tests had a positive predictive value of 91.4% and a negative predictive value of 98.9%. However the sensitivity of this test remains only as good as the product of the sensitivity of the first test, and the sensitivity of the second test given the first was positive. Hence although an improvement in specificity may be achieved, it is likely at the cost of sensitivity. Tuning the cut offs used for both stages of the test may improve performance above standard manufacturers recommendations (which are typically set for one-stage testing).[138] The relatively high negative predictive value of the two-step strategy reflects in part the large numbers of negative

samples tested, rather than simply a particularly good sensitivity of this combination. However GDH positive, toxin EIA negative results, should alert the clinician to the possibility of a false-negative test in a way that EIA testing alone does not. These cases may also represent 'excretors' of *C. difficile* without infection but with diarrhoea from another cause, who may be relevant to infection control efforts.[137]

2.8.2 MOLECULAR TYPING

Molecular typing offers several potential insights into *C. difficile* epidemiology. Importantly for any work studying *C. difficile* transmission, it allows the large number of endemic cases to be subdivided such that the number of cases within a single genotype is small enough to make determining individual transmission events tractable. This is of considerable value in an outbreak setting too where distinct genotypes can rule out transmission.[139] Molecular typing data are also of use at a population level. Trends in the incidence of particular genotypes may reflect changing *C. difficile* epidemiology.[29] Genotypes may be associated with a particular phenotype, for example the increased virulence observed with the ribotype-027/ST1,[3] or a particular environmental niche, for example ribotype-078/ST11 in production animals.[113]

Several competing typing systems exist for *C. difficile* which are based around detecting genetic variation. Schemes in common use include: restriction endonuclease analysis, REA, where DNA is cut into fragments at specific 4-8bp restriction sites by a restriction endonuclease enzyme;[140,141] pulsed field gel electrophoresis, PFGE, also using restriction enzymes, but followed by electrophoresis where the polarity of the current is changed at predetermined

intervals to improve separation of large DNA fragments;[142] PCR-ribotyping, PCR-based amplification of the variable spacer regions between the 16S and 23S ribosomal RNA genes, multiple genes exist per organism and so a banding pattern can be detected on electrophoresis;[143,144] multilocus sequence typing, MLST, PCR based amplification of 7 housekeeping genes, followed by capillary sequencing and comparison with a database of known combinations of alleles to give types[145]); and multilocus variable-number tandem-repeat analysis, MLVA, PCR amplification of 7 variable number tandem repeat loci, the number of repeats present is proportional to the length of the product which is detected by electrophoresis.[146] PFGE typing is widely used in North America, with *C. difficile* assigned North America pulsed-field (NAP) types, for example ribotype-027/ST1 is known as NAP1.[147] PCR ribotyping is the dominant typing system in Europe.[58] PCR-ribotyping, REA, and PFGE all rely on analysis of gel banding patterns, resulting in labour intensive workflows and potential difficulties with inter-laboratory comparisons.[145] In contrast, MLST is a high throughput sequence-based scheme. The original MLST scheme[148] was limited by the fact one locus often failed to amplify in certain strains.[145] A revised *C. difficile* MLST scheme developed in Oxford[145] is increasingly being used and is supported by a central web-based store of ST alleles and STs (pubmlst.org/cdifficile).[149]

The relative discriminatory power of typing schemes can be assessed using Simpson's index of diversity, a numeric index of discrimination (DI), based on the probability that any two unrelated strains sampled from the test population will be placed into different typing groups.[150] The index is dependent both on the performance of the typing scheme and the underlying population sampled. Data

from 3 studies are presented in **Table 2**: Killgore *et al* compared seven typing schemes across 42 isolates from Canada, The Netherlands, the UK and the USA,[147]; Griffiths *et al*[145] compare MLST and PCR-ribotyping within a collection of 102 UK clinical isolates; Tenover *et al* compare 350 isolates from seven laboratories across North America.[151]

Typing Scheme	DI	95% CI	Reference
REA	0.93	0.88-0.98	[147]
PFGE	0.84	0.75-0.93	[147]
PCR-ribotyping	0.70	0.56-0.83	[147]
MLST (previous scheme)	0.70	0.56-0.83	[147]
MLST (Griffiths *et al* scheme)	0.90	-	[145]
PCR-ribotyping	0.92	-	[145]
REA	0.78	0.76-0.81	[151]
PFGE	0.65	0.61-0.69	[151]
PCR-ribotyping	0.83	0.80-0.85	[151]

Table 2. Discriminatory performance of *C. difficile* typing schemes. Comparisons of discrimination can be made within studies, but care should be taken when comparing values across studies as the underlying study population differs.

PCR-ribotyping and MLST perform similarly. The earlier, smaller study by Killgore *et al* suggests REA and PFGE may be more discriminating than PCR-ribotyping, however this is in the context of a global strain collection, rather than routine clinical practice in a local area. In contrast, a recent comparison of REA, PFGE and PCR-ribotyping suggests that although the 3 methods produce broadly congruent results, PCR-ribotyping offers the greatest level of discrimination. When comparing the additional value of adding one typing method to another, the addition of PFGE or REA to PCR-ribotyping results adds little discrimination (Wallace coefficients [152,153] of 0.902 and 0.834 respectively, 1 = no additional discrimination), whereas the addition of PCR-ribotyping to REA and to PFGE in

particular adds more discrimination (Wallace coefficients of 0.658 and 0.488 respectively).

2.9 MANAGEMENT AND TREATMENT

In addition to the prevention measures outlined above, the management of *C. difficile* should include specific treatment with antibiotics with activity against *C. difficile*, judicious supportive care and where possible cessation of provocative antibiotics.[29,100] Provocative antibiotics during or shortly after treatment increase risk of *C. difficile* relapse; in a meta-analysis of 3 studies continued antibiotics was associated with an odds ratio for relapse of 4.23 (95% CI 2.10 – 8.55).[154]

The two principle antibiotic treatments for CDI are metronidazole and vancomycin. There is only a single small randomized trial comparing either of these treatments to placebo; 44 patients, many of whom had severe disease, were randomized to vancomycin or placebo. Vancomycin was associated with resolution of symptoms ($p<0.02$) in those with evidence of *C. difficile* toxins and toxigenic *C. difficile* in the stool (10 of 16 who had evidence of pseudomembranous colitis), and clearance of *C. difficile* from stool cultures ($p<0.01$).[155] There is therefore no evidence on whether treatment offers any benefit over placebo in mild disease above supportive care. In fact, in the case of asymptomatic carriage neither vancomycin nor metronidazole performs better than placebo at clearing *C. difficile* carriage.[156] This may be in part because detectable concentrations of metronidazole in the formed stool in these patients was not achieved,[156] and vancomycin may alter the gut microbiome sufficiently to create an opportunity for

the return of *C. difficile* once treatment is stopped.[157] However given treating mild CDI is established practice, a placebo controlled trial is unlikely to be feasible.[158]

There is also relatively little data directly comparing metronidazole with vancomycin for the treatment of CDI. Three relatively small trials exist [159-161], with a total of 335 participants across the studies. No difference was found between the treatments when comparing symptomatic cure without recurrence in a meta-analysis.[158] However a subgroup analysis in the most recent trial suggests that in patients with severe disease, vancomycin may be more likely to result in clinical cure than metronidazole (97% versus 76%, p=0.02).[161]

Reduced susceptibility of *C. difficile* to treatment antibiotics is more of a concern with metronidazole than vancomycin. Concentrations of metronidazole in the gut are highest in the context of diarrhoea (~10mg/l) compared to formed stool (~1mg/kg).[162] Historically metronidazole minimum inhibitory concentrations (MICs) have typically been 0.25–2 mg/l, however reduced susceptibility has been reported; for example in Leeds, 24% of ribotype-001 isolates studied in 2005-6 had MICs ≥6mg/l, although other ribotypes including 027 and 106, and historical 001, were fully sensitive.[163] Resistant strains with MICs of 16-64mg/l have been isolated from clinical samples in Spain,[164] in a series of 415 isolates obtained during 1993-2000, metronidazole concentrations of 4mg/l were required to inhibit the growth of 90% of isolates (MIC_{90}) with 6.3% having MICs of 32mg/l or higher.[165] In contrast, after two days of oral vancomycin faecal concentrations reach ~350mg/l,[155] and the highest ever reported MIC for vancomycin is 16mg/l.[29]

Several other agents have been used to treat CDI. Data from small clinical trials suggest nitazoxanide and teicoplanin may be alternative options to standard treatment, and rifaximin has been used in uncontrolled trials for management of multiple recurrences.[166] The most promising new agent is fidaxomicin. In a randomized trial rates of clinical cure were non-inferior with fidaxomicin compared to vancomycin in a modified intention to treat analysis (88.2% vs 85.8% respectively).[167] Significantly fewer patients in the fidaxomicin group had a recurrence of infection (15.4% vs 25.3%, p=0.005), a potential advantage given the high rates of recurrence seen following CDI. This may in part be due to the greater specificity of fidaxomicin for *C. difficile* and less disruption of the rest of the gut flora.[157] In a meta-analysis of the two trials of fidaxomicin conducted to date, using a full intention to treat time to event analysis, fidaxomicin was also associated with a 40% (95% CI 26-51%) reduction in persistent diarrhoea, recurrence or death to day 40, and a 37% (95% CI 24-60%) reduction in persistent diarrhoea or death to day 12. Fidaxomicin may therefore also improve short-term and longer-term patient outcomes beyond just preventing recurrence.[168]

Recurrent CDI is significant challenge in the management of individual patients; 6-25% of patients will have at least 1 additional episode.[29] Recurrence is also therefore a significant contributor to the overall burden of *C. difficile* disease and a major potential source of onward transmission. Recurrent disease can be the consequence of either re-infection or relapse of the initial infection. Distinguishing between the two is important to understanding CDI epidemiology. Guidelines recommend treatment of a first recurrence in the same way as an initial episode and tapered or pulsed vancomycin for further relapses.[29] There is no clear

evidence probiotics are of benefit,[29] however restoration of normal gut flora by "faecal transplant" has been successful in several case series,[169] and its benefit was clearly demonstrated in a recent randomised trial.[170] Resolution of diarrhoea without relapse at 10 weeks was achieved in 13/16 receiving vancomycin treatment followed by duodenal infusion of donor stool, compared to 3/13 receiving vancomycin alone, and 3/13 receiving vancomycin and bowel lavage ($p<0.001$ for both comparisons).

2.10 Whole genome sequencing

2.10.1 Sequencing technology

The availability of high-throughput sequencing technology is transforming bacterial genomics and is fundamental to this thesis. Sanger sequencing, first described in 1977[171], made DNA sequencing of fragments up to several hundred base pairs widely available. Sequences were initially radiolabeled, however the use of fluorescently labelled dideoxynucleotides from 1986 onwards allowed optical reading of sequences, increased throughput and supported the development of automated sequencing platforms.[172,173] Such technology was used to sequence the first bacterial genome in 1995.[174] Randomly generated multiple overlapping fragments of bacterial DNA were sequenced and then aligned into longer, contiguous sequences (contigs) in a process known as shotgun sequencing.

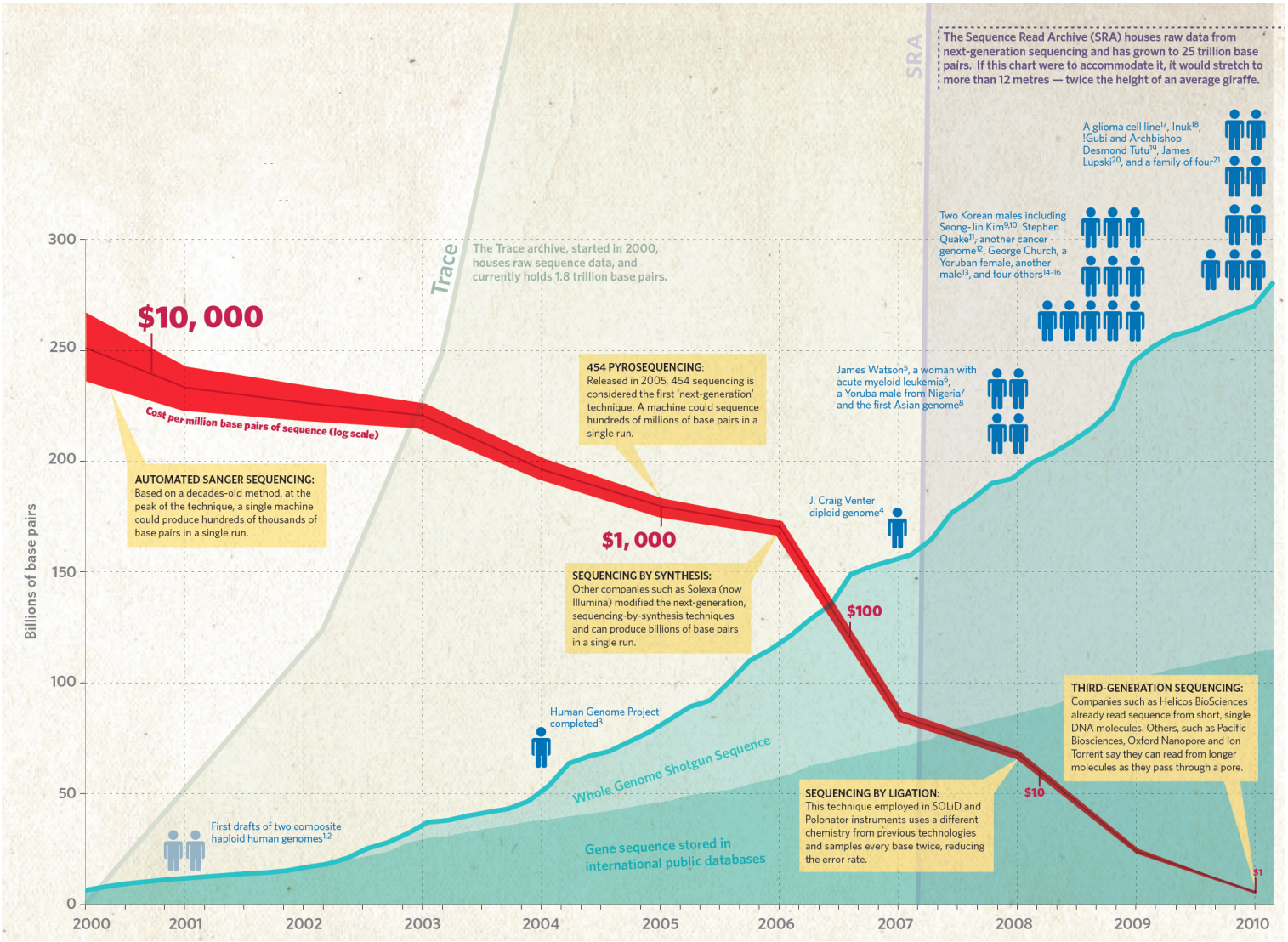


Figure 7. The sequence explosion. Source: [175]

Subsequently the development of 'next-generation' sequencing platforms, initially pyrosequencing[176] and then sequencing by synthesis (of which Illumina technology (Illumina, San Diego, CA, USA) used in this thesis is an example),[177] drove an explosion in the availability of affordable sequence data (**Figure 7**). These methods rely on generation of a DNA library prior to sequencing – DNA is fragmented, the ends repaired, size selected, and adaptors added which allow a universal primer to bind to the DNA, facilitating amplification prior to sequencing.[173,178] DNA is immobilised or attached such that the local amplified DNA is identical (**Figure 8**). When sequencing these co-located clonally amplified libraries, the observed signal is a consensus of the nucleotides or probes added to the cluster or bead of identical templates for a given cycle. The Illumina platform uses a process of cyclic reversible termination: initially a DNA polymerase bound to the template sequence includes an adapted fluorescently labelled nucleotide that terminates DNA synthesis, unincorporated nucleotides are washed away, and an image then taken to identify added nucleotides, finally a cleavage step removes the fluorescent dye and the terminator to allow the process to repeat (**Figure 9**). These technologies require relatively large amounts of input DNA (2-30μg), and generate reads between 100 and 250 bp in length (or up to ~500bp with pyrosequencing). In contrast, emerging single-molecule template based sequencing requires less starting material (<1μg), does not involve amplification prior to sequencing, and offers longer read lengths, but at present with less accuracy.[173,178]

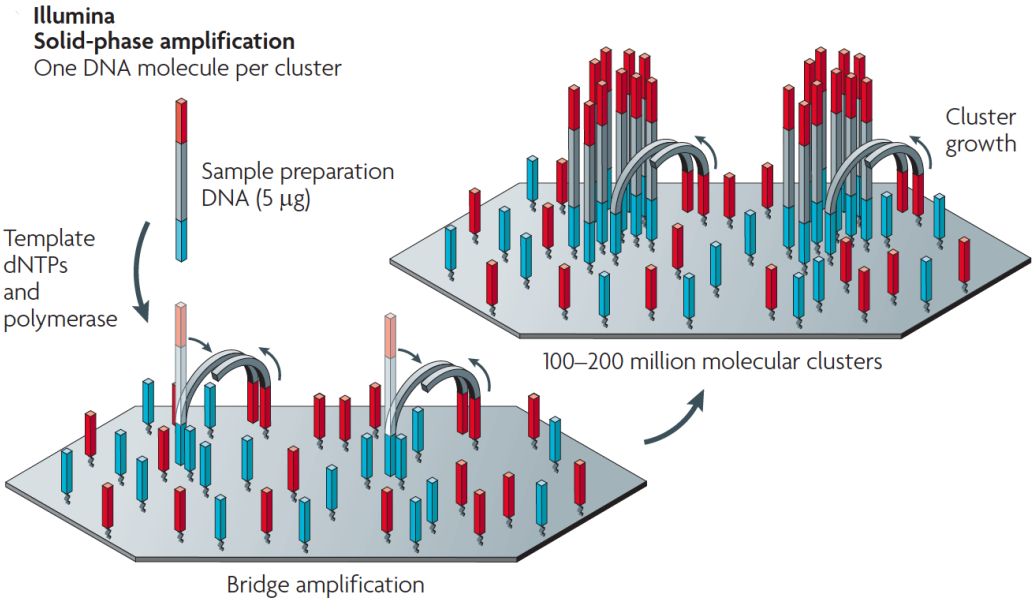


Figure 8. Illumina template amplification. Prior to sequencing the initial DNA template is amplified after local attachment to a cluster through a process known as bridge amplification. Source: [173]

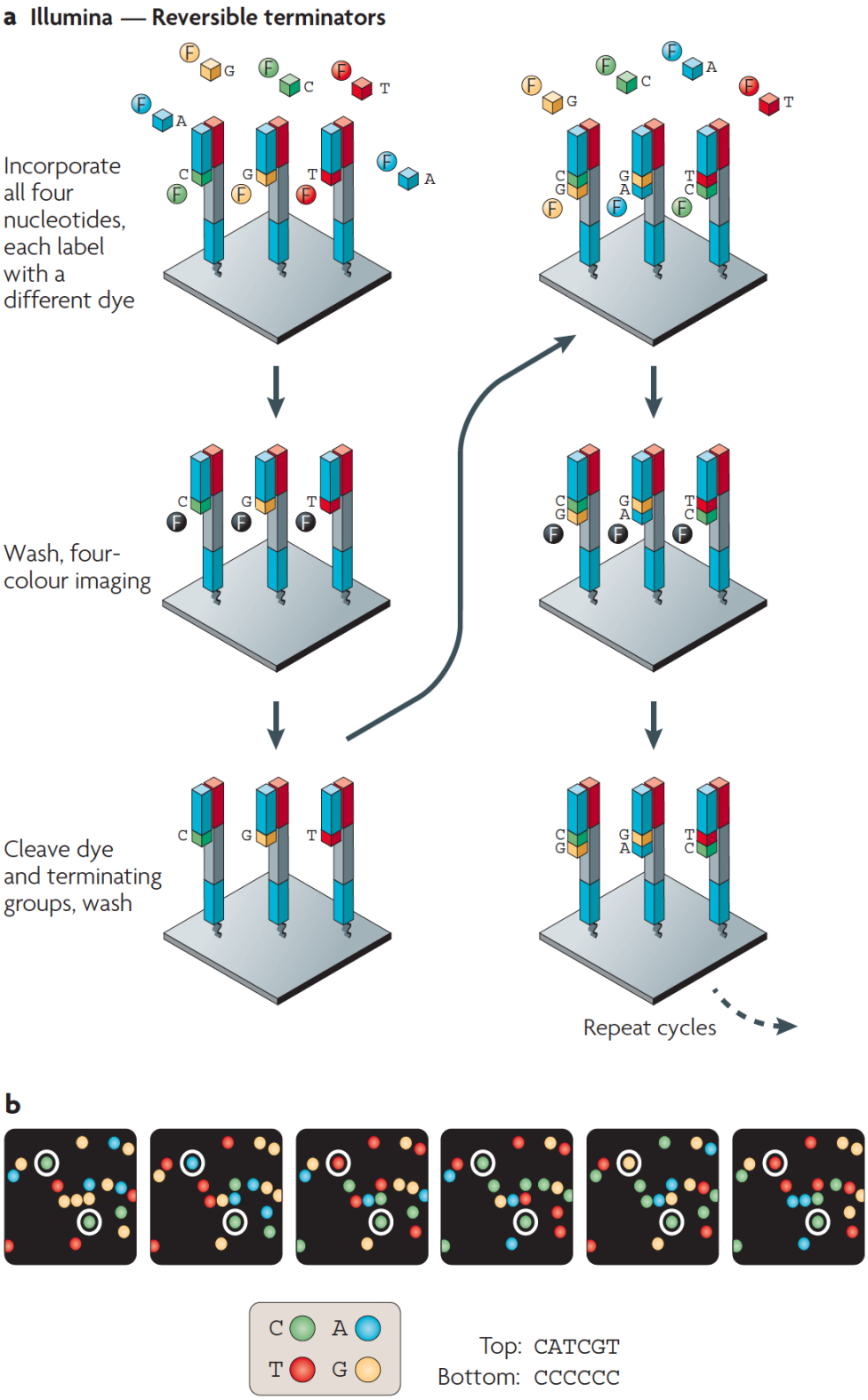


Figure 9. Illumina sequencing and imaging. Panel A shows the sequencing cycle with only a single template shown for each cluster for ease of illustration, and panel B an example of the optical read out and the resulting sequence for 2 clusters. Source: [173]

Differing platforms offer a range of throughputs, runtimes and costs, and have their own error profiles. For example, the Illumina HiSeq 2000 platform used in this thesis is able to generate up to 600 gigabases (Gb) of output (sufficient for ≥1000 *C. difficile* genomes) at 100bp read lengths in 11 days, at a cost of ~£40 per isolate, and the Illumina MiSeq 4.5Gb output (sufficient for ≥12 *C. difficile* genomes) at 150bp read lengths in 24 hours for ~£85 per isolate (Illumina, San Diego, CA, USA and in house costs).

2.10.2 DATA PROCESSING

Short read sequence data generated by next-generation platforms requires further processing before it is suitable for most analysis applications. Two approaches both attempt to reconstruct the genome sequenced as completely as possible: mapping or alignment of reads against a known reference sequence, or *de novo* assembly using overlaps between reads.[178]

A mapping approach requires the data first be aligned to the reference using software such as Stampy[179], and then a summary is produced of the number and quality of reads aligned at each site, for example using SAMTools.[180] The quality of each base within a read is determined during the sequencing process and assigned an error probability, similarly during the mapping process the quality of the alignment of the read to reference is also reported in terms of an error probability. Multiple overlapping reads are present at each site in the reference, this depth of coverage typically ranges from 50-100 fold. High coverage combined with quality metrics allow for sophisticated filtering of output data allowing variants to be identified with high specificity. However genetic material sequenced, but not present in the reference (e.g. mobile genetic elements and other accessory

genome), cannot be analysed, and areas of high divergence from the reference may also be poorly represented. As such this approach is closely dependent on the chosen reference being sufficiently similar to the sample sequenced. Repetitive regions cannot be accurately sequenced as reads with the same or very similar sequence may be generated from multiple regions of the genome, and therefore map equally well to multiple locations, potentially obscuring variants arising in one of the regions. As such these regions are usually excluded from downstream analysis. However these regions are also a problem for *de novo* assemblers as they may assemble into just 1 single very high coverage region.

De novo assembly, for example with Velvet[181], is able to assemble genomes without prior knowledge of their content, addressing some of the limitations of reference based mapping. Overlapping reads are merged together to generate larger contigs, however complete closure of the reference into a single contig is not achieved, in part because of repetitive regions between contigs and regions poorly represented in the original DNA libraries and reads. Instead a bacterial genome is typically represented as a few hundred separate contigs. Current software contains fewer quality metrics than are available with mapped data, relying mostly on the size of contigs generated (larger being better) and the average depth of reads supporting each contig. Hence, designing systems to achieve precisely reproducible results on re-sequencing the same DNA is harder to achieve than with mapped data. Therefore whilst *de novo* data are useful for confirming the presence and absence of genes which may or may not be present within a reference genome, it is currently less well suited to precise comparisons of genomes within outbreak investigations.

2.10.3 CLOSTRIDIUM DIFFICILE SEQUENCING

The first complete genome of *C. difficile* was sequenced in 2006.[182] Strain 630, CD630, is a multi-drug resistant strain, originally obtained from Swiss outbreak in the early 1980s.[183] The major finding from this first whole genome sequence was that a large proportion of the genome (11%) consisted of mobile genetic elements, mainly conjugative transposons (**Figure 10**). These elements are thought to be putatively responsible for *C. difficile* acquiring a range of antimicrobial resistance and virulence genes.[182]

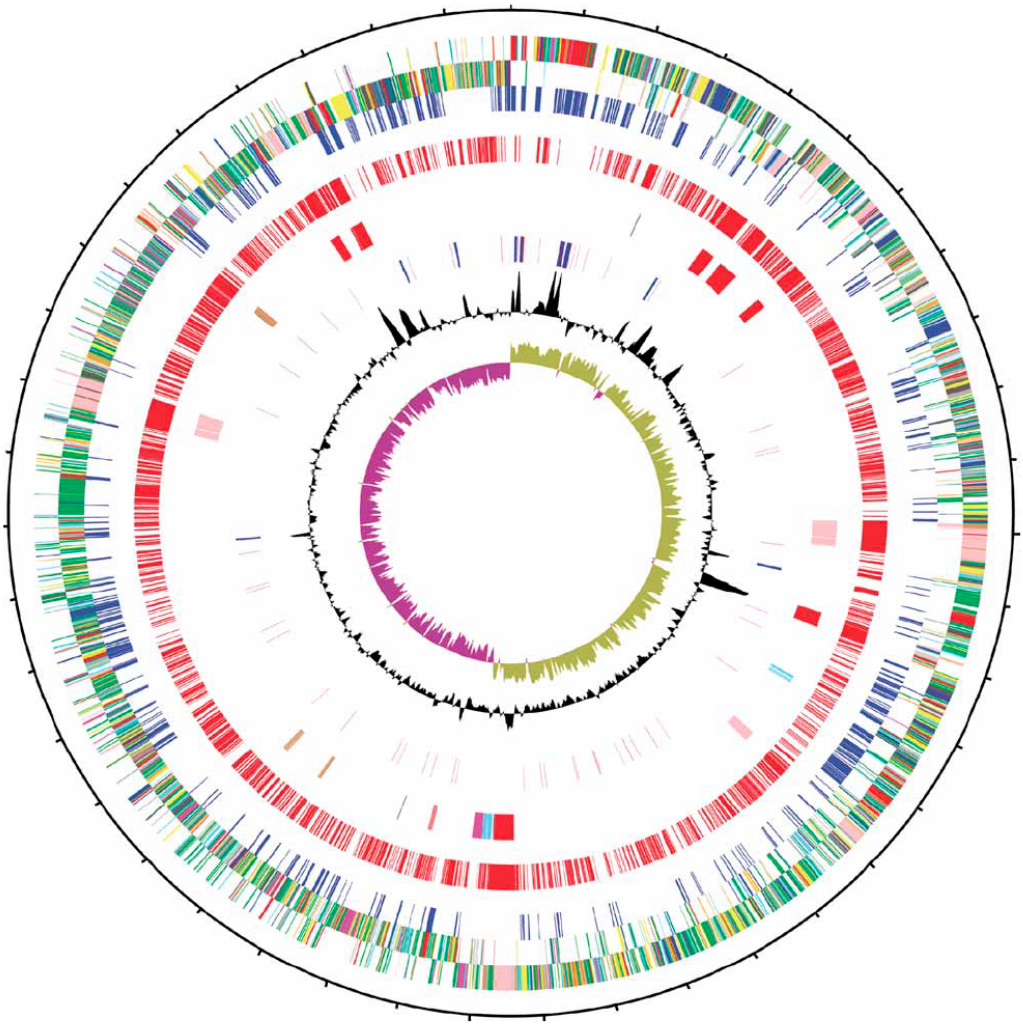


Figure 10. Circular representations of the genome of *C. difficile*. The concentric circles represent (from outside to in): 1 and 2, coding sequences (transcribed clockwise and

counter-clockwise); 3, coding sequences shared with previously sequenced clostridia (blue); 4, coding sequences unique to *C. difficile*, red; 5, mobile genetic elements, (red: conjugative transposons, pink: prophage, blue: skin element and other unknown mobile element, magenta: genomic island, brown: partial transposon or prophage); 6, RNA genes; 7, GC content; 8 GC deviation ((G-C/(G+C)). Source: [182]

Further whole genomes have since been produced, including from ribotype-027 isolates. 234 additional genes were identified within historical 027 isolates when compared to the CD630 strain, and intriguingly 5 novel genetic regions within a recent epidemic ribotype-027 isolate that may underlie its hypervirulence.[184] Analysis of a sequenced collection of diverse strains confirms *C. difficile* is a genetically diverse species, with horizontal gene transfer and recombination important to its evolution.[185] Estimation of long term evolutionary rates suggests *C. difficile* evolved somewhere between 1.1 and 85 million years ago.[185] These studies provide a foundation for applying whole genome sequencing to understanding transmission, and importantly have provided a number of closed reference genomes that can be used for reference based mapping of next-generation sequencing data.

2.11 THESIS OUTLINE

The goal of this thesis is to understand of the transmission and sources of endemic *C. difficile*. Initially a collection of all Oxfordshire *C. difficile* isolates from September 2007 to March 2010 is analysed by combining multi-locus sequence typing of isolates and epidemiological data on hospital admissions and ward movements. The aim of this analysis is to understand the proportion of *C. difficile* infection that can be attributed to ward-based contact with another symptomatic case. It also provides insights into the likely infectious and incubation periods associated with ward-based transmission.

The following chapter develops important foundations for applying whole genome sequencing to *C. difficile* transmission. The accuracy and reproducibility of *C. difficile* sequencing is described and then the degree of within host diversity and evolution investigated. Mixed infection represents a potential challenge to applying any genetic typing or sequencing method to transmission studies. If only 1 of several strains present in a patient is sampled transmission of the other strains may be missed. Therefore, mixed infection is first quantified using multi-locus sequence data, and then a tool developed to test for mixed infection in whole genome sequences.

Whole genome sequencing of all Oxfordshire *C. difficile* from September 2007 to March 2011 is undertaken. Each case is analysed for the presence of a genetically similar previous case, and potential epidemiological relationships between closely genetically-related cases sought. Having sequenced all symptomatic cases in a defined geographical region over a 3.6 year period the total genetic diversity within the symptomatic population is then quantified, in particular evidence for or against introductions of infection from external sources is sought. Trends over time in the proportions of genetically-related and genetically-distinct cases are investigated for evidence of the impact of infection control measures.

Finally, I designed, set up and ran a study to specifically probe the role of asymptomatic carriage of *C. difficile* in disease acquisition.

The thesis concludes with a discussion of the very likely application of sequencing technology to routine control of *C. difficile* at an infection control and public health level.

3 CHARACTERISATION OF WARD-BASED *C. DIFFICILE* TRANSMISSION WITH EPIDEMIOGLOICAL AND GENOTYPING DATA

3.1 INTRODUCTION

The implementation of costly prevention strategies[4] has only been partially successful in controlling CDI. Despite recent falls in its incidence in the UK *C. difficile* continues to cause a considerable burden of disease.[55] Its endemic nature obscures identification of sources and routes of transmission based on shared time and space alone. The relative importance of various routes of *C. difficile* transmission is therefore unknown, and hence the extent to which more rigorous implementation of current infection control measures can control *C. difficile* disease is also unclear.

This chapter investigates all CDI in a large geographical region over a 2.5 year period. MLST genotyping is use to subdivide endemic *C. difficile* into many distinct lineages, with each lineage containing a small enough number of cases to analyse as a separate 'outbreak'. Estimates of the interval between CDI diagnosis and putative onward transmission, incubation periods, and the proportion of cases arising from ward-based transmission from known symptomatic CDI cases are obtained.

3.2 METHODS

3.2.1 SETTING

The Oxford University Hospitals NHS Trust (OUH) (~1,700 beds) provides >90% of hospital care, and all acute services, to ~600000 people residing in Oxfordshire, United Kingdom. The Trust has two large sites in Oxford—the Churchill (medical specialities, cancer centre; two floors) and John Radcliffe (acute services, surgical specialities, women's centre, children's centre; seven floors)—and a smaller district hospital 35 miles north of Oxford in Banbury (two floors). At the time of this analysis the Nuffield Orthopaedic hospital was not part of the OUH. All OUH hospitals have discrete wards containing 20–30 beds in 4–6 bedded bays with shared bathrooms and only 2–4 side rooms: only a few specific wards (notably elderly medicine) are all single rooms. The OUH microbiology laboratory tests all stool samples from the region, including from other healthcare facilities (smaller mental health and specialist orthopaedic trusts, and community hospitals) and general practitioners (primary care). This population-based study included all CDI cases identified from routine clinical samples at the OUH microbiology laboratory from 01 September 2007 to 31 March 2010, regardless of location (inpatient, outpatient, community, other hospital). Hospital admission and ward movement data were available for all patients from the Infections in Oxfordshire Research Database, IORD.

3.2.2 SAMPLES AND MULTILOCUS SEQUENCE TYPING

Throughout this period, the OUH hospitals operated a rigorous infection control policy (**Table 3**), which required samples to be sent for *C. difficile* testing from any admitted patient with diarrhoea (locally defined as ≥3 unformed stools, i.e., taking

the shape of the container, within 24 hours), and for oral vancomycin treatment to be initiated as first-line empiric therapy when samples were sent. Compliance was monitored weekly by infection control staff, with immediate feedback to ward staff. UK Department of Health policy also required that all unformed stool samples from those aged ≥65 years were sent to the laboratory be tested for *C. difficile*, whether or not the patient met the above diarrhoea criteria and *C. difficile* testing had been requested by the clinician sending the sample. All samples were tested by enzyme immunoassay (EIA) for *C. difficile* toxins A and B (Meridian Bioscience, Cincinnati, OH). EIA-positive samples were cultured [145] and *C. difficile* isolates genotyped by multi-locus sequence typing (MLST).[145]

Selective culture of stools was undertaken following alcohol shock. Industrial methylated spirits (0.5 ml) was added to a 0.5 ml faecal sample (pea-sized portion if the stool was formed), and the sample was vortex mixed and incubated at room temperature for 1 hour. A loopful was then transferred onto modified Brazier's cycloserine-cefoxitin-egg yolk (CCEY) agar (CCEY agar base containing cycloserine-cefoxitin supplement and 5% defibrinated horse blood), and the plates incubated anaerobically at 37°C for up to 7 days. A single colony was sub-cultured onto a Columbia blood agar (CBA) plate and incubated for 48 hours, after which colonies giving the characteristic odour and fluorescence under UV illumination were obtained and underwent MLST. For long-term storage, isolates were emulsified in nutrient broth containing 10% glycerol and stored at -80°C.

MLST was performed on cultured isolates. Seven bacterial housekeeping genes were amplified by PCR with forward and reverse primers: *adk*, *atpA*, *dxr*, *glyA*,

recA, *sodA*, *tpi* (**Figure 11**), using the method of Griffiths *et al*.[145] Amplification products were purified and then capillary sequenced.

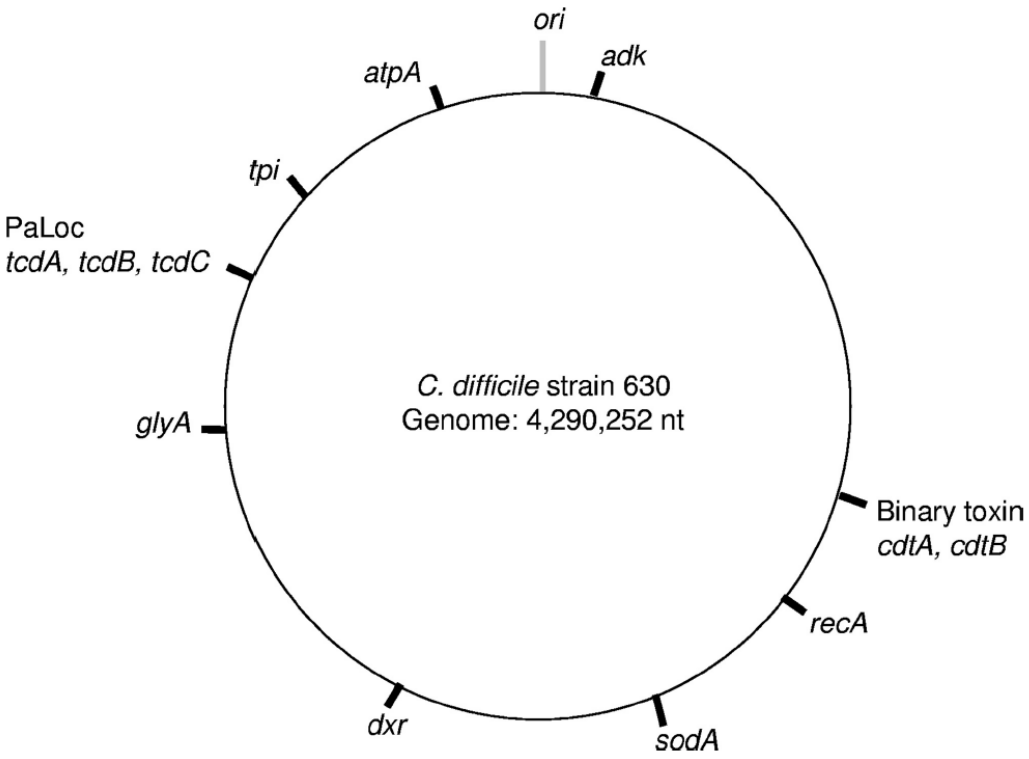


Figure 11. Multilocus sequence typing loci. The relative positions of the seven housekeeping loci chosen for MLST on the *C. difficile* strain 630 chromosome are shown, together with the pathogenicity locus (PaLoc) and the genes encoding binary toxin. Source:[145]

A single colony from each sample was typed, except for morphologically distinct colonies, which were typed separately. EIA-positive, culture-negative isolates were not considered in analysis because toxigenic culture is one of two gold standards (the other being cell cytotoxicity [186]) and false EIA-positive rates up to 20% are well recognised.[187] Therefore these samples most likely represent EIA false positives.

Interventions in place at the start of the study (September 2007)	
Intervention	**Compliance monitoring**
Antibiotic policy: Cephalosporin and fluroquinolone use restricted. Community acquired pneumonia: amoxicillin / co-amoxiclav ± macrolide; urinary tract infection: nitrofurantoin / co-amoxiclav; cellultis: flucloxacillin. Cephalosporin use in penicillin allergy, fluroquinolone for severe beta-lactam allergy.	Pharmacy review of all inpatient prescriptions against hospital-wide antibiotic policy
C. difficile isolation policy: All patients with suspected *C. difficile* infection, including patients with diarrhea of unknown cause (≥3 unformed stools in 24 hours) isolated in side rooms with en suite bathroom or dedicated commode wherever possible. Hand decontamination with soap and water, aprons and gloves worn for contact. Continued in confirmed cases until 48 hours following return to normal bowel habit.	Monitored daily, Monday to Friday by infection control team review of all inpatients with diarrhea. Frequency of monitoring decreased to 3 times a week from 2008 to 2009, and once a week during 2010. Spot checks of patient notes undertaken to ensure infection control service aware of all patients with suspected *C. difficile*, confirmed cases notified by laboratory to infection control team.
Cleaning: Actichor plus combined detergent and Sodium hypochlorite solution (called an "enhanced clean") used for once daily clean of every surface and wall in room with patient with suspected *C. difficile*.	Daily monitoring, Monday to Friday of the number of additional enhanced cleans requested with the number of patients with suspected or confirmed active *C. difficile* cases. Records from procurement monitored Monday to Friday. Frequency reduced over time as above.
C. difficile testing: Three samples to be sent for C. difficile testing from any patients with suspected infectious diarrhea (≥3 unformed stools in 24 hours)	Reviewed as part of infection control monitoring of these patients as above
Mandatory C. difficile testing: Mandatory testing of all diarrheal samples sent from patients aged 65 years or older for *C. difficile*, regardless of whether test requested	Number of EIA tests increases from 400-500 per month before, to 800-1000 per month after introduction in May 2007
C. difficile treatment: Empirical treatment of oral vancomycin 125mg qds, continued for 14 days if diagnosis confirmed	Initiation of treatment confirmed by infection control team for all confirmed cases. Pharmacy monitoring of duration of treatment.
Feedback: Failure to comply with infection control policy on isolation and cleaning escalated to ward and then senior management until compliance achieved.	

Study starts (September 2007)		
Date of Change	**Intervention**	**Compliance Monitoring**
January 2009	***Care pathway:*** *C. difficile* care pathway added to policy Care checklist including communication of diagnosis to patient, family, and medical team; review of medication (antibiotics, acid suppression, laxatives); fluid balance monitoring (fluid and stool chart); nutritional assessment; skin integrity assessment; sepsis monitoring; compliance with isolation as above	All confirmed *C. difficile* cases reviewed by infection control team to ensure care pathway in place
September 2009	***C. difficile isolation policy:*** Isolation as previously, but now within 2hrs of suspicion Exception where patients requiring one-to-one monitoring or care and unsuitable for side room care	Monitored Monday to Friday by infection control service at introduction, then 3 times a week and then weekly as above
December 2009	***Sending of stool specimens:*** Reduction in number of stool samples tested. Sending one diarrheal specimen and if negative then send a second	Monitored weekly from January 2010, over testing reported back weekly to wards

Table 3. *Clostridium difficile* infection control interventions in Oxfordshire. Interventions in place at the start of the study (September 2007), and changes during the study (through March 2010).

3.2.3 Analysis

The first EIA-positive, culture-positive *C. difficile* sample of each sequence type (ST) from each patient was included in the primary analysis (55/927 [6%] patients had multiple CDI with different STs). A network of ward-based contacts between cases, representing potential transmission events, was created for each ST (**Figure 12**). Cases were represented as nodes in the network and contacts between them, potential transmission events, as edges or links.

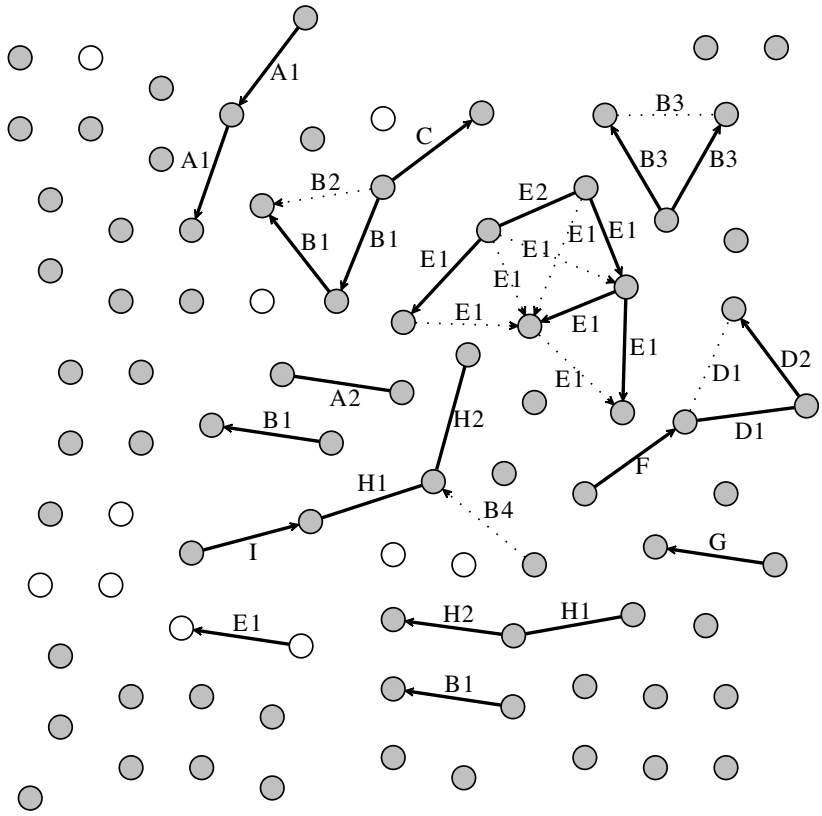


Figure 12. Ward based links. Ward-based links between cases of an example ST, ST2. Cases from the first part of the study, 1 September 2007–29 February 2008 (the run-in period, where the source of the cases may plausibly be before 1 September 2007), are shown as white circles, and cases from the remainder of the study, 1 March 2008–31 March 2010 (the test period) are shown in grey. Where patients with CDI shared time on a common ward before either or both patients' CDI (**Figure 13**), links are indicated with solid or broken lines, representing possible transmission events for example maximum allowable infectious and incubation periods of 8 and 12 weeks, respectively. All potential ward-based transmission links are included; the single most plausible link to each recipient based on minimising infectious period is shown in solid lines. Links arising from ward contamination persisting after discharge of the donor patients are not shown. Letters indicate wards on which contact events occurred; wards that are related by specialty have the same letter but different numbers. Links of known direction (ward contact after potential donor CDI and before recipient CDI) are indicated with an arrowhead. Cases are approximately evenly spaced, with linked cases adjacent for the purpose of illustration (i.e., the distance between circles does not generally correspond to any physical quantity)

Links were made when two CDI cases shared time on a ward, either (i) after the first case's sample (the "donor") and before the second case's sample (the

"recipient") (directional links), or (ii) before both cases' samples were taken (non-directional links) (**Figure 13**).

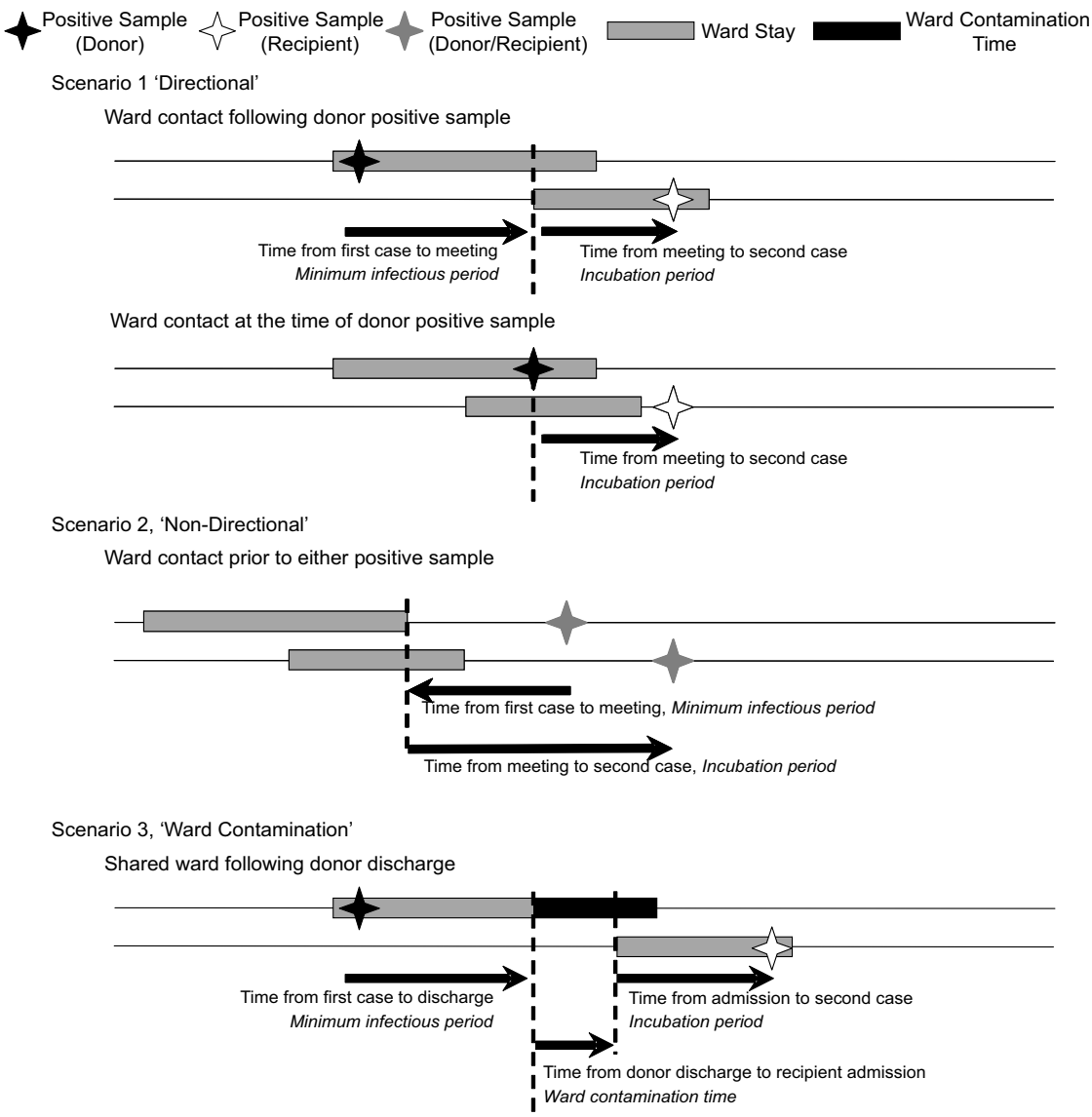


Figure 13. Schema showing different types of ward-based contact.

As defined, ward-based contacts incorporated direct person-to-person spread and indirect transmission via the environment during shared ward exposure. For each

link, the putative “minimum infectious period” was defined as the time between the first sample from the potential donor and the earliest ward contact with the recipient. The putative “incubation period” was defined as the time between this ward contact and the first sample in the recipient. Contacts were defined to occur at a ward level as specific bed/bay information was not available electronically. Day case attendances including haemodialysis were included, but outpatient attendances were not.

Where the ward contact between two patients was greater than one day, the assumed time of transmission was chosen to minimise the infectious period, because donors were expected to be most infectious immediately following a EIA-positive stool sample after which they would have received presumptive treatment and been managed as if diarrhoea was potentially infectious. Therefore where the donor sample was taken prior to the start of the ward contact, the point of transmission was assumed to be the start of the ward contact. Where the donor sample was taken during the contact, the point of transmission was assumed to be the date of the donor sample. Where donor-recipient pairs had more than one ward contact representing plausible transmissions, the contact that minimised the infectious period was chosen to define infection and incubation periods, as above.

Ward contacts indicated the direction of transmission from donor to recipient except where the two patients shared time and space on a ward before either was diagnosed with *C. difficile*. In this scenario, the direction of transmission was indeterminate and could represent a pre-symptomatic transmission event, with either patient considered a possible donor to the other.

Within each cluster of linked cases in a network, the initial links made from ward contacts/contamination could contain biological inconsistencies, for example a recipient with more than one donor. Therefore to avoid double-counting, the characteristics of the most plausible transmission link were retained, chosen assuming that donors were most infectious closest to diagnosis (ie minimising putative infectious period). To achieve this a pruning algorithm was applied in two stages. Firstly, starting with the earliest case in each cluster, directional links representing transmission to secondary cases were identified. Any cases linked to the earliest case by a link of indeterminate direction had this direction set to be consistent with transmission given the existing directional links. This provides an upper bound on the number of cases that can be plausibly linked together in each cluster. In the second stage, the network was examined for any biologically impossible link combinations, e.g. a recipient with two potential 'donors'. Here, as above, only the link with the shortest infectious period was retained. The algorithm was repeated for each secondary case and each cluster until the entire network had been examined. For any remaining indeterminate links, the minimum infectious period was arbitrarily assumed to be the shorter of the two times from ward contact end to EIA-positive sample, and incubation period the longer. Thus any recipient linked to one or more 'donors' in the original network is always counted as a transmission link in the final network, but only the characteristics of the relationship with one donor (the most plausible assuming infectious period is minimised) is presented to avoid double counting. When assessing the duration of incubation and infectious periods only links arising from directional links (where the designation of infectious and incubation periods is unambiguous) were used.

Two separate sets of networks were constructed. In the first, contacts up to 26 weeks from either donor or recipient, longer than the currently accepted infectious and incubation periods,[29,102-104] were allowed to inform the most likely infectious/incubation periods. In the second set, incubation/infectious periods were restricted based on results from the first set of networks, and the possibility of transmission from the ward environment after the donor was discharged was then included. The "potential ward contamination time" was defined as the time between donor ward discharge (or end of donor infectivity) and recipient ward admission, allowing up to 26 weeks, and assuming direct ward contact was a more likely route of transmission than ward contamination (**Figure 13**).

The proportion of cases that could be explained by credible ward-based transmission from symptomatic EIA-positive, culture-positive patients was calculated only for CDIs from 01 March 2008 to 31 March 2010, the "test" cases, to allow for new CDIs resulting from acquisition some time previously. However, all samples during and after a 6 month "run-in" period, 01 September 2007 – 29 February 2008, were considered as possible donors to "test" cases.

3.2.4 SENSITIVITY ANALYSIS

One key assumption is choosing the most plausible potential transmission link based on minimising infectious periods; analyses were therefore repeated with the alternative assumption of minimising incubation periods. Additionally, to test the impact of the assumption that transmission between donor and recipient occurred at the start of any contact between the patients following the date of diagnosis in the donor (i.e. that patients are most infectious immediately following diagnosis) the analysis was repeated stratified by the duration of overlap between patients.

Since the pruning algorithm could also affect the distribution of implied infection and incubation periods in the networks, these distributions were compared between networks both before and after pruning.

Some patients had successive EIA-positive samples with the same ST. Assuming these represent on-going diarrhoea due to the same infection, in a further sensitivity analysis donor infectious periods were calculated relative to their most recent (rather than their first) EIA-positive sample with this ST prior to the ward contact. To estimate the effect of missing cases, not detected as part of routine testing, the analysis was repeated after randomly removing cases from the complete dataset. Networks were regenerated and the proportion of the remaining cases with a credible ward-based donor calculated. Between 5% and 50% of the original cases were removed, with 1000 repetitions undertaken for each 5% increment. Extrapolation of the results was used to estimate the possible effect of missing cases.

3.2.5 CONTROLS

To calculate the proportion of links between cases that might be the result of shared ward contacts occurring by chance rather than actual transmission, ward-based networks were constructed for the same number of patients without CDI for each ST. These controls were patients with EIA-negative diarrhoea matched to have similar hospital exposure, based on the number of unique patients each case shared time and space with during the study. To ensure a similar temporal spacing of cases EIA-negative controls were also matched on tests within ±3, ±15, or ±90 days of the EIA-positive case. The EIA-negative controls are more likely to be similar to patients with CDI than general hospital controls as they also will have

had diarrhoea; hospital controls chosen at random will over-represent groups with lower CDI risks such as maternity cases. However given concern about false negative test results which could mean a proportion of EIA-negative control patients actually had CDI,[187] a second set of controls drawn from all hospital inpatients without diarrhoea was also used, matched for hospital exposure, sex, age decade and admission on the day of the case's EIA-positive sample (or for cases testing EIA-positive in the community, matched for being an inpatient on the day the case was last discharged). Each set of controls was randomly drawn 1000 times, and the median proportion of controls linked by ward-based contact calculated.

Net linked cases, correcting the overall linked CDI cases for chance ward meetings (estimated from controls), was estimated using the law of total probability, namely:

$$\begin{aligned} \mathrm{p(link)} = {} & \mathrm{p(link|transmission)p(transmission)} \\ & + \mathrm{p(link|no\ transmission)p(no\ transmission)} \end{aligned}$$

Assuming p(link|transmission) = 1, estimating p(link|no transmission) from controls, and rearranging using p(no transmission) = 1 - p(transmission) gives

$$\mathrm{p(transmission)} = \frac{\mathrm{p(link)} - \mathrm{p(link|no\ transmission)}}{1 - \mathrm{p(link|no\ transmission)}}$$

Controls were also used to assess the extent to which estimates of incubation and infectious periods were conditional on patient ward movements. For example, a patient may cease to be "infectious" in the context of the study on hospital

discharge, regardless of whether they remain infectious to their household contacts.

3.2.6 Statistics and computation

Graphs and standard statistics were done using STATA 11.1 (http://www.stata.com). Networks were analysed using NetworkX (http://networkx.lanl.gov) and draw with Graphviz (http://www.graphviz.org). Random samples were generated and networks analysed using MySQL 5.1 (http://www.mysql.com).

3.3 Results

3.3.1 Samples and STs

From 01 September 2007 to 31 March 2010, 29299 unformed stool samples from 14858 Oxfordshire patients were tested for *C. difficile*: 102 tests per 10000 overnight stays among OUH inpatients and a mean of 30 and 208 samples per month from day cases/outpatients/emergency department and non-OUH locations, respectively (**Figure 14**). In total, 1803 (6.2%) tests were EIA-positive, a rate of 9.4 CDIs/10000 overnight stays among OUH inpatients and a mean of 2.8 and 12.9 CDIs/month from day cases/outpatients/emergency department and non-OUH locations, respectively (**Figure 14**), with relatively little variation over the study period.

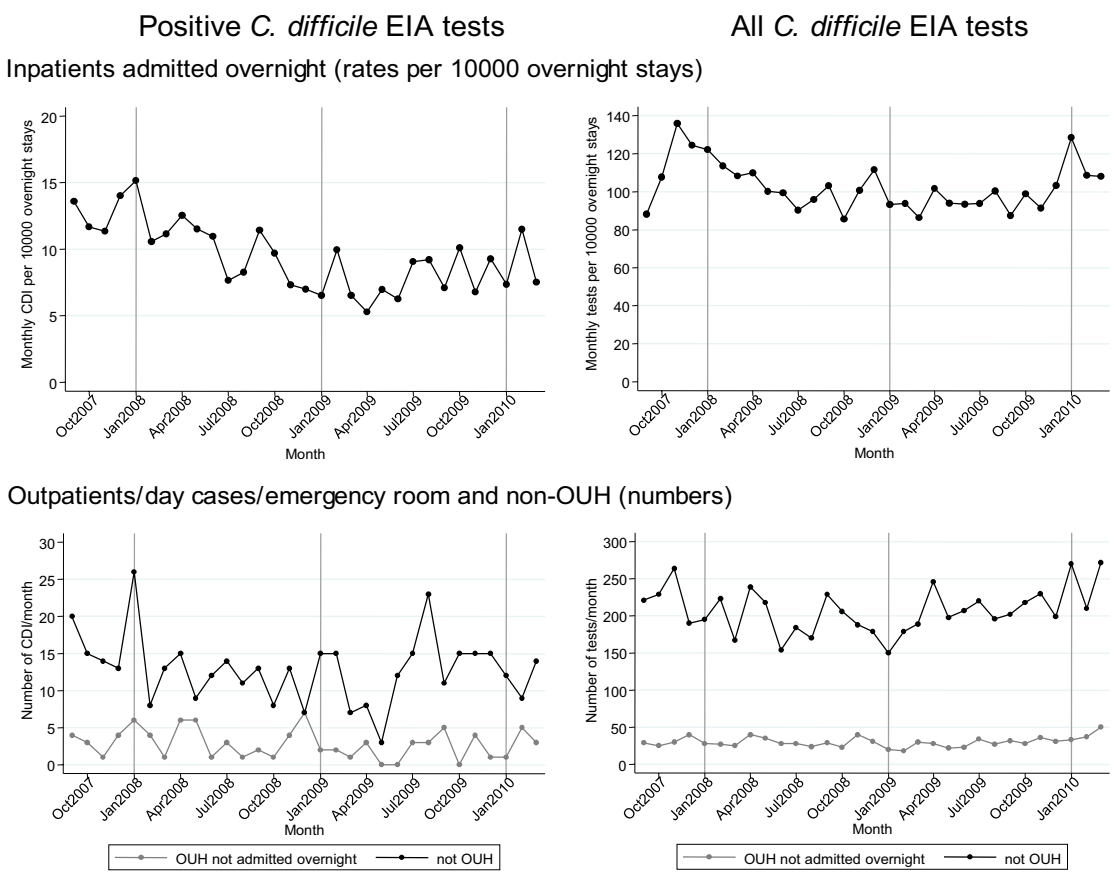
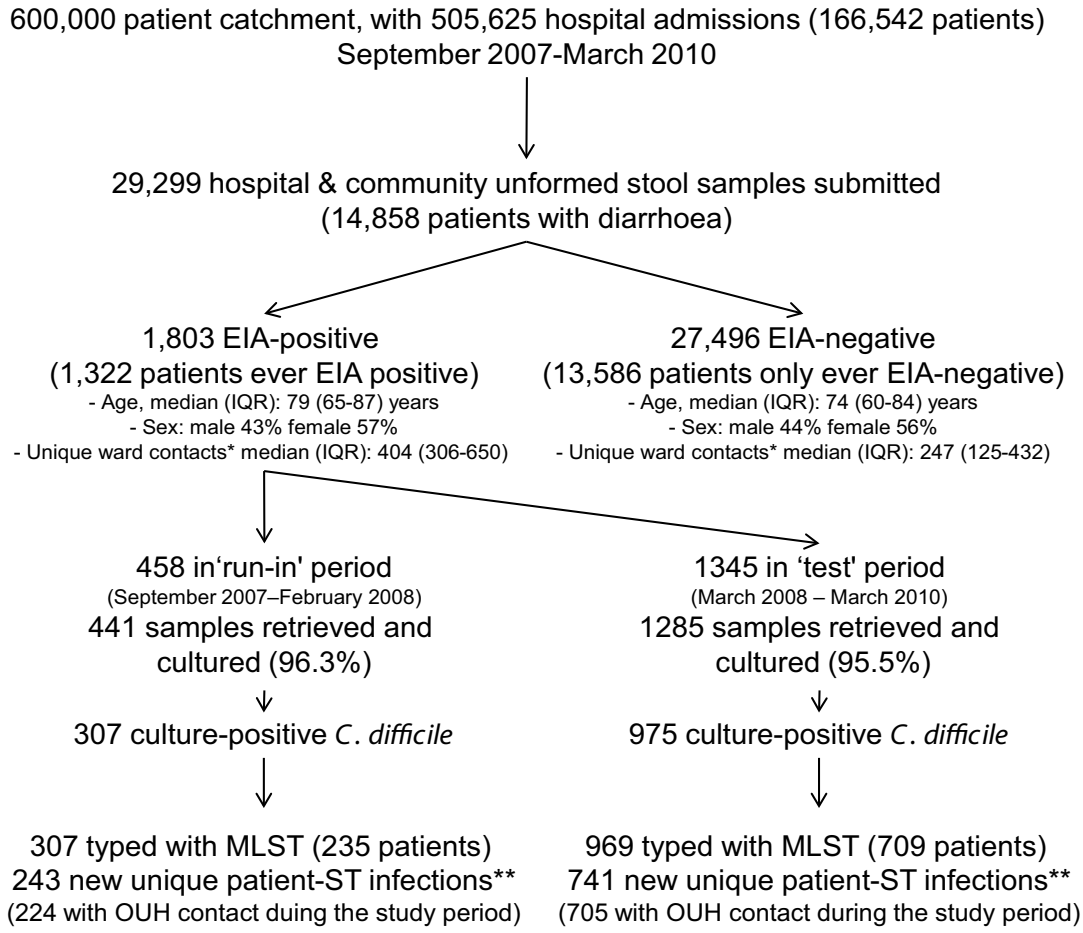


Figure 14. *C. difficile* tests and EIA-positive tests over the study period. The top two panels give rates in inpatients per 10000 overnight stays. Numbers of cases and tests in outpatients are given in the bottom two panels.

Of 1803 EIA positive samples, 1726 (96%) were retrieved and cultured. 1282 samples were culture-positive, of which 1276 (from 927 patients) were genotyped by MLST (**Figure 15**): 69 distinct STs were identified. 893 (70%) cases occurred in OUH inpatients, of which 456 (51%) were admitted under acute/elderly medicine, 152 (17%) general surgery, 68 (8%) renal/transplant, 52 (6%) haematology/oncology, 38 (4%) specialist surgery, 29 (3%) trauma/orthopaedics and the remaining 98 (11%) under other smaller specialities.



ST	Number of 'test' cases (N=741)	ST	Number of 'test' cases (N=741)	ST	Number of 'test' cases (N=741)
1	133	42	25	12, 45	8
2	76	11	20	35	7
8	69	7, 14	19	15	6
6	46	9, 37, 55	13	33, 63	5
44	37	13, 17	12	43, 53	4
5	31	16	11	4, 22, 41, 46	3
3	29	58	10	18, 48, 51, 56, 57, 65, 72, 75, 77	2
10	28	36, 49, 54	9	19, 24, 25, 26, 31, 34, 50, 52, 59, 60, 66, 67, 68, 70, 71, 73, 78, 90, 91, 92, 97	1

* the number of unique patients with whom the index patient shared a ward during the study period.
** 55 patients had multiple ST infections during the study period: 54 patients with 2 STs, and one with 4 STs. 6 isolates were not successfully typed (either due to ambiguous sequences despite repeated testing or loss of sample).
Note: 55 cases had no record of OUH inpatient admission during the study. Some patients had different ST infections in the test and run-in periods and are therefore counted in both.

Figure 15. Samples, patients and sequence types, 01 September 2007 to 31 March 2010.

C. difficile isolates were grouped into a 6 month run-in followed by a 2 year test period. **Figure 16** shows when isolates of each ST occurred over time, and suggests both temporal clustering and a large number of sporadic cases. **Figure 12** uses ST2 (the most common non-PCR-ribotype-027 ST) to illustrate the ward-based links that could be made between same-ST cases. 55 (6%) CDI cases had no OUH exposure during the study and were excluded from further analysis, since, by definition, no ward-based donor or onward transmission can be identified.

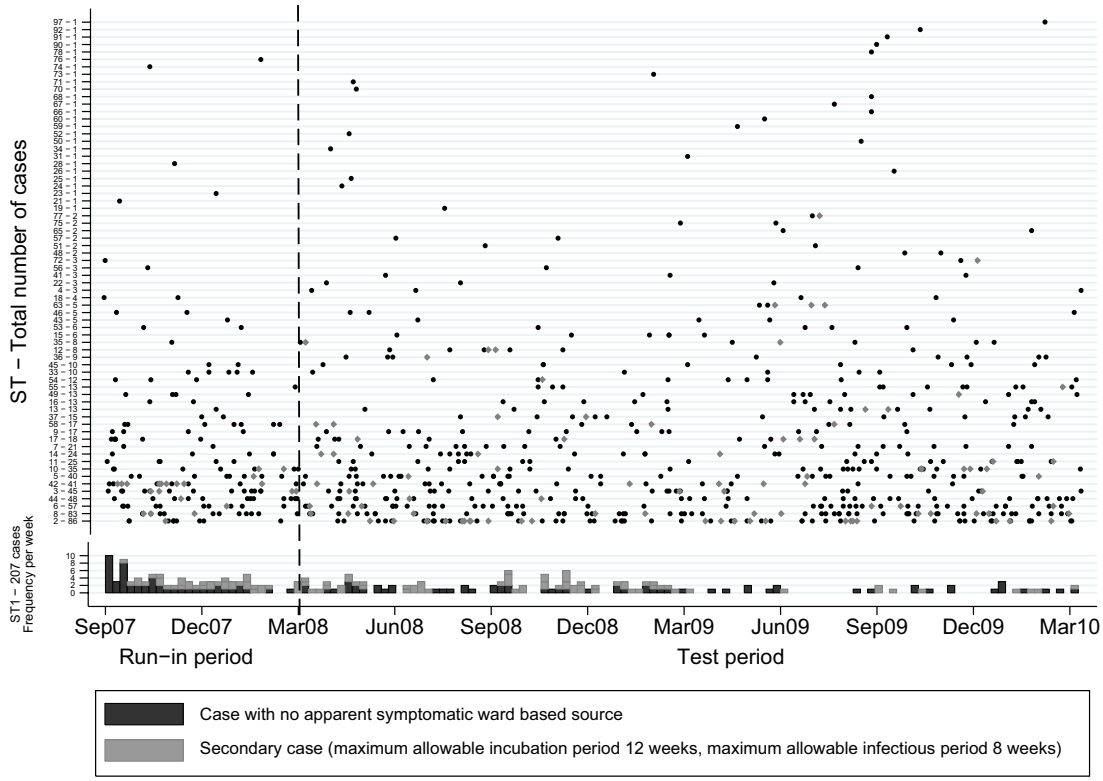


Figure 16. Cases by sequence type. The first EIA positive sample per sequence type per patient is plotted as a point for sequence types other than ST1 (ribotype-027) where 207 cases are instead shown as a histogram per week. Secondary cases are those with an identified credible donor based on incubation periods up to 12 weeks, and infectious periods up to 8 weeks.

3.3.2 INFECTIOUS/INCUBATION PERIODS

To include all links likely to represent true transmission, ward contacts up to 26 weeks from either case, longer than biologically expected,[29] were initially allowed. **Figure 17** shows the distribution of the minimum infectious/incubation periods from the 218 most plausible of the 362 potential transmission links where ward contact occurred between potential donor and recipient CDI (directional links). **Figure 18** shows all possible links, with similar results. With 26 week limits, of 705 test cases from 01 March 2008 to 31 March 2010, 408 (58%) were unlinked, 74 (10%) were donors of infection without an identified source for their own infection, and 223 (32%) had a credible ward-based donor sharing the same ST, assuming no ward contamination persisted after discharge (**Table 4**).

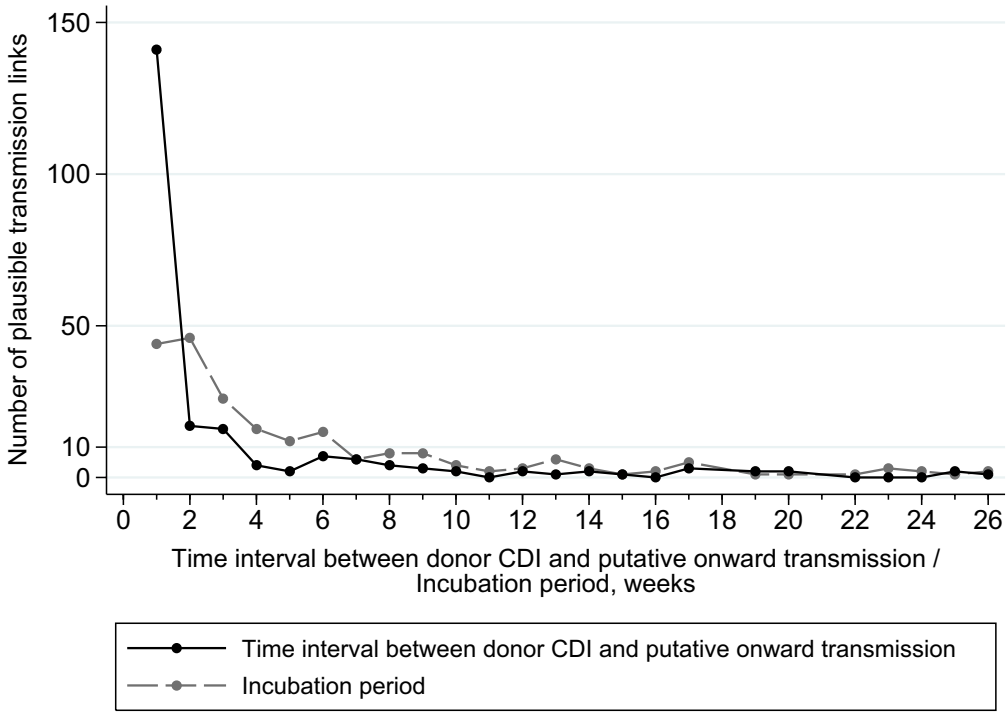


Figure 17. Distribution of infectious and incubation periods for transmissions within 69 sequence types and 705 'test' CDI cases, most plausible links. Only the most plausible transmission links where the direction of transmission is known (ward contact after potential donor CDI and before recipient CDI) are plotted, assuming maximum

allowable infectious and incubation periods of 26 weeks and no ward contamination persisting after donor discharge (see **Figure 18** for distribution from all potential links). Times are plotted rounded up to the nearest week, e.g. days 0 to 6 as week 1.

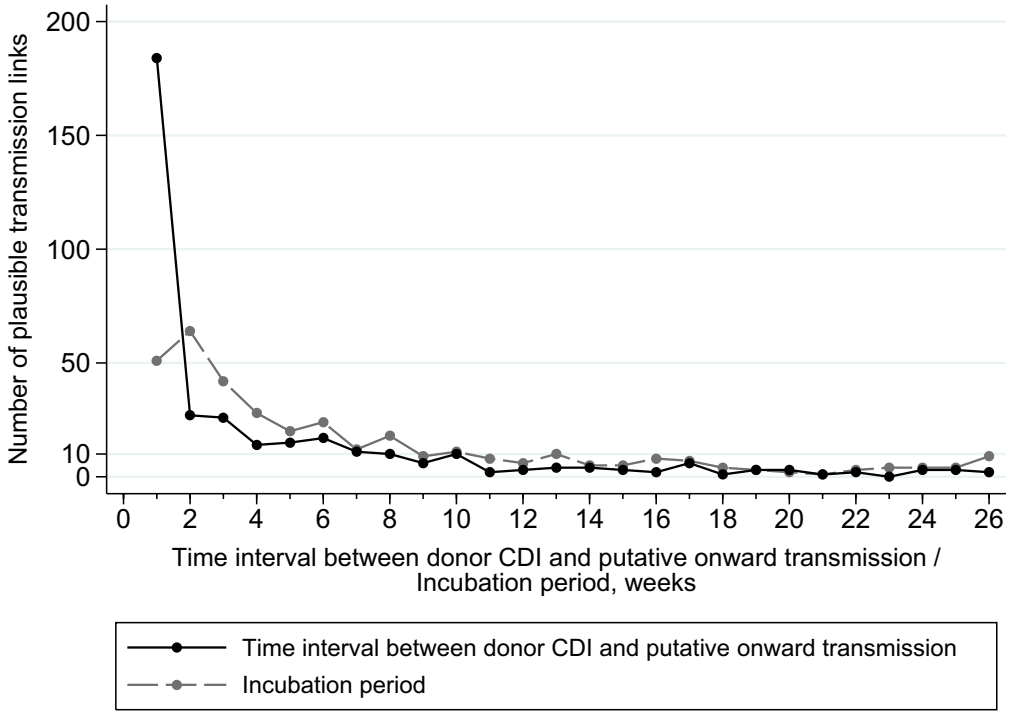


Figure 18. Distribution of infectious and incubation periods for transmissions within 69 sequence types and 705 'test' CDI cases, all links. All potential transmission links where the direction of transmission is known (ward contact after potential donor CDI and before recipient CDI) are plotted, assuming maximum allowable infectious and incubation periods of 26 weeks and no ward contamination persisting after donor discharge (see **Figure 17** for distributions from the most plausible links.) Times are plotted rounded up to the nearest week, e.g. days 0 to 6 as week 1.

	Maximum allowable infectious and incubation periods 26 weeks				Maximum allowable infectious and incubation periods 8 and 12 weeks respectively			
	Most plausible links		All possible links		Most plausible links		All links	
Cluster size	Cluster frequency	Test cases in clusters (%)	Cluster frequency	Test cases in clusters (%)	Cluster frequency	Test cases in clusters (%)	Cluster frequency	Test cases in clusters (%)
1	408	408 (58%)	403	403 (55%)	465	465 (66%)	463	463 (66%)
2	44	85 (12%)	37	71 (10%)	46	90 (13%)	44	86 (12%)
3	21	58 (8%)	16	43 (6%)	22	62 (9%)	19	53 (8%)
4	8	31 (4%)	8	29 (4%)	6	24 (3%)	4	16 (2%)
5	3	15 (2%)	2	10 (1%)	1	5 (1%)	2	10 (1%)
≥6	10	108 (15%)	8	149 (21%)	7	59 (8%)	2	77 (11%)
Cases with a credible donor		223/705 (32%)				165/705 (23%)		

Table 4. Network characteristics. Clusters are included where one or more cases in a cluster fell within the 'test' period (01 March 2008 – 31 March 2010). 'Test' cases may be present in a cluster as a donor, recipient or both.

Not all ward-based links between cases represent true transmissions, since two patients can meet by chance without a transmission occurring. Consistent with this, the number of ward-based links with longer infectious/incubation periods approached a constant background rate after 8–12 weeks. The distribution of links made between matched control patients without CDI (EIA negative patients with diarrhoea, sampled for each ST in the same numbers as cases) was used to assess 1) whether the constant background rate at longer infectious/incubation periods was also observed in control patients without disease and 2) the extent to which the distribution of infectious and incubation periods is conditional on the normal pattern of admissions, discharges, and ward movements. **Figure 19** shows that even in control patients more links are formed at short "infectious" and "incubation" periods, however there is a clear excess of links formed amongst the cases at shorter infectious/incubation periods (**Figure 20**).

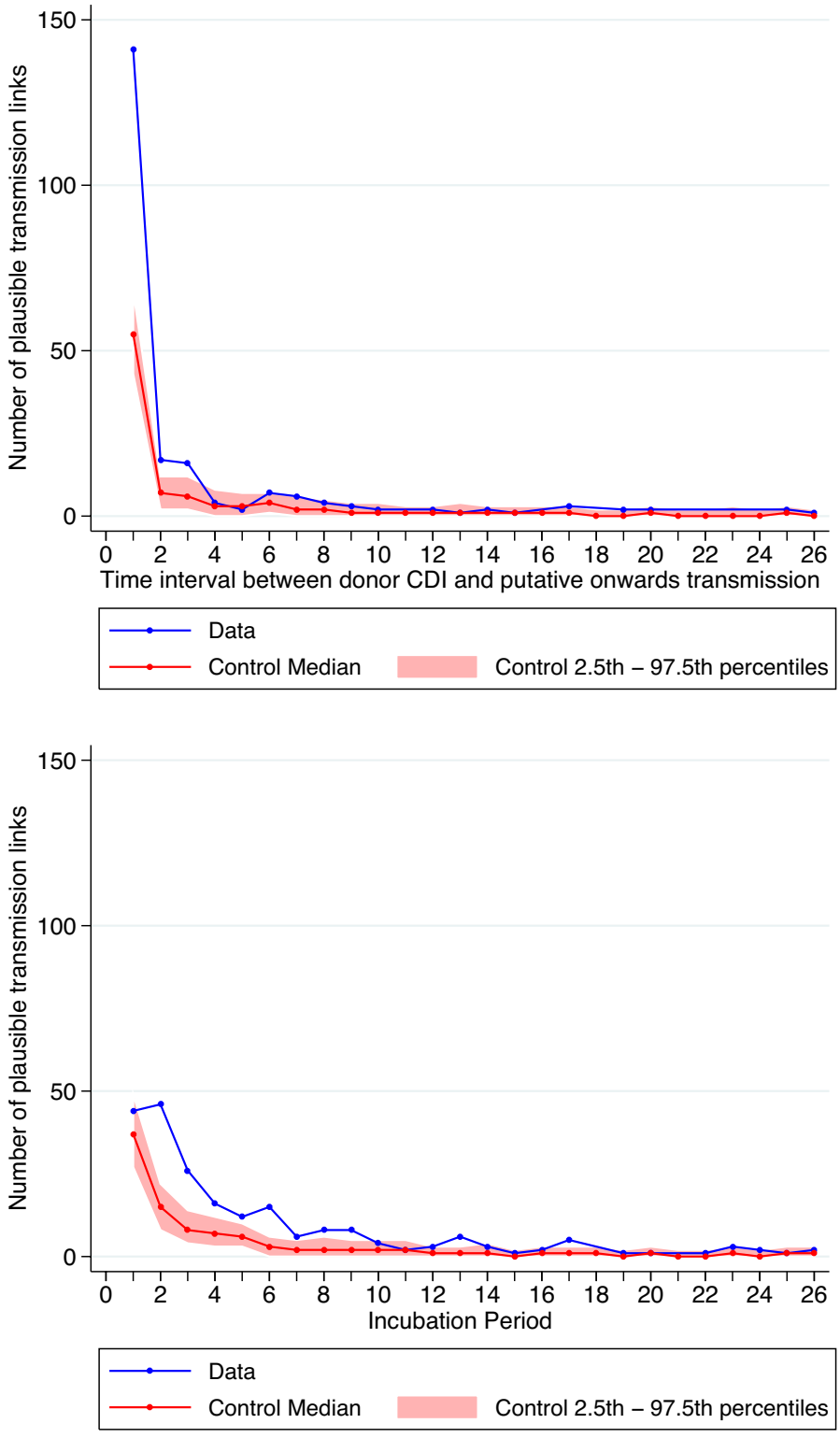


Figure 19. Distribution of infectious and incubation periods for transmissions within 69 sequence types and 705 'test' CDI cases, most plausible links, compared to the distribution of implied "infectious" and "incubation" periods for contacts between 705 matched control patients. Infectious periods are shown in the top panel, and incubation periods in the bottom panel. 1000 sets of control patients were generated with the median and 2.5 – 97.5th percentiles shown.

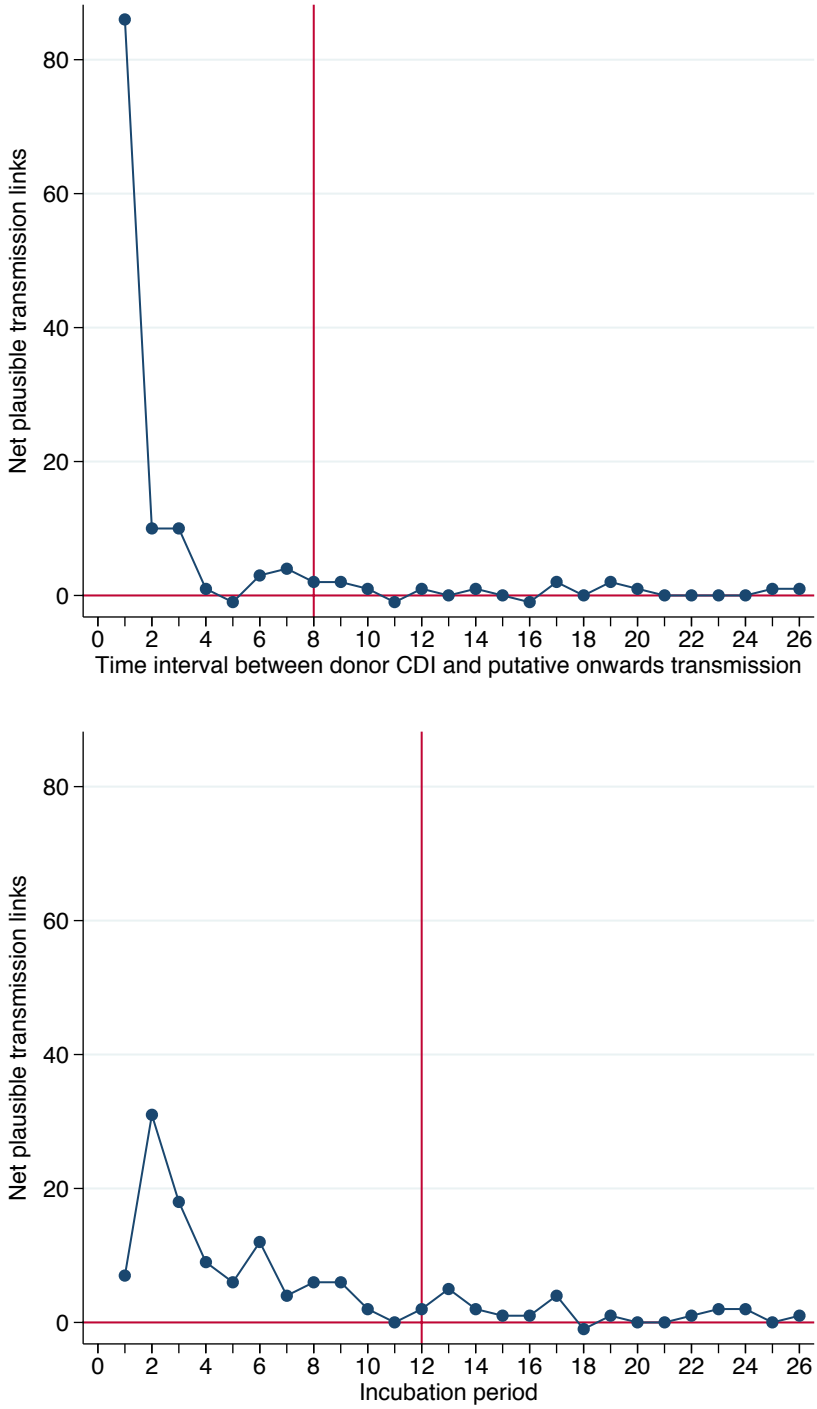


Figure 20. Net plausible transmission links. Distribution of infectious and incubation periods for transmissions within 69 sequence types and 705 'test' CDI cases, most plausible links, after subtraction of median links formed in matched controls (**Figure 19**). Infectious periods are shown in the top panel, and incubation periods in the bottom panel.

Most time intervals between potential donor CDI and putative onward transmission (minimum infectious period) were ≤1 week after the donor was diagnosed (141/218, 65%), with the majority ≤4 weeks (178/218, 82%), and few >8 weeks (21/218, 10%) (**Figure 17**, **Figure 20**). The median (interquartile range) minimum infectious period for the most plausible directional potential transmission link was 1 (0–14) day, compared with 4 (0–20) days and 8 (0–33) days for the most plausible and all potential links, respectively. The most common incubation period (from ward-based contact with donor to recipient CDI) was ≤4 weeks (132/218, 61%), with few >12 weeks (28/218, 13%) after the ward-based contact. The median (interquartile range) incubation periods for the most plausible directional potential transmission links, the most plausible links, and all potential links were 18 (8–42) days, 24 (10–61) days, and 33 (13–74) days, respectively.

3.3.3 Proportion of cases with a credible donor

Figure 21 shows the proportion of CDI cases with a credible ward-based donor from final networks based on a 12 week maximum incubation period with no ward contamination persisting beyond discharge. With an 8 week maximum infectious period, 465 (66%) of the 705 test cases were unlinked, and 165 (23%) had a credible ward-based donor sharing the same ST (**Table 4**). The percentage of cases with a credible donor was highest in renal/transplant (37%), haematology/oncology (29%), and acute/elderly medicine (28%), with fewer linked cases in general surgery (20%), trauma/orthopaedics (16%), and other medical (13%) and surgical (6%) specialities. Increasing the maximum allowable infectious period to 12 week increased the proportion with credible ward-based

donors to ~25%, whilst decreasing the maximum allowable incubation period to 4 week decreased it to ~17%. The proportion of EIA-negative controls, matched for exposure to other hospital patients, with ward-based links to other controls using the same algorithm was ~10% (**Figure 21**). These control data provide an estimate of the proportion of links among the CDI cases that were likely to have arisen by chance meetings between patients as they moved around the hospitals, and suggest that a substantial minority of the possible transmission links seen in cases could actually be chance admissions to the same ward. After adjustment for chance meetings, a net 16% of cases are linked by probable transmission events. Varying other matching criteria, including diarrhoea-free rather than EIA-negative controls, produced results similar to those in **Figure 21**.

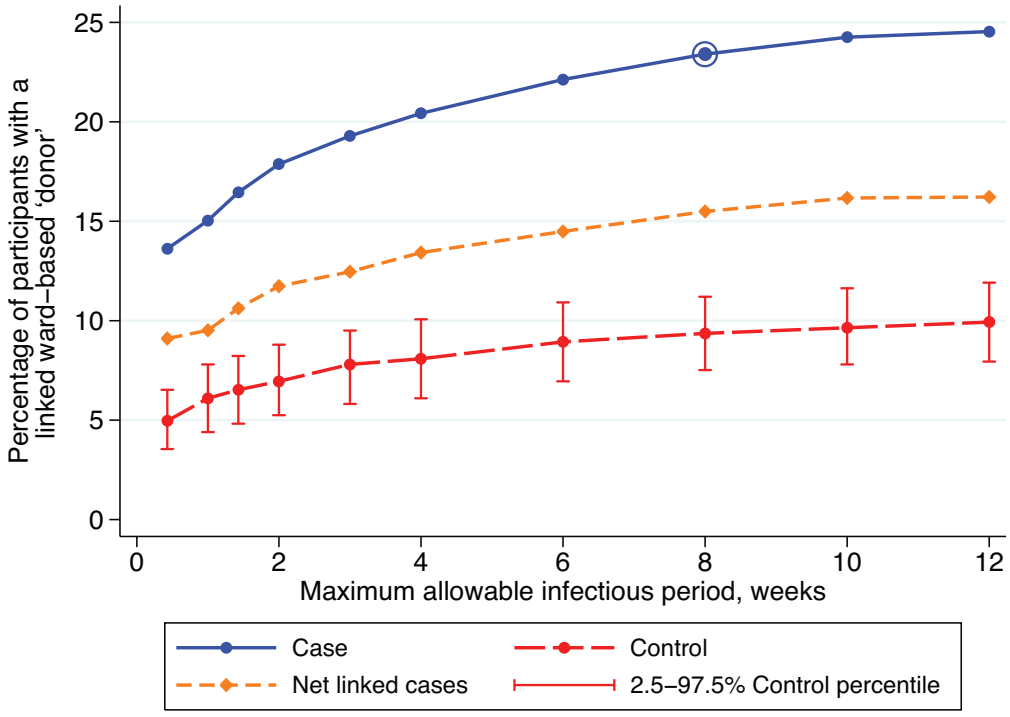


Figure 21. Percentage of cases explained by ward-based contact with a *C. difficile* toxin EIA-positive donor. The circled dot indicates a potential upper bound for the proportion of linked CDI cases. Controls were drawn at random 1000 times from hospital-exposure-matched EIA negative patients; linked ward-based contacts for controls

therefore represent the proportion of ward-based contacts that would likely arise in the CDI cases merely from chance movements around the hospitals. Net linked cases correct the overall total linked CDI cases for these chance ward meetings estimated from controls. Maximum allowable incubation period of 12 weeks, no ward contamination.

3.3.4 WARD CONTAMINATION BEYOND DISCHARGE

Figure 22 shows the proportion of CDI cases with a credible ward-based donor based on an 8 week maximum allowable infectious period and a 12 week maximum allowable incubation period, allowing for transmission from ward contamination after donor discharge up to a maximum of 26 weeks post discharge. The numbers of links arising from putative ward contamination increase similarly in CDI cases and controls without CDI, giving few additional net linked cases due to ward contamination after CDI case discharge. This suggests that such links are more likely due to chance admissions of cases with the same ST to the same ward rather than transmission from persistent ward contamination.

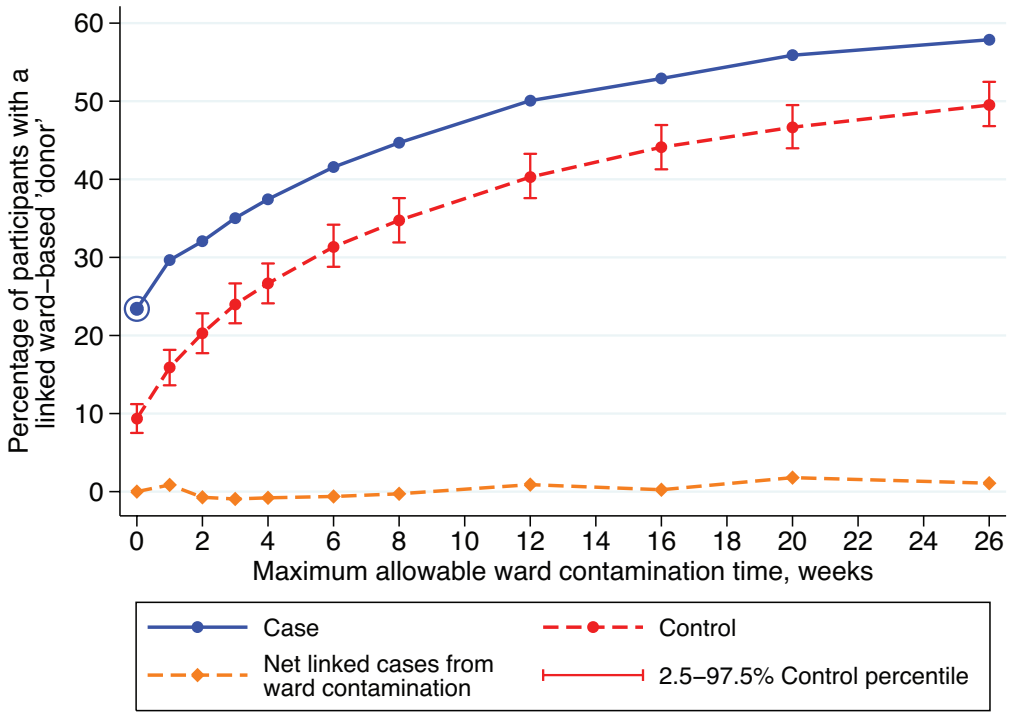


Figure 22. Percentage of cases explained by ward-based contact with a *C. difficile* toxin EIA-positive donor, allowing ward contamination. The circled dot indicates a potential upper bound for the proportion of linked CDI cases excluding ward contamination. Controls were drawn at random 1000 times from hospital-exposure-matched EIA negative patients; linked ward-based contacts for controls there represent the proportion of ward-based contacts that would likely arise in the CDI cases merely from chance movements around the hospitals. Net linked cases correct the overall total linked CDI cases for these chance ward meetings estimated from controls. Maximum allowable infectious period of 8 weeks and incubation period of 12 weeks.

3.3.5 Sensitivity analyses

Several important assumptions were tested in sensitivity analyses. To test the impact of the assumption that transmission between donor and recipient occurred at the start of any contact between the patients following the date of diagnosis in the donor (i.e. that patients are most infectious immediately following diagnosis) the analysis was repeated stratified by the duration of overlap between patients (**Figure 23**, **Figure 24**). The distributions of infectious and incubation periods were similar when restricting links to those where the point of potential

transmission was known more precisely (0-1 day ward contact durations only), and also similar across the other terciles of overlap duration, 2-6 days, and ≥7 days (**Figure 23**, **Figure 24**).

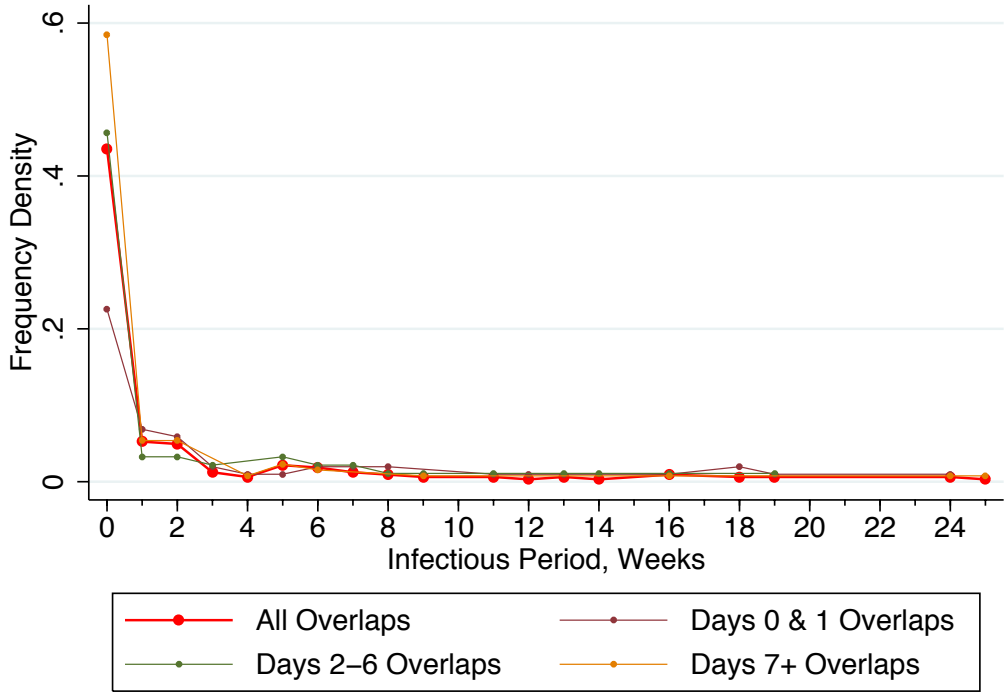


Figure 23. Distribution of infectious periods for transmissions, most plausible links, stratified by terciles of duration of overlap.

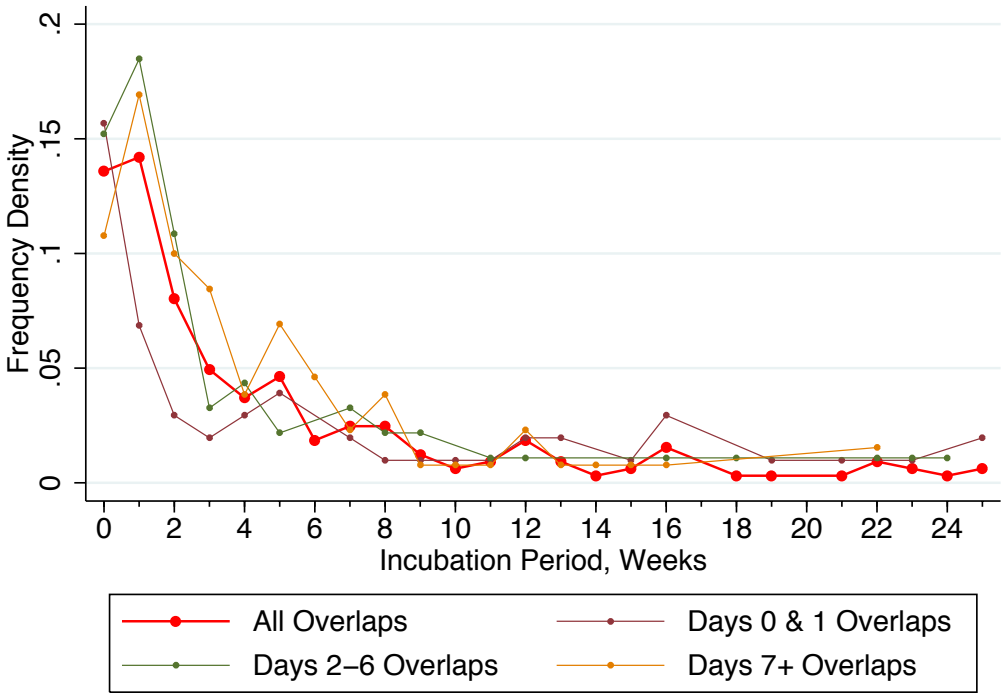


Figure 24. Distribution of incubation periods for transmissions, most plausible links, stratified by terciles of duration of overlap.

The distributions of infectious and incubation periods were also similar when incubation periods rather than infectious periods were minimised (**Figure 25**). As expected this increased the numbers of short incubation periods, but the upper bounds of plausible incubation and infectious periods remained similar. Results were also similar when calculating infectious/incubation periods relative to the most recent, rather than the first, donor sample.

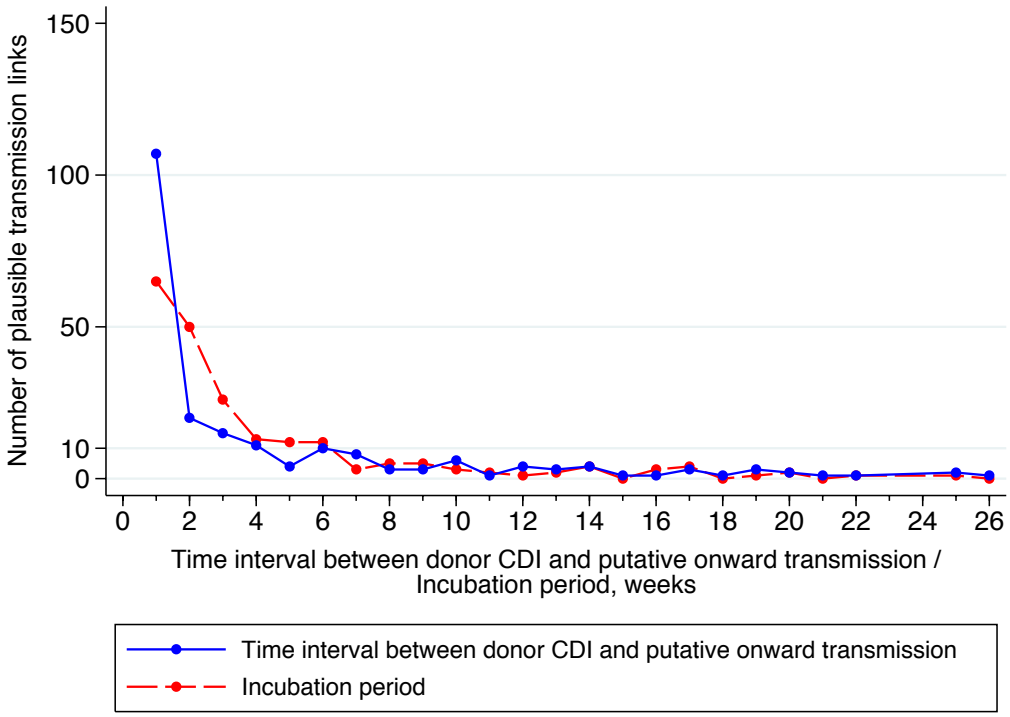


Figure 25. Distribution of infectious and incubation periods for transmissions within 69 sequence types and 705 'test' CDI cases, minimising incubation periods, most plausible links shown.

3.3.6 MISSING DATA SIMULATION

Investigation of the impact of missing data showed that, under plausible upper limits on the maximum allowable infectious/incubation periods as above, the proportion of the remaining cases with a credible donor fell from 23% to 22% if 10% of cases were randomly removed, and to 18% if 30% of cases were removed (**Figure 26**). Extrapolation suggests that had the 4% of EIA-positive samples not retrieved for culture (see **Figure 15**) actually had STs obtained, the proportion of cases with a credible donor might have been ~25%, rather than ~23%. Even if 30% of cases were missed only ~30% of cases are likely to have been related by ward based transmission.

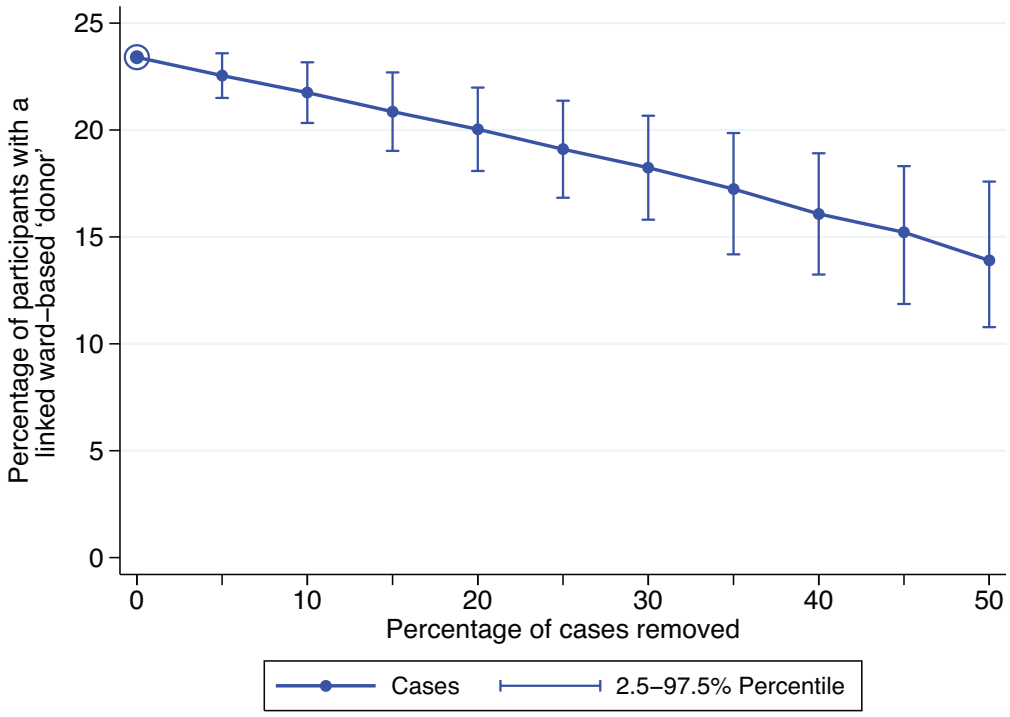


Figure 26. Missing data analysis. Maximum allowable infectious period of 8 weeks and incubation period of 12 week. To estimate the effect of missing cases, where samples were not retrieved for culture or were not detected as part of routine testing, the analysis was repeated after randomly removing cases from the complete dataset. Networks were regenerated and the proportion of the remaining cases with a credible ward-based donor calculated. Between 5% and 50% of the original cases were removed, with 1000 repetitions undertaken for each 5% increment.

3.4 Discussion

This uniquely large dataset of genotyped *C. difficile* isolates, ~20% of which were the hypervirulent PCR-ribotype-027/NAP-1/ST1 strain, was obtained from patients in defined ward locations, in a group of hospitals from a large UK NHS Trust practising infection control in line with published guidance.[4] The study investigates for the first time for symptomatic *C. difficile* EIA-positive, culture-positive cases, (i) the proportion of cases attributable to credible ward-based transmission from other symptomatic cases, (ii) the time between CDI diagnosis and potential onward transmission (denoted the minimum infectious period) and

between putative acquisition and CDI diagnosis (incubation period) based on these putative transmissions, and (iii) the possible impact on transmission of ward contamination persisting following discharge of CDI cases. Crucially, the findings are based on data collected over 2.5 years, and are thus less susceptible to biases inherent during CDI outbreaks. The CDI rates during the study were within the 3.8–9.5/10000 overnight stays typically observed in endemic settings,[29] the local infection control policy was based on widely available guidance,[4] and the hospitals studied contain both specialist and general hospitals typical of those found in the UK. Therefore these findings should be generalisable to other endemic settings.

Plausible ward-based transmissions were most commonly observed during the week following the first EIA-positive sample. This supports the current recommendations for immediate isolation of symptomatic patients, and is consistent with evidence showing that *C. difficile* shedding is most prominent during episodes of diarrhoea.[86,93] Current guidance suggests isolation should continue until 48 hours after diarrhoea resolution;[4,100] these data show that the potential for transmission persisted for up to 8 weeks. There was no evidence of substantial onward transmission from wards after patient discharge; whilst this clearly can occur, these data suggest the straightforward enhanced cleaning protocols used were sufficient to prevent most environment-related transmission occurring after CDI cases left wards. This study does not distinguish between direct person-to-person spread and indirect (spore-based) transmission via the environment during shared ward exposure, classing both as “ward contact”.

The data suggest that incubation periods most commonly last a few days to 4 weeks, but extending these up to 12 weeks can provide plausible ward-based transmission links. Previous studies have been inconclusive. Cohort studies taking serial samples from inpatients (likely receiving antibiotics) suggest a median 2 – 3 days from first positive test to symptoms;[79,86,88] epidemiological investigations showing that hospital exposure over the last 60 days is a risk factor for community-onset disease suggest that incubation periods may be considerably longer.[105] Providing exact limits is complicated by the additional dependence of symptom onset on antibiotic use, data not collected electronically in the OUH hospitals and therefore not available for this study. As well as informing outbreak investigation, this study broadly supports the current surveillance definitions in recent Society for Healthcare Epidemiology of America / Infectious Diseases Society of America guidelines,[29] which define cases occurring 0 – 4 weeks post-discharge as "healthcare associated" and 4 – 12 weeks post-discharge as having a possible healthcare facility origin.

In this study no more than ~25% of patients with detected symptomatic CDI could plausibly have acquired the infection from other patients with EIA-positive symptomatic CDI via ward contact. Chance meetings between patients, without transmission occurring, could mean that this 25% is in itself an overestimate. Similarly, in a study of ten nosocomial pathogens, not including *C. difficile*, in an intensive care unit setting, investigators found that only 15% of infections could be attributed to ward-based patient-to-patient spread based on molecular typing.[188] As expected, the proportion of CDI cases with identifiable donors varied according to location/speciality, likely because of differences in risk of CDI

acquisition by susceptible patients, such as those on renal wards. The fact that patients on OUH elderly medical wards are nursed in single rooms probably lowered the proportion of linked donors and recipients somewhat compared to 4-bedded bay wards. The ~25% of cases linked via ward contact represent a major hospital-acquired infection problem; however, the remaining ~75% of unexplained transmission raises concern about other acquisition routes not captured by this study. These could include transmission from patients with falsely EIA-negative symptomatic CDI,[187] asymptomatic carriers[107] (e.g., patients, relatives, and staff), or significant acquisition within the community (e.g., from food[189] or animals[190]) with importation into the hospitals.

Since (i) 800–1000 EIA tests were performed each month, (ii) there was an OUH policy of universal testing of nosocomial diarrhoea (**Table 3**), (iii) the same laboratory also processed all community samples, and (iv) all unformed stool samples from patients >65 year old were routinely tested for *C. difficile*, large numbers of completely untested inpatient or community symptomatic patients are unlikely. The OUH 6.2% *C. difficile* EIA positivity rate in the study is similar to the average 6.45% for 170 English hospitals in 2008,[191] suggesting the OUH hospitals were unlikely to be identifying many fewer EIA-positive cases than other typical UK hospitals. The particular EIA assay used in the OUH laboratory has an independently reported sensitivity of 91.7% (84.7%–96.1%):[135] this relatively low sensitivity (common to all toxin EIAs[187,192]) makes it possible that ~10% of cases could have been missed because of false-negative EIAs. However, because of widespread concerns about test sensitivity, for most of the study period (to 31 December 2009), multiple diarrhoeal samples were submitted from each patient,

either simultaneously or serially, reducing the chance of completely missing patients with symptomatic CDI. Mixed CDIs could be another reason for missing cases, although it is plausible that the most prevalent cultured strain that was sequence typed would also be the one most likely to be transmitted. Relatively low rates of mixed infection (2%–13%) have been identified in previous investigations, most with limited sample sizes, that examined multiple *C. difficile* colonies from single faecal samples.[193-197] Further investigation of mixed infection is detailed in chapter 4. Extrapolation from **Figure 26** suggests that if the transmission behaviour of ~10%–20% missed cases were similar to that of the observed cases, the proportion of linked cases might be a few percentage points higher than 25%. In a more extreme situation, where 50% of cases had been missed, the proportion of linked cases would still be <40%. Unidentified symptomatic cases might account for more transmission than observed cases if they were more infectious, e.g., because of lack of isolation. However, multiple transmissions over a short period of time from an unidentified donor would result in ward-based clusters of cases, which within the analysis would result in recipients being falsely assigned as transmitting to each other. While the true donor would remain unidentified, the analysis would nevertheless classify any such transmissions as ward-based and thus would capture a large proportion of such outbreaks in the linked cases already reported. The high numbers of unlinked cases thus suggests that increased transmission from unidentified symptomatic cases is likely to be uncommon.

Asymptomatic carriage of *C. difficile*, whilst possibly protective against subsequent CDI in the carrier,[77] may still be a source for transmission.[107] If there were undetected ward-based transmission from asymptomatic carriers, the results

would underestimate the proportion of patients that had acquired CDI as inpatients, similarly to the analysis removing cases at random. However, as above, multiple transmissions from an unidentified asymptomatic donor would likely result in the false attribution of direct ward-based contacts between recipients, and the high numbers of unlinked cases argues against a significant proportion of infection arising from a few highly infectious unidentified carriers. However, as between 7% and 26% of adult inpatients may be asymptomatic carriers of *C. difficile*,[29] even if these patients infect relatively few patients each, asymptomatic carriage may be an important route of transmission, and requires further investigation.

At the start of this study, and in common with the wider clinical community,[29] it was expected that the majority of CDI cases would be attributable to in-hospital transmission from other known symptomatic cases. Whilst the limitations discussed above imply that total in-hospital transmission could be greater than in-hospital transmission from this group, nevertheless, the findings suggest that, contrary to prevailing beliefs, ward-based transmission from known EIA-positive symptomatic cases cannot account for the majority of new CDIs. These are the only cases currently visible to infection control teams, and therefore are the only cases directly targeted by current infection control policies and practice. Improved tests, such as those based on nucleic acid/PCR-based methods, might improve detection rates (although at the expense of specificity),[136] and it should be noted that even infections arising from an unknown source are to a certain extent preventable by measures such as good antimicrobial stewardship.[4]

The fact that 1,276 isolates could be divided into 69 distinct STs in the proportions given in **Figure 15**, with the number of cases of each ST small enough to analyse as separate lineages, suggests that MLST is reasonably discriminatory for identifying transmission. The study was not powered to analyse the transmission properties of each ST independently; therefore, more data are required to comment on the extent to which the findings may have varied with a different mix of prevalent strains.

However, whole genome sequence data offers a significant improvement over MLST in the genetic resolution available to distinguish isolates. For example, providing rates of microevolution within hosts have been estimated over short timescales (chapter 4), it can refute the possibility of transmission between epidemiologically linked cases with the same ST that differ substantially at the whole genome level. Although using MLST offers a significant advantage over no genetic data, multiple chance meetings between patients sharing the same ST still occurred during the study. Therefore while the heuristic approach taken in this chapter can approximate findings over all transmissions, identifying the precise individuals involved in individual transmissions is often not possible. Whole genome data has the potential to reduce the probability that chance ward meetings of patients are misattributed as transmission events based on shared genotypes and therefore to improve the precision of transmission analyses. Additionally, genetically identical but apparently epidemiologically unlinked cases may be identified, suggesting other routes of transmission. The following chapters firstly establish the necessary foundation for applying whole genome data to

transmission, and then apply the technology to an extended version of the dataset used in this study.

In summary, in this endemic setting with well-implemented infection control measures, up to three-quarters of new CDIs are not easily explained by conventional assumptions of ward-based transmission from symptomatic patients and so may not be targeted by current interventions. A better understanding of other routes of transmission and reservoirs is needed to determine what other types of control interventions are required to reduce the spread of *C. difficile.*

4 PREREQUISITES FOR APPLYING WHOLE GENOME DATA

4.1 INTRODUCTION

Whole genome sequencing (WGS) of *C. difficile* offers much greater genetic resolution than has previously been available; for example compare the 4.3Mb whole genome[182] to the 3.5kb sequenced in multilocus sequence typing.[145] WGS therefore provides the prospect of high precision investigation of outbreaks[198] and more detailed understanding of endemic transmission. Close genetic relationships between organisms isolated from different infected cases suggest transmission can be considered a strong possibility, whereas transmission between cases with genetically distinct isolates can be excluded reliably.

However, before such data can be used in transmission studies, its properties must be understood, in particular the error profile of the data and how reproducible it is, and the extent of within host sequence diversity and how it evolves over time. In order for WGS to be useful the amount of variation arising from error and within-host diversity/evolution must be small relative to the background diversity observed in the population of interest. These are addressed in this chapter before considering the challenge to transmission analyses posed by mixed infection, i.e. infection with two or more *C. difficile* strains in the same host.

4.2 WHOLE GENOME SEQUENCING METHODS

4.2.1 SAMPLE PREPARATION AND SEQUENCING

DNA for whole genome sequencing was obtained from the same collection of *C. difficile* clinical isolates described in chapter 3, with an expanded collection of

samples to March 2011. Each previously frozen isolate was inoculated on to a Columbia blood agar plate and incubated anaerobically for 48 hours at 37°C. A single colony was selected from the resulting growth and transferred to a 2nd Columbia blood agar plate and incubated anaerobically for a further 48 hours at 37°c. DNA was extracted using a commercial kit (FastDNA, MP Biomedicals, California, USA; QIAamp, Qiagen, Hilden, Germany; QuickGene, Fujifilm, Tokyo, Japan).

A combination of standard Illumina and adapted protocols was used to produce multiplexed paired-end libraries. Pools of 96 samples were sequenced at the Wellcome Trust Centre for Human Genetics, Oxford, UK, using sequencing-by-synthesis technology, on the Illumina Genome Analyzer II (GAII), GAIIx, and HiSeq2000 platforms, generating 51, 101-108, 100 base paired-end reads respectively.

4.2.2 SEQUENCE DATA PROCESSING

Sequence reads were analysed and assembled using a pipeline developed specifically for bacterial genomes by the Department of Statistics, University of Oxford, and refined as part of this thesis. Pipeline parameters were tuned using a collection of 15 *Staphylococcus aureus* and 51 *C. difficile* isolates sequenced more than once prior to November 2010. Parameter values were iteratively chosen to ensure no false differences were identified between replicate sequences after data processing. Potential false variants were identified and then alignments inspected to determine the most appropriate parameter or filtering adjustment.

Each isolate was mapped using Stampy v1.0.11 (without Burrows-Wheeler Aligner pre-mapping, using an expected substitution rate of 0.01)[179] to the *C. difficile* 630 reference genome (Genbank: AM180355.1), CD630,[182] with the exception of Clade 2 isolates[199] (mostly ST1/ribotype-027) which were mapped to CD196 (Genbank: NC_013315.1) to allow the additional novel sequence in these samples to be compared.[184] Preliminary investigation mapping sequence data to various reference sequences demonstrated that choosing a closer reference improved the sensitivity of variant detection, principally because variants could be detected in novel genetic material present only in the closer reference.

Single nucleotide variants (SNVs) were identified across all mapped non-repetitive core genome sites using SAMtools (version 0.1.12-10)[180] mpileup with the extended base-alignment quality flag, and requiring a minimum PHRED scaled ($-10\log_{10}$ of the error probability) base quality and mapping quality of 30 (with no cap on mapping quality). To identify repetitive regions, BLAST[200] searches of the CD630 and CD196 reference genomes were made using fragments of the same genome. Multiple BLAST matches in the reference genome were marked as repetitive and excluded from analyses, accounting for 9.6% of the CD630 genome and 6.0% of the CD196 genome. SNVs were then filtered using the following criteria: a consensus of ≥75% reads mapped to a site was required to support a SNV; calls were required to be homozygous under a diploid model; only SNVs supported by ≥5 reads, including one in each direction were accepted.

De novo assemblies were created for all genomes, principally to allow for assessment of the presence, absence, and type of particular genes. Assemblies were generated with Velvet run with the default settings and the Velvet Optimiser

(http://www.ebi.ac.uk/~zerbino/velvet/).[181] BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) [200] databases were generated for each sequence from the contigs created by Velvet. BLAST searches for genes of interest were conducted against these databases using an E-value threshold of 0.01 (i.e. 0.01 matches expected by chance). These searches were used to confirm the presence and absence of the *tcdA* and *tcdB* toxin genes, and to extract the 7 MLST genes to generate *in silico* STs. *In silico* STs were compared to the original ST obtained for each sample by PCR. Sequences with a mismatch between the PCR and *in silico* STs represent possible sampling handling errors and were therefore discarded and the isolates re-sequenced and the MLST PCR also repeated.

4.3 Reproducibility

To confirm the reproducible nature of the sequencing and validate the data processing pipeline, an expanded collection of *C. difficile* replicate sequences was developed. The CD630 reference genome was re-sequenced on all plates sent for sequencing. Fifty such technical replicates, obtained across all 3 sequencing platforms used, were all identical after data processing, with a mean read depth of 86, and a mean of 87% of the reference genome called. Interestingly 7 SNVs were consistently identified between our laboratory strain of CD630 and the original capillary sequenced version (**Table 5**).[182] These either represent errors in the original reference sequence or divergent evolution between our laboratory strain and that used in the original sequencing.

Site in CD630 reference genome	Capillary sequenced base[182]	Illumina sequenced base
103226	G	T
690659	A	T
2044515	C	G
2634898	C	T
2832893	G	T
2924656	C	T
3528665	G	T

Table 5. Location of SNVs identified in the local CD630 reference across all Illumina sequences.

Additionally, 85 clinical samples were sequenced multiple times on the same and different sequencing plates. A total of 180 additional sequences were generated such that 38 samples were sequenced twice, 41 three times, and the remainder 4-22 times. A total of 690 pairs of replicates were analysed in an all-against-all comparison comparing each replicate to all other sequences from the same isolate. Three pairs differed by 3 SNVs, 4 pairs by 2 SNVs and 6 pairs by 1 SNV. However a single sequence was involved in all pairs ≥2 SNVs apart, and 2 of the 1 SNV pairs. To avoid the potential for this multiple counting, each replicate was compared only to the most recent sequence obtained from the same isolate. In this analysis only 2/180 replicate sequences differed, each by only 1 SNV (1 SNV error per 90 genomes compared).

4.4 WITHIN HOST DIVERSITY AND EVOLUTION

The high specificity of the data pipeline used for identifying SNVs is critical as differences of only a few SNVs are important in distinguishing potential transmission. The number of SNVs typically observed between sequences obtained from two CDI cases related by transmission is difficult to observe directly, given

the endemic nature of *C. difficile* and therefore the difficulty identifying individual transmission events with certainty. Instead, patients with on-going or recurrent CDI infection with the same strain were used to develop estimates of within host diversity and microevolution that could then be applied to transmission between cases. Supporting data for these estimates was also provided by two *in vitro* experiments. Firstly multiple colonies sequenced from the same patient stool sample at multiple timepoints were used to support estimates of within host diversity. Secondly, isolates were sequenced from a laboratory gut model simulating CDI induction and treatment to assess rates of evolution, including under selective antibiotic pressure.

4.4.1 SERIAL PATIENT SAMPLES

Within the Oxfordshire CDI cohort described above, a number of patients were tested on ≥2 occasions during the study and found to be infected with the same *C. difficile* MLST sequence type, i.e. likely on-going infections or relapses of the same infection.[197] Isolates from the first and last faecal sample received from 148 of these patients, taken a median(IQR)[range] 51(28-105)[0–561] days apart, were sequenced. The mean number of reads mapped to each site was 83, with a mean of 83% of the reference genome called.

Three patients were excluded with samples separated by 60, 60, and 374 days and isolates separated by 20, 823, and 20 SNVs, respectively. Each of these was considered more likely to represent a reinfection, as there was no evidence of a single recombination; furthermore, on both occasions where the isolates were separated by 20 SNVs, a sequence matching the later sample had previously been found in another patient.

Several different models were used to analyse the 145 remaining pairs of samples, with the final model selected on the basis of its goodness of fit. An important consideration is whether transmission represents an evolutionary bottleneck, i.e. is initial infection established by a single clone, such that the within host diversity shortly after acquisition is minimal (**Figure 27**, panel A). If transmission is an evolutionary bottleneck, and most CDI cases become symptomatic shortly after *C. difficile* acquisition (a reasonable assumption from the data presented in chapter 3) and are tested rapidly, the number of SNVs between first and last samples can be modelled such that no SNVs are expected between two samples taken at diagnosis. However, if it is possible for more than one clone to be transmitted and become established then diversity present within a case at diagnosis must be accounted for in the model (**Figure 27**, panel B). Both models are fitted below.

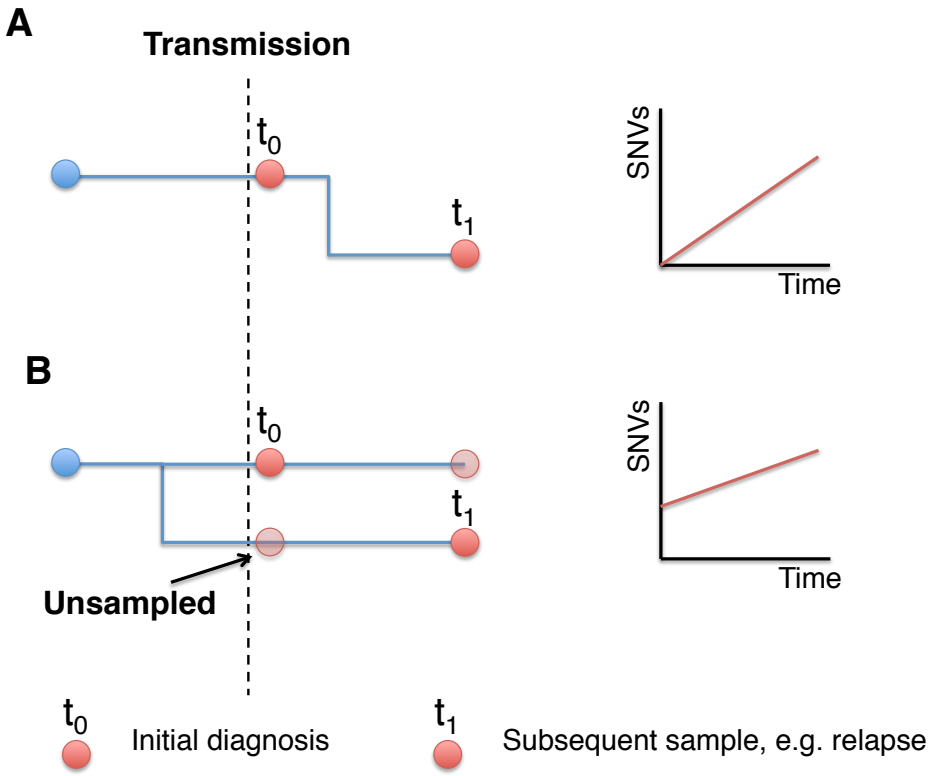


Figure 27. Models of *C. difficile* within host diversity and evolution. Panel A illustrates a bottleneck model, where infection is initiated by a single clone, such that the diversity at initial diagnosis is minimal. Within host diversity and evolution can be modelled in this scenario as starting at zero at time zero. Panel B shows the alternative where the diversity present in the source of transmission is transferred to the new case. Therefore diversity may exist at initial diagnosis, and this must be accounted for when modelling the number of SNVs between first and subsequent samples.

4.4.1.1 Poisson regression model

From the discussion above two Poisson regression models can be fitted. If μ is the rate of microevolution per unit time, in the bottleneck scenario the number of SNVs, s, at a given time since the baseline sample, t, can be assumed to follow a Poisson distribution:

$$s \sim Pois(\mu t)$$

In the model with baseline diversity, this can be modelled as θ, such that:

$$s \sim Pois(\mu t + \theta)$$

4.4.1.2 Coalescent-based model

Coalescent theory can be applied to the model of baseline diversity above.[201] The amount of diversity observed between a pair of samples can be thought of as the sum of 2 Poisson processes, firstly the variation that arose in the time, *t*, between the samples being taken, and secondly the variation that has arose in the time, *u*, between the most recent common ancestor of the two samples and time t_0 (**Figure 28**).

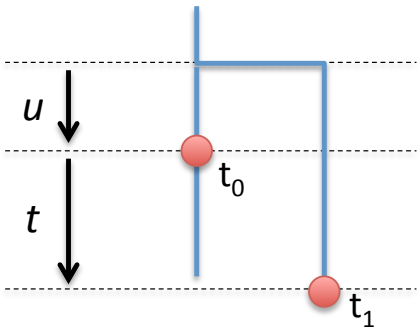


Figure 28. Coalescent-based model. t_0 denotes the time of the first sample, t_1 the time of the second sample. *t* is the time between the samples, and *u* the time between the first sample and the common ancestor of the samples.

The sum of the two processes is itself a Poisson process, however the value of *u* is unknown. Given a constant rate of evolution, μ, and for a given value of *u*, the number of variants between the samples, *s*, is given by:

$$s|u \sim Pois(\mu t) + Pois(\mu 2u)$$

Which can be written:

$$s|u \sim Pois(\mu t + \mu 2u)$$

Under coalescent theory, with neutral selection and a fixed population size, u is exponentially distributed with mean N_e, the effective population size:

$$u \sim Exp(1/N_e)$$

The value of N_e and μ can be jointly estimated by maximum likelihood. To do this the likelihood of N_e and μ given the data is defined as follows. Firstly the probability mass function for $s|u$ can be written:

$$\Pr(s|u) = \frac{e^{-\mu(t+2u)}(\mu(t+2u))^s}{s!}$$

The probability of s, unconditional on u, i.e. the likelihood of μ and N_e for a single observation, can be expressed by integrating over all possible values of u:

$$\Pr(s) = \int_0^\infty \Pr(s|u)\frac{1}{N_e}e^{-\frac{1}{N_e}u}\mathrm{du}$$

$$\Pr(s) = \frac{e^{-\mu t}}{N_e s!}\int_0^\infty (\mu(t+2u))^s e^{-u(2\mu+\frac{1}{N_e})}\,\mathrm{du}$$

Maple (http://www.maplesoft.com/products/maple/) was used to solve the integral. The data can be thought of as n observed pairs of samples separated by time t_i and with SNVs s_i between them. Therefore to calculate the likelihood of μ and N_e given the data, the product of the $\Pr(s_i)$ across all the n pairs of data was calculated:

$$\mathrm{L}(\mu, Ne \mid \text{data}) = \prod_{i=1}^{i=n} \Pr(s_i)$$

Maximum likelihood values of μ and N_e were found by numerical optimisation.

4.4.1.3 Model comparisons

Each model was fitted by maximum likelihood using R (http://www.r-project.org). 95% confidence intervals for parameter estimates were calculated by parametric bootstrap, using 1000 iterations. Model fit was compared using Akaike information criterion, AIC.[202] Parameter estimates and model fits are given in **Table 6**.

Model	Rate of microevolution, μ SNVs / called genome / year (95%CI)	Within host diversity, $2\mu N_e$ or θ SNVs / called genome (95% CI)	Model $\log_{10}$ likelihood	Akaike information criterion
Poisson regression, bottleneck model	2.08 (1.58-2.60)		-183.5	368.9
Poisson regression, baseline diversity model	0.86 (0.23-1.50)	0.26 (0.13-0.39)	-125.1	254.2
Coalescent-based model	0.74 (0.22-1.40)	0.30 (0.13-0.42)	-124.6	253.3

Table 6. Parameter estimates for *C. difficile* within-host diversity and evolution across 3 potential models.

Models allowing for baseline diversity were a considerably better fit to the data suggesting more than one closely related clone might become established following a transmission event. The coalescent-based model and Poisson regression allowing baseline diversity were similar in terms of parameter estimates and model fit, with the coalescent-based model having a marginally smaller AIC.

Using the coalescent-based model 95% prediction intervals were calculated for the expected number of SNVs for a given time difference between samples (**Figure 29**). 0-2 SNVs are expected between isolates <124 days apart, rising to 0-3 SNVs for 124-364 days. Although at short time intervals the 95% prediction interval includes up to 2 SNVs, 0 SNVs is still the most likely observation. See **Figure 30** for the probability of a given number of SNVs for a range of time intervals between first and subsequent samples. Although the evolutionary rate point estimate was lower in ST1/ribotype-027 versus non-ST1s, the difference did not reach statistical significance (p=0.23).

These estimates can be used to judge whether transmission between two hosts sampled a given time apart is plausible. If the cases are related by a transmission event the expected number of SNVs between their isolates should vary in the same way as within host evolution.

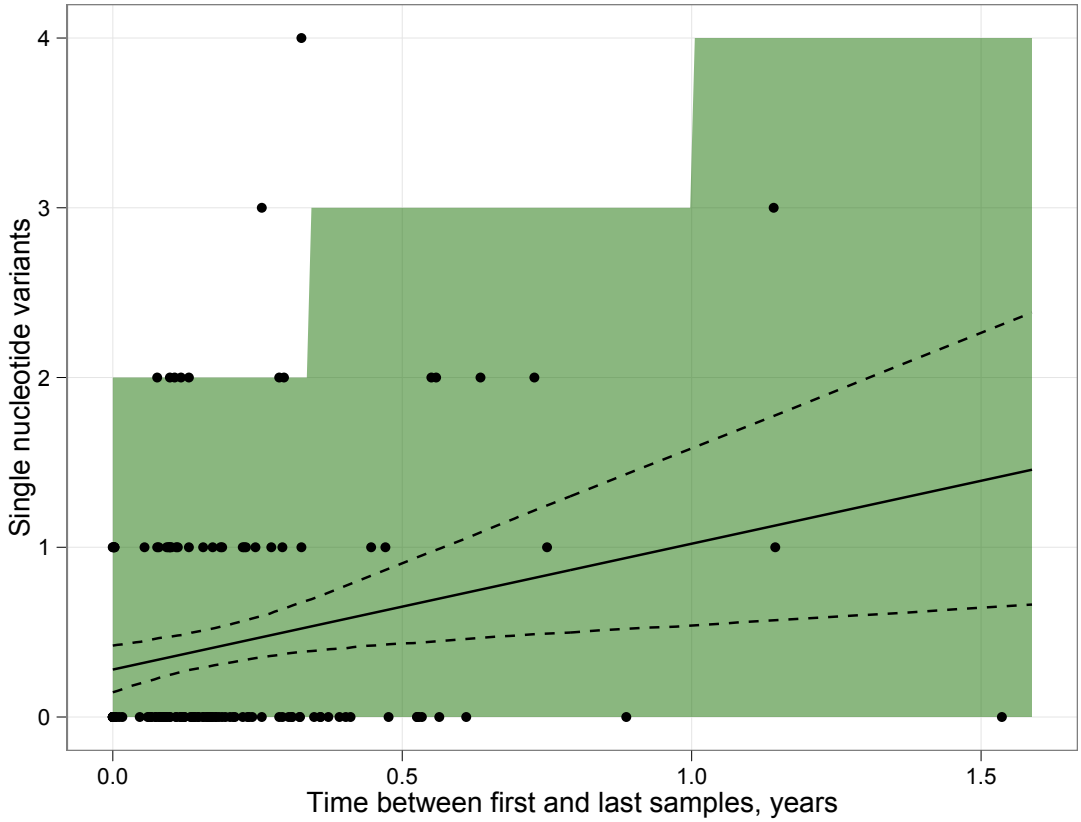


Figure 29. Coalescent-based model for *C. difficile* within-host diversity and evolution. The expected number of SNVs for a given time between samples is shown as a solid line, the dotted lines show the 95% confidence intervals. The 95% prediction interval based on the expected number of SNVs is shown shaded in green.

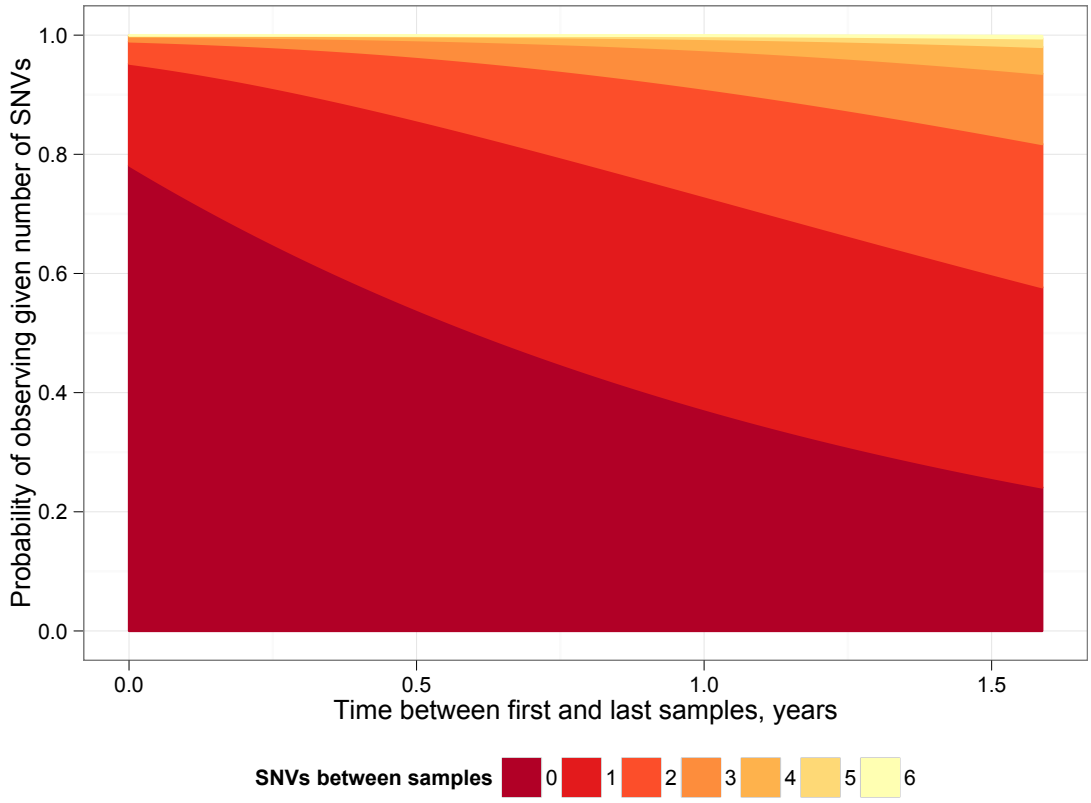


Figure 30. Probability of observing a given number of SNVs over a range of times between first and last samples. Based on coalescent model and parameter estimates above. The probability of a given number of SNVs is shown by the height of the area of that colour.

This analysis makes several important assumptions. The extent to which the population studied, patients with on-going or recurrent disease, represents all CDI is unclear. Such patients tend to receive multiple and prolonged courses of CDI treatment, and may also have received multiple courses of provocative antibiotics, which may exert greater selective pressure on the host's *C. difficile*. Conversely the failure to clear *C. difficile* following initial infection may be associated with a poorer immune response and less selective pressure.[79] The gut model data presented below provides some reassurance the findings are generalizable.

Although the baseline diversity models provided a better fit to the data, than the alternative with transmission representing an absolute evolutionary bottleneck, it is still possible that transmission represents a partial bottleneck. This could have the effect of accelerating rates of evolution as infection passes from one host to another. This has been observed in viral infections in experimentally infected animals.[203] It is postulated that minority variants present in the donor of infection are more likely to reach fixation in the recipient if relatively few infectious particles are transmitted. However, given conducting such human-to-human transmission experiments in *C. difficile* is unlikely to be feasible and endemic transmission events are difficult to identify, the extent to which this effect may be important in bacteria and *C. difficile* specifically is difficult to quantify directly.

The models fitted assume a constant rate of *C. difficile* evolution. However, rates of evolution in reproducing vegetative forms may differ in spores. Interestingly in another organism with a potential genetically latent and active state, *Mycobacterium tuberculosis*, work in macaque monkeys suggests similar rates of evolution in latent and active disease.[204] If spores do have a slower 'molecular clock' this could have two differing effects. If the estimated rate of evolution above is actually based on mix of vegetative and spore forms then, firstly, the estimated rate of evolution is likely to be an underestimate for active forms. This could in theory result in isolates from cases with active disease being falsely thought to be more genetically distant than in fact they are, although **Figure 30** suggests that very large SNV differences, e.g. >10, are unlikely regardless. Additionally the degree of underestimation of the evolutionary rate will depend on the prevalence

of vegetative versus spores in the cases studied and given many of the cases had on-going or recurrence disease a dominance of vegetative forms seems likely. Secondly, where cases' *C. difficile* has spent most time in the spore state, their most recent common ancestor may actually have existed longer ago than the estimated clock rate would suggest. In contrast to the first possibility this effect would tend to make more cases seem plausibly genetically related.

The assumptions behind the coalescent-based model are also simplifications. A constant effective population size is assumed, whereas this is likely to change within the host during induction of infection, symptomatic disease, and then subsequent resolution. Changes in the *C. difficile* population in terms of colony forming units are observed during these phases in the gut model below. The population is also assumed not to be under selective pressure, whereas in fact exposure to new host immune systems and antibiotics are very likely between samples being obtained from a transmission donor and recipient. However similarities in estimates compared to a simpler Poisson model suggest these assumptions are likely reasonable.

Finally, recombination needs to be considered as a possible source of variants observed between cases. If multiple SNVs are introduced in a single recombination event between sampling 2 cases, these cases may be thought to be more genetically distant than in fact they are. This is considered in more detail in the analysis in section 5.2.3.2. However, no clear clustering of variants suggestive of recombination was seen in the serially sampled patient samples.

4.4.2 WITHIN SAMPLE DIVERSITY

To assess if estimates of within host diversity obtained from the evolutionary model in serially sampled patients were reliable, multiple individual bacterial colonies were sequenced from 3 longitudinally sampled cases. One case had 3 samples taken 46 and 118 days after initial diagnosis, and another 2 cases had 2 samples taken 192 and 193 days after initial diagnosis. At each of the 7 timepoints 12 individual colonies were selected from primary stool cultures (see chapter 3.2.2), sub-cultured, and then sequenced.

The results are summarised in **Figure 31**. No variants were observed 34 individual colonies taken over 3 timepoints in a patient with ST1 CDI. In a second patient with ST42 CDI a single colony differed by 1 SNV at the first time point, and 2 of 9 colonies differed by a different single SNV at the second time point. Interestingly a previously undetected mixed infection with ST8 was discovered at the second time point, raising the possibility that acquisition of a new ST was responsible for this patient's relapse. The third patient had evidence of a single SNV in 5 samples at the first timepoint (two of which were heterozygous for the wildtype and the SNV, suggesting the colony picked contained both). By the second timepoint all sequences contained the SNV. Although relatively few multiple colony samples were processed, there was no evidence of more than 1 SNV emerging within a host. Across 79 *C. difficile* colonies sequenced a single SNV difference from the dominant clone was detected in 9 samples (11%, 95%CI, assuming each colony independent, 5-21%). This overlaps with the confidence interval obtained for within host diversity from the coalescent model above (0.30 SNVs (95% CI 0.13-0.42)), supporting its use.

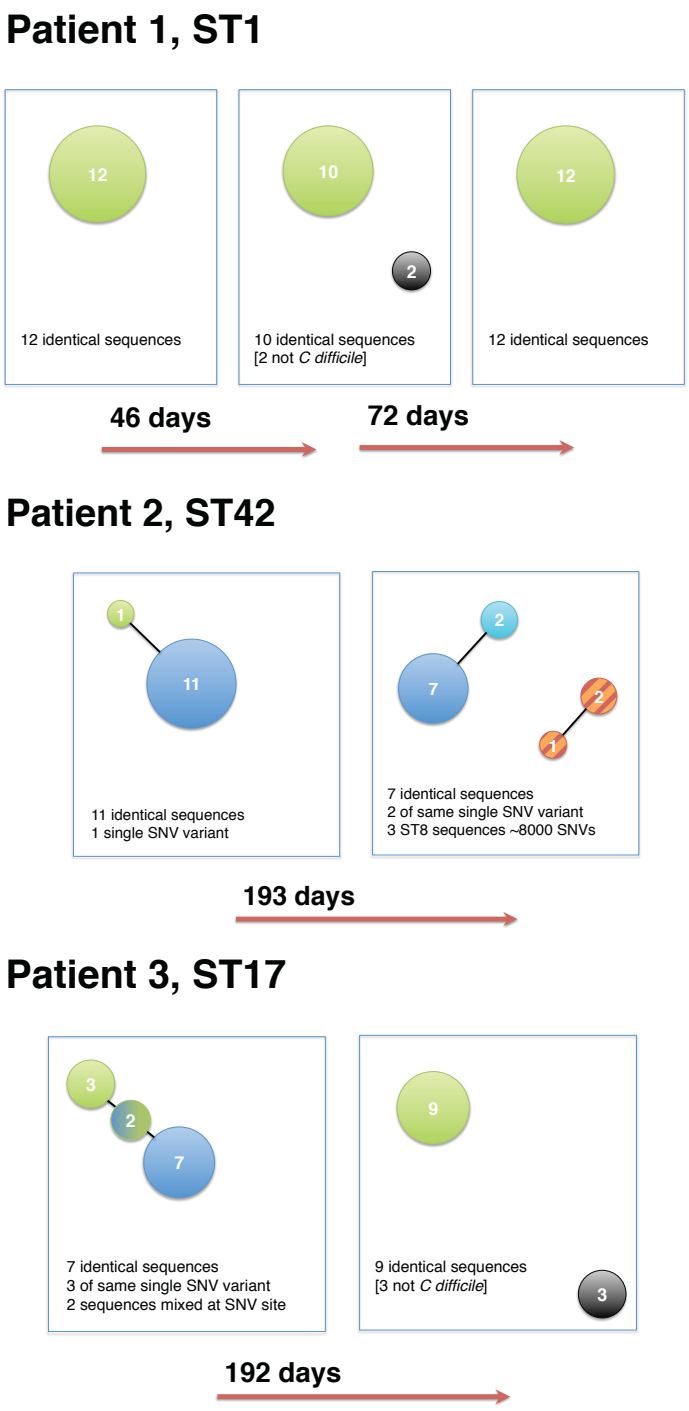


Figure 31. Assessing within-host diversity with multiple bacterial colony sub-culture. Serial samples obtained from 3 patients are shown. Each timepoint is shown as a rectangle, and the number of days between samples indicated. 12 individual colonies were sub-cultured and sequenced at each timepoint. Colonies subsequently found not to be *C. difficile* are shown in black. *C. difficile* sequences are coloured according to any variants found. The number within each circle indicates the number of individual colonies with that particular variant.

4.4.3 GUT MODEL

To further assess the genomic stability of *C. difficile*, in particular to be able to study the effect of interventions and exposures such as CDI induction, and antibiotic treatment in a controlled environment, samples were sequenced from an *in vitro* gut model of CDI. [205,206]

Serial samples were taken from 3 human gut CDI model experiments.[206-208] One model simulated CDI resulting from co-amoxiclav exposure (without CDI treatment), and the other two, CDI following clindamycin exposure, then subsequently treated either with vancomycin for 4 days or cadazolid (a novel oxazolidinone/quinolone hybrid[209]) for 7 days. Briefly, the gut model consists of three vessels in a weir cascade system, held under anaerobic conditions, and designed to reflect the increasing alkalinity (pH 5.5-6.8) and reduced substrate availability of the proximal to distal colon. In each experiment the system was inoculated with a pooled faecal slurry of *C. difficile* culture-negative faeces that were obtained from five healthy (*C. difficile* negative) volunteers aged ≥60 years. The model was top-fed with a growth medium at a controlled rate. The following time periods key time periods were observed (**Figure 32**): Period A, steady state, when the model was left to equilibrate for two weeks prior to interventions; Period B, *C. difficile* spore inoculation, when the model was instilled with ~10^7 cfu *C. difficile* ST1 spores and monitored for seven days; Period C, *C. difficile* spore inoculation and CDI initiation, when the model was instilled with a further ~10^7cfu *C. difficile* ST1 spores and CDI-provoking antibiotics (co-amoxiclav 8mg/L, 3 times daily for 7 days; or clindamycin, 33.9 mg/L, 4 times daily for 7 days); Period D, CDI, i.e. *C. difficile* cytotoxin induction, following provocative antibiotic treatment;

Period E, In two of the three gut models, CDI treatment (vancomycin 125mg/L, 4 times daily for 4 days, or cadazolid 250mg/L twice daily for 7 days [208,209]) following high level cytotoxin production (≥4 relative units) for ≥2 days; and finally period F, a rest period when no further interventions took place. *C. difficile* isolates were obtained from the gut model at the timepoints indicated in **Figure 33**, **Figure 34**, and **Figure 35**. Following anaerobic culture for 24 hours on Columbia blood agar, DNA was extracted from sub-culture of a representative sweep taken across the primary culture plate, except for 3 occasions where 12 individual colonies were sub-cultured and DNA extracted from the resulting growth to assess for within-'host' diversity (co-amoxiclav model, final timepoint and clindamycin/vancomycin model immediately prior to CDI treatment and at the final timepoint).

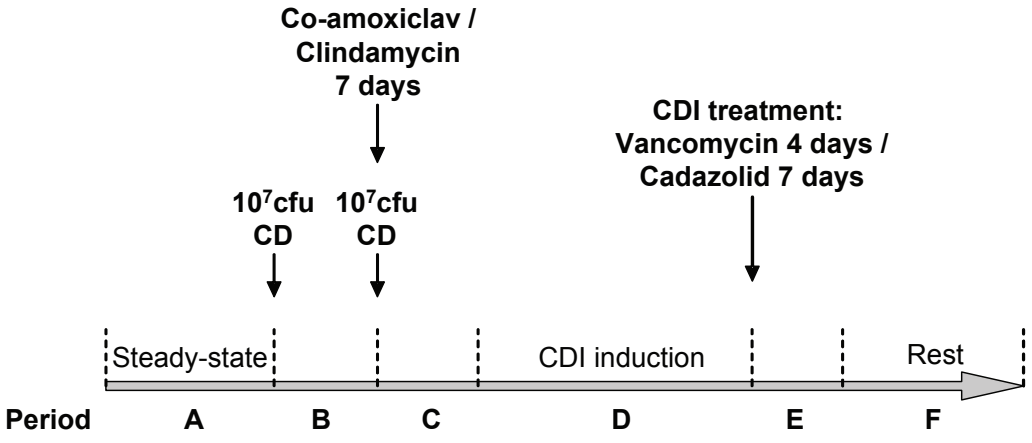


Figure 32. Gut model samples. A schematic diagram of the gut model experiments to induce and treat *Clostridium difficile* infection. CD, *Clostridium difficile*. CDI, *C. difficile infection*. Cfu, colony forming units.

DNA was sequenced as above, with reads mapped to an ST1 reference genome, CD196 (Genbank: FN538970).[184] Single nucleotide variants were identified after quality filtering (see section 4.2.2).

A total of 62 *C. difficile* isolates were obtained and successfully sequenced from 29 samples collected from the 3 gut models: 18 isolates from the co-amoxiclav induction model (7 timepoints over 35 days, 12 isolates at last timepoint, **Figure 33**), 33 isolates from the clindamycin induction, vancomycin treatment model (11 timepoints over 70 days, 12 isolates at two timepoints, just prior to, and following vancomycin treatment, **Figure 34**), and 11 isolates from the clindamycin/cadazolid model (11 timepoints over 32 days, **Figure 35**). A mean 90.3% of the CD196 reference genome was called, with a mean read depth of 62.3.

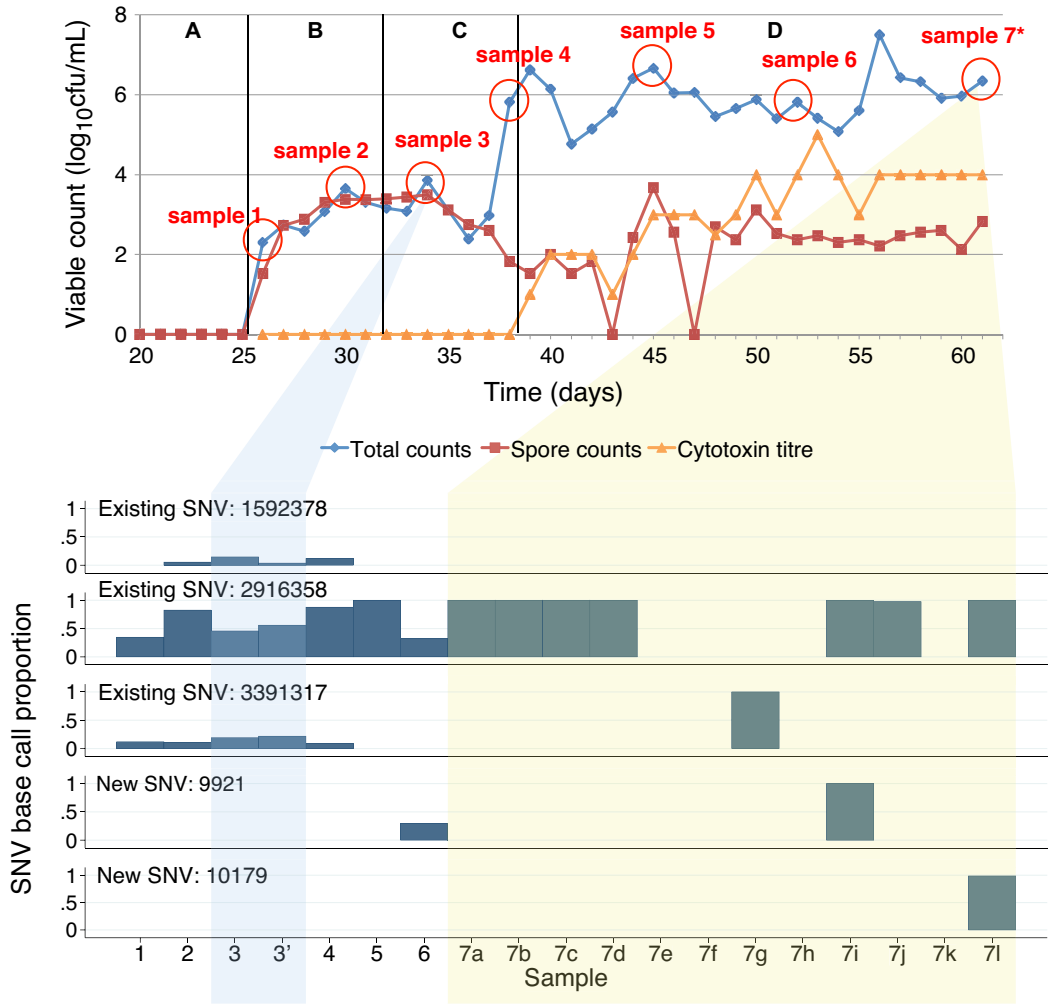


Figure 33. Co-amoxiclav induction gut model samples and variable site base counts. Panel A shows the total viable counts, spore counts and cytotoxin titres. Sampling points are shown as circles. Sampling points where 12 colonies were sub-cultured and sequenced separately are shown with an *. Panel B shows the base counts obtained at the variable sites identified between sequences. Each sample is represented in a different column. Samples taken at different timepoints are numbered sequentially. Where 12 colonies were sub-cultured at a given timepoint these are indicated by different letters following the sample number and yellow background shading. Replicate samples sequenced to assess the accuracy of the sequencing are shown with a ', e.g. 3' and blue background shading. The proportion of high quality (base quality PHRED scaled score ≥30, mapping quality ≥30) base counts supporting the SNV identified is shown for each variable site as a separate row. Median (range) calls per site 49 (23-99). No site with >2 different base calls.

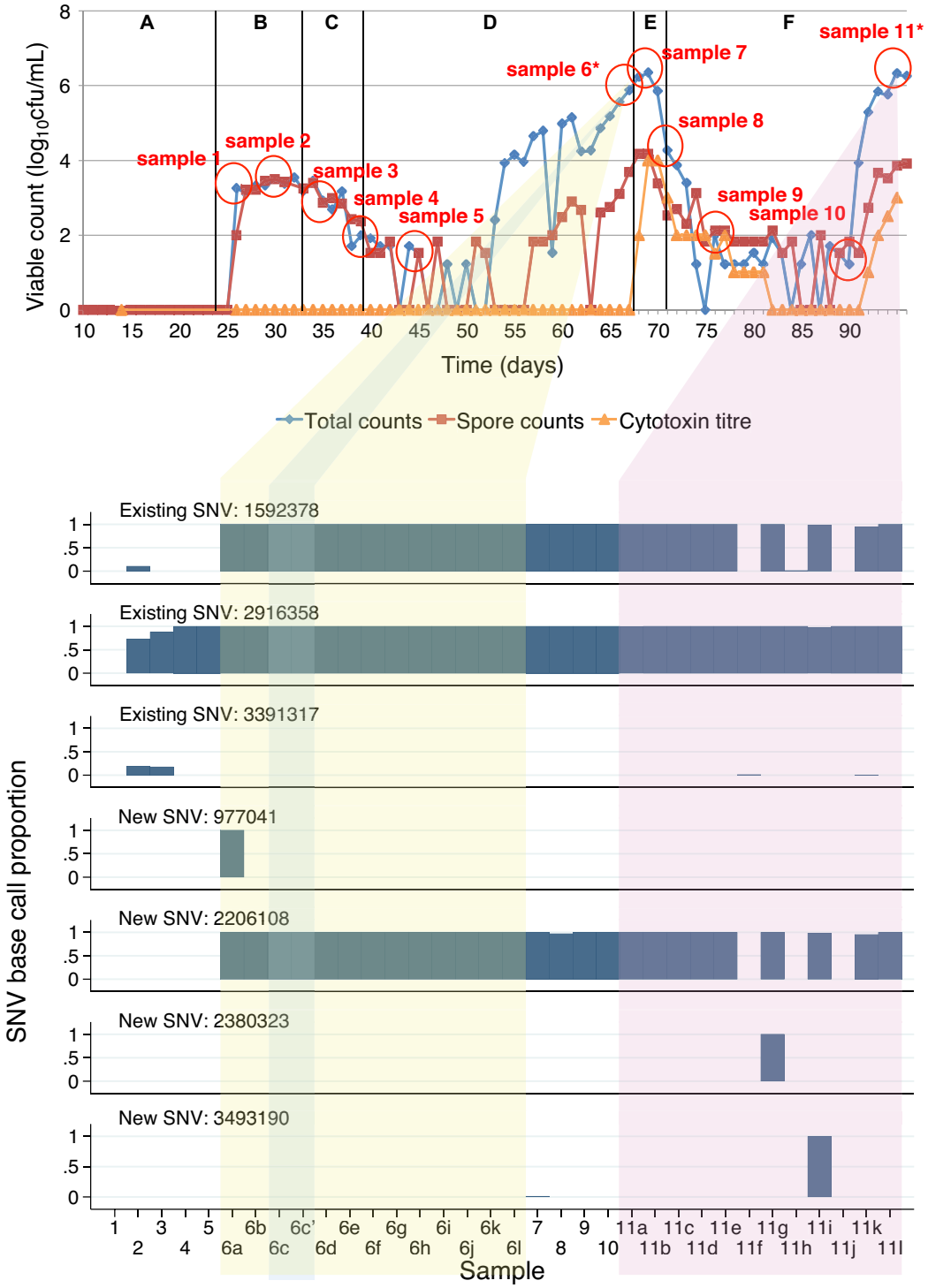


Figure 34. Clindamycin-induction, vancomycin-treatment gut model samples and variable site base counts. Panel A shows the total viable counts, spore counts and cytotoxin titres. Sampling points are shown as circles. Sampling points where 12 colonies

were sub-cultured and sequenced separately are shown with an *. Panel B shows the base counts obtained at the variable sites identified between sequences. Each sample is represented in a different column. Samples taken at different timepoints are numbered sequentially. Where 12 colonies were sub-cultured at a given timepoint these are indicated by different letters following the sample number and yellow background shading (pre-treatment) and purple background shading (post-treatment). Replicate samples sequenced to assess the accuracy of the sequencing are shown with a ‘, e.g. 6c’ and blue background shading. The proportion of high quality (base quality ≥30, mapping quality ≥30) base counts supporting the SNV identified is shown for each variable site as a separate row. Median (range) calls per site 71 (28-198). No site with >2 different base calls.

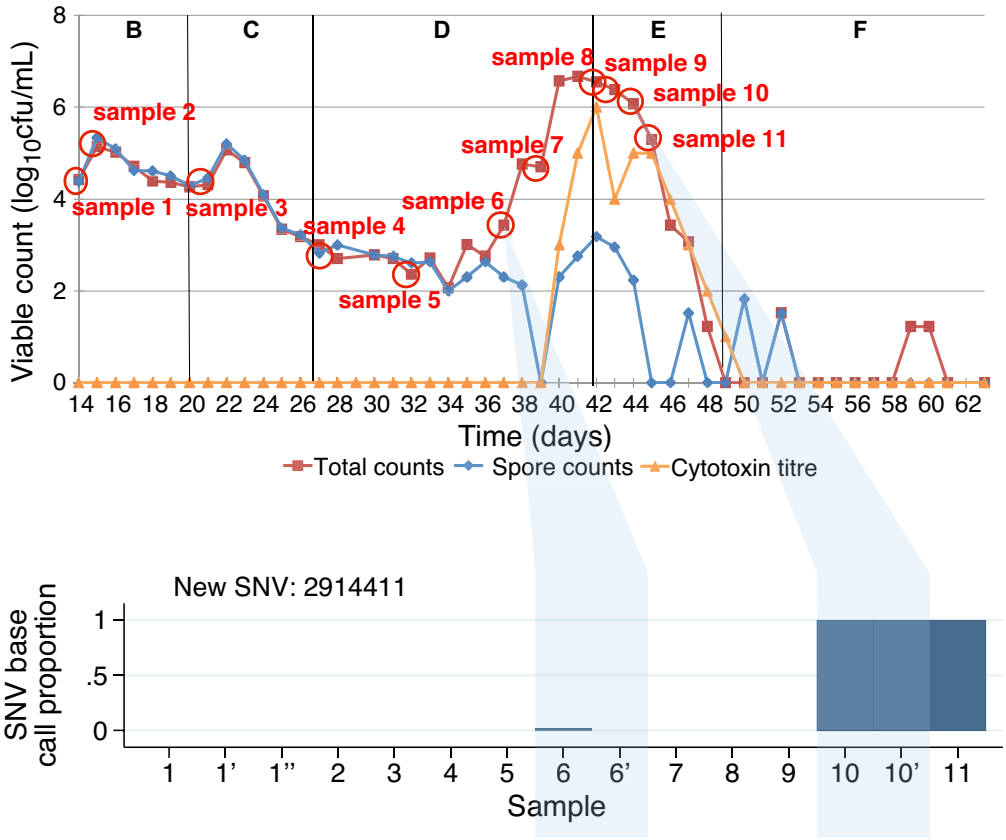


Figure 35. Clindamycin-induction, cadazolid-treatment gut model samples and variable site base counts. Panel A shows the total viable counts, spore counts and cytotoxin titres. Sampling points are shown as circles. Panel B shows the base counts obtained at the variable sites identified between sequences. Each sample is represented in a different column. Samples taken at different timepoints are numbered sequentially. Replicate samples sequenced to assess the accuracy of the sequencing are shown with a ‘, e.g. 6’ and blue shading. The proportion of high quality (base quality ≥30, mapping quality ≥30) base counts supporting the SNV identified is shown for each variable site as a separate row. Median (range) calls per site 57 (28-89). No site with >2 different base calls.

A total of 10 variable sites were identified across all 3 gut models. Both the co-amoxiclav and clindamycin-vancomycin models were inoculated with the same *C. difficile* ST1 isolate. Sequences obtained from whole culture plate sweeps at the first 2 timepoints in these models show evidence of heterozygousity at 3 of the variant sites in both experiments (sites 1592378, 2916358, 3391317) suggesting that a mixture of *C. difficile* with and without these SNVs was present in the initial inoculum. In the co-amoxiclav model one of these SNVs (2916358) persisted throughout and was present in 7 of the 12 colonies sequenced at the final timepoint. In contrast the proportion of reads containing the other 2 SNVs diminished during the population expansion associated with CDI cytotoxin induction (phase D), and were present in 0/12 and 1/12 colonies at the end of the model. In the clindamycin/vancomycin model, 2 of the pre-existing SNVs became established (sites 1592378, 2916358), as well as a new SNV which arose during induction and persisted during treatment (2206108). In both models, a number of minority variants emerged, evidence of which could be seen in multiple picks taken following CDI cytotoxin induction (co-amoxiclav model, sites 9921 and 10179; clindamycin/vancomycin model, site 977041) and CDI treatment (clindamycin/vancomycin model, 2380323, 3493190) (**Figure 33**, **Figure 34**). In the final model, clindamycin/cadazolid, a single SNV was observed at the final 2 timepoints (2914411) following treatment.

Overall, over a total 137 days of follow up, including three CDI inductions and two CDI treatments, only two new SNVs emerged that became established in the majority of sequences. Pre-existing SNVs present in the minority of the within-model inoculum population became dominant twice in one model, and once in

another. Five additional SNVs were detected that were only present in the minority of colonies at one or two timepoints.

As with the clinical isolates, the gut model isolates contained within host variation at the initial sample. This contributed significantly to SNV differences observed between baseline and subsequent samples. The 2 new SNVs observed over 137 days of gut model follow up are consistent with the estimates for *C. difficile* evolution from serially sampled patients (**Figure 29**, **Figure 30**).

The presence of minority variants detected by sequencing sub-cultures of multiple colonies, whilst not unexpected, is a reminder that diversity may be detected between transmitted cases as a result of sampling a minority variant in either the donor or the recipient. Similarly, if only a single colony is sub-cultured, potential mixed infection may also be missed. However sequencing sweeps of primary culture plates using existing data pipelines may also miss minority variants if these are attributed to sequencing error, or result in a heterozygous base calls, which may be converted to a null call by sequencing quality filtering algorithms. Therefore, sequencing the full diversity present on primary culture plates together with bioinformatic approaches specifically designed to detect mixed infection may provide an optimal approach. This is discussed further in section 4.6.

In summary, the consistency of whole genome sequence data from gut model *C. difficile* isolates adds to the data presented on the high stability of genomic sequences in isolates from patients and supports the use of sequence data in detailed transmission investigations.

4.5 POPULATION DIVERSITY

Given the genomic stability of *C. difficile* and its relatively recent emergence as a major clinical problem[210] relatively little diversity in strains causing human disease might be expected. However long-term evolutionary analysis of *C. difficile* reveals it evolved at least a million years ago, with subsequent evolution generating thousands of SNVs.[185] *C. difficile*'s recent emergence as a pathogen likely reflects the relatively recent widespread use of antibiotics. It still remains important to define how much variation exists in a defined geographic region over a relatively short period of time, as this will determine the discriminatory power of sequencing in an outbreak or endemic setting.

To examine the diversity present in the Oxfordshire CDI cohort, the numbers of SNVs between all possible pairs of the 1223 sequenced cases diagnosed between September 2007 and March 2011 were calculated. The distribution of these pairwise differences is shown in **Figure 36**. In the top panel the major peaks correspond to comparisons made across clades.[199] Focusing on SNV differences ≤100, a relative excess of 0, 1, and 2 SNVs suggests these cases may be related by transmission, which would be consistent with the within host diversity and evolution estimates above (**Figure 29**, **Figure 30**).

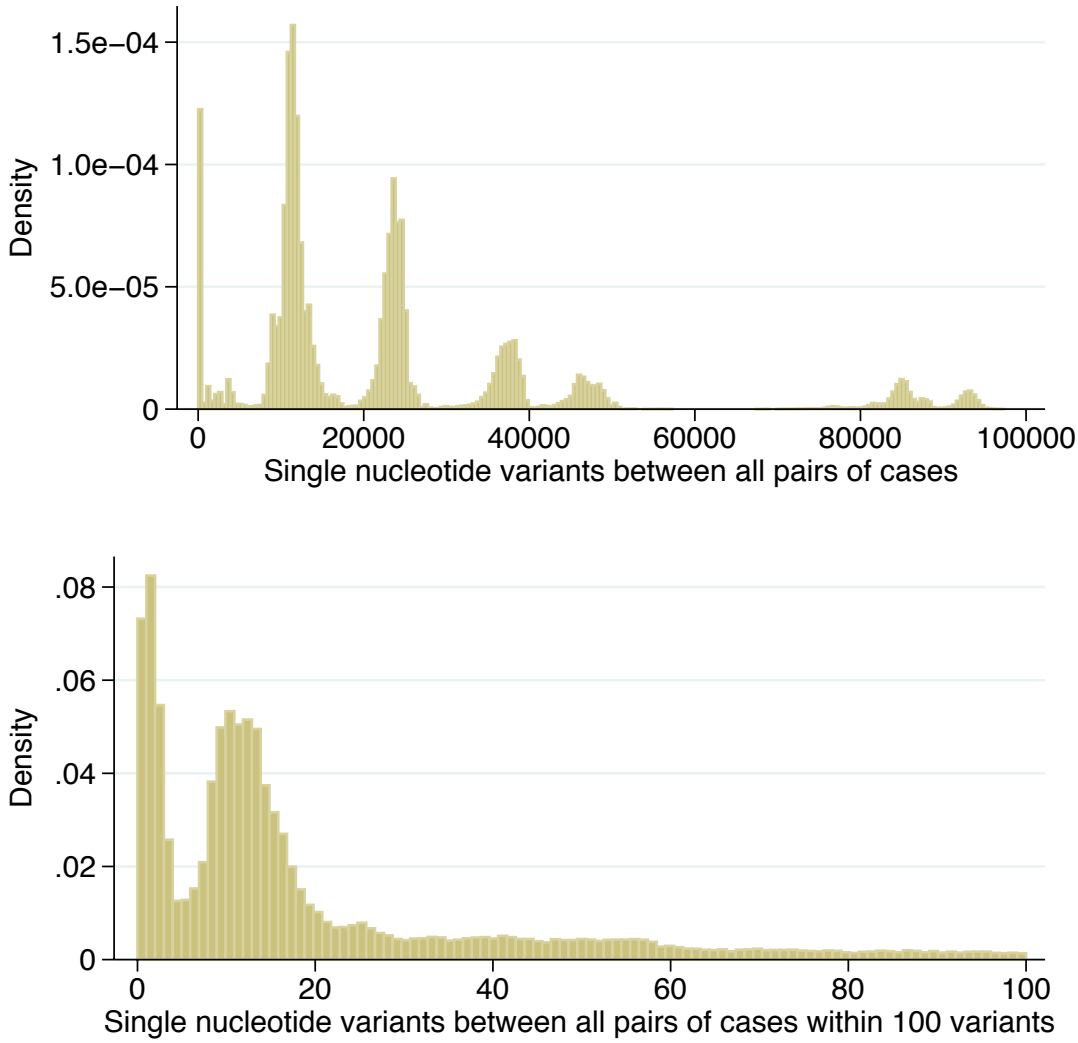


Figure 36. Distribution of single nucleotide variant differences between all pairs of Oxfordshire CDI cases, September 2007 – March 2011. The top panel shows all possible pairs. The bottom panel all pairs of cases with 100 SNVs.

The Simpson Diversity Index, DI[150] can be used compare the additional resolution offered by WGS over MLST. The DI is the probability that two cases taken at random do not share the same type. In the case of MLST, this is the probability two cases do not share the same ST. Across the 1223 cases, 78 distinct STs were present, with a DI of 0.935. The value of DI, defining types using differing WGS SNV thresholds is shown in **Figure 37**. Using a SNV threshold of 2 to define

types the DI was 0.989, and 0.977 using a threshold of 10 SNVs, demonstrating the additional discrimination with WGS.

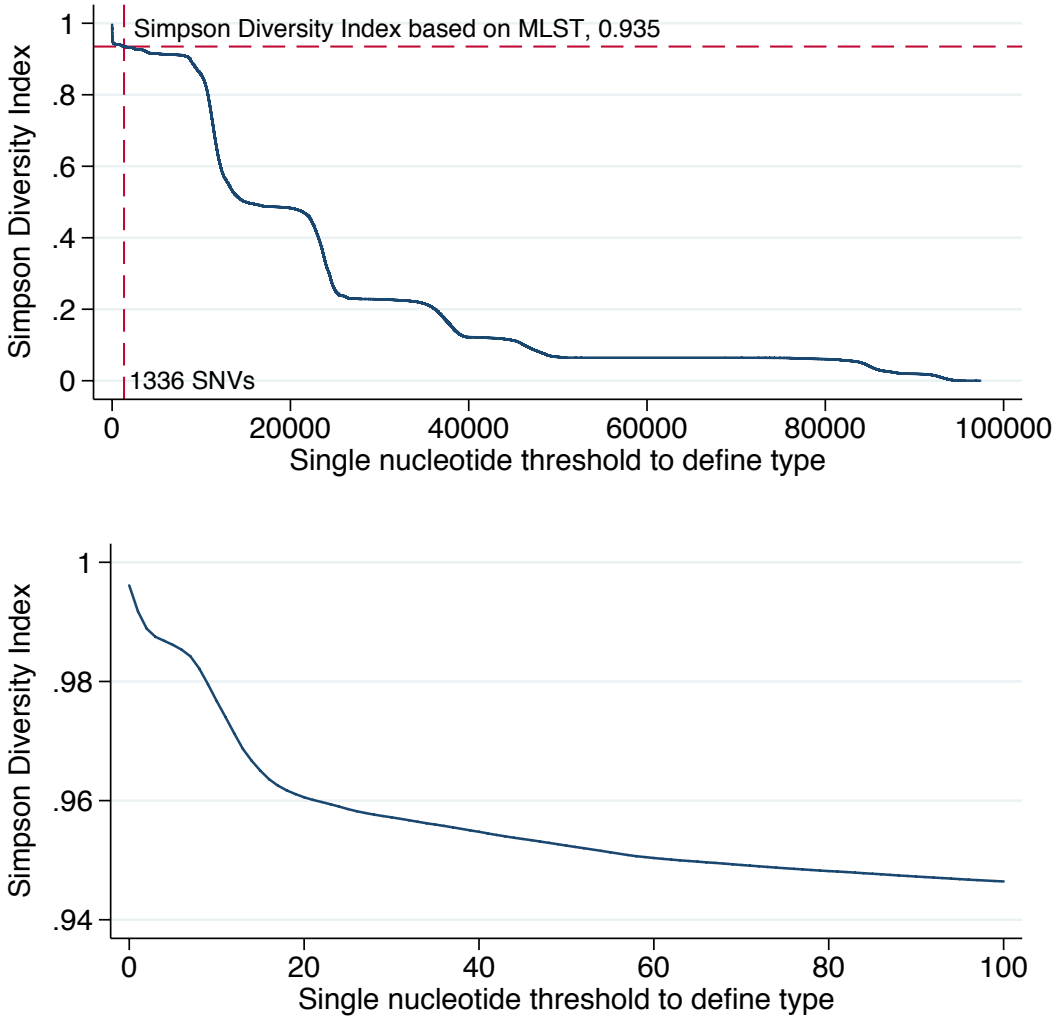


Figure 37. Simpson Diversity Index, DI, defining types using varying SNV thresholds across 1223 Oxfordshire CDI cases. The top panel shows the DI across all thresholds, and the bottom panel the DI for SNV thresholds up to 100. The DI based on MLST is shown as a dotted line on the top panel, which corresponds to the DI based on a SNV threshold of 1336.

4.6 MIXED INFECTION

Genotyping and whole genome sequencing aim to rule out transmission between genetically distinct isolates, however, potentially undetected mixed infection impairs exclusion of transmission with certainty, and may therefore limit the utility of this technique.[195,211] If a mixed infection is present in a transmission donor or recipient, but only one isolate sampled from each, as has been done in the majority of bacterial whole genome sequencing studies to date[198,212-216] (but not all, see [217]), it is possible the sequences may differ despite the same strain actually being present in both cases. In this scenario transmission will be incorrectly excluded, and the recipient may falsely appear not to have an identified donor. Traditional approaches to investigating mixed infection are expensive because they involve separate sub-culture of multiple bacterial colonies, a process in which multiple individual colonies are transferred to a separate culture plate and re-incubated.[195] Because this approach is cost and labour intensive, it is not used in routine clinical laboratories or in large-scale transmission studies.

Two pieces of work are presented that address the importance of mixed infection in *C. difficile*. Firstly the prevalence of mixed infection is assessed using multilocus sequence typing of patients tested more than once over a short time period. Secondly a tool is developed to screen whole genome sequence data for evidence of mixed infection. It is applied to selected transmission events enriched for the possibility of mixed infection, to assess if there is evidence of mixed infection playing an important role in the undetected transmission described in chapter 3.

4.6.1 DETECTING MIXED INFECTION WITH GENOTYPING

The population based cohort of all Oxfordshire CDI (described in section 3.2.1) was used to estimate rates of re-infection and mixed infection using samples obtained and multilocus sequenced typed (see section 3.2.2) from September 2006 to March 2011. The early period (September 2006 to August 2007) was not included in the transmission analyses in chapter 3 as the number of samples retrieved for culture during this period was less complete (~70%, compared to >95% subsequently). During the study, local infection control policy was that any inpatient with diarrhoea (≥3 unformed (taking the shape of the container) stools in 24 hours) should be tested for *C. difficile*. From May 2007, laboratory policy was also to test all unformed stool samples from those aged ≥65 years for *C. difficile*, whether or not this had been requested. Patients were only considered CDI-negative after 3 EIA-negative samples. Given the delay (albeit short) between sample submission and test results becoming available, multiple positive samples were sometimes obtained from the same patient on the same day (or shortly afterwards), allowing an estimate of mixed infection rates from MLST. Similarly, samples taken further apart from patients with symptomatic recurrences or persistent diarrhoea allow rates of re-infection due to new strains to be estimated. This is important as 6-25% of patients experience a further episode of CDI after initial diagnosis.[29] If CDI recurrence in these patients is the result of re-infection with a new strain this needs to be considered in transmission analyses, whereas patients with relapse of the same strain should be analysed differently.

All patients with ≥2 samples from which genotyped isolates were obtained were included in analyses, which were stratified according to whether or not the first

CDI was with the epidemic ST1. Logistic regression was used to estimate the relationship between pairs of consecutive isolates having different genotypes and the elapsed days between samples, using fractional polynomial models to reflect non-linearity.[218] Stata 11.2 was used for all analyses.

Of 3258 EIA-positive samples, 2906 (89%) were retrieved for culture. *C. difficile* was recovered in 2235/2906 (77%), and successfully genotyped in 2231 (1549 patients). 79 distinct STs were identified, with ST1 the most prevalent (30%); other STs each accounted for <8% of isolates, and 22 STs were isolated once. Most (1356, 61%) isolates were from clade 1.[199]

On 109 occasions more than one EIA-positive, culture-positive isolate was obtained from different faecal samples from the same patient on the same day (97 patients with 2 samples, 11 with 3 samples, 1 with 4 samples). On 3/109 (2.8%, 95% CI 0.6-7.8%) occasions, involving 2 different patients, >1 ST was isolated from the same patient on the same day. If multiple samples taken within 0-7 days were considered as the same CDI, >1 genotype was isolated on 21/292 (7.2%; 95% CI 4.5 - 10.8%) occasions; similarly, rates for samples 14 days apart (21/305) were 6.9% (4.3 - 10.3%).

All 560 consecutive same-patient pairs of EIA-positive, culture-positive samples with STs taken ≥1 day apart (from 388 patients) were analysed for evidence of re-infection with a new strain. Consecutive samples were taken a median (IQR) [range] 24 (1-52) [1-1414] days apart, 203 (36%) within 7 days. 88/560 (16%) pairs had different genotypes, with the samples taken a median (IQR) 54 (30-132) days apart; compared with 21 (1-43) days for pairs sharing the same ST

(p<0.0001). 31 pairs had >6 months between consecutive samples, of which 13/31 (42%) shared the same ST, as did 5/14 (36%) pairs with >1 year between samples. 18/194 (9%) pairs where the first isolate was the most common ST, ST1, had a second isolate of a different ST (median 70 days) versus 70/366 (19%) where the first isolate was not ST1 (median 53 days) (p=0.002) (**Figure 38**). The chance of a subsequent sample having a different ST increased 5-fold from 1 to 11 days, and overall increased by 58% (odds ratio, 95% CI 40-79%) for every doubling of the time between samples (p<0.0001), allowing for differences in the underlying prevalence of ST1 versus other STs (**Figure 39**).

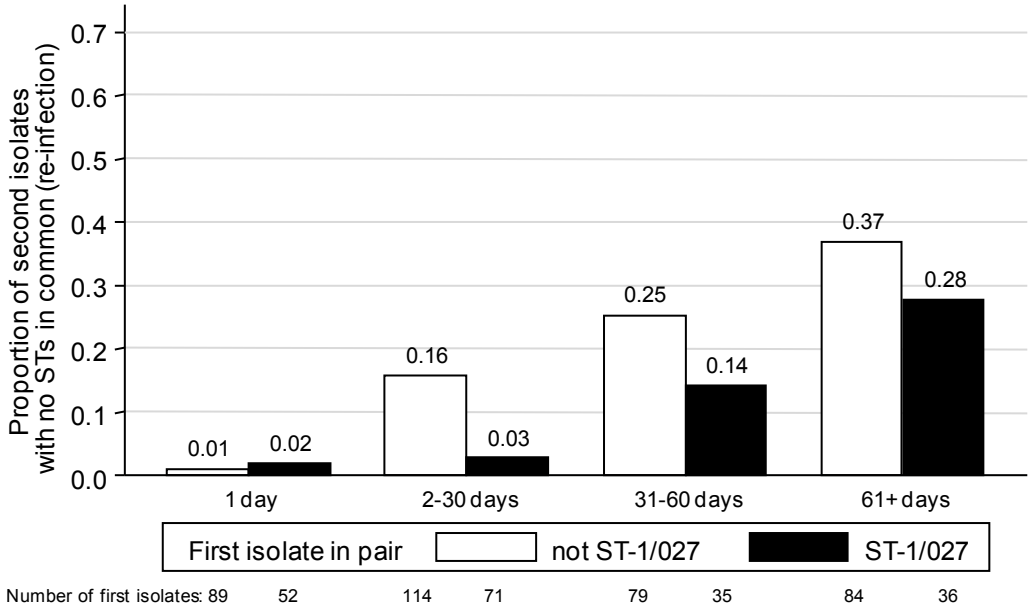


Figure 38. Chance of consecutive *C. difficile* isolates having no sequence types (STs) in common according to time between samples. Observed data, categorising time between consecutive faecal samples into approximate quartiles.

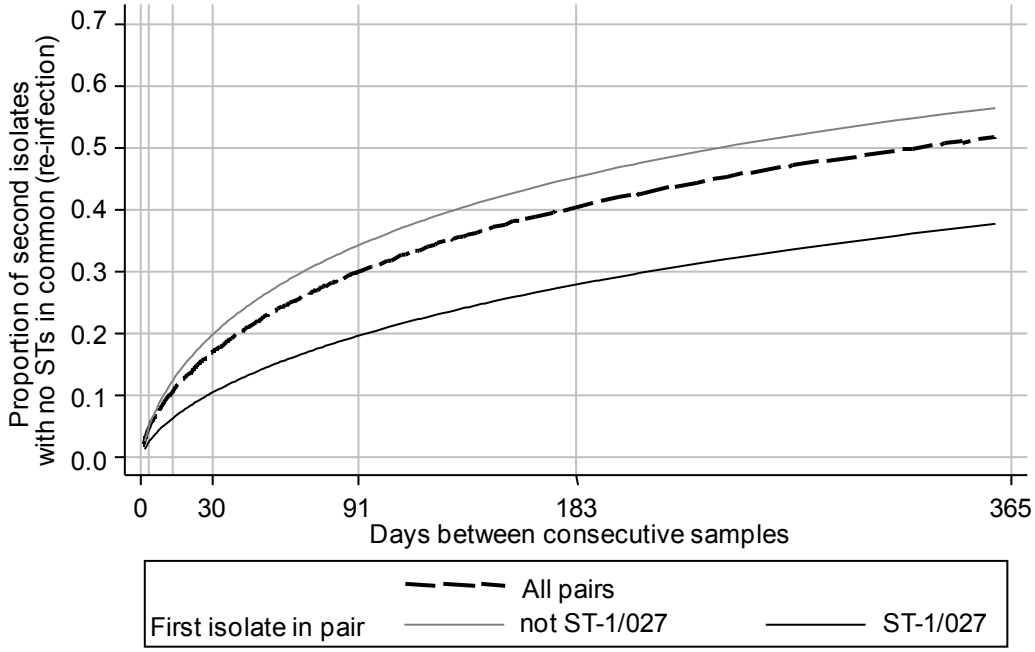


Figure 39. Chance of consecutive *C. difficile* isolates having no sequence types (STs) in common according to time between samples. Fitted logistic regression model. Data split by whether the first ST is the most common ST1 or not, because re-infection with a different lineage of the same ST is more likely where the first ST is relatively common, e.g. ST1. Therefore, the estimated re-infection rate where the first ST in a pair is ST1 is likely under-estimated. Best fitting fractional polynomial model included log(days between consecutive samples).

Assessment of CDI associated with the presence of multiple strains by opportunistic sampling of patients on the same day found rates of 2.8%. These data suggest the chance of identifying a new ST infection increases most rapidly during the first days after the initial positive sample, and indicates that strains identified during the first 7-14 days are most likely to reflect original, possibly mixed, CDIs. The measured mixed CDI rate (~7% based on multiple EIA positive, culture positive isolates obtained over up to 7 days) is consistent with those from previous investigations (9-13%) that examined multiple *C. difficile* colonies from single faecal samples.[193-196]

These data also provide clear evidence that as the time between faecal samples increases there is a significant and marked increase in the rate of isolation of different *C. difficile* strains from individual patients. However, as with other infectious diseases such as tuberculosis,[219] the timings of same strain recurrence and new re-infections overlap to some extent. Indeed, there were a reasonable minority of patients in whom relatively uncommon STs were re-isolated >2 months after the original CDI, in some cases over 1 year later. Similarly, whilst MLST offers reasonable discriminatory power to detect different strain infections,[145] the dominance of common STs, particularly ST1 which accounted for 30% of cases, may result in an under-estimate of re-infections and mixed infections with different lineages of the same ST.

Overall, mixed infection rates ~7% suggest this needs to be considered before transmission can be excluded. The extent to which mixed infection plays a role in onward transmission is explored in the next section. As re-infection plays a considerable role in symptomatic recurrence, re-testing of patients with recurrence and analysis of their isolates should be undertaken in transmission studies, and this is done in the analysis in chapter 5.

4.6.2 Estimating mixed infection with whole genome data

The properties of WGS data can potentially be exploited to detect mixed infection without the need for separate sequencing of multiple bacterial colonies. In this section a method for such analysis is developed and then applied to look for evidence of transmission of *C. difficile* arising from mixed infection.

WGS in bacteria typically yields generous depth of coverage (measured by the number of reads mapping to any particular site in the sequenced genome[178,220]). Interrogation of these reads offers the prospect of detecting mixed infection by sequencing an aggregate of colonies at the same cost as sequencing an individual colony. In this approach, the short reads produced by next generation sequencers are mapped to a reference genome using a standard method.[178] The composition of bases mapping to any given nucleotide position can then be analysed to detect evidence of multiple strains. Whereas bacterial genomes should normally be haploid, a pattern of bases that resembles a heterozygous base call in a diploid genome is symptomatic of mixed infection.[221] This idea has been used to detect viral genetic diversity within individual hosts.[222,223] In viral sequences, the density of SNVs may be sufficient to allow common SNVs to be identified between overlapping reads and haplotypes to be reconstructed.[222] Clearly this ability is dependent on the within-host viral diversity, the sequencing technology and depth of sequencing coverage. In contrast, the density of SNVs between sequences in potential mixed bacterial infections is much lower; in *C. difficile*, there may be 100-10000 SNVs over a total genome of 4.3 million base pairs, which corresponds to just 1 SNV in 400-40000 base pairs.[185] At this density, SNVs are sufficiently sparse that complete haplotype reconstruction is not possible from current short-read sequencing with read lengths of the order of 100 base pairs. Many if not all SNVs are likely to lack adjacent variants closer than the maximum read length, making it impossible to associate these reads with the correct haplotype. The one exception to this is the scenario in which the haplotypes make up markedly different proportions of the sample.

In chapter 3, <25% of *C. difficile* infections in Oxfordshire, UK, over a 2.5 year period could be linked to a previous case with the same strain type via hospital ward contact, suggesting a substantial unsampled reservoir for human infections. However a potential limitation was that only one strain was sequence typed per case (with the exception of morphological variants) and therefore mixed infections could in theory comprise some or much of the unsampled reservoir. The analysis above in section 4.6.1 suggests *C. difficile* mixed infection rates of ~7%, consistent with other studies over the last decade.[193-197] However the significance of mixed infection in transmission has never been investigated. Therefore a method for detecting mixed infection from bacterial WGS data was developed, that exploits frequency differences between the dominant and minor strain making up the sample and compares putative base calls to a database of known sequences to assist in determining the haplotypes present. The algorithm developed is tested using *in silico* and *in vitro* mixed infection experiments and then applied to estimate the extent of transmission arising from mixed *C. difficile* infections in order to determine its relevance to hospital transmission and outbreak investigations.

4.6.2.1 METHODS

4.6.2.1.1 SETTING AND CASES

Hospital admission and ward movement data on the 1276 *C. difficile* infections from Oxfordshire used in chapter 3 were analysed in a separate study using a stochastic compartmental transmission model.[224] Low-resolution genetic data

from MLST was available for each case,[145] obtained from standard sub-culture of a single bacterial colony.[211] Initially the stochastic transmission model was fitted without incorporating genetic data of any kind (neither MLST nor WGS). This preliminary analysis was then cross-referenced to the MLST data in order to identify 15 pairs of cases enriched for likely transmission events associated with mixed infection. Putative transmission between these cases had been inferred by the stochastic transmission model on the basis of close epidemiological linkage, meaning that they either shared space and time on the same hospital ward around the diagnosis of the first case and before the second, or shared the same hospital ward in quick succession such that transmission via *C. difficile* spores was possible (**Figure 40**). All 15 pairs were associated with a high posterior probability of transmission ($p>0.45$), and had only a single highly likely donor for each recipient. However the single isolate typed from the potential donor and recipient in each pair had different STs. As MLST is a low resolution typing method based on sequencing conserved bacterial housekeeping genes, differing STs are distinct at the whole genome level, and therefore not compatible with transmission.[145,225] However, one important and normally overlooked explanation for the ST-mismatch is that the donor or the recipient could have had mixed infection, i.e. that genuine transmission between the two patients was masked by only genotyping one of several strains present in either or both patients. If mixed infection contributes significantly to transmission, it is likely this group of donor-recipient pairs will be enriched for undetected transmissions between multiply infected cases. Therefore these cases form a more sensitive test for mixed infection than random sampling from the total 1276 cases.

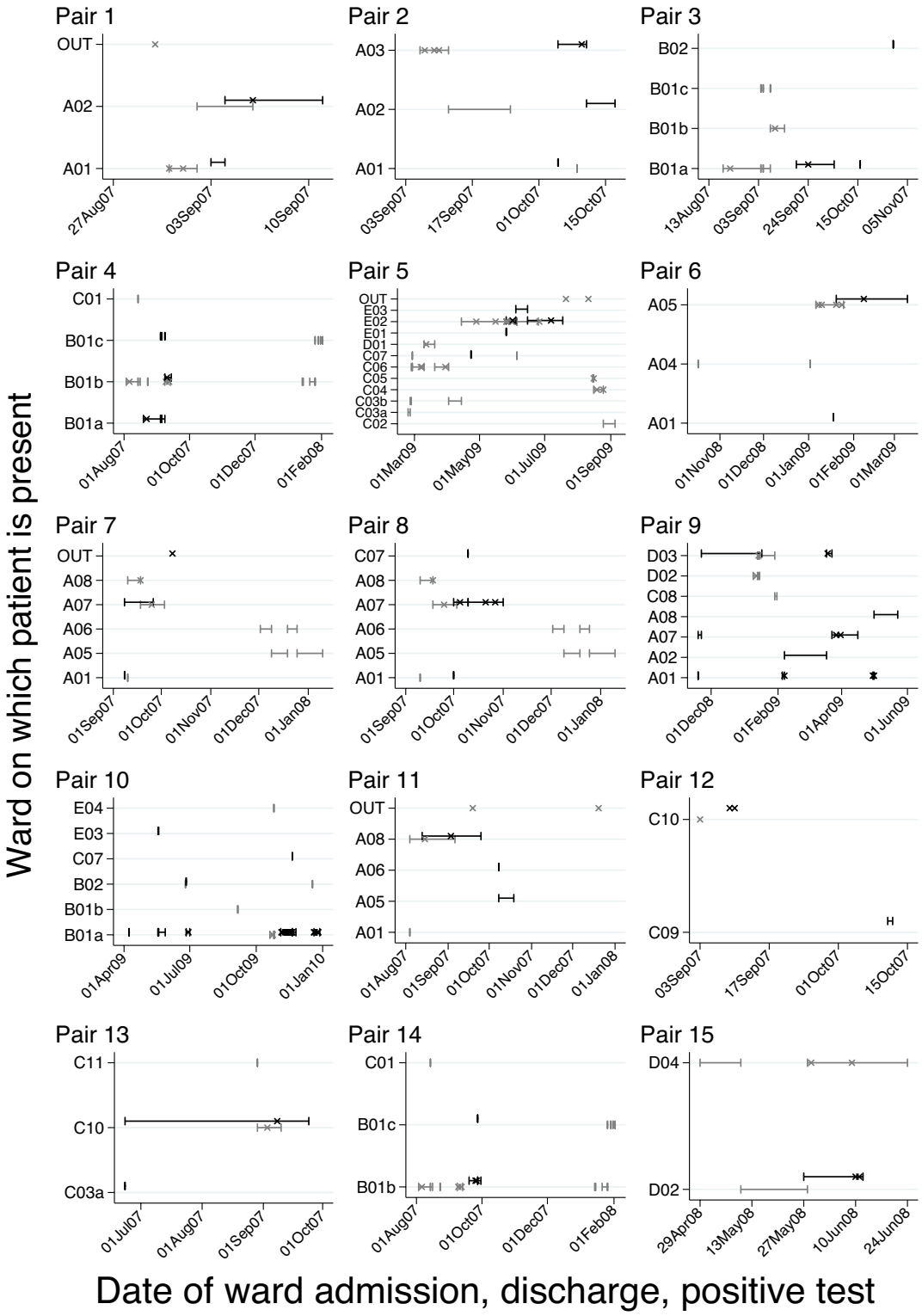


Figure 40. Epidemiological relationships between 15 potential donors and recipients of mixed infection transmission. Potential donors are shown in grey, and potential recipients in black. Time on a hospital ward around the time of diagnosis is shown as a horizontal line bounded by short vertical lines. Each hospital / hospital area is given a distinct letter, each ward a number, and groups of similar wards are given the

same number followed by a lower case letter. Positive samples for *C. difficile* are shown as crosses.

4.6.2.1.2 SAMPLE PREPARATION AND SEQUENCING

Twenty-six faecal samples from 12 putative donors and 15 putative recipients (3 donors had 2 putative recipients, and 1 recipient was also a putative donor) representing all 15 potential mixed-infection transmissions were cultured. Selective culture for *C. difficile* was undertaken following an alcohol-shock on modified Brazier's cycloserine-cefoxitin-egg yolk agar. Plates were incubated anaerobically at 37°C for up to 7 days.[145] DNA was extracted directly from a sweep taken across each primary culture plate to capture the complete genetic diversity present: a 5μl loop was passed through an area of confluent growth, and the loopful of growth then suspended in saline prior to DNA extraction with a commercial kit (QuickGene, Fujifilm, Tokoyo, Japan). All growth was morphologically consistent with *C. difficile*, exhibited a characteristic odour and fluoresced under ultraviolet light. Extracted DNA underwent whole genome sequencing using the Illumina HiSeq 2000 platform generating 100 base-pair reads. Sequence reads were mapped using two aligners, Stampy[179] (with an expected substitution rate of 0.01) and Burrows-Wheeler Aligner (BWA, with default settings)[226] to the *C. difficile* 630 reference genome (Genbank:AM180355), CD630.[182] High quality base counts were extracted from mapped data for variable sites using SAMtools,[180] retaining bases with a base quality score ≥30 and a mapping quality score ≥30. As the initial algorithm was designed to detect mixed ST infection, the variable sites analysed were first restricted to the 150 single nucleotide variants (SNVs) within the 7 MLST loci based on all published alleles.[227] The variable sites studied were subsequently

extended to make full use of the whole genome data, see results. To allow extraction of 16S ribosomal RNA genes, reads were also assembled *de novo* using Velvet with the Velvet Optimiser.[181]

4.6.2.1.3 ESTIMATION OF MIXTURE PROPORTION AND HAPLOTYPE DIVERGENCE

A maximum likelihood based method to detect mixed infection in bacterial WGS data was developed. This mixed infection estimator, MIE, is based on high quality base counts at sites known to vary on the basis of available previously sequenced isolates. Each sample was assumed to be a mixture of 2 haplotypes, resulting in one dominant and one minor haplotype, with the proportion of the total sequence present made up by the dominant haplotype denoted μ, the mixture proportion. For each sample analysed let:

N = total number of variable sites considered

n_j = total number of reads at a variable site j = 1…N

b_{ij} = an observed nucleotide from a single read i = 1…n_j mapped to variable site j, from the set {A, C, G, T}

B_j = a vector of the n_j nucleotides from the reads mapped to site j, $\left\{b_{1j}, b_{2j}, \ldots, b_{n_j j}\right\}$

ε = Pr(sequencing error in a base call). Assumed constant across all bases calls, having filtered the data to exclude low quality bases and reads

μ = mixture proportion, the proportion of the sample from the dominant haplotype ($0.5 \leq \mu \leq 1$)

a_{1j} = nucleotide in the dominant haplotype at site j

a_{2j} = nucleotide in the minor haplotype at site j

A_j = the combination of nucleotides in the dominant and minor haplotypes respectively, a_{1j} and a_{2j}, one of the set of all 16 possible pairs of nucleotides, A:

$$A \in \{\text{AA, AC, AG, AT, CA, ..., TT}\} \quad (1)$$

The probability of observing a particular nucleotide in a given read mapped to site j can be expressed in terms of the underlying dominant and minor haplotypes, the mixture proportion and error probability. The probability of a sequence error, ε, is assumed to be constant across all variable sites, and if an error occurs it is assumed to be equally likely to result in any of the three alternative nucleotides (i.e. if the true nucleotide is A, then a read containing C, G, or T is equally likely):

$$\Pr(b_{ij} \mid A_j, \mu, \varepsilon) = \begin{cases} \mu(1-\varepsilon) + (1-\mu)(1-\varepsilon) & \text{if } b_{ij} = a_{1j} = a_{2j} \\ \mu(1-\varepsilon) + (1-\mu)\frac{\varepsilon}{3} & \text{if } b_{ij} = a_{1j} \neq a_{2j} \\ \mu\frac{\varepsilon}{3} + (1-\mu)(1-\varepsilon) & \text{if } b_{ij} = a_{2j} \neq a_{1j} \\ \mu\frac{\varepsilon}{3} + (1-\mu)\frac{\varepsilon}{3} & \text{otherwise} \end{cases} \quad (2)$$

As ε is treated as a known constant, the probability of observing the n_j nucleotides mapped to site j, for a given value of A_j is:

$$\Pr(B_j \mid A_j, \mu) = \prod_{i=1}^{n_j} \Pr\left(b_{ij} \mid A_j, \mu\right) \quad (3)$$

Where A_j is unknown, summing over all possible values of A_j gives:

$$\Pr(B_j \mid \mu) = \sum_{A_j \in A} \prod_{i=1}^{n_j} \Pr\left(b_{ij} \mid A_j, \mu\right) \Pr\left(A_j\right) \quad (4)$$

To define $\Pr(A_j)$ for each possible value of A_j, let d be the proportion of all variable sites included in the analysis that are divergent between the dominant and minor haplotypes. At sites divergent between the haplotypes 12 possible pairs of nucleotides could be present, and at non-divergent sites 4 pairs of nucleotides are possible, such that:

$$\Pr(A_j \mid d) = \begin{cases} \dfrac{1-d}{4} & \text{if } A_j \in \text{homozygous: } \{\text{AA, CC, GG, TT}\} \\ \dfrac{d}{12} & \text{if } A_j \in \text{heterozygous: } \{\text{AC, AG, AT, ..., TG}\} \end{cases} \quad (5)$$

Combining (4) and (5), the probability of observing the n_j nucleotides mapped to site j, for a given mixture proportion and divergence between the dominant and minor haplotypes is:

$$\Pr(B_j \mid \mu, d) = \sum_{A_j \in A} \prod_{i=1}^{n_j} \Pr(b_{ij} \mid A_j, \mu) \Pr(A_j \mid d) \quad (6)$$

The values of μ and d were then jointly estimated from their likelihood:

$$L_{\mu,d} \propto \prod_{j=1}^{N} \Pr(B_j \mid \mu, d) \quad (7)$$

4.6.2.1.4 CONFIRMATION OF MIXED INFECTION

A major question is how to determine whether the data are consistent with a mixed infection. A simple but arbitrary approach would be fix thresholds of μ and d, e.g. $\hat{\mu} < 95\%, \hat{d} > 1$ SNV. However this ignores the certainty with which these parameters are estimated. Here an alternative approach was adopted: comparing the maximum likelihood obtained under the mixed infection model with the likelihood of the data without mixed infection, i.e. μ = 1, d=0. Comparing the log-likelihood ratio statistic to a chi-squared distribution is problematic, as the null hypothesis is on the edge of the parameter space. Therefore a calibration set of 100 samples known not to contain mixed infection (because they were sequenced from a single colony) was to determine a deviance (-2 log likelihood ratio) threshold for confirming mixed infection. The value of the threshold was chosen to achieve a 5% false positive rate.

This empirical approach of using a calibration set of actual sequences also has the advantage that it accounts for low-level sample contamination that may occur during sequencing. This would not be easily accounted for in another alternative for determining mixed infection, for example simulating under the null hypothesis in a bootstrapping approach.

Confidence intervals for μ and d were generated by non-parametric bootstrap sampling. The variable sites were sampled with replacement 1000 times keeping each B_j constant.

4.6.2.1.5 ESTIMATION OF HAPLOTYPES

A database of known sequences was used to estimate the dominant and minor haplotypes present, using the estimate of μ obtained above. From (6), for each potential haplotype pair the value of A_j and d are known, and therefore the probability of observing all the nucleotides mapped to site j is:

$$\Pr(B_j \mid \hat{\mu}) = \prod_{i=1}^{n_j} \Pr(b_{ij} \mid \hat{\mu}) \qquad (8)$$

Therefore the likelihood of a given pair of dominant and minor haplotypes given the sequence data is:

$$L_{pair} \propto \prod_{j=1}^{N} \Pr(B_j \mid \hat{\mu}) \qquad (9)$$

Finally, the posterior probability of each pair of haplotypes was obtained using the product of the prevalence of the dominant and minor haplotypes as their prior probability and the likelihood of each pair above. For MLST analyses the prevalence of each ST in Oxfordshire from September 2007 to March 2010 was used.[211] STs not found during the study but present in pubMLST[227] were assumed to be as prevalent as the least common ST. For whole genome analyses the prevalence of each unique sequence during the study was used.

When using whole genome data to assess whether the dominant and minor pair selected were consistent with the data, the estimated divergence and 95% confidence intervals were compared with the actual divergence between the database-matched dominant and minor sequences. Where the database-matched pairwise divergence fell within the estimated 95% confidence interval the pair was considered a good match. However when the divergence fell outside of the confidence interval this was interpreted as evidence that the true dominant or minor sequence was not present in the database.

For all analyses the value of ε was set to the sum of the base and error mapping rates, assuming each had a PHRED score of 30 (the thresholds used for determining high quality bases to retain in the analysis), i.e. $\varepsilon = 2 \times 10^{-3}$. This represents the upper bound on the value of ε after filtering. However results were similar with lower values of ε tested up to $\varepsilon = 2 \times 10^{-4}$. In fact, comparison of samples sequenced from a single colony, and therefore expected to be pure, with simulated read data using $\varepsilon = 2 \times 10^{-3}$ and assuming a pure sample suggests that the actual error rate may be greater than 2×10^{-3} despite filtering, probably reflecting low level sample contamination which technically has sequenced well. This is accounted for empirically by use of a calibration set to determine a threshold for mixed infection.

4.6.2.1.6 SIMULATED MIXED INFECTIONS

A combination of *in silico* and *in vitro* simulated mixed infections were used to test the performance of the method. To generate *in silico* simulated mixed infections two samples known not to be mixed themselves were mixed in varying proportions. Firstly, read depths were normalised across the two samples by

multiplying base counts in the sample with lower coverage to match the coverage in the other sample. Nucleotides were then sampled from reads mapped to each variable site (with replacement). The input sequence each read was sampled from was determined using a binomial distribution with parameters of normalised read depth, and input mixture proportion.

In vitro mixed infections were generated by mixing DNA extracted from single colonies. DNA from each sample was carefully quantified using picogreen technology (Tecan, Männedorf, Switzerland) and mixed in known proportions prior to sequencing.

4.6.2.2 RESULTS

4.6.2.2.1 CALIBRATION OF MIXED INFECTION DETECTION

To calibrate a threshold for identifying mixed infection using the MIE all reads from 100 whole genome sequences derived from a single colony, and thus expected not to be mixed, were initially analysed, investigating the 150 variable sites within the MLST loci for evidence of mixed ST infection. The ST previously obtained by PCR was recovered as the dominant strain using the MIE on all occasions and accounted for a median of 100% (interquartile range [IQR] 99.9-100%, range 93.5-100%) of the sample based on a median (IQR) read depth of 80 (67 – 93) (**Figure 41**). The divergence between the dominant and minor haplotypes was estimated at a median 0 SNVs (IQR 0 – 0 SNVs, range 0 – 15 SNVs). A likelihood ratio statistic was used to compare the maximum likelihood obtained under the mixed infection model, with the likelihood of the data without mixed infection. For each sample in the calibration set the deviance (-2 times the log

likelihood ratio) was calculated and the quantiles of its distribution used to set a threshold for calling mixed infection of ≥19.4 in order to achieve a 5% false-positive rate. This empirical approach to choosing the significance threshold avoids making unrealistic assumptions about the statistical distribution of the deviance under the null hypothesis of single infection.

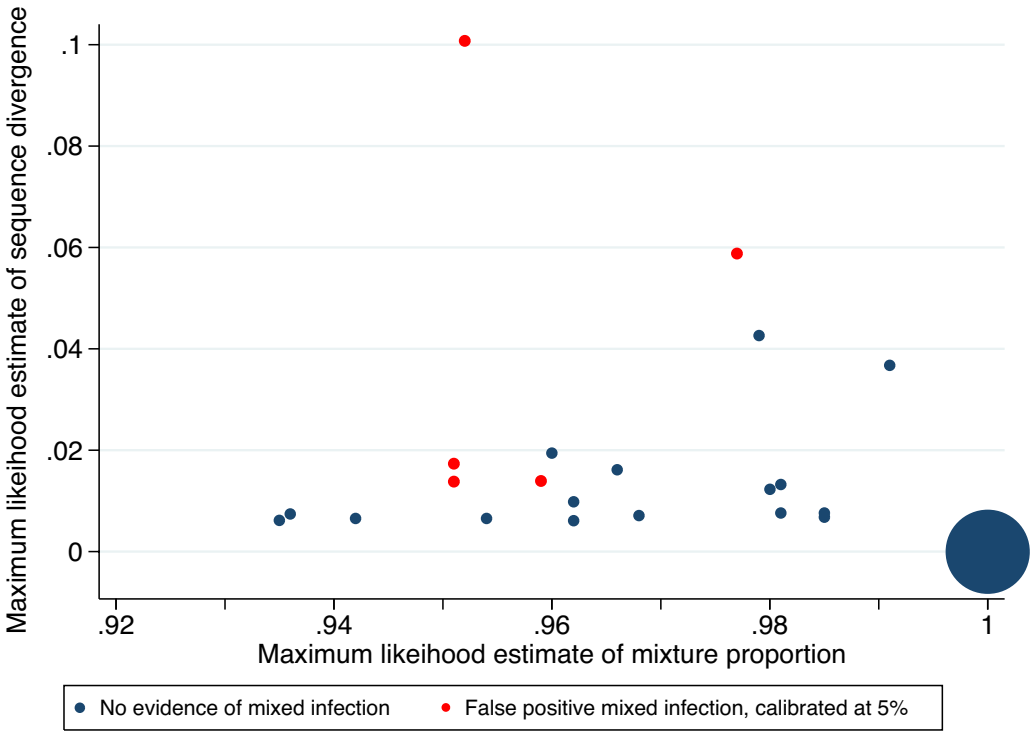


Figure 41. Estimated mixture proportion and sequence divergence for 100 sequences derived from a single colony. Markers are weighted by frequency. A likelihood ratio test was used to compare the maximum likelihood obtained under the mixed infection model, with the likelihood of the data without mixed infection. For each sample in the calibration set -2 times the log likelihood ratio was calculated and the quantiles of the distribution used to set a threshold for calling mixed infection of ≥19.4 in order to approximate a 5% false-positive rate. Under the model the mixture proportion, between sequence divergence and the read depth contribute to the likelihood of a mixed infection, hence the pattern seen in the red, false-positive, dots where lower mixture proportions, increased sequence divergence, and increased read depth (not shown) contribute to a significant result.

4.6.2.2.2 IN SILICO SIMULATED MIXED INFECTION

Simulated mixed infections were generated to test the ability of the MIE to detect mixed infections, and the constituent strains. Reads obtained from the unmixed samples above were mixed *in silico* to create 10000 simulated mixed ST infections with median (IQR) read depth 78 (67 - 90) and mixture proportions from 0.5 to 0.95. The known input mixture proportion was estimated with a root mean square error, RMSE, of 0.086 (see **Figure 42a** for the distribution of estimated mixed proportions across the 10 input proportions). Accurate mixture proportion estimates were obtained even when the simulated sequence divergence was as low as 3 SNVs between dominant and minor sequences. Mixture proportions closer to 1 were associated with a smaller variance and RMSE. The divergence between sequences at the MLST loci was estimated with a RMSE of 0.079, consistently across varying mixture proportions (**Figure 42b**). Having already set the specificity of the algorithm to 95% with the empirical calibration procedure above, the sensitivity of the MIE in this dataset for detection of simulated mixed infections was 99.0%.

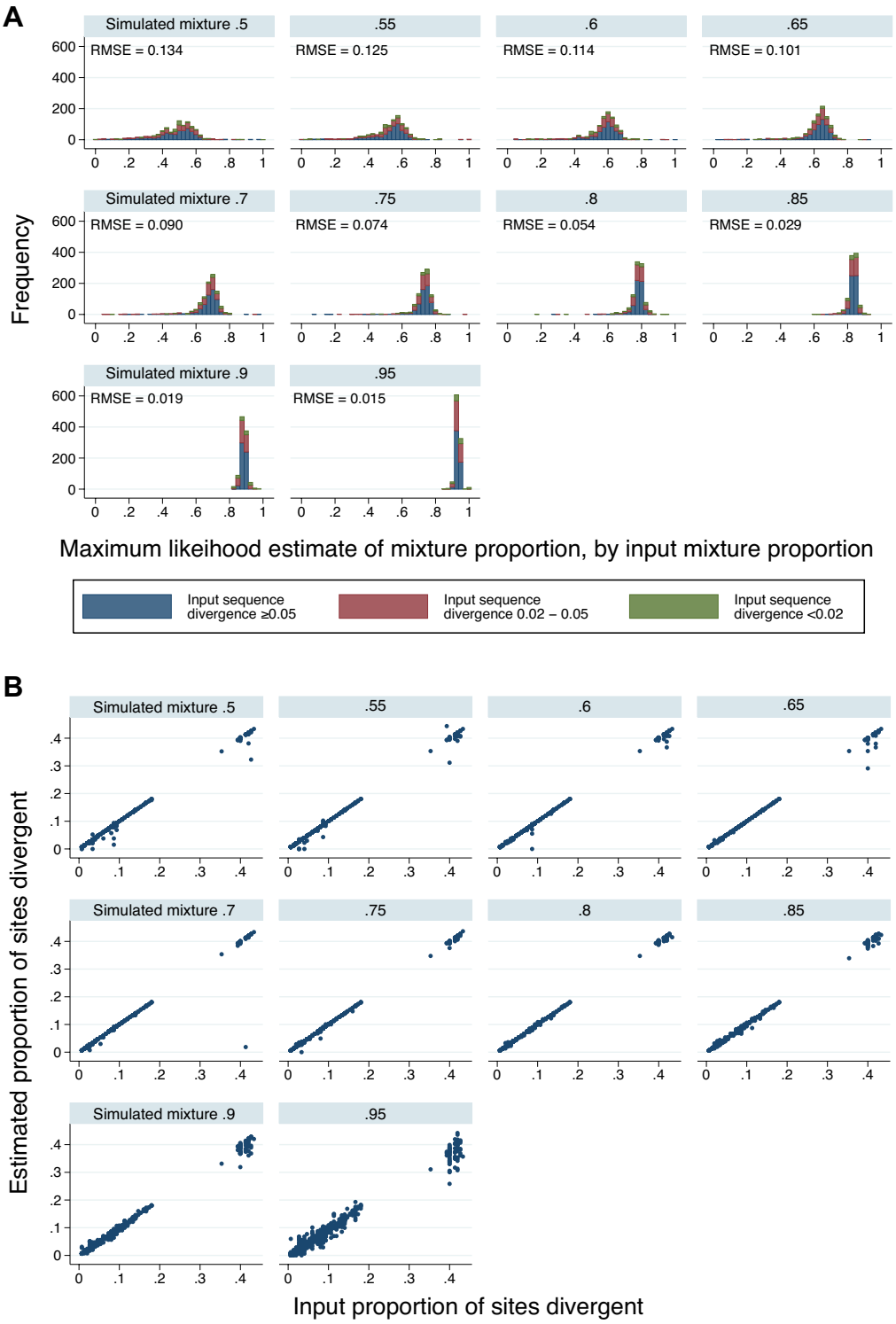


Figure 42. Estimation of the mixture proportion and sequence divergence across simulated mixtures. Panel A shows distribution of estimated mixture proportions over 1000 different pairs of input sequences, separately for 10 different input mixture proportions. The estimated mixture proportion shown is for the input dominant sequence. Panel B shows the relationship between the true proportion of sites divergent and the estimated proportion for the same simulations. RMSE, root mean square error.

The two input STs were recovered as the most likely pair on 9704/10000 occasions, 9151 with the correct ordering of the dominant and minor STs (**Figure 43**). As expected, the minor ST was less likely to be recovered when it made up a smaller proportion of the overall sequence. Recovery of neither input ST was associated with mixture proportions near 0.5 and relatively low divergence between input sequences (median input divergence 0.033 where neither ST recovered versus 0.073 in all other samples, p<0.001).

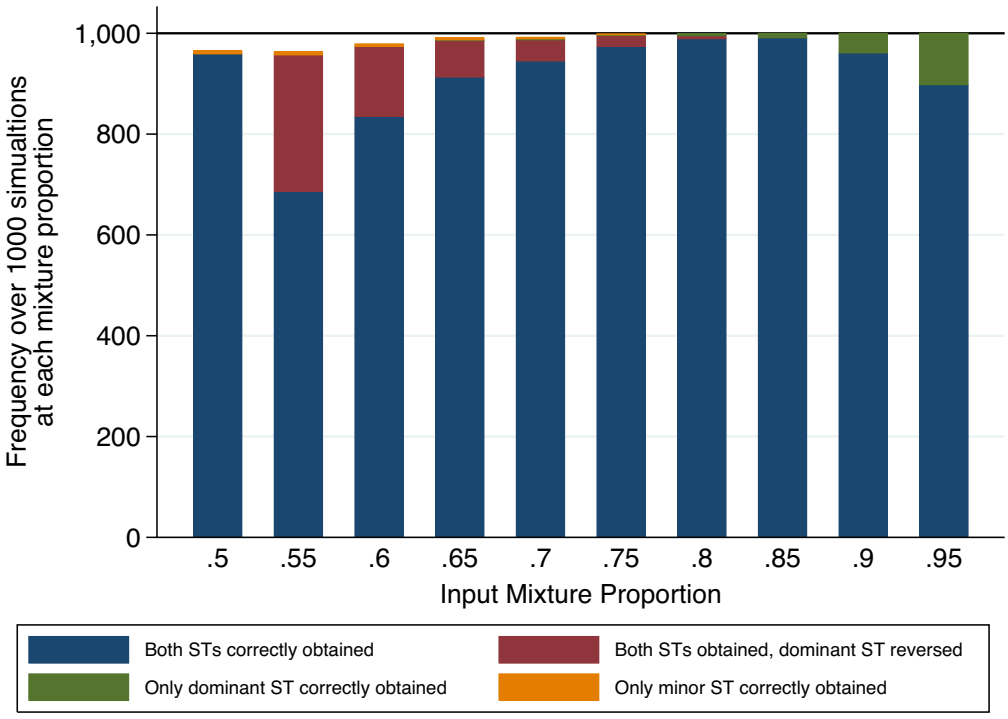


Figure 43. Estimation of dominant and minor ST across simulated mixtures. Estimation based on analysis of 150 variable sites within the MLST loci. At input mixture proportions of 0.5 either order of dominant and minor ST was considered correct.

4.6.2.2.3 IN VITRO SIMULATED MIXED INFECTION

To confirm the performance of the MIE *in vitro*, DNA extracted from single colonies was mixed in known proportions prior to sequencing. Thirty-six mixed ST infections were simulated: DNA from 12 single ST infections was mixed with DNA

from 12 different single ST infections, at 3 different mixture proportions – 50/50%, 70/30% and 90/10%. Using the MIE and whole genome data the input pair of dominant and minor haplotypes was obtained as the most likely on all occasions and the mixture proportion and divergence estimated with RMSEs of 0.032 and 0.002 respectively (**Figure 44**, **Table 7**).

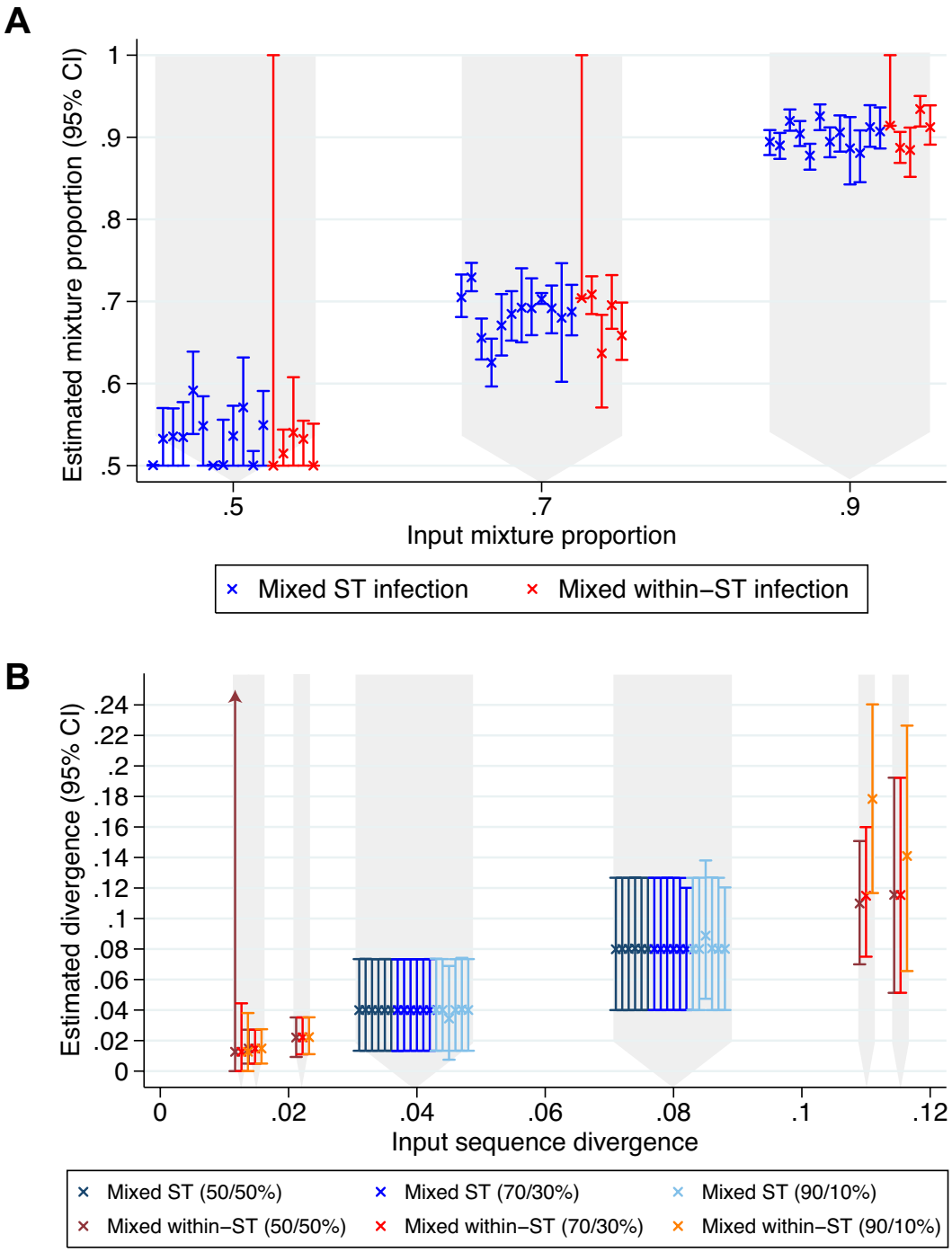


Figure 44. *In vitro* simulated mixed infections. Panel A shows the estimated mixture proportion for 3 input DNA mixture proportions. For ease of visualisation individual data points have different x-axis values, but correspond to the 3 x-axis values as indicated by the grey background. Points obtained from mixes of two differing STs are shown in red, and points from mixing two isolates of the same ST in blue. The large confidence intervals for the leftmost of each red group of samples is a sample with only a single variant site between the two input sequences, which when excluded in bootstrap sampling makes the sample appear unmixed. Panel B shows the estimated divergence between sequences for differing input divergence and mixture proportions. The leftmost group of 3 points

represent a sample with a single variant site, which when excluded in bootstrap sampling makes the sample appear unmixed and estimates of *d* unstable between 0 and 1.

Type	ST sample 1	ST sample 2	Input *d*, SNVs	Input μ	Maximum likelihood estimate μ (95% confidence interval)	Maximum likelihood estimate *d* in SNVs (95% confidence interval)	-2 log likelihood ratio (mixed vs. unmixed infection)	Estimated probability of input pair
Mixed-ST infection	1	2	12	0.5	0.50 (0.50 - 0.50)	12.0 (6.0 - 19.0)	6239.8	1.000
Mixed-ST infection	14	67	12	0.5	0.53 (0.50 - 0.57)	12.0 (6.0 - 19.0)	4515.1	1.000
Mixed-ST infection	1	44	12	0.5	0.54 (0.50 - 0.57)	12.0 (6.0 - 19.0)	3275.3	0.997
Mixed-ST infection	1	16	12	0.5	0.53 (0.50 - 0.58)	12.1 (6.0 - 19.0)	4537.8	1.000
Mixed-ST infection	1	44	12	0.5	0.59 (0.54 - 0.64)	12.0 (6.0 - 19.0)	3209.6	1.000
Mixed-ST infection	1	2	12	0.5	0.55 (0.50 - 0.58)	12.0 (6.0 - 19.0)	3392.6	1.000
Mixed-ST infection	34	33	6	0.5	0.50 (0.50 - 0.50)	6.0 (2.0 - 11.0)	2136.4	1.000
Mixed-ST infection	14	44	6	0.5	0.50 (0.50 - 0.56)	6.0 (2.0 - 11.0)	2487.8	1.000
Mixed-ST infection	8	13	6	0.5	0.54 (0.50 - 0.57)	6.0 (2.0 - 11.0)	2476.4	1.000
Mixed-ST infection	14	44	6	0.5	0.57 (0.50 - 0.63)	6.0 (2.0 - 11.0)	2828.1	1.000
Mixed-ST infection	8	49	6	0.5	0.50 (0.50 - 0.52)	6.0 (2.0 - 11.0)	3639.6	0.333
Mixed-ST infection	2	54	6	0.5	0.55 (0.50 - 0.59)	6.0 (2.0 - 11.0)	3423.9	0.958
Mixed-ST infection	2	1	12	0.7	0.71 (0.68 - 0.73)	12.0 (6.0 - 19.0)	3444.5	1.000
Mixed-ST infection	67	14	12	0.7	0.73 (0.71 - 0.75)	12.0 (6.0 - 19.0)	3417.9	1.000
Mixed-ST infection	44	1	12	0.7	0.66 (0.63 - 0.68)	12.0 (6.0 - 19.0)	3668.2	1.000
Mixed-ST infection	16	1	12	0.7	0.63 (0.60 - 0.65)	12.0 (6.0 - 19.0)	3483.5	1.000
Mixed-ST infection	44	1	12	0.7	0.67 (0.63 - 0.71)	12.0 (6.0 - 19.0)	2918.8	1.000
Mixed-ST infection	2	1	12	0.7	0.68 (0.65 - 0.71)	12.0 (6.0 - 18.0)	3401.8	1.000
Mixed-ST infection	33	34	6	0.7	0.69 (0.65 - 0.74)	6.0 (2.0 - 11.0)	1847.1	1.000
Mixed-ST infection	44	14	6	0.7	0.69 (0.66 - 0.73)	6.0 (2.0 - 11.0)	1784.5	1.000
Mixed-ST infection	13	8	6	0.7	0.70 (0.70 - 0.71)	6.0 (2.0 - 11.0)	1305.9	1.000
Mixed-ST infection	44	14	6	0.7	0.69 (0.66 - 0.72)	6.0 (2.0 - 11.0)	2326.4	1.000

Mixed-ST infection	49	8	6	0.7	0.68 (0.60 - 0.75)	6.0 (2.0 - 11.0)	1553.4	1.000
Mixed-ST infection	54	2	6	0.7	0.69 (0.66 - 0.72)	6.0 (2.0 - 11.0)	2120.8	1.000
Mixed-ST infection	1	2	12	0.9	0.89 (0.88 - 0.91)	12.0 (6.0 - 19.0)	1112.4	1.000
Mixed-ST infection	14	67	12	0.9	0.89 (0.87 - 0.91)	12.0 (6.0 - 19.0)	866.4	1.000
Mixed-ST infection	1	44	12	0.9	0.92 (0.91 - 0.93)	13.3 (7.1 - 20.7)	526.0	1.000
Mixed-ST infection	1	16	12	0.9	0.90 (0.89 - 0.92)	12.1 (6.0 - 19.0)	785.2	1.000
Mixed-ST infection	1	44	12	0.9	0.88 (0.86 - 0.89)	12.0 (6.0 - 19.0)	1072.7	1.000
Mixed-ST infection	1	2	12	0.9	0.93 (0.91 - 0.94)	12.0 (6.0 - 18.1)	576.1	0.995
Mixed-ST infection	34	33	6	0.9	0.89 (0.88 - 0.91)	6.0 (2.0 - 11.0)	382.4	1.000
Mixed-ST infection	14	44	6	0.9	0.91 (0.88 - 0.93)	6.0 (2.0 - 11.0)	401.5	1.000
Mixed-ST infection	8	13	6	0.9	0.89 (0.84 - 0.92)	5.2 (1.1 - 10.3)	317.5	1.000
Mixed-ST infection	14	44	6	0.9	0.88 (0.85 - 0.91)	6.0 (2.0 - 11.0)	540.9	1.000
Mixed-ST infection	8	49	6	0.9	0.91 (0.89 - 0.94)	6.1 (2.0 - 11.1)	228.5	1.000
Mixed-ST infection	2	54	6	0.9	0.91 (0.89 - 0.94)	6.0 (2.0 - 11.0)	360.1	1.000

Table 7. Estimated mixture proportion and sequence divergence for 36 *in vitro* mixed-ST infections. DNA from two previously sequenced isolates of differing sequence types (ST) was mixed in 3 proportions 50/50%, 70/30%, 90/10% for 12 pairs of isolates to created 36 mixed ST infections. At input mixture proportions of 0.5 either order of dominant and minor ST was considered correct. A likelihood ratio statistic was used to compare the maximum likelihood obtained under the mixed infection model, with the likelihood of the data without mixed infection. Samples with a -2 log likelihood ratio ≥19.4 were considered mixed (see calibration set results).

In order to demonstrate the MIE is also able to detect mixed infections where the two infecting strains are of the same ST, but differ at a whole genome level, mixed infections of the same ST were simulated. A database of all sequenced Oxfordshire isolates was used to determine the variable sites within each ST across the rest of the whole genome. These variable sites were analysed for evidence of within-ST mixed infection, using the same algorithm as for mixed ST infection, determining the most likely dominant and minor sequences from the unique whole genome

sequences within each ST in the database. Fifteen within-ST mixed infections were generated with DNA from 5 pairs of isolates sharing the same ST (STs 1, 3, 8, 14, 46), but with differing whole genome sequences, at 3 different mixture proportions (50/50%, 70/30% and 90/10%). The correct dominant and minor sequences were obtained on all occasions (**Table 8**). Mixture proportions and between sequence divergence were accurately estimated with a single exception where the within sequence divergence was somewhat over-estimated in a 90/10% mix (**Figure 44**). Accurate estimation of mixture proportions was possible even in a mixed infection where the samples differed only by a single site, with estimated mixture proportions of 0.50, 0.70 and 0.92 for input values of 0.50, 0.70 and 0.90. The median (IQR) read depths for these simulations were 82 (72-91).

Type	ST sample 1	ST sample 2	Input *d*, SNVs	Input μ	Maximum likelihood estimate μ (95% confidence interval)	Maximum likelihood estimate *d* in SNVs (95% confidence interval)	-2 log likelihood ratio (mixed vs. unmixed infection)	Estimated probability of input pair
Within-ST mixed infection	1	1	1	0.5	0.50 (0.50 - 1.00)	1.0 (0.0 - 78.7)	510.7	1.000
Within-ST mixed infection	3	3	12	0.5	0.51 (0.50 - 0.54)	11.9 (5.0 - 19.0)	4874.9	1.000
Within-ST mixed infection	8	8	6	0.5	0.54 (0.50 - 0.61)	6.0 (2.0 - 11.0)	1695.9	0.995
Within-ST mixed infection	14	14	22	0.5	0.53 (0.50 - 0.55)	22.0 (14.0 - 30.2)	11975.4	1.000
Within-ST mixed infection	46	46	9	0.5	0.50 (0.50 - 0.55)	9.0 (4.0 - 15.0)	3374.6	1.000
Within-ST mixed infection	1	1	1	0.7	0.70 (0.70 - 1.00)	1.0 (0.0 - 3.5)	323.9	1.000
Within-ST mixed infection	3	3	12	0.7	0.71 (0.68 - 0.73)	12.0 (6.0 - 19.0)	3298.9	1.000
Within-ST mixed infection	8	8	6	0.7	0.64 (0.57 - 0.68)	6.0 (2.0 - 11.0)	1631.9	1.000
Within-ST mixed infection	14	14	22	0.7	0.70 (0.67 - 0.73)	23.0 (15.0 - 32.0)	6839.6	1.000
Within-ST mixed infection	46	46	9	0.7	0.66 (0.63 - 0.70)	9.0 (4.0 - 15.0)	2760.8	1.000
Within-ST mixed infection	1	1	1	0.9	0.91 (0.91 - 1.00)	1.0 (0.0 - 3.0)	57.6	1.000
Within-ST mixed infection	3	3	12	0.9	0.89 (0.87 - 0.91)	12.1 (6.0 - 19.1)	841.6	1.000
Within-ST mixed infection	8	8	6	0.9	0.88 (0.85 - 0.91)	6.0 (2.0 - 11.1)	279.4	1.000
Within-ST mixed infection	14	14	22	0.9	0.93 (0.91 - 0.95)	35.7 (23.3 - 48.1)	1136.4	1.000
Within-ST mixed infection	46	46	9	0.9	0.91 (0.89 - 0.94)	11.0 (5.1 - 17.7)	488.9	1.000

Table 8. Estimated mixture proportion and sequence divergence for 15 *in vitro* within-ST mixed infections. DNA from two previously sequenced isolates of the same sequence type was mixed in the same proportions for 5 pairs of isolates to create 15 within-ST mixed infections. At input mixture proportions of 0.5 either order of dominant and minor ST was considered correct. A likelihood ratio statistic was used to compare the maximum likelihood obtained under the mixed infection model, with the likelihood of the data without mixed infection. Samples with a -2 log likelihood ratio ≥19.4 were considered mixed (see calibration set results).

4.6.2.2.4 MIXED INFECTION TRANSMISSION SAMPLES

Having confirmed the accurate performance of the MIE, it was applied to the 15 potential mixed infection transmissions described above where transmission was

highly plausible based on hospital contacts but the STs obtained from sequencing single colonies differed (**Figure 40**).

When sequence data from the 26 samples was aligned, the Burrows Wheeler Aligner, BWA[226], outperformed Stampy[179] because a number of samples also included sequence from non-*C. difficile* anaerobic bacteria. DNA for sequencing was obtained from an area of confluent growth on primary culture plates, and despite the use of selective agar and individual colonies resembling *C. difficile*, other similar antibiotic-resistant anaerobic gut bacteria were detected in some samples, by extracting 16S ribosomal RNA genes using BLAST[200] from *de novo* assemblies[181] of the sequences and comparing these with the Ribosomal Database Project[228]. Sequence reads from these other species do not map or map poorly to the reference genome, therefore the percentage of reads mapped with Stampy to the CD630 reference ranged from 11.0% – 95.4%, with 15/26 samples having <60% of reads mapped. As Stampy is designed to perform well with relatively large sequence variation relative to the reference, in the more contaminated samples markedly divergent reads were mapped to the MLST loci. These reads must have arisen from other species as such divergence would not be expected within the highly conserved housekeeping genes of the MLST loci within *C. difficile*. These divergent reads were interpreted by the MIE as mixed infection, such that a clear relationship was seen between samples estimated to contain mixed infection based on Stampy mapping and those with low percentages of reads mapped to the reference (**Figure 45a**). Therefore all samples were remapped with BWA to increase the penalties associated with insertions and deletions relative to the reference such that only reads arising from *C. difficile*

would map to the MLST loci. This allowed assessment of the proportion of mixed infection across all 26 samples (**Figure 45b**). Reductions in read depth were modest with BWA compared to Stampy (overall median (IQR) read depth was 21 (15 – 80) with Stampy across 26 samples, versus 18 (13 -75) with BWA).

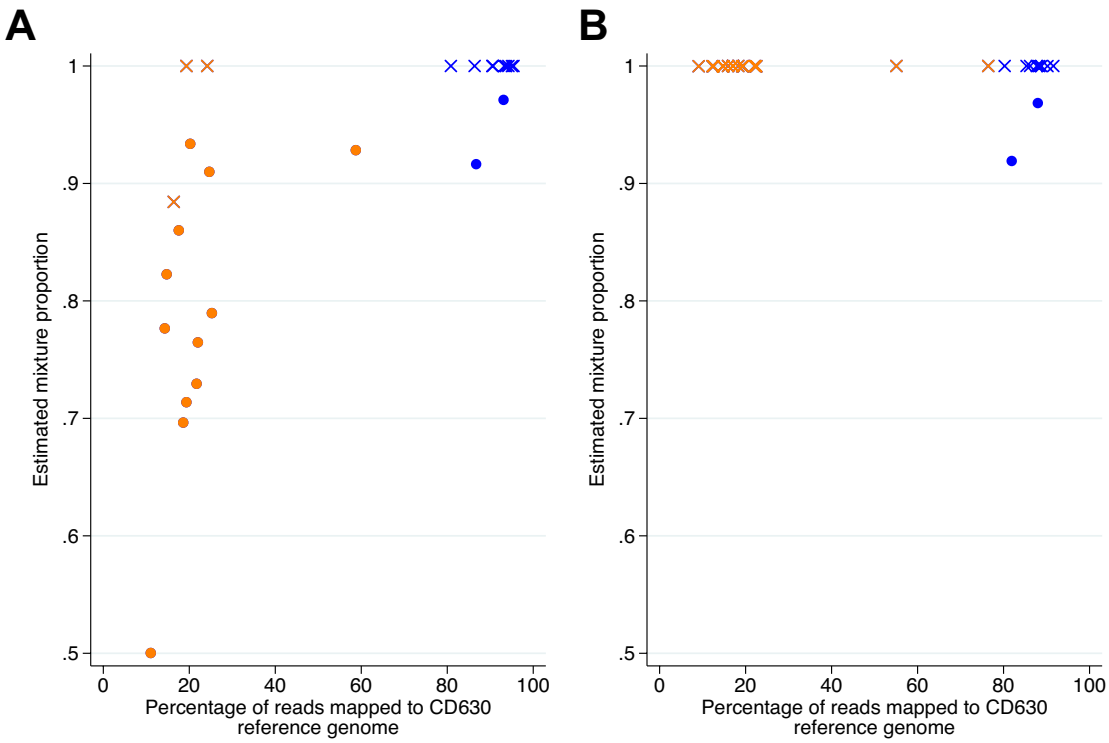


Figure 45. Reads mapped and estimated mixture proportion for possible mixed-ST clinical infections, across two alignment programs. Points in orange show evidence of contamination with other bacteria (i.e. <80% of reads mapped to reference genome), other points are shown in blue. Mixed infections detected using a -2 log likelihood ratio statistic threshold of ≥19.4 (as defined in the calibration samples) are shown as filled circles, other points are shown as crosses. Panel A shows the data obtained from alignments generated using Stampy. Stampy is designed to perform well with relatively high sequence variation relative to the reference, in particular insertions or deletions. In the more contaminated samples markedly divergent reads from other species were observed mapped to the highly conserved MLST loci resulting in falsely identifying mixed infections. Panel B show the data obtained from alignments generated with Burrows Wheeler Aligner. Two mixed infections were detected.

Using the MIE with whole genome data, overall 2 of 26 cases (8% of cases, 95% confidence interval, CI, 1-25%) had evidence of mixed ST infection, coincidently in

the same transmission model pair, pair 13 (see **Figure 40**). The estimated dominant ST matched the original ST from MLST PCR in all 26 cases. The putative donor with a mixed ST infection (donor A with a ST1 dominant infection) had a minor ST46 infection that accounted for 3% (95%CI 2-4%) of the sample. However, this did not concord with the dominant ST17 infection found in the putative recipient (recipient A). This recipient in turn had a minor ST recovered, ST1, with sample frequency 8% (95%CI 6-10%), which was compatible with acquisition from donor A. Therefore one of the donor-recipient ST matches predicted by the stochastic transmission model on the basis of shared time and space in the hospital but ruled out by single colony sequencing appeared to be explained by mixed infection (**Figure 46**).

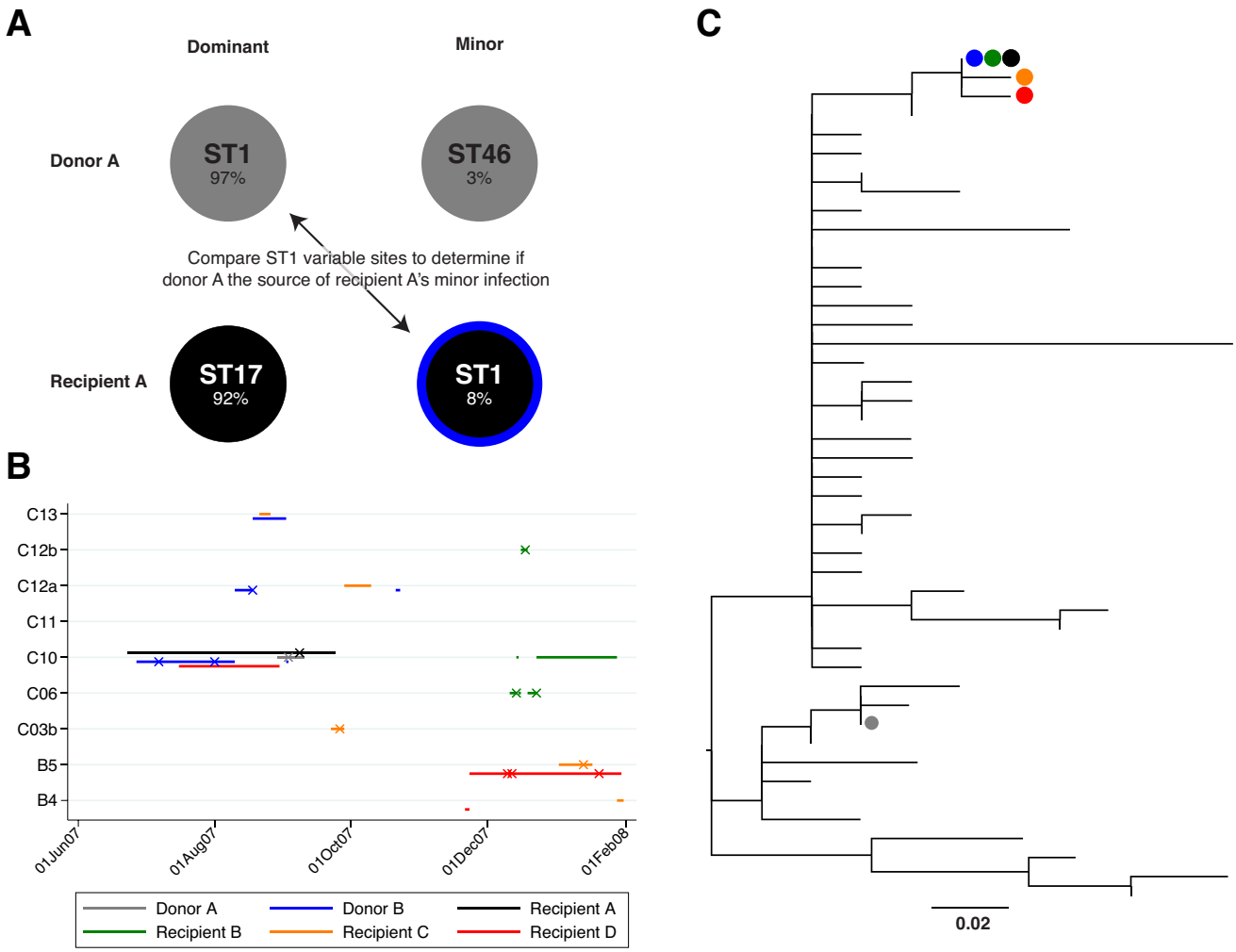


Figure 46. Phylogenetic and epidemiological relationships between cases related to a detected mixed infection. Panel A shows a depiction of the 2 mixed infections identified in donor A and recipient A. A transmission event from donor A to recipient A was predicted by a stochastic transmission model based on ward admission data. However donor A and recipient A had differing multilocus sequence types (STs) on initial testing of a single isolate from each case, suggesting a possible undetected mixed infection. Using the mixed infection estimator a minor ST infection in recipient A was found sharing the same ST, ST1 as donor A. However, applying the MIE to variable sites within ST1, the minor sequence in recipient A was most likely to have arisen from another case, donor B, shown in blue. Panel B shows the epidemiological relationships between donor A, recipient A, donor B and cases sharing similar sequences. Ward stays are shown as horizontal lines and positive tests as crosses. Panel C shows a phylogenetic tree of 45 distinct whole genome sequences from Oxfordshire patients with ST1 *Clostridium difficile* infection. Maximum likelihood tree based on 79 variable sites identified drawn using PhyML.[229] The donor proposed by the transmission model is shown in grey (donor A). The minor sequence in recipient A is shown in black, matching the sequence found in donor B, in blue. Recipient B shared an identical sequence to recipient A. Recipients C and D are two cases phylogenetically descended from the donor B, recipient A, recipient B sequences. Note only donor A and recipient A were analysed for the presence of mixed infection.

To scrutinize in more detail whether the WGS data were compatible with transmission of the dominant ST1 infection in donor A to recipient A as a minor infection, a panel of 45 unique Oxfordshire ST1 genomes was exploited to assist in whole genome prediction of the dominant and minor haplotypes in both cases (**Figure 46a**). Informally, the method compared recipient A's minor ST1 sequence to all 45 ST1 whole genome sequences in the Oxfordshire database (which included an ST1 genome sequenced from a single colony from donor A) using a total of 79 ST1-specific SNVs across the whole genome. The most likely recipient A minor sequence (posterior probability=0.9997) was from another patient (donor B), and differed by 8 SNVs scattered throughout the genome from the sequence found in donor A (**Figure 46c**). In fact, donor B represents a substantially more plausible donor than donor A identified by the stochastic transmission model on the basis of epidemiological data alone, because of the short-term rate of evolution in *C. difficile*, estimated at <1 SNV/genome/year in section 4.4.1. Donor B was also epidemiologically linked to recipient A, albeit less strongly than donor A. Recipient A was diagnosed on day 77 of a 93-day admission on a surgical ward. Donor B was diagnosed 63 days earlier and spent 34 days after diagnosis on the same ward as the recipient, and was also readmitted for 2 days to the same ward, 6 days before the recipient's diagnosis (**Figure 46b**). Not only does this reiterate the power of WGS for differentiating potential transmission donors that appear identical on the basis of low-resolution genotyping alone, it also demonstrates that the method is able to extend the approach to mixed infections and identify the source of the minor strain.

There was no strong evidence for onward transmission from the minor sequence in the mixed infection in recipient A. A single further case (Figure 4b,c, recipient B) with the identical sequence was identified, but the patient had not shared time or space in hospital with donor B or recipient A prior to diagnosis. Given the relatively high prevalence of ST1 and its relatively low diversity even at the whole genome level, indirect transmission via community contact or an undiagnosed third party is the most likely explanation. Additionally, only two descendant sequences (**Figure 46**, recipient C, recipient D) were identified from the phylogenetic tree of all Oxfordshire ST1s (**Figure 46c**). Both patients shared time on the same ward with donor B, but not with the mixed infection case (recipient A) after this case's diagnosis. Therefore donor B may have been the source of onward transmissions, but probably not the mixed infection recipient A.

Having found evidence of mixed infections with differing STs, the MIE was applied to search for previously undetected mixed infections of the same ST in the 24 putative donors and recipients without evidence of mixed-ST infection (**Table 9**). Five samples contained evidence of mixed infection according to the MIE. In three cases the divergence estimated between dominant and minor sequences was consistent with levels of within host diversity observed in serially sampled patients where up to 2 SNVs were expected between samples taken on the same day (95% prediction interval, section 4.4.1). As such these cases might not have arisen from two transmission events, but from evolution within a host of the same strain. Discounting these 3 cases, 2 mixed infections of the same ST were identified, to add to the 2 mixed infection cases identified with different STs. The two within-ST mixed infection cases had an estimated divergence between the

dominant and minor sequences that differed substantially from the best matching sequences in the database, 542 SNVs and 56 SNVs compared to best matches in the database of 1022 SNVs and ≤4 SNVs respectively (**Table 9**). This suggests the true minor sequence was not present in the database, highlighting that whilst the method developed works best with an established database of local sequences, nevertheless it still is able to identify when novel sequences arise.

Sequence type	Variable sites analysed	Database unique genomes	-2 log likelihood ratio (mixed vs. unmixed infection)	Maximum likelihood estimate mixture proportion, μ (95% confidence interval)	Maximum likelihood estimate divergence, d in SNVs (95% confidence interval)	Database matched dominant / minor pairs within 95% cumulative probability	SNVs between database-matched dominant and minor pairs
1	78	45	<0.1	1.00 (1.00 - 1.00)	0.0 (0.0 - 0.0)		
	79		<0.1	1.00 (1.00 - 1.00)	0.0 (0.0 - 0.0)		
	79		<0.1	1.00 (1.00 - 1.00)	0.0 (0.0 - 0.0)		
	79		<0.1	1.00 (1.00 - 1.00)	0.0 (0.0 - 0.0)		
	79		<0.1	1.00 (1.00 - 1.00)	0.0 (0.0 - 0.0)		
2	1613	68	56.8	0.68 (0.56 - 1.00)	3.1 (0.0 - 6.4)	1	6
	1621		<0.1	1.00 (1.00 - 1.00)	0.0 (0.0 - 0.0)		
3	541	38	<0.1	1.00 (1.00 - 1.00)	0.0 (0.0 - 0.0)		
	541		<0.1	1.00 (1.00 - 1.00)	0.0 (0.0 - 0.0)		
	541		<0.1	1.00 (1.00 - 1.00)	0.0 (0.0 - 0.0)		
5	1310	41	568.6	0.97 (0.96 - 0.98)	541.6 (411.0 - 763.7)	1	1022
6	828	65	<0.1	1.00 (1.00 - 1.00)	0.0 (0.0 - 0.0)		
	828		<0.1	1.00 (1.00 - 1.00)	0.0 (0.0 - 0.0)		
	828		<0.1	1.00 (1.00 - 1.00)	0.0 (0.0 - 0.0)		
9	289	19	<0.1	1.00 (1.00 - 1.00)	0.0 (0.0 - 0.0)		
10	213	39	15.4	0.95 (0.91 - 1.00)	8.6 (0.0 - 34.0)		
16	3012	13	8.4	0.99 (0.97 - 1.00)	449.4 (32.4 - 3011.9)		
17	133	14	133.2	0.84 (0.69 - 1.00)	3.8 (0.0 - 8.5)	1	41
	133		75.9	0.86 (0.86 - 1.00)	1.2 (0.0 - 3.8)	1	69
18	17	5	<0.1	1.00 (1.00 - 1.00)	0.0 (0.0 - 17.0)		
35	586	9	10.8	0.83 (0.83 - 1.00)	1.0 (0.0 - 3.0)		
42	163	23	<0.1	1.00 (1.00 - 1.00)	157.2 (0.0 - 163.0)		
44	3513	55	24.4	0.98 (0.98 - 0.99)	55.9 (27.0 - 91.9)	13	1 - 4
63	38	3	<0.1	1.00 (1.00 - 1.00)	0.0 (0.0 - 0.0)		

Table 9. Within ST mixed infection. Analysis for mixed infections with two lineages of the same sequence type is shown in 24 donors/recipients without mixed sequence type infection. The number of variable sites within each sequence type contributing to the analysis is shown together with the number of unique sequences of the sequence type present in the database. A likelihood ratio statistic was used to compare the maximum likelihood obtained under the mixed infection model, with the likelihood of the data without mixed infection. Samples with a -2 log likelihood statistic ≥19.4 were considered mixed (see calibration set results). The probability of each possible dominant/minor sequence pair from the database being the true pair was calculated, and pairs arranged in descending probability. The number of pairs accounting for ≥95% probability is shown in the penultimate column, with the range of SNVs present between the most likely dominant-minor pairs (within 95% cumulative probability) in the last column. This can be compared to the estimated divergence between dominant and minor sequences which is estimated independently and provides a check on whether the pairs identified within the database are consistent with the divergence between the dominant and minor sequences.

4.6.2.3 DISCUSSION

This is a novel approach for detecting mixed infection from bacterial whole genome sequence data with low SNV density, that utilizes a computationally efficient algorithm that performs well in *in silico* and *in vitro* simulations. This offers the prospect of screening for mixed infection in transmission studies and routine outbreak surveillance without labour and cost-intensive individual sub-culture of multiple colony picks and the expense of typing or sequencing these isolates separately. The utility of the approach, which is generalizable to any bacterial pathogen/loci for which a database of known sequences exists, is demonstrated using both WGS and MLST in *C. difficile*.

This new approach revealed a number of biologically meaningful findings. In the sample of clinical cases significantly enriched for the possibility of mixed ST infection due to the high prior probability of transmission based on epidemiological data but without matching sequence types, only 2/26 (8%) cases had evidence of a mixed ST infection. This is consistent with previous estimates for mixed genotype infections of ~7-13% [193-197], although a higher prevalence

might have been expected had mixed infection genuinely been contributing to transmission. These previously undetected mixed infection events could not account for any transmission between the 15 putative donor-recipient pairs sequenced. The fact that these pairs were highly selected based on hospital exposure suggests that mixed infection is unlikely to explain a large proportion of the ~75% CDI cases that could not be linked to a previous case based on hospital ward exposure in chapter 3. However, by interrogating a larger database of >1200 genomes representing potential donors that were previously sequenced from single colonies, an example of transmission leading to mixed infection was found. Using the extra resolution afforded by whole genome data it was possible to refine the estimate of the likely donor. This revealed a previously undetected transmission event, reflecting additional transmission from this likely donor, and another transmission event on the ward in question. The significance of the infection for the recipient is unclear; but as ST1 is a virulent strain, it is possible that whilst it only accounted for the minority of the *C. difficile* sequenced it may nevertheless have been the cause of the patient's illness. Therefore, use of the MIE in outbreak investigation demonstrably offers the ability to detect additional transmission as well as robust determination that true transmission events are not being missed. However, as no evidence was found for onward transmission of *C. difficile* from a mixed infection case, use of this method may not be required in large transmission studies where missing a few uncommon transmission events would not affect the overall inference. It was also possible to detect mixed infections where both infections shared the same sequence type in 2/24 (8%) cases without mixed ST infections and detect likely within host variants in 3 further cases.

In order to capture the full diversity of *C. difficile* present on the primary culture plate a sweep was taken across all the growth. In around half the samples this resulted in contamination of the sequenced reads with other bacterial species, despite morphological appearances consistent with *C. difficile*. This necessitated use of a restrictive mapping algorithm (BWA) favouring mapping reads closely related to the reference. Differences in the proportion of mixed infections estimated by the same algorithm from these two mapping methods highlight the impact of such choices on inferences made from whole genome data. The principle advantage of the primary culture sweep is the ability to capture the full diversity present on the plate, rather than selecting a limited number of colonies for sub-culture. However if contamination with other bacterial species is a concern, one possible refinement still enabling an assessment of mixed infection, without expensive sequencing of multiple single picks, might be to sample multiple individual colonies from the primary culture plate, and sub-culture these together on a single plate prior to sequencing. Paradoxically such approaches may actually increase sensitivity even if only relatively modest numbers of colonies are sampled (e.g. 10 – 20 colonies). This is because high levels of contamination result in reduced read depths for a given sequencing effort, as evidenced by the lower median depth achieved in samples with <60% of reads mapped, 14, compared to 77 in samples with ≥60% of reads mapped. However, further sub-culture does risk increasing any biases introduced by differential growth of strains on culture media, relative to their original frequency within the host. The sequencing process itself can also potentially introduce read frequency biases, however this did not appear to have a significant impact in the *in vitro* simulations performed.

Although the MIE assumes mixtures contain only 2 sequences, it still should detect the presence of mixed infection where there is a dominant sequence and several minor sequences. The estimated minor sequence would be expected to be a hybrid of the true minor sequences. This might be apparent where the minor sequence did not match a known sequence, or where >2 nucleotides were found at a single SNV site. A possible hybrid minor sequence could then prompt further more detailed investigation including sub-culture of individual isolates. In the case of *C. difficile*, mixed infection with more than 2 genotypes is reported, but the majority of mixed infections are with 2 genotypes.[194,195] When comparing bacterial sequences, SNVs are often sparsely distributed throughout the genome, making it likely that a mixture with several minor sequences would still only contain biallelic SNVs. Therefore instead of the two-stage approach of estimating the mixture proportions followed by haplotype matching used here, any future approach to detect mixtures of more than 2 bacterial sequences would have to jointly estimate mixture proportions and haplotypes. Such an approach would still have to make use of a library of known haplotypes, given the limited numbers of SNVs relative to read lengths. As the current approach also depends on access to a database of known haplotypes this emphasises the benefits of read archives which could enable sequence data generated by different researchers to be incorporated into such a database. Future availability of long read sequencing, including "strand sequencing" with no theoretical read length limit,[178] may allow approaches taken in viral sequencing to be applied using SNVs identified at the ends of overlapping sequence fragments to reconstruct haplotypes[222,223] or may simplify the identification of mixed infection to identifying individual genomes sequenced in a single read. However until then, next-generation whole genome

sequencing offers the potential for high-throughput, labour- and cost-effective screening for mixed infection, and such approaches should become the standard when investigating potential outbreaks. In an outbreak situation the MIE offers the ability to accurately detect transmission from mixed infection, and to rule out transmission with increased certainty, which is important to provide a complete and accurate assessment of the source of each and every case under investigation. In the case of larger transmission studies involving *C. difficile,* as no evidence was found for onward transmission from a mixed infection case, approaches assessing overall patterns that do not account for mixed infection when assessing the source of each case are still likely to be good approximations as missing a few uncommon transmission events should not affect the overall inference. Therefore in the analysis of Oxfordshire *C. difficile* transmission based on whole genome sequencing that follows in chapter 5, it was decided it was reasonable to assess the origin of cases based on dominant infections alone.

5 CHARACTERISING OXFORDSHIRE *C. DIFFICILE* TRANSMISSION USING WHOLE GENOME SEQUENCING AND EPIDEMIOLOGICAL DATA

5.1 INTRODUCTION

The combined analysis of genotyping and ward movement data presented in chapter 3 shows ward-based contact with known CDI cases only accounts for the minority of disease.[211] This contrasts with the prior view that most episodes of *Clostridium difficile* infection (CDI) result from recent acquisition from symptomatic cases within a healthcare setting. The rationale underlying this belief is well supported: CDI case-to-case transmission and surrounding contamination is well documented.[4,76,86,88] However, there are multiple other potential sources, including asymptomatically colonised patients[87,230] and the wider environment, including water, farm/domestic animals, and food.[113] Their contribution to the overall burden of disease is unclear, particularly with increasing reports of community-associated CDI.[55]

A limitation of the analysis in chapter 3 is that only ward-based links between cases were sought. With the method used it was not possible to consider transmission across hospital wards, because so many chance contacts occur when considering time in the same hospital as a potential mechanism by which transmission could occur. Therefore, cases may have other forms of contact that were not captured, either within the hospital, for example by contact with healthcare staff moving between wards, or contact in the community. Such missing contacts might mean that the estimate of up to 25% of CDI being due to ward-

based transmission from symptomatic cases underestimates the true CDI resulting from prior symptomatic cases. In contrast, as many chance contacts occurred between CDI cases sharing the same ST, it is equally plausible that the extent of transmission on wards was over-estimated using MLST and that many same-ST donor-recipient pairs have substantial variation between them at the whole genome level. As demonstrated in chapter 4, whole genome data have huge potential to inform on transmission of *C. difficile*. True transmission events can be supported by close genetic relationships, and importantly transmission can be refuted between cases that are genetically distinct, even those that may share the same ST. The greater resolution offered increases discrimination between cases and therefore reduces the number of chance meetings between cases that may have been falsely attributed as transmission in chapter 3. Additionally, given the increased genetic discrimination available, cases that are genetically closely related can be investigated carefully for possible transmission, even where the route is not immediately obvious, potentially revealing novel routes of transmission.

To better quantify the role symptomatic patients play in onward transmission of *C. difficile*, all CDI cases from a defined geographic region, Oxfordshire, over a 3.6 year period were whole genome sequenced. Having sequenced all the cases, the maximum number of cases that could possibly have been acquired from other symptomatic cases can be estimated from WGS, and independently of epidemiological data. Additionally, by assessing the level of genetic diversity in disease-causing strains, the number of exposures to the reservoir from which disease arises can be assessed, as each novel genetic subtype found in the

symptomatic population represents a transmission event from the reservoir. Finally changes in transmission from symptomatic cases over time (as opposed to crude incidence) can be investigated.

5.2 METHODS

5.2.1 SETTING

This study used an extended collection of all Oxfordshire hospital and community CDI cases from September 2007 to March 2011 inclusive. The setting is described in section 3.2.1.

From September 2007 to March 2011 inclusive all faecal samples received by the OUH laboratory for *C. difficile* testing underwent enzyme immunoassay (EIA) for toxins A and B (Meridian Bioscience, Cincinnati, OH). During this period, hospital policy was to send faecal samples from all inpatients with diarrhoea (≥3 stools in 24 hours taking the shape of a container). EIA-positive samples were cultured and multilocus sequenced typed (MLST).[145,211] Culture-positive isolates underwent WGS. Repeat isolates of the same sequence type (ST) from the same patient were not routinely sequenced, except 148 randomly selected pairs of samples used to estimate rates of within-host diversity and evolution (see section 4.4.1). Samples from the same patient with different STs were sequenced allowing detected mixed infections and re-infections to be accounted for.[197]

Hospital admission and ward movement data, patients' home postcode district (28 distinct locations), and patients' general practice location were available for all patients.

5.2.2 SEQUENCING

DNA was extracted from a single sub-cultured colony, sequenced using Illumina technology, and the data processed using the pipeline described in section 4.2. Sequence reads were mapped to the *C. difficile* 630 reference genome (Genbank:AM180355), CD630,[182] with the exception of Clade 2 isolates[199] mapped to a clade-specific reference, CD196 (Genbank:FN538970).[184] Across all sequences a mean 83.4% of the CD630 reference genome was called after filtering for non-clade 2 samples, and 93.4% of the CD196 reference genome for clade 2 samples, i.e. 92.3% and 99.4% of the non-repetitive genome for CD630 and CD196 respectively. The remainder of the CD630 genome uncalled mostly represents the absence of CD630 mobile genetic material in many of the samples sequenced.

5.2.3 GENETIC ANALYSES

Single nucleotide variant differences between samples were compared. Given each sequence contains missing data, i.e. not 100% of the reference genome sites are called in every sequence, a methodology that robustly accounts for missing data is required. Two types of missing data are present. Firstly, repetitive regions are systematically masked as they cannot be resolved with the read lengths provided by current sequencing technology. These regions accounting for 9.6% of CD630 and 6.0% of CD196 are not called in any of the samples. As there are no data on the nature of any differences between samples in this region of the genome it is best excluded for any analysis, hence why evolutionary rates are presented in chapter 4 in terms of SNVs per called genome. It is plausible that repetitive elements may have very different patterns and rates of evolution, so simply scaling the known SNV differences across these regions may be misleading.

The other type of missing data is only missing in some of the samples; this arises for example from variable presence of mobile genetic elements, low level sample contamination, sequencing error, and failure to represent the whole genome at the library and cluster preparation stage. True variable presence and absence of mobile genetic elements represents a specific challenge as most phylogenetic methods and genetic distances are based on comparing core genetic material common to all samples. Therefore known annotated mobile genetic elements were also excluded from comparisons of genomes.

One possible approach is to conduct comparisons based on a 'core genome' that is consistently called across all samples. However when large numbers of samples are compared this 'core genome' may become small enough that large amounts of potentially useful data are discarded. For example, across the 248208 variable sites identified across all sequences in this study, no individual site was called in all sequences after filtering.

Maximum likelihood tree drawing algorithms such as PhyML (http://www.atgc-montpellier.fr/phyml/)[229] address missing data by averaging over the uncertainty at each missing site. Information on the most likely nucleotide at each missing site is inferred using the rest of the tree, with the missing nucleotide most likely to take on the value of adjacent sequences in the tree. However performing such calculations over >1000 samples over the whole *C. difficile* genome is too computationally intensive to be viable. Therefore 2 strategies were employed.

Firstly, a set of 248208 variable sites were established which had at least 2 known different nucleotides across all the genomes. Sites which had only a single wild

type nucleotide or missing data across all the genomes were assumed to be invariant, i.e. all missing data at these sites was assumed to have the same value as the reference. The whole genomes (excluding filtered reparative sites and mobile elements), not just the variable sites, were then used to create maximum likelihood trees. Retention of the complete genome sequence allows for more accurate estimates of key parameters, such as substitution rates, while removing uncertainty from sites not known to be variable increases speed considerably.

Secondly, sequences within 100 observed SNVs were grouped and analysed separately. To group the sequences, the observed pairwise difference between all genomes was calculated. Sites where data were missing in one or both sequences did not contribute to the pairwise difference. A graph was created using NetworkX (http://networkx.lanl.gov), with all genomes represented as nodes, and edges drawn between nodes with pairwise differences of ≤100 SNVs. Thus, each connected component* of the graph contained a group of sequences within 100 SNVs of each other. As defined each sequence was within 100 SNVs of at least one other sequence, but potentially >100 SNVs from some sequences in the group.

Finally maximum likelihood trees were estimated using PhyML with a generalised time reversible substitution model for each group. SNV distances between genomes for subsequent analyses were obtained from the estimated trees, i.e. SNV distances corrected for the impact of missing data. For larger distances (>100 SNVs) between genomes, where precision was less important as transmission could be excluded, the observed SNV difference was used.

* A subgraph in which any two nodes are connected by a path of edges

5.2.3.1 GENETIC SUBTYPES

Given the diversity observed within a host over time (see section 4.4.1), the probability of observing >10 SNVs through evolution during the 3.6-year study is <0.001. Therefore, cases >10 SNVs apart from any other were considered distinct subtypes to assess how many genetically-distinct clusters of cases occurred during the study. Subtypes >100 SNVs from any other case are also presented for comparison. The total number of distinct subtypes present in Oxfordshire was estimated using the number of distinct subtypes identified for a given sampling effort using models described by Lin[231,232] and Clench.[232,233]

5.2.3.2 IMPACT OF RECOMBINATION

Recombination is potentially important in genetic transmission analyses. Recombination can introduce multiple substitutions in a single event, and therefore may make genomes appear more genetically distant than they actually are. Analysis of the first *C. difficile* whole genome[182] revealed multiple mobile elements with potential for recombination. Estimates of the relative impact of recombination versus mutation, r/m, undertaken on a subset of the genomes in this study[225] using ClonalFrame[234] ranged from 0.0 to 7.4 across different lineages, suggesting up to 7.4 times more variation might have arisen in some lineages from recombination compared to mutation. However these estimates were based on the entire evolutionary history of 15 groups of sequences with most recent common ancestors dating back to between 1582 and 2005, i.e. considerably longer time scales than are directly relevant to individual transmission events. Interestingly the newest lineage, ST1, with a most recent common ancestor in 2005 contained a minimal amount of recombination.

To assess the impact of recombination in the current study, a sensitivity analysis was conducted to assess the extent to which observed SNVs between samples could be accounted for by recombination rather than mutation. From each of the groups of samples within 100 SNVs used in the PhyML analysis, the observed pairwise SNV difference between all possible pairs of samples within the groups was obtained. For each pair of sequences, a recombination adjusted SNV difference was calculated.

The model underlying ClonalFrame[234] was used to adjust the observed pairwise SNV differences (the *pairwise differences*) for the effect of homologous recombination. ClonalFrame analyses the clustering within the genome of substitutions introduced by recombination to simultaneously estimate which substitutions are the result of recombination or mutation, the underlying phylogeny, and important parameters underlying the process. In the ClonalFrame model, mutation occurs at rate θ_s per nucleotide and recombination is initiated at rate ρ_s per nucleotide. When recombination occurs, the length of homologous DNA imported is assumed to follow a geometric distribution with mean δ nucleotides. The nucleotide divergence of the imported DNA is specified by the parameter ν.

To estimate the number of pairwise differences attributable to mutation, firstly T, the length of the branch separating the two genomes in a two-taxa tree, measured in units of the expected number of point mutations per nucleotide was estimated. To estimate T in a computationally efficient manner, the ClonalFrame model was used, which can be thought of as a hidden Markov model (HMM, see, e.g. [235]) when there are only two genomes. The hidden state of the HMM records whether each nucleotide was subject to recombination or not on the branch connecting the

two genomes. Nucleotides unaffected by recombination are said to be *unimported* and nucleotides subject to recombination are said to be *imported*.[234] Based on the ClonalFrame model, the following transition probability matrix for the hidden variable between adjacent sites, H_i and H_{i+1} was defined:

$$\Pr\left(H_{i+1} \mid H_i\right) = \begin{cases} e^{-T\rho_s/\theta_s} & H_i = unimported \text{ and } H_{i+1} = unimported \\ 1 - e^{-T\rho_s/\theta_s} & H_i = unimported \text{ and } H_{i+1} = imported \\ 1/\delta & H_i = imported \text{ and } H_{i+1} = unimported \\ 1 - 1/\delta & H_i = imported \text{ and } H_{i+1} = imported \end{cases}$$

The emission probability defines the likelihood for the observed data conditional on the underlying hidden variable. At each nucleotide, *i*, the observed data, D_i, was summarised as either the *same* or *different*, depending on whether the two genomes matched or not. If the site was not called in either or both genomes, it was defined it to be the *same*. Following the ClonalFrame model, the following emission probabilities for the data at nucleotide *i* were defined:

$$\Pr\left(D_i \mid H_i\right) = \begin{cases} e^{-T} & D_i = same \text{ and } H_i = unimported \\ 1 - e^{-T} & D_i = different \text{ and } H_i = unimported \\ 1 - \nu & D_i = same \text{ and } H_i = imported \\ \nu & D_i = different \text{ and } H_i = imported \end{cases}$$

The forward algorithm[235] was used to calculate the likelihood $\Pr(D \mid T, \rho_s/\theta_s, \delta, \nu)$. T was estimated using maximum likelihood, implemented as a unidirectional optimization routine, taking as the values of the auxiliary parameters $\{\rho_s/\theta_s, \delta, \nu\}$ the point estimates from the full ClonalFrame analyses, which have previously been performed separately for 15 *C. difficile* sequence types.[225] Where parameter estimates were not available for the sequence types of a pair of sequences, parameter values from the genetically closest sequence type with estimates available were used.

Unlike the PhyML analysis, as this analysis is conducted in a pairwise manner the method is unable to exploit the overall phylogeny when handling missing data. It is therefore an approximation, and as such was conducted as a sensitivity analysis to assess the impact of recombination, rather than as the main analysis in the study.

5.2.4 EPIDEMIOLOGICAL ANALYSIS

Epidemiological relationships between genetically-related cases were classified using similar definitions to chapter 3, and using time thresholds obtained in chapter 3. The definitions were as follows: 'ward contact', were cases sharing time and space on the same hospital ward, consistent with infectious periods of -1 to +8 weeks from diagnosis[103,104,211] and incubation periods of 0-12 weeks.[29,211] Cases with multiple samples of the same ST were considered infectious for 8 weeks after their last positive test. If no ward contact existed, cases could be linked by time (as above) within the same hospital ('hospital-wide contact'), or by shared exposure to the same ward, but with up to 28 days separating the discharge (or end of infectivity) of the first case and admission of

the subsequent case (to allow for 'ward contamination'). Hospital wide contact was divided into cases sharing the same specialty, and those without a specialty link. 'Community contact' cases either shared a general practice, or had the same home postcode district. Finally, a sensitivity analysis defined contacts without any restriction on the duration of infectious/incubation/ward contamination periods.

Whether patients with undiagnosed CDI/asymptomatic colonization could explain genetically-related cases without hospital contact was also investigated. For each such case the most recent prior genetically-related case was identified. The number of hospital contacts the two cases shared between their diagnoses was calculated, firstly, including only intermediate contacts with a negative EIA-test (to include possible falsely-negative CDIs) and then with all hospital in-patients without an EIA-test during the study. The number of intermediate contacts between pairs of genetically distant cases (>10 SNVs, randomly chosen, matched on the time between samples) was also calculated to estimate the background proportion of unrelated cases that shared intermediate patients.

5.3 RESULTS

In Oxfordshire September 2007 to March 2011, 40924 faecal samples were submitted for *C. difficile* testing. 2377 EIA-positive samples were identified; overall inpatient rates 8.5/10000 overnight stays, 13.0/month outpatient/community cases. Of 2283/2377 (96%) EIA-positive samples retrieved for culture, 1714 (75%) were *C. difficile* culture-positive, representing 1250 infections after accounting for repeated same-ST samples. Isolates from 1223 (98%) cases were successfully sequenced (**Figure 47**); 862/1223 (71%) were OUH inpatients, 246

(20%) from general practices, 60 (5%) OUH outpatients, and 55 (4%) from other hospitals. The median age (IQR) at CDI was 78 (66-86) years, 78 years in OUH inpatients, and 77, 81 and 85 years in general practice, OUH outpatients and other hospitals respectively. The most prevalent STs/ribotypes were ST1/ribotype-027 (214/1223, 17%), ST8/ribotype-002/159 (110, 9%) and ST2/ribotype-014/020/076/220 (103, 8%).

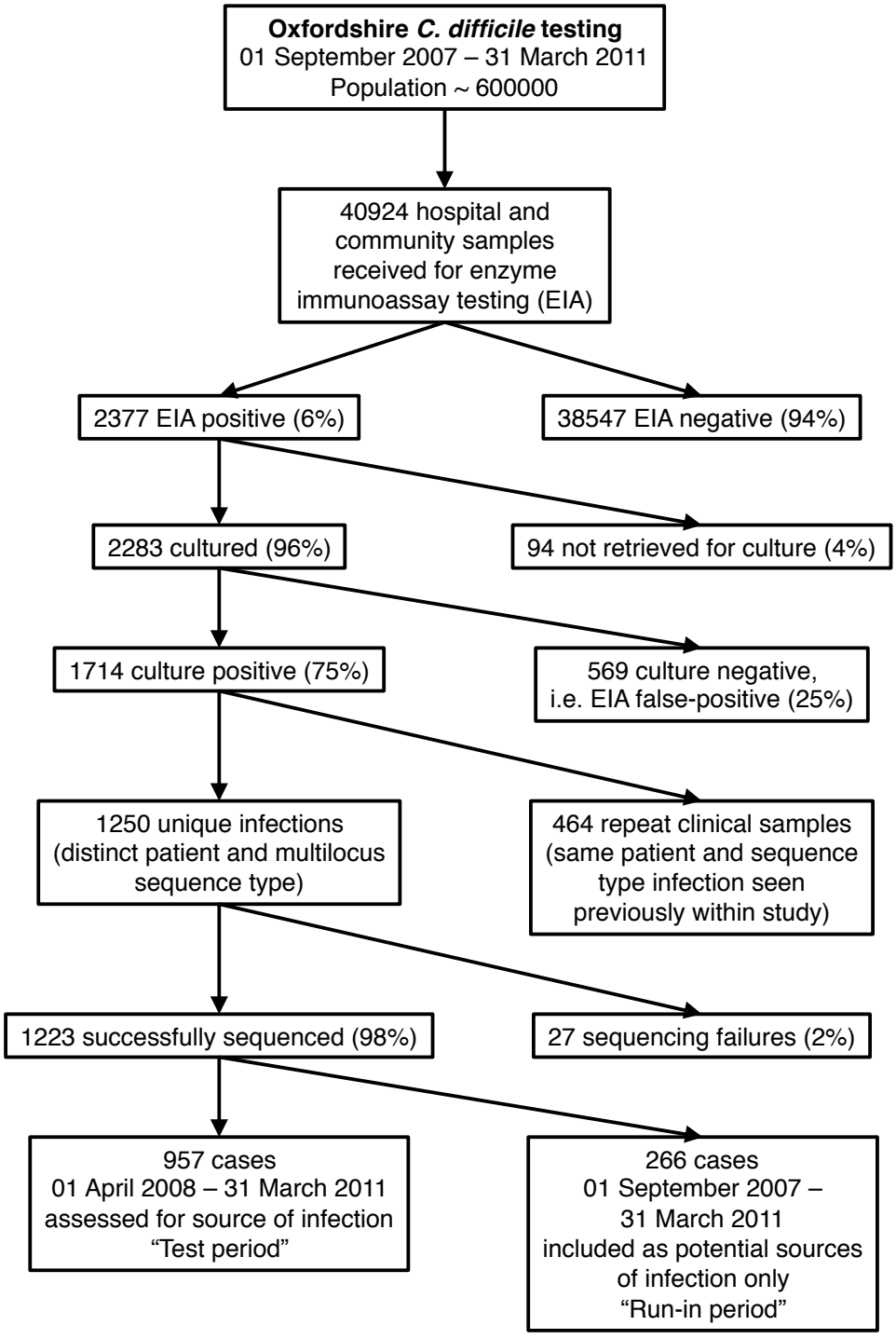


Figure 47. Samples and patients, 01 September 2007 to 31 March 2011.

5.3.1 GENETIC DIVERSITY WITHIN OXFORDSHIRE CASES

Of 957 sequenced isolates from CDIs between April 2008–March 2011, 240 (25%) were genetically-indistinguishable from ≥1 previous case since September 2007 (266 cases, September 2007–March 2008, were included only as potential sources for later CDIs). Using the prediction intervals from section 4.4.1, 0-2 SNVs are expected between isolates <124 days apart, a timescale consistent with transmission given infectious periods <8 weeks and incubation periods <12 weeks. Only 333/957 (35%) were genetically-related to ≥1 previous case (within ≤2 SNVs, i.e. closely related enough to support transmission). Therefore, having sequenced nearly all regional CDI cases over 3.6 years, striking genetic diversity was observed; 428/957 (45%) CDI cases were >10 SNVs different from all previous sequenced symptomatic cases since September 2007 (**Figure 48**).

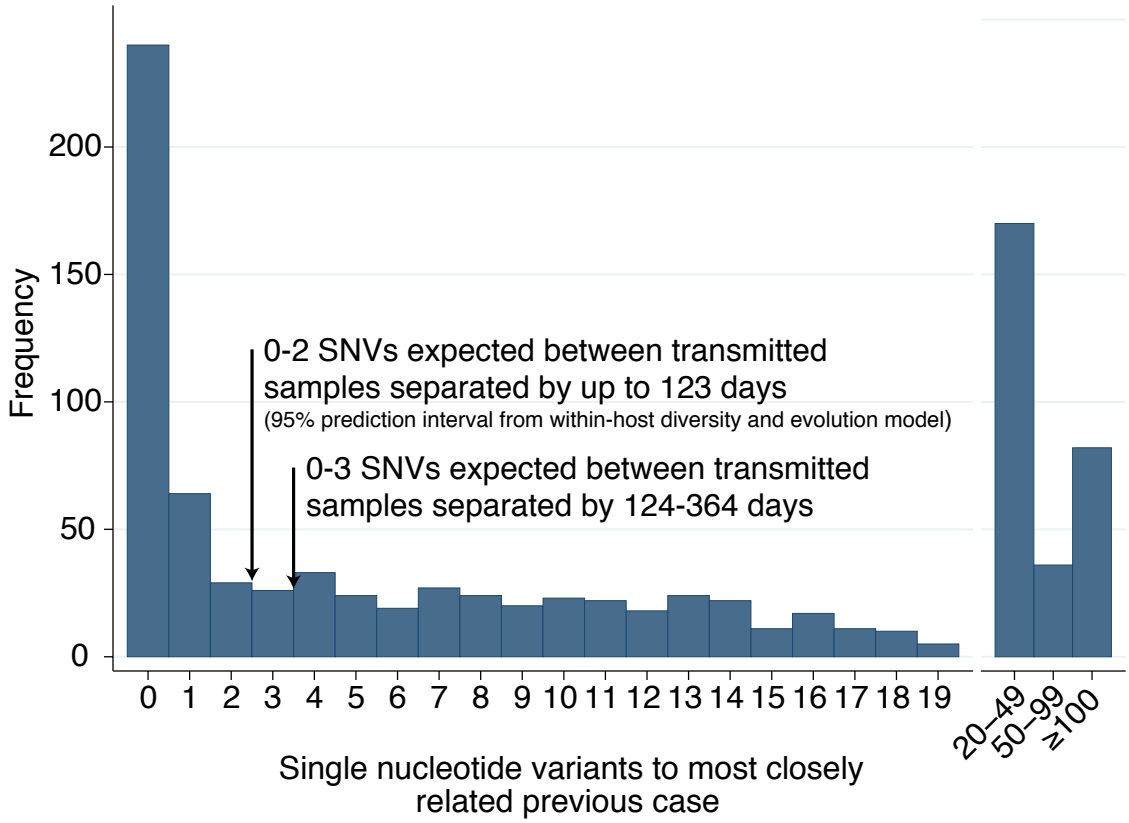


Figure 48. Genetic distances between 957 cases. The number of single nucleotide variants between each case from 01 April 2008 – 31 March 2011 and the most closely related previous case since 01 September 2007 is shown. 95% prediction intervals for the number of SNVs expected between transmitted isolates are shown (based on rates of within-host diversity and evolution from 145 serially sampled patients, see section 4.4.1).

Results were similar in a sensitivity analysis making an adjustment for recombination. Across the groupings of samples within 100 SNVs, only 66/42805 pairs (0.15%) had >2 observed SNVs and ≤2 adjusted SNVs (i.e. variants arising from mutation), suggesting relatively few very closely related sequences are erroneously described as more distant in the main analysis. Similarly, 27817/30057 (93%) of pairs >10 SNVs apart remained more than 10 variants different after removing variants introduced by recombination. The total number of adjusted variants for each pair of sequences, is shown in **Table 10** and **Figure 49**.

		Adjusted variants (SNVs arising from mutation)						
		1	2	3-4	5-7	8-10	>10	Total
Observed SNVs	1	2417						2417
	2	6	2134					2140
	3-4	0	62	1663				1725
	5-7	0	1	78	1529			1608
	8-10	0	0	1	947	3910		4858
	>10	1	3	4	90	2142	27817	30057
	Total	2424	2200	1746	2566	6052	27817	42805

Table 10. Comparison of SNV differences between pairs of samples, with and without adjustment for recombination.

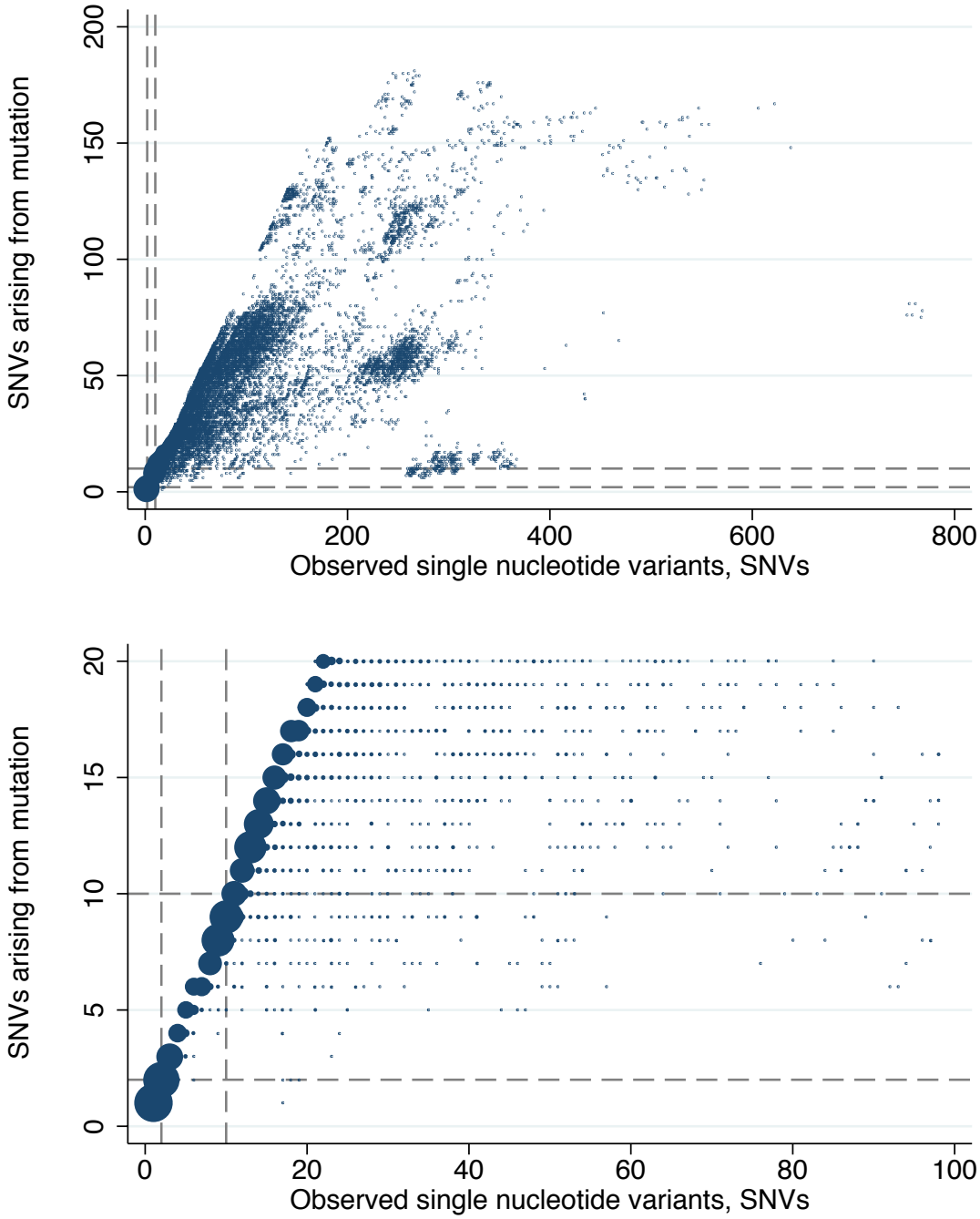


Figure 49. Comparison of SNV differences between pairs of samples, with and without adjustment for recombination. The number of variants after adjusting for recombination (i.e. SNVs arising from mutation) is plotted against the number of observed SNVs between samples. Markers are weighted by the number of pairs of samples with each value. The top panel shows all pairs, for ease of visualisation excluding 3 pairs with >1000 observed SNVs and recombination adjusted variants of 46, 47, 49. The bottom panel shows the same plot in more detail for lower observed and adjusted SNV values.

5.3.2 EPIDEMIOLOGICAL RELATIONSHIPS BETWEEN GENETICALLY-RELATED CASES

Of the 333 genetically-related cases, 126 (38%) shared ward-based contact, 13% of the total 957 cases. 5/333 (2%) cases were linked only by possible ward-based contamination persisting after infectious patient discharge/recovery, a further 29 (9%) shared time in the same hospital (but never on the same ward) and 21 (6%) had both ward-contamination and hospital-wide contact (**Table 11**, **Figure 50**). No hospital-based contact could be established in 152/333 (46%). Few genetically-related cases without hospital contact had possible community contacts: 15/152 (10%) shared a general practice, and a further 17/152 (11%) a district postcode. Overall, 120/333 (36%) of cases within 2 SNVs had no record of any previous hospital or community contact. 68/333 (20%) still had no hospital/community contact allowing implausible unlimited infectious/incubation/ward contamination periods (**Table 12**). Results were consistent using other SNV thresholds (**Figure 50**). In keeping with direct person-to-person transmission likely being associated with ≤2 SNVs (as suggested by the rate of within-host diversity and evolution) and with direct ward contact being the most plausible route for hospital transmission, the number of cases genetically-related and linked by ward contact remained relatively constant as the threshold for calling samples genetically-related was increased from 2 to 10 SNVs. In contrast, as the threshold was increased, the number of cases related within the threshold, but without an epidemiological contact increased.

		0 SNVs		0-2 SNVs		0-10 SNVs	
Total new cases, 01 April 2008 - 31 March 2011		**957**		**957**		**957**	
Genetically distinct	**No previous Oxfordshire case since 01 Sep 2007 within SNV threshold**	**717**	**(75%)**	**624**	**(65%)**	**428**	**(45%)**
Genetically linked to any previous case	**Previous Oxfordshire case since 01 Sep 2007 within SNV threshold**	**240**	**(25%)**	**333**	**(35%)**	**529**	**(55%)**
Closest genetic link epidemiologically related through hospital contact	**Hospital Contact Total** ***Within plausible limits***	**144**	**(15%)**	**181**	**(19%)**	**224**	**(23%)**
	Ward contact *(Infectious -1/+8 weeks from test, Incubation 12 weeks)*	98	(10%)	126	(13%)	136	(14%)
	Ward contamination only *(Contamination persisting up to 4 weeks from ward discharge)*	7	(1%)	5	(1%)	8	(1%)
	Shared specialty only *(Infectious -1/+8 weeks from test, Incubation 12 weeks)*	15	(2%)	17	(2%)	28	(3%)
	Other hospital-wide only *(Infectious -1/+8 weeks from test, Incubation 12 weeks)*	9	(1%)	12	(1%)	22	(2%)
	Mixed ward contamination & hospital-wide	15	(2%)	21	(2%)	30	(3%)
Closest genetic link epidemiologically related through community & not hospital contact	**Community Contact Total** ***Without shared hospital exposure***	**23**	**(2%)**	**32**	**(3%)**	**63**	**(7%)**
	Shared general practice *Without shared hospital exposure*	10	(1%)	15	(2%)	37	(4%)
	Shared postcode district *Without shared hospital exposure or general practice*	13	(1%)	17	(2%)	26	(3%)
Genetically linked, but no hospital or community contact	No known epidemiological link *Within hospital exposure limits above*	73	(8%)	120	(13%)	242	(25%)

Table 11. Epidemiological relationships between each new CDI, 01 April 2008 to 31 March 2011, and the most genetically similar previous case. For each case the nature of the closest epidemiological links with all previous cases within a given SNV limit is shown. Cases sharing time and space on the same hospital ward within plausible limits on infectious (-7 to +56 days of a positive test) and incubation (up to 84 days) periods are denoted "ward-contact". Cases sharing a common ward, but up to 28 days between discharge of the first case (or end of infectivity of the first case) and admission of the second, and no other form of shared hospital exposure are denoted "ward contamination only". Cases present at the same time in the same hospital, but not sharing the same ward, are denoted "shared specialty only" if they shared a common specialty, and "other hospital-wide only" if not. Where patients have a mix of "ward contamination contact" and "hospital-wide contact", but without a direct "ward-contact" this is denoted "mixed spore and hospital-wide". Community links are made when two cases share the same general practice or where two cases share the same home postcode district. Cases without any

form of shared hospital exposure or community contact as defined above are denoted "no known epidemiological link". Where multiple samples of the same ST from the same patient were not sequenced infectious periods went from 1 week before the first isolation of this ST to 8 weeks beyond the last isolation of this ST.

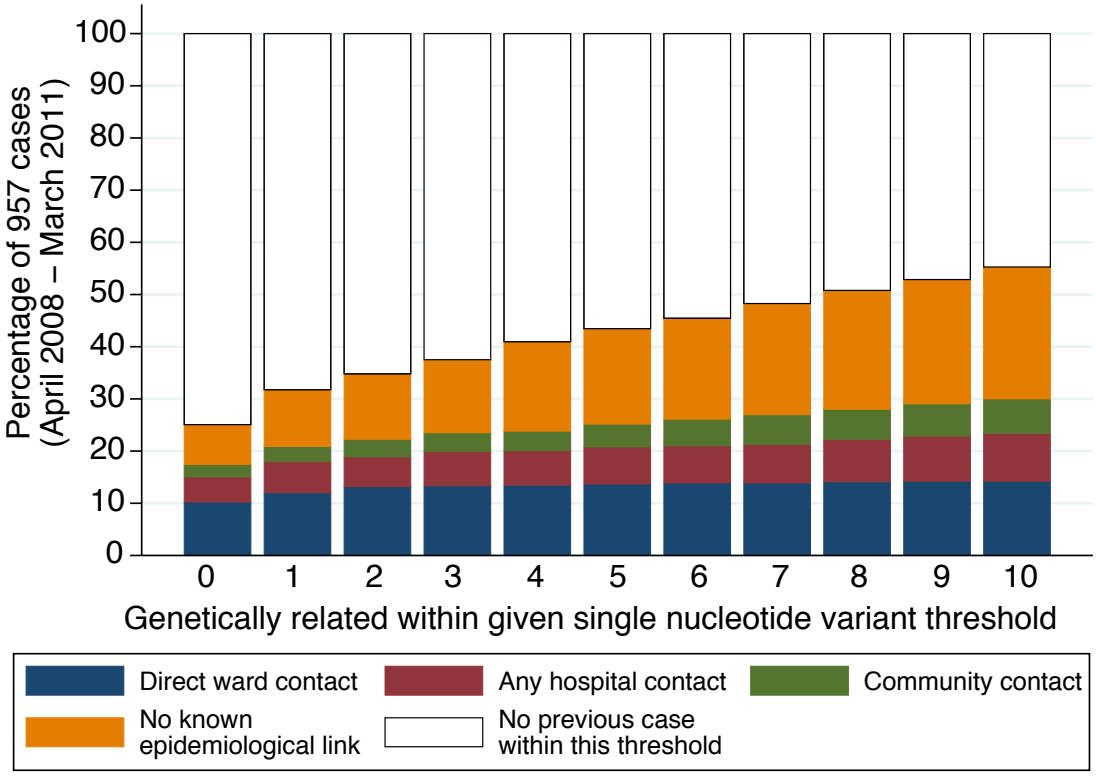


Figure 50. Genetic distances and epidemiological relationships between 957 cases. The percentage of cases genetically-related to a previous case using SNV thresholds from 0 to 10 is shown. The epidemiological links made between genetically-related cases are shown in different colours. Cases denoted "direct ward contact" shared time on the same ward within plausible upper limits on infectious (-1 to +8 weeks of a positive test) and incubation (0-12 weeks) periods (as in table 1). Any hospital contact encompasses cases sharing time in hospital, but not on the same ward, within the same time restrictions and cases sharing a common ward, but up to 28 days between discharge (or end of infectivity) of the first case and admission of the second, and no other form of shared hospital exposure. Community contact cases shared either a home postcode district or a general practice location.

		0 SNVs		0-2 SNVs		0-10 SNVs	
Total new cases, 01 April 2008 - 31 March 2011		957		957		957	
Genetically linked to any previous case	Previous Oxfordshire case since 01 Sep 2007 within SNV threshold	240	(25%)	333	(35%)	529	(55%)
Closest genetic link epidemiologically related through any hospital contact	Hospital Contact *Unlimited infectious, incubation and ward contamination periods*	190	(20%)	256	(27%)	376	(39%)
Closest genetic link epidemiologically related through community & not hospital contact	Shared general practice or postcode district *Without any shared hospital exposure with unlimited infectious, incubation and ward contamination periods*	9	(1%)	9	(1%)	16	(2%)
Genetically linked, but no hospital or community contact	No known epidemiological link *Unlimited infectious, incubation and ward contamination periods*	41	(4%)	68	(7%)	137	(14%)

Table 12. Sensitivity Analysis: Epidemiological relationships between each new CDI, 01 April 2008 to 31 March 2011, and the most genetically similar previous case. For each case the nature of the closest epidemiological links with all previous cases within a given SNV limit is shown. The duration of infectious/incubation/ward contamination periods were unlimited in this sensitivity analysis. Other definitions as **Table 11**.

5.3.3 EPIDEMIOLOGICALLY UNEXPLAINED GENETICALLY-RELATED CASES

There are several possible explanations for the source of the 120 epidemiologically unexplained genetically-related cases (**Table 13**). The median (IQR) time since the most recent prior case within 2 SNVs was 113 (28-281) days, suggesting point sources or laboratory contamination (samples processed in weekly batches) were unlikely. Healthcare worker-mediated transmission is likely to be captured by genetically-related cases sharing the same specialty (17/333 (5%)) already excluded from the 120. 32 (27%) cases had ≥1 EIA-negative contacts in common with their most recent prior genetically-related case (possible transmission via falsely-EIA-negative cases). However 27/120 (23%) date-matched controls >10 SNVs apart also had ≥1 EIA-negative contact (p=0.55). 36/120 (30%) cases had a

hospital contact without an EIA test (possible asymptomatic carrier) in common (32 also had an EIA-negative contact), compared with 35/120 (30%) controls >10 SNVs apart (p=1.00). Allowing unlimited ward contamination duration (to account for possible long-term spores on wards) explained a further 13/120 (11%) cases, allowing unlimited infectious and incubation periods a further 12 (10%). No pair of cases within 2 SNVs attended the same outpatient clinic on the same day. Overall, none of these potential transmission routes could link 48/120(40%) genetically-related cases without any hospital/community contact.

		0 SNVs		0-2 SNVs		0-10 SNVs	
Possible explanations for cases without a clear epidemiological link	No known epidemiological link *within hospital exposure limits above*	73		120		242	
	Samples obtained within ≤7 days *Possible point source / Laboratory contamination*	7	(10%)	11	(9%)	18	(7%)
	EIA negative third party contact in common *(Samples >7 days apart)*	19	(26%)	32	(27%)	43	(18%)
	Other inpatient contact in common *(No EIA negative contact, samples >7 days apart)*	3	(4%)	4	(3%)	7	(3%)
	Ward contamination if assumed to persist infinitely *(and none of above)*	9	(12%)	13	(11%)	44	(18%)
	Hospital contact with another symptomatic case if CDI assumed to be indefinitely infectious, and incubate for an unlimited period *(and none of the above)*	6	(8%)	12	(10%)	24	(10%)
	Genetically related cases remaining without any plausible epidemiological link found	29	(40%)	48	(40%)	106	(44%)

Table 13. Additional epidemiological explanations for cases with a previous case within various SNV thresholds without a clear epidemiological link.

5.3.4 COMPARISON OF DISCRIMINATORY POWER OF WHOLE GENOME SEQUENCING VERSUS CONVENTIONAL GENOTYPING RESULTS

As discussed in section 4.5 whole genome data has greater resolution than genotyping schemes such as MLST, which translates into WGS being more discriminatory as measured at a population level by the Simpson Diversity Index. However combining epidemiological information with genotyping data improves its discrimination. Here the additional value of WGS over genotyping combined with epidemiological information is examined. In the previous sections cases were classified using genetic sequences, then their epidemiological relationships determined. In contrast, in **Figure 51** the results of an analysis similar to chapter 3, based on combining genotyping (MLST) and epidemiological data, are compared with the SNVs between cases. Amongst pairs of cases with ward contact sharing the same ST, 98/190 (52%) had 0 SNVs, and 126/190 (66%) were within 2 SNVs; 54/190 (28%) had >10 SNVs between them. Therefore, as many as a third of cases thought to be related by transmission on the basis of shared genotype and ward contact are in fact sufficiently genetically different that transmission is unlikely. In comparison, only 41/302 (14%) of cases linked by other forms of hospital contact, predominantly shared time, were within 2 SNVs and 230/302 (76%) were ≥10 SNVs apart, ruling out transmission previously thought plausible. This supports the notion these forms of contact are less plausible routes of transmission.

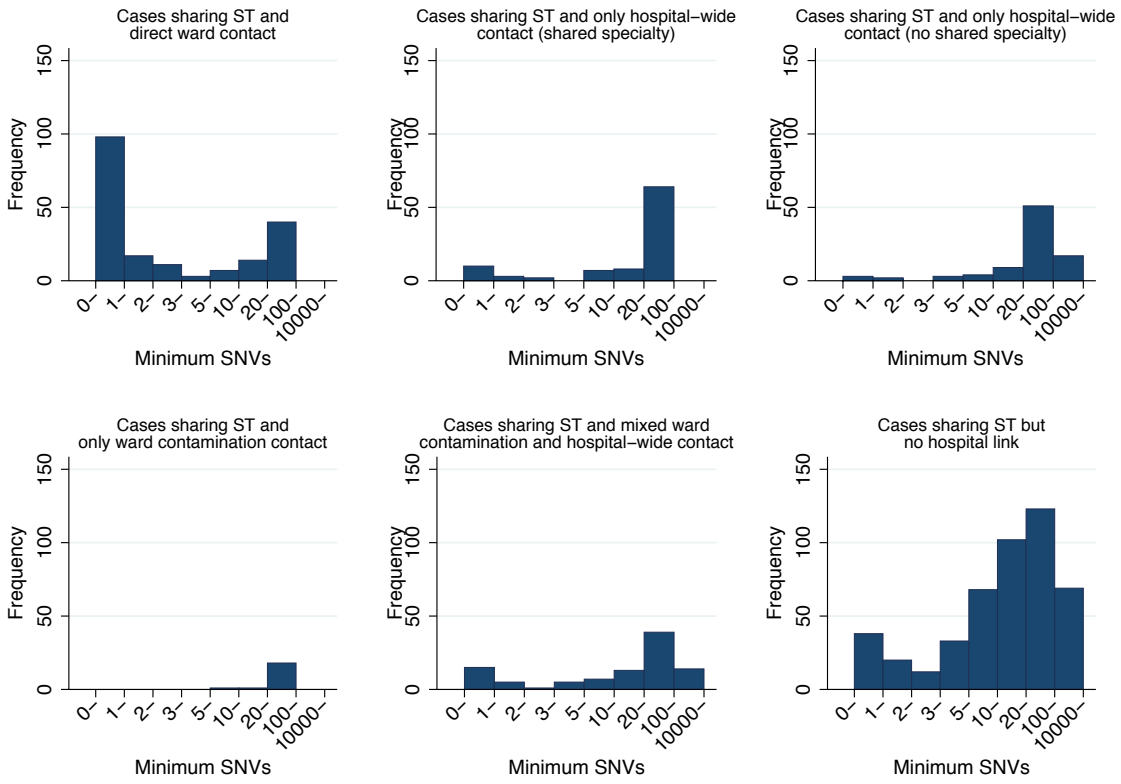


Figure 51. Distribution of SNVs between cases sharing STs and different types of epidemiological contact. The results of analyses combining genotyping (MLST) and epidemiological data from chapter 3 are compared with the SNVs between cases. In contrast to **Table 11** where the data are initially categorised by SNV thresholds and then investigated for epidemiological links, in this figure the data are categorised initially by genotype/epidemiological information. Direct ward contact is defined as shared time on the same ward within plausible upper limits on infectious (-1 to +8 weeks of a positive test) and incubation (0-12 weeks) periods (as in table 1). Hospital-wide contact encompasses cases sharing time in hospital, but not on the same ward, within the same time restrictions. Cases sharing a common ward, but up to 28 days between discharge (or end of infectivity) of the first case and admission of the second, and no other form of shared hospital exposure, are denoted ward contamination contact. If multiple contacts of the same type occur with different potential donors of infection, the lowest SNVs associated with that type of epidemiological contact is shown.

5.3.5 POPULATION DIVERSITY

Given the diversity observed within a host over time, the probability of observing >10 SNVs through evolution during the 3.6-year study is <0.001. Therefore, cases >10 SNVs apart from any other were considered as distinct subtypes to assess how many genetically-distinct clusters of cases occurred during the study (**Figure 52**). From September 2007–March 2011, the number of cases of each subtype varied markedly. As well as clear clustering of cases in time, there are also multiple instances of cases from the same lineage separated by long periods without a symptomatic case; on 12 occasions genetically indistinguishable strains were recovered more than a year apart without any intervening case. Distinct subtypes continued to be identified consistently throughout the study, suggesting cases arise from a considerable reservoir of *C. difficile*. Over the course of the study the overall incidence of CDI fell (per year rate ratio, change in incidence, 0.91 (95% CI 0.90-0.92), therefore to assess the trend in isolation of new subtypes per sampling effort the number of subtypes per cases sequenced is shown in **Figure 53**. Two models estimating the total population diversity suggest that, using a 10 SNV threshold, the number of distinct genetic subtypes present in the reservoir infecting subsequently symptomatic cases is >10000.

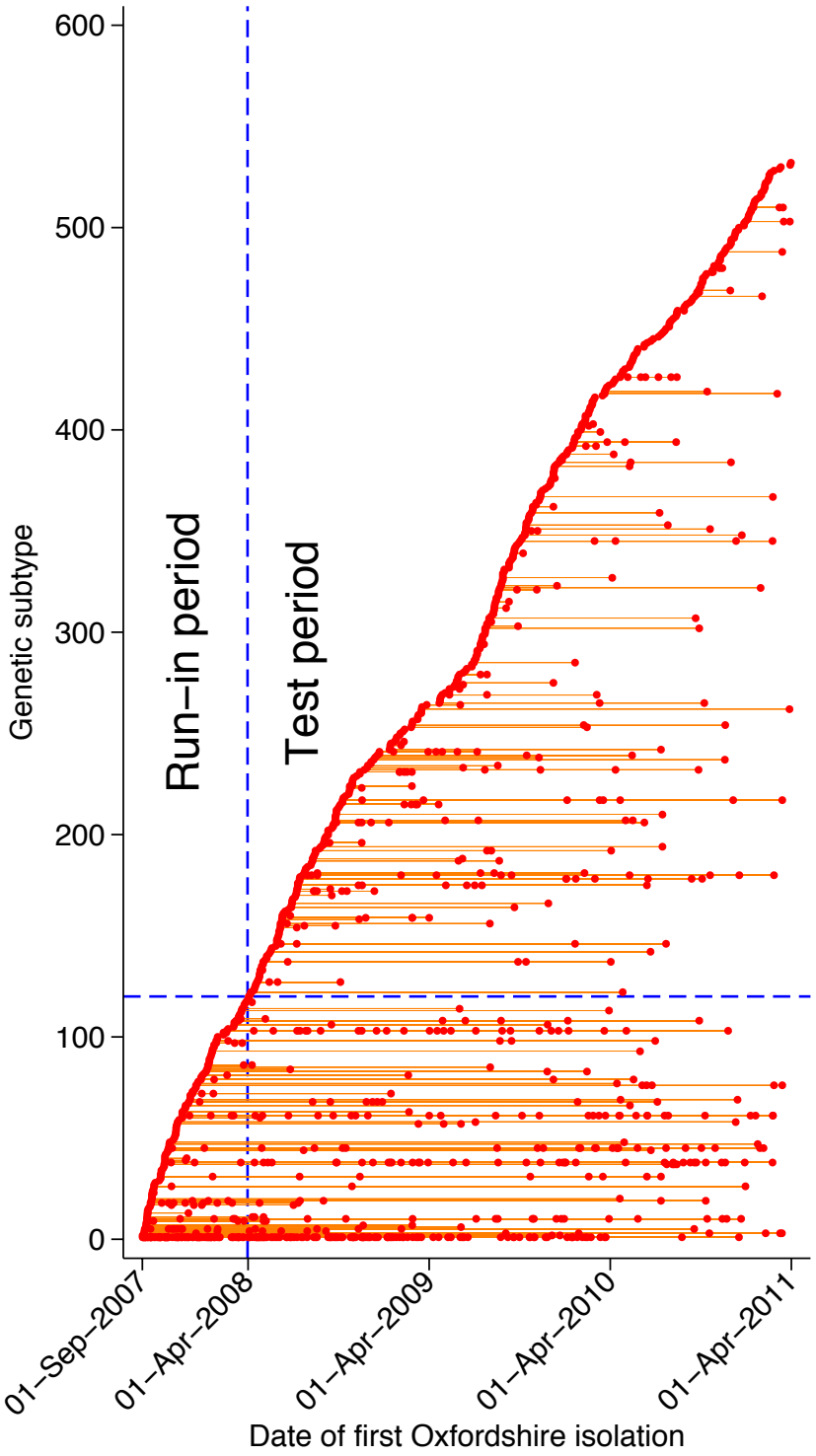


Figure 52. Timing and size of *C. difficile* genetic clusters. Isolates >10 SNVs apart from any prior sample were defined as a distinct genetic subtype. The date of first *C. difficile* isolate of each genetic subtype is plotted on a different horizontal line; subsequent cases of a previously isolated subtype are shown on the same line. Cases 01 September 2007 to 31 March 2008 denoted run-in period. See **Figure 53** for models estimating the total population diversity.

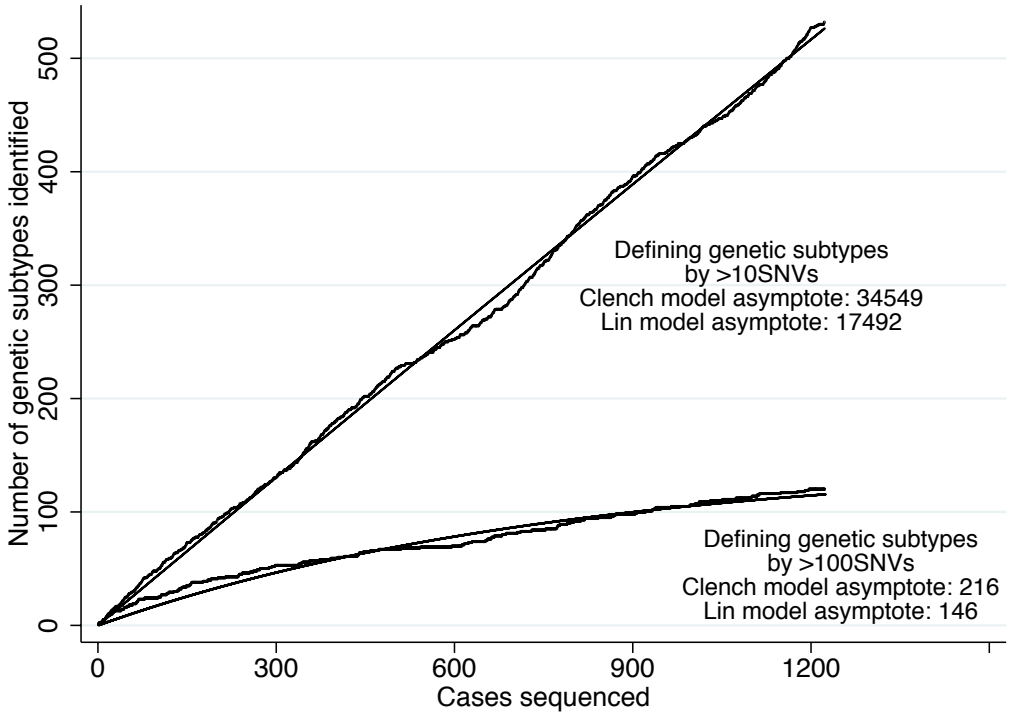


Figure 53. Total population diversity models. The number of distinct genetic types of *C. difficile* identified for a given sampling effort over calendar time is plotted for 2 different SNV thresholds. Two models estimating the total population diversity are plotted shown in bold.

5.3.6 TEMPORAL TRENDS IN TRANSMISSION

The incidence of CDI caused by genetically distinct isolates (>10 SNVs from any previous case, 'introductions' to the symptomatic Oxfordshire population), was compared to the incidence of cases genetically-related to a previous case (≤2 SNVs, possible secondary cases) (**Figure 54**), to test the hypothesis that interventions aimed at reducing transmission might have preferentially affected genetically-related cases. Despite marked month-to-month variation in the incidence of genetically distinct and related cases, a downward trend was observed in both, with no evidence for difference in declines (heterogeneity p=0.44) (**Figure 54a,b**). Within the genetically-related cases a marked reduction was seen in cases with hospital contact, with no change in other cases over time (heterogeneity p<0.001)

(**Figure 54c,d**). Notably variability was substantially smaller in monthly secondary cases with a possible hospital link. Genetically-related cases caused by ST1/ribotype-027/NAP1 accounted for much of the decrease, with a smaller reduction in genetically-related hospital-contact cases due to other *C. difficile* types (although these still declined significantly) (**Figure 55**). Interestingly incidence of genetically-related cases caused by ST1 without hospital contact also fell (albeit more slowly than ST1 with hospital contact), suggesting a possible strain specific effect was important in the within-hospital decline. Genetically related non-ST1 cases without hospital contact showed a non-significant upward trend in incidence, that was significantly different to both the fall seen in ST1 cases without hospital contact and non-ST1 cases with hospital contact.

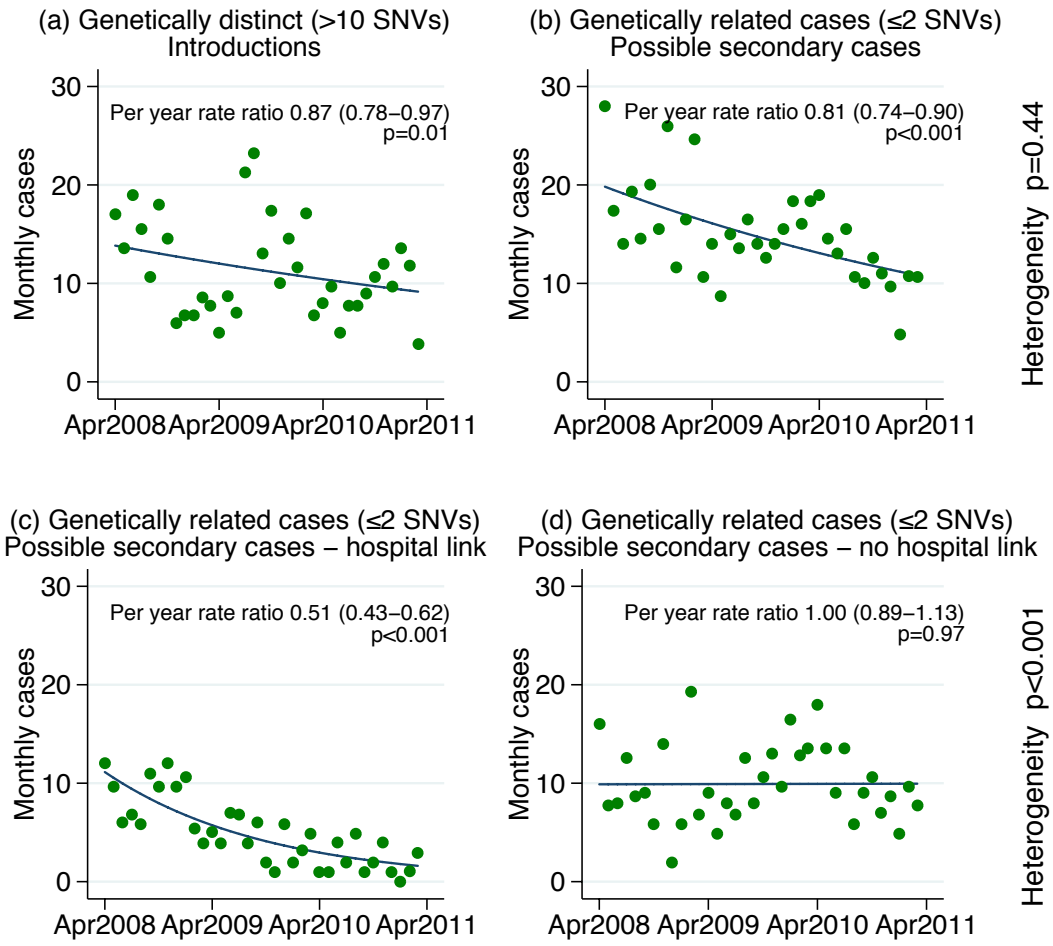


Figure 54. Introductions and secondary cases in the Oxfordshire symptomatic *C. difficile* population. CDI cases caused by an isolate that was >10 SNVs apart from a previous case since 01 September 2007 (novel genetic subtypes) were denoted "introductions" to the symptomatic Oxfordshire population, and cases ≤2 SNVs as possible "secondary cases". Only cases from 01 April 2008 are shown as the first seven months of the study were denoted a run-in period to capture the increased rate of novel subtypes expected initially; thus, temporal patterns are analysed over the final 3 years of the study. The study population during the study was 600000, therefore rates of 30 cases per month correspond to 5 per 100000 population per month. Per year rate ratio (change in incidence), determined by Poisson regression. Heterogeneity p values were determined by stacked Poisson regression. Note, during the study period sample submission rates also fell but by a lesser extent than positive cases (per year rate ratio 0.91 (95% CI 0.90-0.92), heterogeneity p<0.001 versus overall CDI incidence). Plotted monthly cases are scaled 30 times the mean daily cases in the month, to allow for accurate comparison across months. See **Figure 55** for a breakdown of trends in genetically-related cases grouped by ribotype-027 / non-ribotype-027.

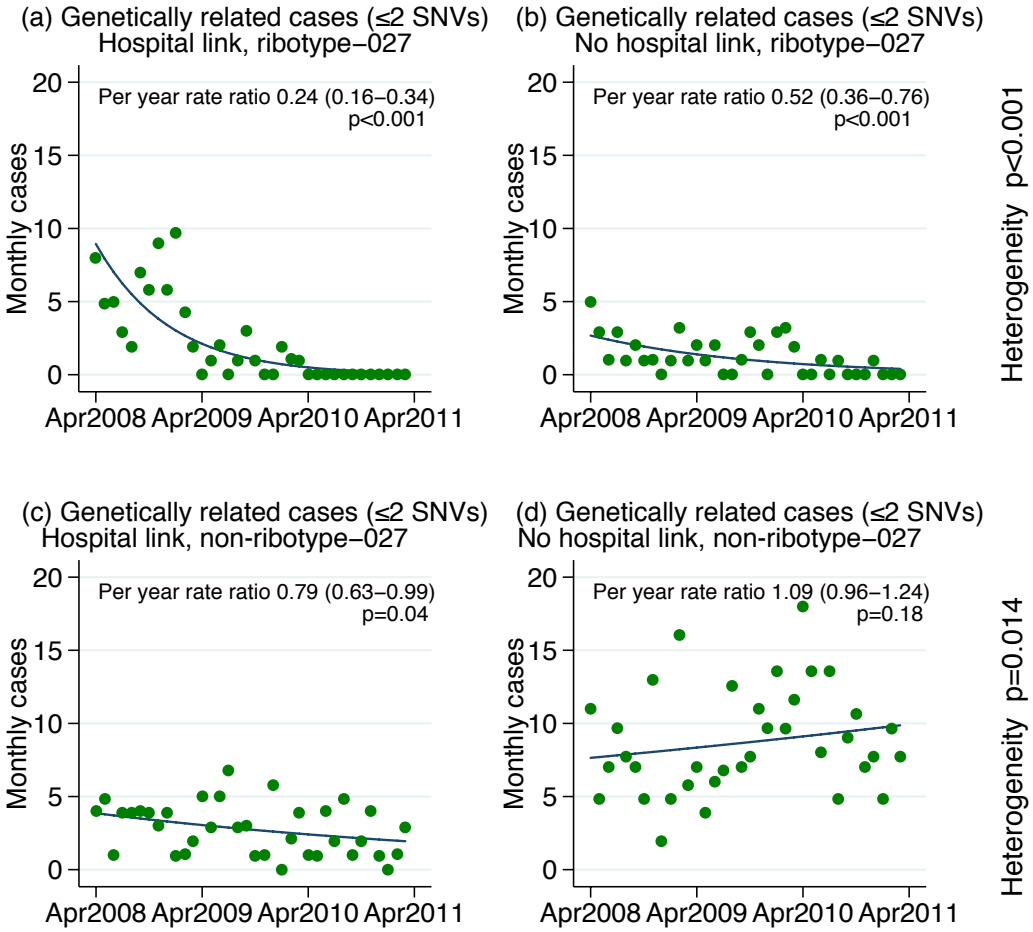


Figure 55. Secondary cases in the Oxfordshire symptomatic *C. difficile* population by ribotype-027 / non-ribotype-027. CDI cases caused by an isolate that was >10 SNVs apart from a previous case since 01 September 2007 (novel genetic subtypes) were denoted "introductions" to the symptomatic Oxfordshire population, and cases ≤2 SNVs as possible "secondary cases". Only cases from 01 April 2008 are shown as the first seven months of the study were denoted a run-in period to capture the increased rate of novel subtypes expected initially; thus, temporal patterns are analysed over the final 3 years of the study. The study population during the study was 600000, therefore rates of 20 cases per month correspond to 3.3 per 100000 population per month. Per year rate ratio (change in incidence), determined by Poisson regression. Heterogeneity p values were determined by stacked Poisson regression. Plotted monthly cases are scaled 30 times the mean daily cases in the month, to allow for accurate comparison across months.

Genetically-related possible secondary cases tended to be older, median (IQR) age 81 (70-87) years compared to introductions 76 (64-85 years) (p<0.001). As expected possible secondary cases were more likely to be classified as healthcare-

onset or healthcare-associated,[29] 253/333 (76%) compared with 282/428 (66%) of introductions (p=0.003).

5.4 Discussion

This 3.6 year study of whole genome sequences from over 1200 CDI cases, in a defined geographic area with typical CDI incidence,[236] shows that significant transmission arises from outside the symptomatic toxin EIA-positive population. Only 35% of cases had a previous case within 2 SNVs from whom *C. difficile* could have plausibly been acquired, irrespective of transmission route. 45% were sufficiently genetically diverse (no previous case ≤10 SNVs) to represent transmission originating outside the symptomatic EIA-positive cases studied, e.g. from asymptomatic carriers.[237] Of note, this conclusion can be made from high-resolution genetic data independently of epidemiological data that may not always be complete. Some genetically distinct cases may have acquired CDI from symptomatic patients outside Oxfordshire, but given the median age, 78 years, it is more likely that most are from an asymptomatic or environmental reservoir. The on-going identification of genetically distinct cases throughout the study points to diverse alternative sources of *C. difficile*. Each diverse subtype identified represents a separate transmission event from a reservoir/asymptomatic carrier to a subsequently symptomatic patient, with typically one or very few cases arising from each exposure.

Using genotyping data in chapter 3 only up to 25% of all cases could be explained by recent ward-based contact with another symptomatic patient.[211] Applying the greater discrimination provided by WGS in this study, 13% of cases were

genetically related and shared ward contact, 19% including all forms of hospital contact, demonstrating that a substantial minority of cases with the same genotype are distinct considering the whole genome. However, WGS was also able to identify a substantial proportion of cases that were genetically-related (≤2 SNVs) without any evidence of plausible previous hospital or community contact (13%). Exploring possible explanations for these 'cryptic' cases, we found no evidence for substantial contributions from third-party transmission within hospital by healthcare workers or patients. Only 5% of genetically-related cases shared the same specialty and team of healthcare workers, but no other form of hospital contact. Of 120 genetically-related cases without any hospital/community contact, 27% had an EIA-negative contact in common with their most recent prior case within 2 SNVs, and an additional 3% had a hospital inpatient never tested for CDI in common. Although these third-party contacts may be responsible for some transmission, genetically-distinct cases had similar numbers of inpatient intermediates suggesting many of these contacts may be chance meetings without transmission. Even so, onward transmission from falsely-EIA-negative cases should reduce with the availability of more sensitive/two-step testing, which may also identify *C. difficile* excretors without disease, another possible CDI source.[238] Genetically-related cases without hospital/community contact may alternatively both have been exposed to a common source; the median 113 days between such cases makes a short-term point source such as fresh food or water less likely, unless it is persistently contaminated. Future detailed investigation of these patients in real-time may provide insights into the location and nature of these sources.

The incidence of cases genetically-distinct from any previous case (introductions to the symptomatic population) and genetically-related to previous cases (possible secondary cases) fell similarly over the study, suggesting that interventions not only targeting transmission from symptomatic cases, such as antimicrobial stewardship, have played a major role in declines in CDI in the region over the last 5 years. Interventions to prevent secondary cases, in place within our hospitals from 2007 (described earlier in **Table 3**) may underlie the substantially larger fall in genetically-related cases associated with hospital contact, particularly those caused by ST1/ribotype-027/NAP1. Genotyping and sequencing were done retrospectively, so did not affect clinical or infection control practice. Notably large decreases in the prevalence of this epidemic strain in England have been described recently, and are believed to be due to targeted infection prevention measures.[239] Alternatively, ST1/ribotype-027/NAP1 may have declined independently of infection control efforts, as has been observed with MRSA strains,[240] possibly due to changes in strain fitness, host immunity, or changes in antimicrobial use combined with strain-specific antibiotic resistance.[184] Declines in fluroquinolone use may have been particularly important. The unchanged incidence of possible secondary cases in the community, particularly due to strains other than ST1, demonstrates an on-going challenge.

This study has several limitations. Initial case ascertainment was by EIA testing which has limited sensitivity, however multiple samples were submitted for many patients during the study. Provided the missing cases (a minority) were missed at random, the diversity observed and trends in introductions and secondary cases are likely to be robustly estimated. As described previously the molecular clock

used to assess potential transmissions is an approximation, particularly as mutation rates in reproducing vegetative forms may be different to spores. However slower or arrested evolution in spores would make genetically-distinct cases even more so. Mixed infections detected by genotyping multiple samples from patients were accounted for, however, undetected mixed infections, and subsequent re-infections with the same ST, may provide an explanation for some apparently unlinked cases. As described in chapter 4, detailed investigation of the most plausible scenarios in which this could theoretically have occurred in this study found no evidence to support this playing a major role. Further, the widely reported prevalence of mixed infections (<10%)[197] could not explain the diversity identified.

This study shows conclusively that symptomatic cases only account for limited CDI transmission in a region with typical CDI incidence,[236] and demonstrates many cases arise from genetically diverse sources. It suggests that antimicrobial stewardship has played a major role in the recent control of *C. difficile* arising from these sources in our region, supporting wide emphasis on these policies. Use of rapid benchtop sequencing allows genetically-related cases to be identified in near real-time,[198] such that cases clearly linked by hospital or community contact can be targeted to prevent further spread, and institutional performance monitored more sensitively by using counts of genetically-related cases, rather than simple case counts. Whole genome data are also sufficiently discriminatory that genetically-related cases without a clear epidemiological link can be investigated in a highly focused way for novel routes of transmission. Such investigations are

likely to shed new light on the source of currently unexplained infections and the diverse sources of *C. difficile* infection.

6 ASYMPTOMATIC CARRIAGE STUDY

6.1 INTRODUCTION

The data presented in chapters 3 and 5 suggest only a minority of Oxfordshire CDI cases acquired *C. difficile* from other symptomatic cases. The remaining cases were infected with genetically diverse *C. difficile*, implying multiple exposures to a varied reservoir, with each exposure resulting in relatively few secondary cases. Data from previous studies suggest *C. difficile* is typically acquired shortly before onset of disease,[39] with long term carriage protective against CDI.[77,78] Therefore exposure to asymptomatic individuals, particularly in a healthcare setting, represents a compelling potential source of disease. Up to 3% of healthy adults may carry *C. difficile* asymptomatically[37] and between 4 and 11% of those admitted hospital are carriers.[39,40] With CDI rates of ~5 cases per 10000 bed days, [29,55,56,58] and for example typical ward stays of ~3 days, there are likely to be 10-20 asymptomatic carriers admitted to hospital for every case. If diverse *C. difficile* strains are carried asymptomatically, and the chance of any one carrier infecting a patient is small, this could generate the pattern of genetic diversity seen in chapter 5.

If asymptomatic carriers are an important source of disease this has substantial implications for hospital infection control, as screening and isolation or eradication in these carriers may interrupt transmission. Nevertheless their relatively large numbers will pose significant challenges. In a small study, vancomycin treatment did not reduce long-term carriage, but did transiently render subjects culture-negative during and immediately after treatment.[156]

Vancomycin was more effective than metronidazole, possibly because only vancomycin achieved detectable drug levels in the formed stools of asymptomatic subjects.[156] Transient suppression of carriage would probably be sufficient to reduce risk of onward transmission throughout most patients' hospital admission as *C. difficile* was not detected until 20 ±8 days after the end of a 10 day vancomycin course. However there is no direct evidence that treating asymptomatic carriers or instituting more stringent isolation of carriers leads to less transmission.[29] The closest to this kind of evidence is a study in a Belgian leukaemia unit, where a combination of environmental cleaning and renovation together with treatment of asymptomatic carriers with vancomycin reduced CDI rates markedly.[241]

To investigate the feasibility of studying asymptomatic carriage in Oxfordshire hospital inpatients a prospective longitudinal cohort study was conducted in several hospital wards within the OUH. This aimed to estimate the prevalence of asymptomatic *C. difficile* carriage in Oxfordshire medical inpatients, assess the incidence of acquisition, and also compare the genetic subtypes of *C. difficile* found in asymptomatic cases with those causing disease.

6.2 METHODS

This longitudinal cohort study of asymptomatic hospital inpatients was designed to screen patients for asymptomatic carriage of *C. difficile* at admission, at regular intervals during admission and around the time of discharge. The study focused on acute general medicine and geratology wards as these specialities had the greatest CDI prevalence. The choice of high prevalence wards was intended to provide the

maximum number of CDI cases on the same ward as asymptomatic carriers, in order to investigate the potential for carrier to case transmissions.

6.2.1 PARTICIPANTS

The study inclusion criteria were all inpatients aged over 18 years admitted to a study ward. Patients diagnosed with CDI in the last 28 days, or with a current clinical suspicion of CDI (≥3 unformed stools in a 24 hour period) were excluded from this specific study (but samples / sequences from these cases were available for comparison with study participants, see below).

Ethical approval was granted to enrol patients without capacity to consent as a significant number of patients affected by *C. difficile* have a degree of transient or permanent cognitive impairment. This was important, because to assess potential routes of acquisition from asymptomatic carriers, good coverage of the whole hospital population is required. Additionally the potential harm of inconvenience for study participants without capacity was low (see samples below).

Recruitment was rotated around 8 study wards on a weekly to fortnightly basis (between 02 February 2012 and 10 June 2012) to reduce the nursing burden of sending stool samples from patients. Following informed consent (or the advice of an appropriate consultee in the case of patients without capacity) where possible patients were recruited within 24 hours of ward admission, following review of ward occupancy lists.

6.2.2 SAMPLES

In order to maximize participation, the study used stool samples, rather than a specific screen using a rectal swab. Patients were asked provide a sample of their first stool after enrolment, and then a sample at approximately 3 day intervals until discharge, and where possible a sample shortly prior to discharge.

Samples were stored at 4˚C and cultured in weekly batches on *C. difficile* selective media as in section 3.2.2. The culture method used is very similar to that in largest study of asymptomatic carriage to date.[39] Samples exhibiting a characteristic appearance, odour, and fluorescence consistent with *C. difficile* were emulsified in nutrient broth containing 10% glycerol and frozen at -80˚C. Samples underwent whole genome sequencing as described in section 4.2.1. For each sequence the presence or absence of toxin genes and an *in silico* MLST was determined (see section 4.2.2).

All Oxfordshire clinical CDI continued to be sequenced throughout the study, such that sequences from Oxfordshire CDI cases from April 2011 to August 2012 inclusive were available for comparison. Between April 2011 and March 2012 samples sent to the OUH microbiology laboratory underwent toxin EIA-testing. EIA-positive samples were cultured. Between April 2012 and August 2012 samples were screened with a GDH test, with a follow up confirmatory EIA test in GHD-positive samples. All GDH-positive samples were cultured, regardless of the subsequent EIA result. This enabled potential *C. difficile* 'excretors' carrying *C. difficile*, but not currently producing toxin, to be included. Between April 2011 and August 2012 a total of 505 symptomatic *C. difficile* culture-positive samples were identified, 481 (95%) of which were successfully sequenced and available for

comparison. A further 137 sequences were available from a *C. difficile* diagnostic study, from between April 2011 and September 2011, that were culture positive, but EIA-negative.

6.2.3 EPIDEMIOLOGICAL DATA

Consent was obtained for use of hospital electronic data, including participant demographics (age, sex, postcode district (first 3-4 characters of the postcode)), hospital admission, discharge and ward movement data within the OUH, and previous and future routine *C. difficile* testing. Additional data on potential risks factors for *C. difficile* were collected from patients including:

- History of diarrhoea in last 6 months
- Nature of normal place of residence (own home, residential home, nursing home)
- Past medical history (including factors that have been associated with *C. difficile* infection in previous studies: a history of inflammatory bowel disease, chronic kidney disease, haematological malignancy, solid tumour, previous gastrointestinal surgery, nasogastric feeding)
- Healthcare exposure (including admission to any hospital in last 6 months, attendance at outpatients, visits to other inpatients, visits to nursing and residential homes, employment in a healthcare setting, contact with healthcare workers at home)

- Known previous symptomatic *C. difficile* infection

- Known exposure to previous *C. difficile* infection

- Antibiotic use within the last 6 months (including agent, duration and date of course)

- Proton pump inhibitor and immunosuppressive medication use

- Other putative risk factors (pets, physical contact with children at least once a month (ages specified in questionnaire), a diet that includes meat, travel outside the UK in the last 6 months, smoking)

6.2.4 ANALYSIS

Data obtained were used to estimate the point prevalence of asymptomatic carriage of *C. difficile*. Complete assessment of prevalence at admission and discharge, and therefore the incidence of acquisition, was not possible given the time delay between many participants recruitment and first bowel motion, and similarly relatively few participants providing more than one sample (see results).

Logistic regression was used to assess associations between the presence of *C. difficile* carriage in participants first stool sample and risk factors for asymptomatic carriage from questionnaire data. Univariate and multivariate models were fitted, with independent risk factors identified using backwards selection with the Akaike information criterion.[202]

Whole genome sequences obtained from asymptomatic carriers were compared to a database of Oxfordshire CDI cases' sequences obtained from 1 year prior to the

study, during the study and for 3 months afterwards. Ward movements of cases and carriers sharing closely related *C. difficile* genomes were analysed for evidence of possible transmission events.

6.3 RESULTS

6.3.1 PARTICIPANTS AND SAMPLES

Two hundred and twenty-seven participants were recruited of 243 inpatients approached (13 declined to participate and a further 3 were ineligible due to a diagnosis of CDI in the last 28 days, **Figure 56**). A total of 296 stools samples were obtained from 132 participants (58% of recruited participants), 71 with a single sample and 61 with more than one sample, median (IQR) [range] samples per person 1 (1-2) [1-11]. A total of 28 samples from 18 participants tested positive for *C. difficile* (14% of those tested). 14 participants (11%) tested positive on their first sample, with a further 4 participants (3%) testing positive after at least 1 initially negative sample (**Figure 57**). Participant 3 had two preceding negative samples before a positive sample, participant 7 had one, participant 13 had two, and participant 14 had three, providing reasonable evidence of a true acquisition event in at least 3 cases, rather than an initial false-negative result. Participant 1 had the only apparent loss of carriage with a positive sample followed by a negative sample.

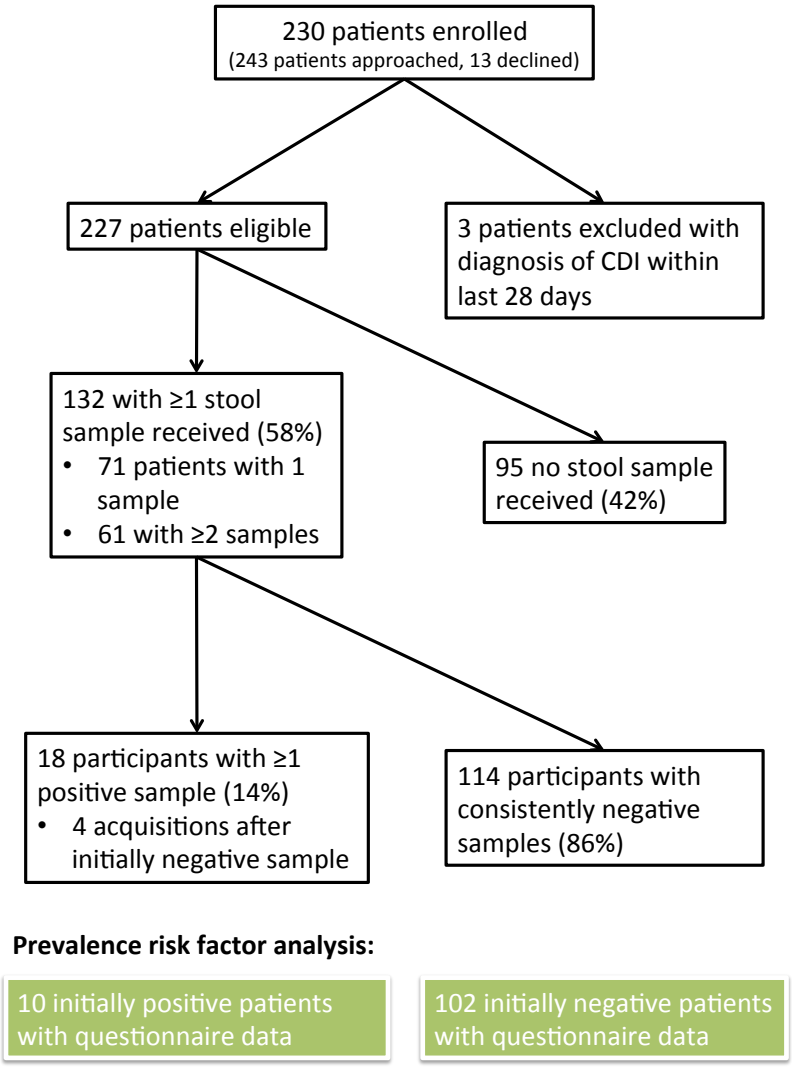


Figure 56. Asymptomatic study participants and samples.

6.3.2 Relationship between carriage and previous / subsequent disease

Participant data were linked with hospital admission and microbiology data to identify participants previously or subsequently testing positive for *C. difficile* as a result of clinical testing in symptomatic patients (**Figure 57**). Four such participants were identified. Participants 4, 7 and 13 had a routine microbiology laboratory diagnosis of CDI 1, 35 and 8 days respectively after a positive carriage test in the study. In all 3 cases no SNVs were found between the study sample and

the subsequent positive sample. Participant 7 also had a positive routine laboratory CDI sample from 178 days prior to enrolment which was also genetically indistinguishable to the participant's later samples, suggesting the single negative sample obtained from this participant may have been falsely negative, or that this strain was archived undetectably within the participant's intestinal flora and selected for by antibiotics received in hospital. In contrast, participant 3 tested positive 86 days following a positive study sample, but with a different infecting *C. difficile* lineage, 11892 SNVs different from the carriage isolate. Participants 3 and 4 reported loose or more frequent stool at the time of enrolment, but less than the ≥3 unformed stools in 24 hours required to meet criteria for routine *C. difficile* testing. Participants 7 and 13 had no diarrhoeal symptoms at enrolment.

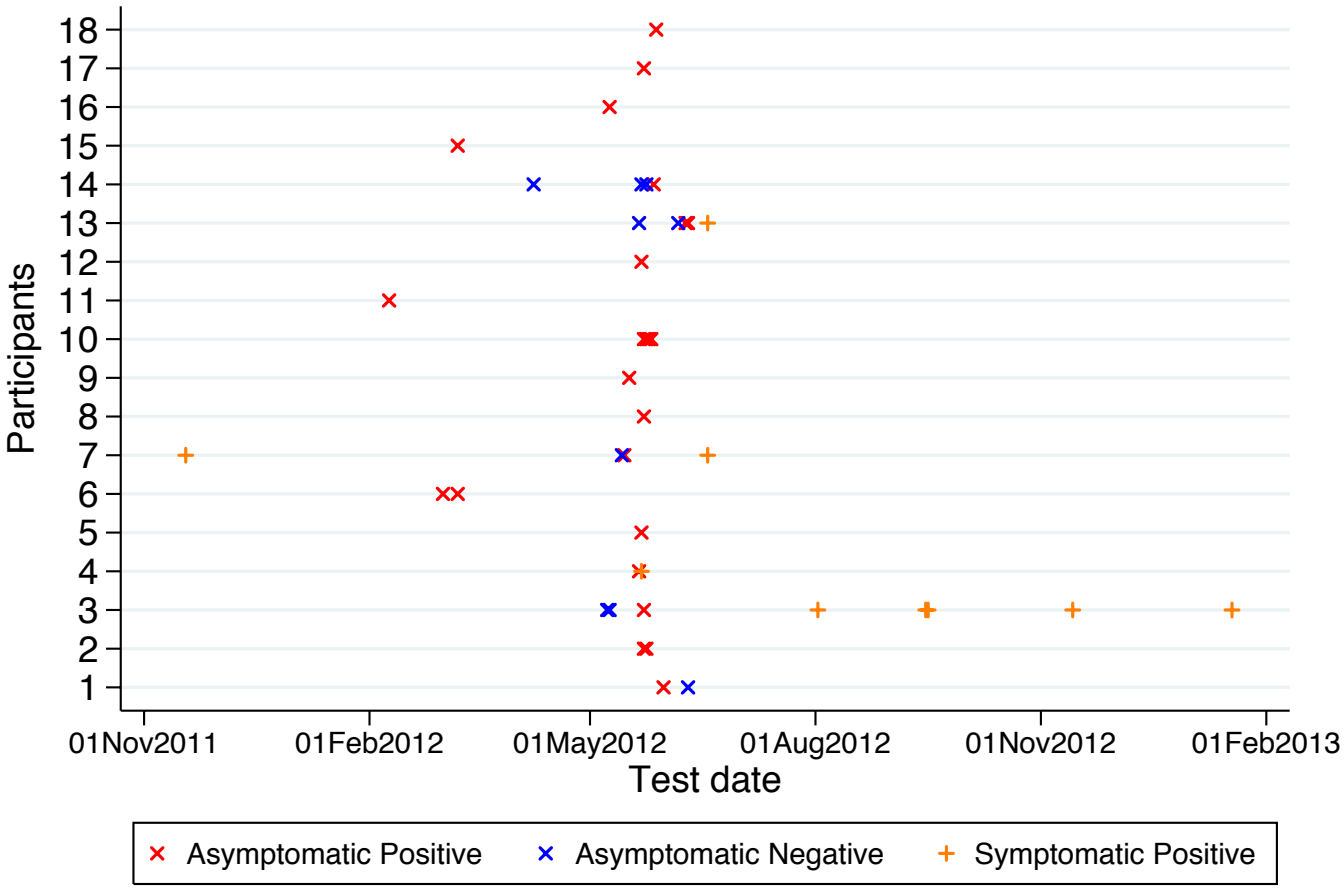


Figure 57. Temporal pattern of samples in participants with ≥1 positive sample. Asymptomatic positive and negative relate to samples obtained during the carriage study. Matching of study participants with the hospital admission and microbiology data allowed 4 participants with a subsequent symptomatic CDI positive sample to be identified, denoted symptomatic positive. 118 participants had consistently negative samples and are not plotted.

6.3.3 Risk factors for carriage

Questionnaire data obtained at participant enrolment were used to assess risk factors for *C. difficile* carriage in participants' first stool sample. A total of 112 (85%) of the 132 participants with ≥1 stool samples were able to complete the study questionnaire. Ten participants initially tested positive, a 9% point prevalence for asymptomatic *C. difficile* carriage in the population tested.

Univariate odds ratios for the risk factors assessed are presented in **Table 14**. The risk factor for carriage with the strongest statistical evidence was current diarrhoea not meeting diagnostic criteria for CDI. Patients with ≥3 unformed stools in a 24 hour period, i.e. meeting OUH criteria for CDI testing, were excluded from the study. However, 22 participants reported loose or more frequent stool than usual, but did not meet the OUH criteria for CDI testing. It is therefore possible that the five of these 22 patients testing positive for *C. difficile* actually should have had a diagnosis of CDI. Non-significant trends towards increased risk of *C. difficile* carriage were seen with nursing or residential home residence, an overnight stay in hospital in the last six months, steroids or another immunosuppressant in the last 6 months, and gastric acid suppressant medication in the last six months. Intriguingly antibiotic exposure in the last six months had a trend towards lower risk of *C. difficile* carriage in univariate analysis.

Risk factor	*C. difficile* carriers (% total)	Non-carriers (% total)	Odds ratio, carriers:non-carriers	95% CI	p value
Current diarrhoea, but not meeting CDI criteria	5 (50%)	17 (17%)	4.94	(1.29, 19.0)	0.02
Nursing or residential home resident	2 (20%)	4 (4%)	6.13	(0.97, 38.7)	0.08
Chronic kidney disease	2 (20%)	22 (22%)	1.10	(0.21, 5.56)	0.90
Non-haematological malignancy	3 (30%)	14 (14%)	2.69	(0.62, 11.66)	0.21
Overnight hospital stay in last 6 months	6 (60%)	37 (36%)	2.59	(0.68, 9,79)	0.15
Outpatient hospital attendance in last 6 months	2 (20%)	31 (31%)	0.56	(0.11, 2.81)	0.47
GP visits in last 6 months					
0	2 (20%)	7 (7%)	1.77	(0.28, 11.1)	0.23
1-2	5 (50%)	31 (30%)	1		
3-9	2 (20%)	44 (43%)	0.28	(0.05, 1.55)	
10+	1 (10%)	20 (20%)	0.31	(0.03, 2.85)	
Healthcare worker in household	1 (10%)	8 (8%)	1.29	(0.14, 11.52)	0.82
Antibiotics in last 6 months	6 (60%)	84 (82%)	0.32	(0.08, 1.25)	0.12
Steroids / Immunosuppressant in last 6 months	5 (50%)	23 (23%)	3.43	(0.91, 12.9)	0.07
Gastric acid suppressant in last 6 months	7 (70%)	47 (46%)	2.73	(0.67, 11.2)	0.14
Household pet	3 (33%)	24 (24%)	1.62	(0.38, 7.00)	0.52
Physical contact with children ≥1 per month	3 (33%)	51 (51%)	0.49	(0.11, 2.07)	0.32
Current smoker	1 (10%)	10 (10%)	1.02	(0.11, 8.92)	0.98

Table 14. Univariate risk factors for asymptomatic carriage of *C. difficile*. There were no *C. difficile* carriers with the following risk factors: inflammatory bowel disease (present in 1/112 (1%) participants), haematological malignancy (5/112 (4%)), gastrointestinal surgery (5/112 (4%) with surgery in current admission, 12/112 (11%) with past surgery), current nasogastric tube placement (3/112 (3%)), previous CDI (5/112 (4%)), contact with a CDI case (2/112 (2%)), vegetarian diet (7/112 (6%)), and overseas travel (4/112 (4%)).

Independent risk factors in a multivariate model are shown in **Table 15** and include current diarrhoea, but not meeting CDI criteria, overnight hospital stay in the last 6 months and steroids or another immunosuppressant in the last six months. As in the univariate analysis, antibiotics in the last six months was associated with lower odds of *C. difficile* carriage.

Risk factor	Odds ratio	95% CI	p value
Current diarrhoea, but not meeting CDI criteria	10.0	(1.96, 50.9)	0.006
Overnight hospital stay in last 6 months	5.53	(1.06, 28.8)	0.02
Antibiotics in last 6 months	0.07	(0.01, 0.45)	0.005
Steroids / Immunosuppressant in last 6 months	7.19	(1.32, 39.1)	0.02

Table 15. Multivariate risk factors for asymptomatic carriage of *C. difficile.* No significant pairwise interactions were identified, but given the relatively small sample size the power to detect these is low.

The lower odds of *C. difficile* carriage with prior antibiotics in the multivariate model is particularly driven by the low number of carriers in the group of participants receiving antibiotics without hospital exposure in the last six months (1/51, 2%, **Table 16**). However, as illustrated, the small number of carriers means that a single change in the number of carriers could have changed odds ratio estimates substantially.

Risk factor combination	*C. difficile* carriers (% row total)	Non-carriers (% row total)
Hospital + Antibiotics	5 (13%)	34 (87%)
Overnight hospital only	1 (25%)	3 (75%)
Antibiotics only	1 (2%)	50 (98%)
Neither	3 (18%)	14 (82%)

Table 16. Relationship between overnight hospital stays and antibiotic exposure in the last six months amongst *C. difficile* carriers and non-carriers.

In a sensitivity analysis, 22 participants reporting loose or more frequent stool, but not meeting CDI diagnosis criteria, were excluded. Of the remaining participants 5/90 (6%) were *C. difficile* carriers. The same risk factors remained in the multivariate model, with broadly similar point estimates for the odds ratios, however with greater uncertainty around these estimates given the smaller numbers of participants in the analysis (**Table 17**).

Risk factor	Odds ratio	95% CI	p value
Overnight hospital stay in last 6 months	5.19	(0.58, 46.0)	0.14
Antibiotics in last 6 months	0.04	(0.002, 0.51)	0.01
Steroids / Immunosuppressant in last 6 months	16.5	(1.25, 215)	0.03

Table 17. Multivariate risk factors for asymptomatic carriage of *C. difficile* excluding 22 patients with loose or more frequent stool, but not meeting CDI diagnosis criteria.

6.3.4 GENETIC RELATIONSHIPS BETWEEN CARRIAGE AND SYMPTOMATIC ISOLATES

The first positive isolate from the 18 participants with ≥1 positive sample(s) underwent whole genome sequencing. Positive samples were obtained between 09 February 2012 and 09 June 2012 (1 in February, 2 in March, 14 in May, and 1 in June 2012). These sequences were compared with sequences obtained from all symptomatic *C. difficile* positive samples processed by the routine OUH microbiology laboratory between April 2011 and August 2012. All samples were mapped to the CD630 reference genome, with a mean of 85.1% of the reference genome called in the 18 carriage samples, and 84.3% across the 618 comparison samples. Mean mapped read depths were 88.9 and 84.3 respectively.

Thirteen of the asymptomatic carriage isolates sequenced (72%) were toxigenic, predominantly from STs in clade 1, the clade responsible for the majority of cases of CDI in Oxfordshire (**Table 18**). No isolates from the epidemic ST1 lineage were seen, but this is consistent with ST1 having largely ceased to be a cause of symptomatic disease in Oxfordshire since 2009 (see chapters 3 and 5). Interestingly asymptomatic carriage of an ST11, an emerging hypervirulent strain from clade 5, was seen in one participant.

ST	Clade	Toxin A/B	Frequency	Participant Id
2	1	+/+	2	1, 16
3	1	+/+	2	14, 18
6	1	+/+	2	6, 10
7	1	-/-	1	15
11	5	+/+	1	9
13	1	+/+	1	17
18	1	+/+	1	13
26	1	-/-	1	8
29	1	-/-	2	2, 3
35	1	+/+	2	7, 11
46	1	+/+	1	5
49	1	+/+	1	4
125	1	-/-	1	12

Table 18. Asymptomatic carriage isolates: multilocus sequence types, clades and toxin status. Participant id follows that used in **Figure 57**.

The number of SNVs between each of the carriage samples and the closest prior and subsequent symptomatic or asymptomatic sample is shown in **Table 19**. In keeping with the significant diversity observed in symptomatic patients (see section 5.3.5) 9/18 (50%) of carriage isolates were >10 SNVs different from any other sample. However, several isolates fell within the diversity observed in common disease-causing sequence types. Five participants had a sample that was ≤2 SNVs from at least one other symptomatic or asymptomatic sample (i.e. consistent with transmission, see section 4.4). Participants 2 and 3 had non-toxigenic samples within 1 SNV of each other. Participants 7, 18, and 14 had samples which were ≤1 SNV from at least one previous symptomatic toxin A/B positive sample. Epidemiological relationships between CDI cases and asymptomatic participants with all samples genetically related within ≤2 SNVs are shown in **Figure 58**.

Participant	ST	Toxin A/B	SNVs to most closely related sample	SNVs to most closely related prior sample	SNVs to most closely related subsequent sample
7 *§	35	+/+	0	0	21
2	29	-/-	1	1	4840
18	3	+/+	1	1	13
3 *†	29	-/-	1	1	4817
14 *	3	+/+	1	1	278
10	6	+/+	5	5	21
11	35	+/+	6	10	6
6	6	+/+	8	14	8
1	2	+/+	10	10	14
16	2	+/+	18	21	18
8	26	-/-	21	21	37
5	46	+/+	24	24	48
17	13	+/+	25	37	25
15	7	-/-	57	4558	57
12	125	-/-	61	3006	61
9	11	+/+	461	461	1294
4 §	49	+/+	944	1627	944
13 *§	18	+/+	1656	2299	1656

Table 19. Single nucleotide variants, SNVs, between 18 asymptomatic carriage samples and most closely genetically related prior and subsequent symptomatic / asymptomatic sample. Participants are ordered by the number of SNVs to the most closely related sample. The participant numbering follows the same scheme used in **Figure 57**. * indicates acquisition following an initially negative sample. § indicates subsequently developed disease with the same strain, † indicates subsequently developed disease with a different strain.

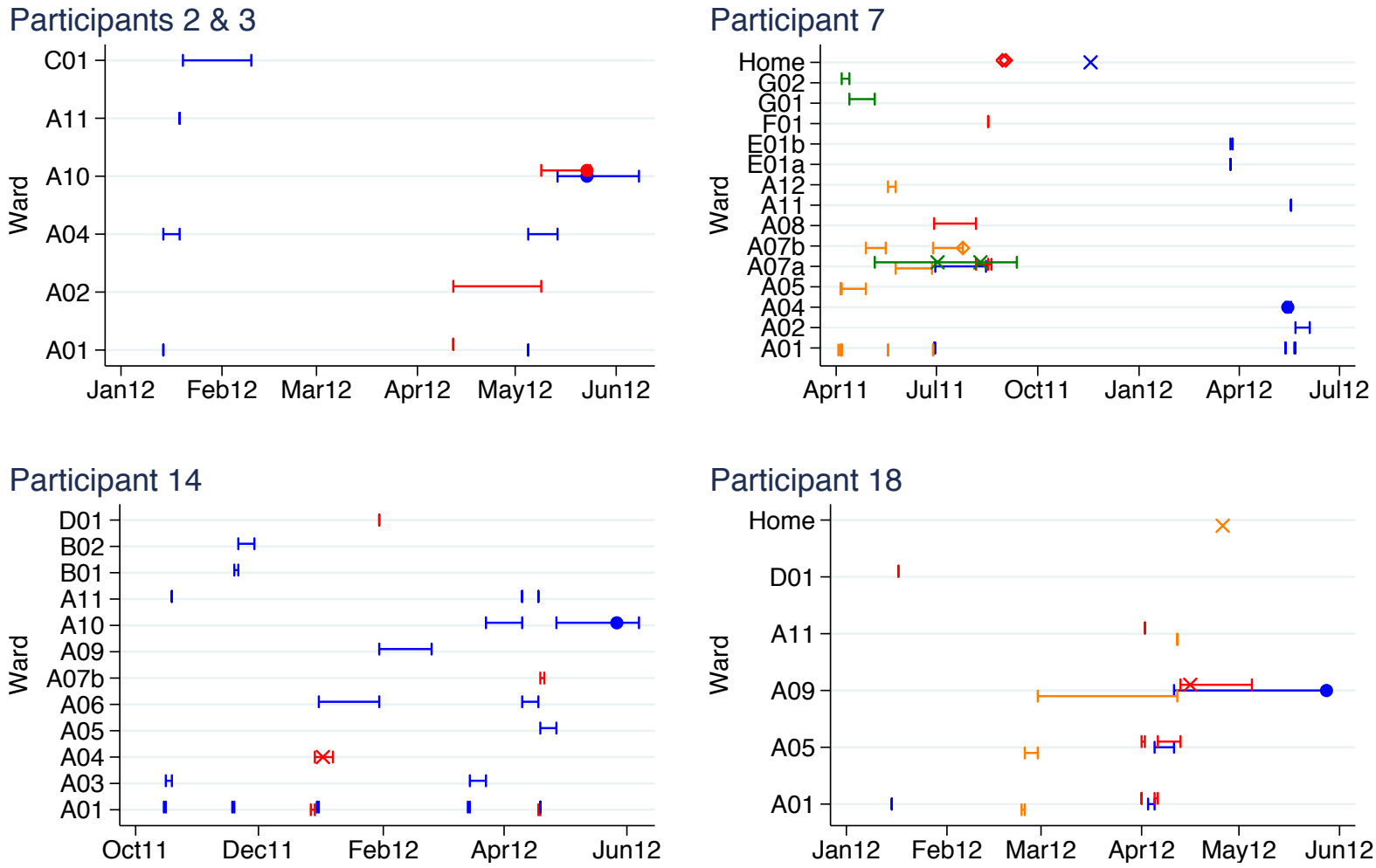


Figure 58. Epidemiological relationships between 4 asymptomatic study participants and genetically related cases. Study participants are shown in blue, with the exception of participant 3, shown in red in the first panel. Symptomatic cases are shown as different colours, and are distinct across different panels. Positive asymptomatic samples from study participants are shown as filled circles. Positive symptomatic samples are shown as crosses. EIA-negative culture-positive samples are shown as diamonds. Ward stays are shown as horizontal lines with capped ends. Wards sharing the same specialty and hospital share the same initial letter; adjacent wards forming a single unit have the same letter and number and are followed by a lower case letter.

Several plausible transmission events involving asymptomatic *C. difficile* carriers can be seen from the combined genetic and epidemiological data. Participants 2 and 3 shared time on a ward together before both were found to be asymptomatic carriers with isolates 1 SNV different, either representing transmission from one asymptomatic participant to another, or both having been exposed to a common source. Participant 18 shared time on a ward with a symptomatic patient before being found to carry *C. difficile* only 1 SNV different to the symptomatic case. The epidemiological linkage between participant 14 and the most closely genetically related case is less strong, with only a short overlap between the two patients prior to the participant's sample. Participant 7's asymptomatic carriage strain appears to have been acquired nearly a year previously, with an intervening symptomatic episode, while sharing time on a ward with 3 other cases, 2 with identical sequences, and 1 a single SNV different. Interestingly 2 of these 3 cases were EIA-negative on an initial screen, although toxin genes were present in the sequences. Although it can be difficult to infer the direction of transmission with often only a single sample from asymptomatic carriers and symptomatic cases, in this relatively small study, no clear evidence of onward transmission from an asymptomatic case was seen despite sequencing nearly all of the subsequent CDI cases for 3 months.

6.4 Discussion

This study provides several useful insights into the role of asymptomatic carriage in transmission of *C. difficile* in Oxfordshire. The estimated point prevalence of asymptomatic carriage in medical inpatients of 6% or 9% (depending on the definitions used to define cases as asymptomatic) is consistent with previous

estimates in other populations.[39] The majority of asymptomatically carried strains were toxigenic and from disease-causing STs. Examples were seen of asymptomatic carriage followed by symptomatic disease due to an identical strain after 1, 8, and 35 days, supporting previous data that incubation periods may be up to a month or more[105] as well occasionally being just days.[86,88] An example was also seen of prolonged asymptomatic carriage following a symptomatic episode 6 months previously. This is in keeping with the data presented in section 4.6.1 suggesting half or more of recurrences up to a year later are due to the same strain. It raises the question whether patients with previous CDI should be isolated at admission to hospital, in a way analogous to patients with previous methicillin-resistant *Staphylococcus aureus* carriage or infection.

The relatively small size of the study means it has limited power to detect risk factors for asymptomatic *C. difficile* carriage. Previous hospital exposure and steroid/immunosuppressant medication were identified as independent risk factors. Patient reported loose or increased stool frequency also increased the risk of carriage, although arguably these patients may have CDI not captured by the relatively restrictive definition of diarrhoea (≥3 unformed stools in 24 hours) currently in use. The finding that patient-reported antibiotic use in the last six months was associated with decreased odds of carriage is unexpected. There are several possible explanations. Enrolment to the study was conditional on participants not having had a recent CDI; it is therefore possible that a proportion of patients with *C. difficile* carriage receiving antibiotics were selected out, as they developed CDI. Recall bias is a potential concern, with more active participants with less healthcare exposure possibly more likely to recall antibiotic use and have

lower intrinsic risk of carriage. *C. difficile* strains sensitive to antibiotics taken by participants may also have been removed by antibiotic use. Trends towards gastric acid suppression and nursing/residential home residence increasing carriage were seen on univariate analysis and may merit further investigation in larger studies in future.

By using whole genome sequencing combined with ward movement data several potential acquisitions of asymptomatic carriage from symptomatic cases and other carriers were observed, demonstrating the utility of this technique in this setting. Interestingly 3 of 4 acquisitions observed were ≤1SNV from another case/carrier and 2 of 3 such acquisitions developed disease. However, despite sequencing all positive cases in the region for ≥3 months after the 18 asymptomatic cases were sampled, no evidence of subsequent genetically-related cases was seen. This questions the extent to which onward transmission from asymptomatic cases contributes to disease, despite the similarities observed between disease causing strains and the asymptomatic carriage isolates. However if there are 10-20 asymptomatic carriers for every case, as suggested in the introduction to this chapter, only relatively few asymptomatic carriers are likely to transmit *C. difficile* resulting in a symptomatic case. Therefore much larger studies are likely to be required to quantify the extent to which asymptomatic carriers account for transmission.

A real challenge in conducting this study was obtaining any samples from patients, and obtaining several samples from patients throughout their hospital stay proved extremely difficult. As designed, opportunistically sampling patients when they open their bowels, the study was highly acceptable to patients with 230/243

(95%) patients approached agreeing to participate. However obtaining samples therefore depended on the goodwill of ward nursing staff to assist patients in returning samples, and on participants' bowel habit. Both factors contributed to relatively low rates of sample return, with only just over half of participants recruited returning one or more samples. An alternative approach would have been to take rectal or perianal swabs from patients. As this is more invasive, participation rates are likely to be lower; in a large Canadian study using rectal swabs and taking blood samples from patients participation rates were 57%.[39] However this approach enables samples to be obtained by dedicated research staff, and provides control over the timing of samples from consenting patients. Although overall either approach results in similar proportions of patients being sampled at least once, if sufficient resources are available, by direct swabbing much more complete serial sampling of participants is likely to be possible. Sampling by dedicated research staff is also likely to ensure sample return is much more sustainable.

Another important limitation in any carriage study is how closely the limit of detection of the organism corresponds to complete absence of the organism. In this study the culture methods used are well supported,[39,242] but a single possible false negative result was observed. Undetected carriage may also contribute to transmission; although it seems plausible that *C. difficile* in detectable quantities is more likely to be transmitted, and therefore likely to be a more significant potential source of infection than undetectable carriage.

Overall, this study supports data from other settings on the prevalence of asymptomatic carriage, and provides clear evidence for an overlap between

carried strains and those causing disease. It also suggests potential patient populations that may be at increased risk of carriage. Whole genome sequencing provides a powerful tool for studying potential transmission events involving asymptomatic carriers, however much larger studies are required to quantify the contribution of asymptomatic carriage to onward transmission of *C. difficile* leading to disease.

7 CONCLUSIONS AND FUTURE WORK

7.1 UNDERSTANDING TRANSMISSION

Clostridium difficile continues to be a major source of disease in patients, particularly following antibiotic exposure. At the start of the work undertaken in this thesis, the prevailing understanding was that most CDI cases were the result of recent acquisition from another symptomatic case in a healthcare setting. However the studies presented demonstrate that within a large UK teaching hospital group, with typical CDI incidence rates, and infection control in line with standard published guidelines, only the minority of CDI is attributable to transmission from other symptomatic cases. In chapter 3, using MLST and ward movement data, ward-based transmission from symptomatic cases could only account for a maximum of 25% of disease. With the much greater genetic resolution available from whole genome sequencing, in chapter 5 it was possible to show that the majority of Oxfordshire cases were sufficiently genetically distinct from any previous symptomatic case that – regardless of the route of transmission – most cases acquire their *C. difficile* from reservoirs outside those sampled in symptomatic cases.

However, importantly work in this thesis does not discount ward-based transmission of *C. difficile*. Amongst CDI cases with isolates that are genetically related to one or more previous cases, most have shared time and space on a hospital ward with the earlier case, suggesting ward-based transmission does account for the majority of genetically-related cases. There also appears to be potential for spread across hospital wards, for occasional transmission through

residual ward contamination remaining after patient discharge, and through contact in the community. Intriguingly around a third of genetically related cases have no clear hospital or community link in common. More detailed investigation of these cases did not reveal an obvious single source of infection, but suggested the sources may be multifactorial. A number of these genetically related cases without obvious epidemiological links might be the result of longer transmission chains, with several intermediate hosts. As such it may not be possible to directly link the two cases sampled. Investigation of such cases in real-time may potentially yield significant insights in the future. Traditional epidemiological fieldwork may establish factors in common that reveal novel routes of transmission. Intensive sampling of patients and the environment around cases may enable other links to be found between genetically related cases, for example a ward previously shared by both patients at different times may still contain the spores that infected both patients, or sampling an asymptomatic patient, visitor, or staff member in contact with both cases may yield the infecting isolate. However, as sampling by definition only occurs after diagnosis of both cases, it may be difficult to separate potential sources of transmission from secondary spread or contamination that has occurred as a result of the cases. Therefore the ability of such investigations to shed light on the routes of transmission responsible for the majority of *C. difficile* infections that are genetically distinct for each other may be limited and will critically depend on turnaround time for culturing and sequencing pathogen DNA.

Although this thesis provides compelling evidence in the setting studied that only a minority of *C. difficile* is acquired from other cases, it raises the major new challenge of where the majority of *C. difficile* infections come from. New sampling

frames targeting potential reservoirs of disease are likely to be required to investigate this question further. Reservoirs of potential interest include asymptomatic adults, children, healthcare workers, and the wider environment, water and the food chain in particular. Concurrent sampling of possible reservoirs and CDI cases is helpful, particularly with the high level of genetic discrimination provided by whole genome sequencing. However, simultaneous serial sampling of patients, their contacts, and environment is needed to demonstrate not simply that overlapping strains exist in different populations/environments, but to demonstrate the direction of transmission. For example, convincing evidence of transmission might first establish the presence of a given strain in a patient's diet, with the patient initially testing negative for this strain, then subsequently acquiring it and progressing to disease as a result. Phylogenetic studies of infecting strains may also prove informative, with more ancestral strains likely to be found in disease reservoirs, and contamination arising secondary to infected cases or carriers more likely to contain novel mutations reflecting recent unique evolution. However the rate of *C. difficile* evolution, over the parts of the genome routinely resolved by current sequencing technology, is slow enough that reservoir, case, and subsequent isolates may in many cases all have identical sequences. Further, the number of potential reservoirs to be sampled longitudinally for each case raises significant logistical and financial barriers.

The feasibility study conducted in chapter 6 demonstrates some of the challenges in investigating asymptomatic individuals who are likely to be an important part of future sampling frames, given rates of asymptomatic carriage of ≥3% in the general population and acquisition of carriage in hospitals.[39] Regular sampling

of participants in a hospital setting is only likely to be possible with dedicated research staff and the use of relatively invasive rectal or perianal swabs. Clearly this has implications for the numbers of people willing to participate and therefore the proportion of patients in a particular ward or hospital that can be sampled at any one time. Therefore if asymptomatic patients are a source of onward transmission, only some of these transmission events will be captured, and the total proportion of cases attributable to asymptomatic carriers will need to be estimated from participation rates. A sub-group of patients with *C. difficile* carriage, but without infection, are those with diarrhoea due to other causes. These patients, *C. difficile* 'excretors', may carry a toxigenic strain of *C. difficile*, but not produce sufficient toxin to cause disease. They may have enhanced potential to transmit *C. difficile* given the increased environmental contamination associated with their diarrhoea. Such patients may be suspected of having CDI and tested by clinicians. They are identified by two-step testing strategies as carrying *C. difficile*, but without producing toxin. Isolates from such cases in Oxfordshire have been stored since April 2012, alongside those from symptomatic cases and allowing this route of transmission to be investigated in the near future.

7.2 Infection control implications

This thesis also has significant implications for strategies to control *C. difficile* infection. However, it is important to note that the findings are contingent on the infection control and antimicrobial policies in place in the hospitals and region studied. As such the limited amount of transmission from symptomatic cases should probably be regarded as evidence of the benefits of infection control measures rather than evidence they are not required.

The similar falls in the incidence of genetically-related and genetically-distinct cases seen in Oxfordshire (section 5.3.6) suggest strategies, such as antimicrobial stewardship, that prevent *C. difficile* regardless of the source of transmission, have been particularly important in falls in CDI incidence seen in the UK over the last 5 years. In fact, given *C. difficile* arises from genetically diverse sources, and not simply just symptomatic patients, further tightening of hospital infection control is unlikely to have much impact beyond the <20% of cases that may be the result of transmission from known cases in hospital. This should be considered when developing policies to promote further reductions in *C. difficile* incidence.

Repeating a similar study in a setting of static or rising *C. difficile* incidence, such as in parts of North America, would be of significant interest. It should be able to determine if the increased incidence is the result of more transmission from symptomatic cases, suggesting scope for improved hospital infection control. Alternatively, and perhaps more likely given the infection control already in place is similar to the UK, a similar pattern of genetically diverse infecting strains may be seen. This should prompt a review of antimicrobial use, particularly antimicrobials that local *C. difficile* strains are known to carry resistance to.

It may be that *C. difficile* is sufficiently environmentally ubiquitous that preventing all exposure is not feasible. Therefore in addition to antimicrobial stewardship, future research should include other measures to prevent progression from colonisation to disease. There are current efforts to develop a vaccine (e.g. [243], a bivalent vaccine containing inactivated toxoids A and B), while others have investigated the potential of human monoclonal antibodies to toxins A and B to prevent disease.[244] Other approaches might be to prophylactically protect or

restore the faecal microbiome in patients receiving antibiotics, or to give specific anti-*C. difficile* prophylaxis alongside antibiotics, for example using fidaxomicin which causes little disruption to the faecal microbiome itself.[245] Of note, current costs of fidaxomicin (£1350 per 10 day course) make this strategy prohibitively costly.

If asymptomatic individuals are found to be an important source of transmission this also has potentially major implications for control. For example screening all admissions to hospital with perianal/rectal swabs is likely to be regarded by many as undignified. However there is a precedent for this, since large-scale use of rectal screening for vancomycin-resistant enterococci in high risk patients is common practice in some hospitals (e.g. [246]). If patients are screened, then isolation of patients found to be carriers might be problematic in settings with limited numbers of single patient rooms. Further studies on the impact of treatment of asymptomatic carriage would also be required. Although there is little evidence at present for transmission from asymptomatic visitors and staff, this would pose an even greater challenge.

Whole genome sequencing of *C. difficile* is also likely to result in wider infection control benefits. Sequencing cases in real-time with available benchtop sequencers can deliver much more focused outbreak investigation.[198] Clusters of cases can be rapidly classified into those that are genetically distinct, i.e. not the result of transmission, and those that are genetically related, where transmission is a strong possibility. This allows infection control efforts to be targeted to prevent further transmission in the latter case, and to focus on general risk factors for *C. difficile* such as antimicrobial use in the former.

Currently identification of potential outbreaks within hospitals relies on ward-based monitoring for increased incidence. However as patients are frequently moved around hospital wards (e.g. see **Figure 58**), several transmission events from a case may occur on different wards, such that detection of an outbreak is obscured. One potential solution is real-time monitoring of contact networks between patients, but this is still imperfect as certain contacts may be missed. Genomically driven epidemiological investigation provides an alternative solution, where all cases are sequenced as they occur: genetic matches between cases are identified, and then flagged for more detailed investigation as potential outbreaks. Such an approach is much more likely to detect outbreaks mediated by novel routes, for example contaminated food, instruments, or a member of staff.

Whole genome sequencing also offers the prospect of improved metrics of infection control performance in institutions and healthcare systems. Rather than simply assessing crude incidence, the rates of genetically-related secondary cases can be monitored as an approximation to the amount of transmission occurring. This metric may also be useful in clinical trials of interventions designed to interrupt transmission.

7.3 SEQUENCING TECHNOLOGY DEVELOPMENTS

As sequencing technology and associated knowledge develops, several other applications relevant to *C. difficile* are likely to become possible. Sequence data can be used to assess for the presence of known virulence and antimicrobial resistance determinants. This has already be shown to be possible for a number of organisms, for example *Staphylococcus aureus*.[198,216,247] As antimicrobial stewardship

appears to be central to control of *C. difficile*, monitoring for the emergence of resistance to frequently used antibiotics may be relevant, as well as monitoring for resistance to *C. difficile* treatments, e.g. to the new drug fidaxomicin.[248]

As the volume of data provided by sequencing increases with advances in technology, sequencing may also become a useful diagnostic tool. Sequencing total DNA present in stool samples may allow identification of multiple potential pathogens, and may even identify novel pathogens. It may also allow assessment of the current microbiome in patients, which could be of prognostically relevant in predicting risk of *C. difficile* infection, relapse, or more severe disease. The challenge may be identifying a causative pathogen amongst multiple carried microorganisms, e.g. if norovirus and *C. difficile* are detected, both in sufficient frequency to cause disease.

Strand sequencing is currently being developed to offer much greater read lengths than are possible with current technology.[178] In future the 10-15% of the *C. difficile* genome not current called may be resolved, including elements with different patterns of change over time, such as repetitive and mobile elements. These may offer further improved discrimination between isolates over short time scales providing even finer resolution for transmission studies. Alternatively, they may prove to be hyper-variable and not useful (e.g. if significant back mutation occurs).

More broadly, beyond just *C. difficile*, the advent of whole genome sequencing is likely to transform research and service microbiology.[220] Traditional tools are likely to be increasingly replaced by sequencing-based species identification, strain

typing, and antimicrobial resistance testing. This offers the prospect of more efficient and rapid workflows, which should, amongst many other improvements, make the work undertaken in this thesis possible across a whole range of pathogens.

7.4 Conclusions

Given current UK infection control and antimicrobial usage, this thesis establishes that the majority of *C. difficile* infections do not arise as a result of transmission from other symptomatic cases. Therefore strategies to prevent *C. difficile* regardless of the source of infection should receive particular attention, both in terms of clinical practice, where antimicrobial stewardship is especially important, and future research. We must also better understand where *C. difficile* comes from, and the work presented should motivate a detailed search for additional sources from which *C. difficile* transmission occurs.

8 REFERENCES

1. Bartlett JG. Narrative review: the new epidemic of Clostridium difficile-associated enteric disease. Ann Intern Med **2006**; 145:758–764.

2. Bartlett JG, Chang TW, Gurwith M, Gorbach SL, Onderdonk AB. Antibiotic-associated pseudomembranous colitis due to toxin-producing clostridia. N Engl J Med **1978**; 298:531–534.

3. Loo VG, Poirier L, Miller MA, et al. A predominantly clonal multi-institutional outbreak of Clostridium difficile-associated diarrhea with high morbidity and mortality. N Engl J Med **2005**; 353:2442–2449.

4. Vonberg R-P, Kuijper EJ, Wilcox MH, et al. Infection control measures to limit the spread of Clostridium difficile. Clin Microbiol Infect **2008**; 14 Suppl 5:2–20.

5. Muto CA, Blank MK, Marsh JW, et al. Control of an outbreak of infection with the hypervirulent Clostridium difficile BI strain in a university hospital using a comprehensive 'bundle' approach. Clin Infect Dis **2007**; 45:1266–1273.

6. Valiquette L, Cossette B, Garant M-P, Diab H, Pepin J. Impact of a reduction in the use of high-risk antibiotics on the course of an epidemic of Clostridium difficile-associated disease caused by the hypervirulent NAP1/027 strain. Clin Infect Dis **2007**; 45 Suppl 2:S112–21.

7. Carroll KC, Bartlett JG. Biology of Clostridium difficile: implications for epidemiology and diagnosis. Annu Rev Microbiol **2011**; 65:501–521.

8. Hall IC, O'Toole E. Interstinal flora in new-bown infants with a description of a new pathogen anaerobe, bacillus difficilis. Am J Dis Child **1935**; 49:390–402.

9. Hall IC, O'Toole E. Bacterial flora of first specimens of meconium passed by fifty new-born infants. Arch Pediatr Adolesc Med **1934**; 47:1279–1285.

10. Sebald M, Hauser D. Pasteur, oxygen and the anaerobes revisited. Anaerobe **1995**; 1:11–16.

11. Hall IC. Differentiation and identification of the sporulating anaerobes. J Infect Dis **1922**; 30:445–504.

12. Surawicz C, McFarland L. Pseudomembranous Colitis: Causes and Cures. Digestion **1999**; 60:91–100.

13. Mogg GA, Keighley MR, Burdon DW, et al. Antibiotic-associated colitis - a review of 66 cases. Br J Surg **1979**; 66:738–742.

14. Khan M, Hall W. Staphylococcal enterocolitis - treatment with oral vancomycin. Ann Intern Med **1966**; 65:1–8.

15. Tedesco FJ, Barton RW, Alpers DH. Clindamycin-associated colitis. A prospective study. Ann Intern Med **1974**; 81:429–433.

16. Marr JJ, Sans MD, Tedesco FJ. Bacterial studies of Clindamycin-associated colitis. A preliminary report. Gastroenterology **1975**; 69:352–358.

17. Steer HW. The pseudomembranous colitis associated with clindamycin therapy--a viral colitis. Gut **1975**; 16:695–706.

18. Larson HE, Parry JV, Price AB, Davies DR, Dolby J, Tyrrell DA. Undescribed toxin in pseudomembranous colitis. BMJ **1977**; 1:1246–1248.

19. Bartlett JG, Onderdonk AB, Cisneros RL, Kasper DL. Clindamycin-Associated Colitis Due to a Toxin-Producing Species of Clostridium in Hamsters. J Infect Dis **1977**; 136:701–705.

20. Ballantyne EE. Gas Gangrene Antitoxin Production. Can J Comp Med Vet Sci **1944**; 8:108.

21. Blevins SM, Bronze MS. Robert Koch and the 'golden age' of bacteriology. Int J Infect Dis **2010**; 14:e744–51.

22. Kaufmann SHE, Schaible UE. 100th anniversary of Robert Koch's Nobel Prize for the discovery of the tubercle bacillus. Trends Microbiol **2005**; 13:469–475.

23. Bartlett JG, Taylor NS, Chang T, Dzink J. Clinical and laboratory observations in Clostridium difficile colitis. Am J Clin Nutr **1980**; 33:2521–2526.

24. Bartlett JG. Antibiotic-associated diarrhea. Clin Infect Dis **1992**; 15:573–581.

25. McFarland LV. Epidemiology, risk factors and treatments for antibiotic-associated diarrhea. Dig Dis **1998**; 16:292–307.

26. Hempel S, Newberry SJ, Maher AR, et al. Probiotics for the prevention and treatment of antibiotic-associated diarrhea: a systematic review and meta-analysis. JAMA **2012**; 307:1959–1969.

27. Sorg JA, Sonenshein AL. Bile salts and glycine as cogerminants for Clostridium difficile spores. J Bacteriol **2008**; 190:2505–2512.

28. Lawley TD, Croucher NJ, Yu L, et al. Proteomic and genomic characterization of highly infectious Clostridium difficile 630 spores. J Bacteriol **2009**; 191:5377–5386.

29. Cohen SH, Gerding DN, Johnson S, et al. Clinical practice guidelines for Clostridium difficile infection in adults: 2010 update by the society for healthcare epidemiology of America (SHEA) and the infectious diseases society of America (IDSA). Infect Control Hosp Epidemiol **2010**; 31:431–455.

30. Kelly CP, Pothoulakis C, LaMont JT. Clostridium difficile colitis. N Engl J Med **1994**; 330:257–262.

31. Jump RLP, Pultz MJ, Donskey CJ. Vegetative Clostridium difficile survives in room air on moist surfaces and in gastric contents with reduced acidity: a potential mechanism to explain the association between proton pump inhibitors and C. difficile-associated diarrhea? Antimicrob Agents Chemother **2007**; 51:2883–2887.

32. Tleyjeh IM, Bin Abdulhak AA, Riaz M, et al. Association between Proton Pump Inhibitor Therapy and Clostridium difficile Infection: A Contemporary Systematic Review and Meta-Analysis. PLoS ONE **2012**; 7:e50836.

33. Leonard J, Marshall JK, Moayyedi P. Systematic review of the risk of enteric infection in patients taking acid suppression. Am J Gastroenterol **2007**; 102:2047–56– quiz 2057.

34. Panessa-Warren BJ, Tortora GT, Warren JB. High resolution FESEM and TEM reveal bacterial spore attachment. Microsc Microanal **2007**; 13:251–266.

35. Hurley BW, Nguyen CC. The spectrum of pseudomembranous enterocolitis and antibiotic-associated diarrhea. Arch Intern Med **2002**; 162:2177–2184.

36. Dubberke ER, Wertheimer AI. Review of current literature on the economic burden of Clostridium difficile infection. Infect Control Hosp Epidemiol **2009**; 30:57–66.

37. Bartlett JG. Clostridium difficile: history of its role as an enteric pathogen and the current state of knowledge about the organism. Clin Infect Dis **1994**; 18 Suppl 4:S265–72.

38. Enoch DA, Butler MJ, Pai S, Aliyu SH, Karas JA. Clostridium difficile in children: colonisation and disease. J Infect **2011**; 63:105–113.

39. Loo VG, Bourgault A-M, Poirier L, et al. Host and pathogen factors for Clostridium difficile infection and colonization. N Engl J Med **2011**; 365:1693–1703.

40. Barbut F, Petit JC. Epidemiology of Clostridium difficile-associated infections. Clin Microbiol Infect **2001**; 7:405–410.

41. Rea MC, O'Sullivan O, Shanahan F, et al. Clostridium difficile carriage in elderly subjects and associated changes in the intestinal microbiota. J Clin Microbiol **2012**; 50:867–875.

42. Dethlefsen L, Huse S, Sogin ML, Relman DA. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. PLoS Biol **2008**; 6:e280.

43. Jernberg C, Löfmark S, Edlund C, Jansson JK. Long-term impacts of antibiotic exposure on the human intestinal microbiota. Microbiology (Reading, Engl.) **2010**; 156:3216–3223.

44. Cohen SH, Tang YJ, Silva J. Analysis of the pathogenicity locus in Clostridium difficile strains. J Infect Dis **2000**; 181:659–663.

45. Voth DE, Ballard JD. Clostridium difficile toxins: mechanism of action and role in disease. Clin Microbiol Rev **2005**; 18:247–263.

46. Kuehne SA, Cartman ST, Heap JT, Kelly ML, Cockayne A, Minton NP. The role of toxin A and toxin B in Clostridium difficile infection. Nature **2010**; 467:711–713.

47. Lyras D, O'Connor JR, Howarth PM, et al. Toxin B is essential for virulence of Clostridium difficile. Nature **2009**; 458:1176–1179.

48. Alfa MJ, Kabani A, Lyerly D, et al. Characterization of a toxin A-negative, toxin B-positive strain of Clostridium difficile responsible for a nosocomial outbreak of Clostridium difficile-associated diarrhea. J Clin Microbiol **2000**; 38:2706–2714.

49. Drudy D, Fanning S, Kyne L. Toxin A-negative, toxin B-positive Clostridium difficile. Int J Infect Dis **2007**; 11:5–10.

50. Drudy D, Harnedy N, Fanning S, Hannan M, Kyne L. Emergence and Control of Fluoroquinolone-Resistant, Toxin A–Negative, Toxin B–Positive Clostridium difficile. Infect Control Hosp Epidemiol **2007**; 28:932–940.

51. McDonald LC, Killgore GE, Thompson A, et al. An epidemic, toxin gene-variant strain of Clostridium difficile. N Engl J Med **2005**; 353:2433–2441.

52. Gülke I, Pfeifer G, Liese J, et al. Characterization of the enzymatic component of the ADP-ribosyltransferase toxin CDTa from Clostridium difficile. Infect. Immun. **2001**; 69:6004–6011.

53. Curry SR, Marsh JW, Muto CA, O'Leary MM, Pasculle AW, Harrison LH. tcdC genotypes associated with severe TcdC truncation in an epidemic clone and other strains of Clostridium difficile. J Clin Microbiol **2007**; 45:215–221.

54. Cartman ST, Kelly ML, Heeg D, Heap JT, Minton NP. Precise manipulation of the Clostridium difficile chromosome reveals a lack of association between the tcdC genotype and toxin production. Appl Environ Microbiol **2012**; 78:4683–4690.

55. Health Protection Agency. Quarterly Epidemiological Commentary: Mandatory MRSA bacteraemia, *Clostridium difficile* infection & MSSA bacteraemia (up to January – March 2012). 2012. Available at: http://www.hpa.org.uk/webc/HPAwebFile/HPAweb_C/1284473407318. Accessed 12 July 2012.

56. Health Protection Agency. Quarterly Epidemiological Commentary: Mandatory MRSA bacteraemia, *Clostridium difficile* infection & MSSA bacteraemia (up to January – March 2011). 2011. Available at: http://www.hpa.org.uk/web/HPAwebFile/HPAweb_C/1284473407318. Accessed 12 August 2011.

57. Lucado J, Gould C, Elixhauser A. *Clostridium difficile* infections (CDI) in hospital stays, 2009. 2012. Available at: http://www.hcup-us.ahrq.gov/reports/statbriefs/sb124.pdf. Accessed 4 December 2012.

58. Freeman J, Bauer MP, Baines SD, et al. The Changing Epidemiology of Clostridium difficile Infections. Clin Microbiol Rev **2010**; 23:529–549.

59. Goorhuis A, van der Kooi T, Vaessen N, et al. Spread and epidemiology of Clostridium difficile polymerase chain reaction ribotype 027/toxinotype III in The Netherlands. Clin Infect Dis **2007**; 45:695–703.

60. Office for National Statistics. Deaths involving *Clostridium difficile*: England and Wales - 2009. 2010. Available at: http://www.statistics.gov.uk/pdfdir/cdif0810.pdf. Accessed 12 August 2011.

61. Office for National Statistics. Death involving *Clostridium difficile*: England and Wales 2006 to 2010. 2011. Available at: http://www.ons.gov.uk/ons/rel/subnational-health2/deaths-involving-clostridium-difficile/2006-to-2010/statistical-bulletin.html. Accessed 12 July 2012.

62. McGlone SM, Bailey RR, Zimmer SM, et al. The economic burden of Clostridium difficile. Clin Microbiol Infect **2012**; 18:282–289.

63. Ghantoji SS, Sail K, Lairson DR, DuPont HL, Garey KW. Economic healthcare costs of Clostridium difficile infection: a systematic review. J Hosp Infect **2010**; 74:309–318.

64. Dubberke ER, Butler AM, Reske KA, et al. Attributable outcomes of endemic Clostridium difficile-associated disease in nonsurgical patients. Emerging Infect Dis **2008**; 14:1031–1038.

65. Owens RC, Donskey CJ, Gaynes RP, Loo VG, Muto CA. Antimicrobial-associated risk factors for Clostridium difficile infection. Clin Infect Dis **2008**; 46 Suppl 1:S19–31.

66. Bignardi GE. Risk factors for Clostridium difficile infection. J Hosp Infect **1998**; 40:1–15.

67. Pepin J, Saheb N, Coulombe M-A, et al. Emergence of fluoroquinolones as the predominant risk factor for Clostridium difficile-associated diarrhea: a cohort study during an epidemic in Quebec. Clin Infect Dis **2005**; 41:1254–1260.

68. Gerding DN, Olson MM, Peterson LR, et al. Clostridium difficile-associated diarrhea and colitis in adults. A prospective case-controlled epidemiologic study. Arch Intern Med **1986**; 146:95–100.

69. Hensgens MPM, Goorhuis A, Dekkers OM, Kuijper EJ. Time interval of increased risk for Clostridium difficile infection after exposure to antibiotics. J Antimicrob Chemother **2012**; 67:742–748.

70. McFarland LV, Surawicz CM, Stamm WE. Risk factors for Clostridium difficile carriage and C. difficile-associated diarrhea in a cohort of hospitalized patients. J Infect Dis **1990**; 162:678–684.

71. Safdar N, Maki DG. The commonality of risk factors for nosocomial colonization and infection with antimicrobial-resistant Staphylococcus aureus, enterococcus, gram-negative bacilli, Clostridium difficile, and Candida. Ann Intern Med **2002**; 136:834–844.

72. Brown E, Talbot GH, Axelrod P, Provencher M, Hoegg C. Risk factors for Clostridium difficile toxin-associated diarrhea. Infect Control Hosp Epidemiol **1990**; 11:283–290.

73. McDonald LC, Owings M, Jernigan DB. Clostridium difficile infection in patients discharged from US short-stay hospitals, 1996-2003. Emerging Infect Dis **2006**; 12:409–415.

74. Pant C, Madonia P, Minocha A. Does PPI therapy predispose to Clostridium difficile infection? Nat Rev Gastroenterol Hepatol **2009**; 6:555–557.

75. Asha NJ, Tompkins D, Wilcox MH. Comparative Analysis of Prevalence, Risk Factors, and Molecular Epidemiology of Antibiotic-Associated Diarrhea Due to

Clostridium difficile, Clostridium perfringens, and Staphylococcus aureus. J Clin Microbiol **2006**; 44:2785–2791.

76. Dubberke ER, Reske KA, Olsen MA, et al. Evaluation of Clostridium difficile-associated disease pressure as a risk factor for C difficile-associated disease. Arch Intern Med **2007**; 167:1092–1097.

77. Shim JK, Johnson S, Samore MH, Bliss DZ, Gerding DN. Primary symptomless colonisation by Clostridium difficile and decreased risk of subsequent diarrhoea. Lancet **1998**; 351:633–636.

78. Johnson S, Clabots CR, Linn FV, Olson MM, Peterson LR, Gerding DN. Nosocomial Clostridium difficile colonisation and disease. Lancet **1990**; 336:97–100.

79. Kyne L, Warny M, Qamar A, Kelly CP. Asymptomatic carriage of Clostridium difficile and serum levels of IgG antibody against toxin A. N Engl J Med **2000**; 342:390–397.

80. Zelner JL, King AA, Moe CL, Eisenberg JNS. How infections propagate after point-source outbreaks: an analysis of secondary norovirus transmission. Epidemiology **2010**; 21:711–718.

81. Glynn JR, Palmer SR. Incubation period, severity of disease, and infecting dose: evidence from a Salmonella outbreak. Am J Epidemiol **1992**; 136:1369–1377.

82. Medley GF, Anderson RM, Cox DR, Billard L. Incubation period of AIDS in patients infected via blood transfusion. Nature **1987**; 328:719–721.

83. Hornick RB, Greisman SE, Woodward TE, DuPont HL, Dawkins AT, Snyder MJ. Typhoid fever: pathogenesis and immunologic control. N Engl J Med **1970**; 283:686–691.

84. Magnuson HJ, Thomas EW, Olansky S, Kaplan BI, de Mello L, Cutler JC. Inoculation syphilis in human volunteers. Medicine (Baltimore) **1956**; 35:33–82.

85. Cook G, Zumla A. Manson's Tropical Diseases. 22nd ed. Philadelphia, PA: Saunders Ltd, 2008.

86. McFarland LV, Mulligan ME, Kwok RY, Stamm WE. Nosocomial acquisition of Clostridium difficile infection. N Engl J Med **1989**; 320:204–210.

87. Clabots CR, Johnson S, Olson MM, Peterson LR, Gerding DN. Acquisition of Clostridium difficile by hospitalized patients: evidence for colonized new admissions as a source of infection. J Infect Dis **1992**; 166:561–567.

88. Samore MH, DeGirolami PC, Tlucko A, Lichtenberg DA, Melvin ZA, Karchmer AW. Clostridium difficile colonization and diarrhea at a tertiary care hospital. Clin Infect Dis **1994**; 18:181–187.

89. Chang VT, Nelson K. The role of physical proximity in nosocomial diarrhea. Clin Infect Dis **2000**; 31:717–722.

90. Samore MH, Venkataraman L, DeGirolami PC, Arbeit RD, Karchmer AW. Clinical and molecular epidemiology of sporadic and clustered cases of nosocomial Clostridium difficile diarrhea. Am J Med **1996**; 100:32–40.

91. McDonald LC, Coignard B, Dubberke E, et al. Recommendations for surveillance of Clostridium difficile-associated disease. Infect Control Hosp Epidemiol **2007**; 28:140–145.

92. Norén T, Akerlund T, Bäck E, et al. Molecular epidemiology of hospital-associated and community-acquired Clostridium difficile infection in a Swedish county. J Clin Microbiol **2004**; 42:3635–3643.

93. Kim KH, Fekety R, Batts DH, et al. Isolation of Clostridium difficile from the environment and contacts of patients with antibiotic-associated colitis. J Infect Dis **1981**; 143:42–50.

94. Struelens MJ, Maas A, Nonhoff C, et al. Control of nosocomial transmission of Clostridium difficile based on sporadic case surveillance. Am J Med **1991**; 91:138S–144S.

95. Dumford DM, Nerandzic MM, Eckstein BC, Donskey CJ. What is on that keyboard? Detecting hidden environmental reservoirs of Clostridium difficile during an outbreak associated with North American pulsed-field gel electrophoresis type 1 strains. Am J Infect Control **2009**; 37:15–19.

96. Dubberke ER, Reske KA, Noble-Wang J, et al. Prevalence of Clostridium difficile environmental contamination and strain variability in multiple health care facilities. Am J Infect Control **2007**; 35:315–318.

97. Perry C, Marshall R, Jones E. Bacterial contamination of uniforms. J Hosp Infect **2001**; 48:238–241.

98. Best EL, Fawley WN, Parnell P, Wilcox MH. The potential for airborne dispersal of Clostridium difficile from symptomatic patients. Clin Infect Dis **2010**; 50:1450–1457.

99. Donskey CJ. Preventing transmission of Clostridium difficile: is the answer blowing in the wind? Clin Infect Dis **2010**; 50:1458–1461.

100. Health Protection Agency, Department of Health. Clostridium difficile infection: How to deal with the problem. 2009. Available at: http://www.hpa.org.uk/webc/HPAwebFile/HPAweb_C/1232006607827. Accessed 6 March 2011.

101. al-Barrak A, Embil J, Dyck B, et al. An outbreak of toxin A negative, toxin B positive Clostridium difficile-associated diarrhea in a Canadian tertiary-care hospital. Can Commun Dis Rep **1999**; 25:65–69.

102. Bobulsky GS, Al-Nassir WN, Riggs MM, Sethi AK, Donskey CJ. Clostridium difficile skin contamination in patients with C. difficile-associated disease. Clin Infect Dis **2008**; 46:447–450.

103. Sethi AK, Al-Nassir WN, Nerandzic MM, Bobulsky GS, Donskey CJ. Persistence of skin contamination and environmental shedding of Clostridium difficile during and after treatment of C. difficile infection. Infect Control Hosp Epidemiol **2010**; 31:21–27.

104. Jinno S, Kundrapu S, Guerrero DM, Jury LA, Nerandzic MM, Donskey CJ. Potential for transmission of Clostridium difficile by asymptomatic acute care patients and long-term care facility residents with prior C. difficile infection. Infect Control Hosp Epidemiol **2012**; 33:638–639.

105. Palmore TN, Sohn S, Malak SF, Eagan J, Sepkowitz KA. Risk factors for acquisition of Clostridium difficile-associated diarrhea among outpatients at a cancer hospital. Infect Control Hosp Epidemiol **2005**; 26:680–684.

106. McFarland LV, Clarridge JE, Beneda HW, Raugi GJ. Fluoroquinolone Use and Risk Factors for Clostridium difficile-Associated Disease within a Veterans Administration Health Care System. Clin Infect Dis **2007**; 45:1141–1151.

107. Riggs MM, Sethi AK, Zabarsky TF, Eckstein EC, Jump RLP, Donskey CJ. Asymptomatic carriers are a potential source for transmission of epidemic and nonepidemic Clostridium difficile strains among long-term care facility residents. Clin Infect Dis **2007**; 45:992–998.

108. Simor AE, Bradley SF, Strausbaugh LJ, Crossley K, Nicolle LE, SHEA Long-Term-Care Committee. Clostridium difficile in long-term-care facilities for the elderly. Infect Control Hosp Epidemiol **2002**; 23:696–703.

109. Arvand M, Moser V, Schwehn C, Bettge-Weller G, Hensgens MP, Kuijper EJ. High prevalence of Clostridium difficile colonization among nursing home residents in Hesse, Germany. PLoS ONE **2012**; 7:e30183.

110. Simor AE, Yake SL, Tsimidis K. Infection due to Clostridium difficile among elderly residents of a long-term-care facility. Clin Infect Dis **1993**; 17:672–678.

111. Birgand G, Blanckaert K, Carbonne A, et al. Investigation of a large outbreak of Clostridium difficile PCR-ribotype 027 infections in northern France, 2006-2007 and associated clusters in 2008-2009. Euro Surveill **2010**; 15.

112. Burns K, Morris-Downes M, Fawley WN, Smyth E, Wilcox MH, Fitzpatrick F. Infection due to C. difficile ribotype 078: first report of cases in the Republic of Ireland. J Hosp Infect **2010**; 75:287–291.

113. Hensgens MPM, Keessen EC, Squire MM, et al. Clostridium difficile infection in the community: a zoonotic disease? Clin Microbiol Infect **2012**; 18:635–645.

114. Weese JS, Avery BP, Rousseau J, Reid-Smith RJ. Detection and enumeration of Clostridium difficile spores in retail beef and pork. Appl Environ Microbiol **2009**; 75:5009–5011.

115. Saif al N, Brazier JS. The distribution of Clostridium difficile in the environment of South Wales. J Med Microbiol **1996**; 45:133–137.

116. Bakri MM, Brown DJ, Butcher JP, Sutherland AD. Clostridium difficile in ready-to-eat salads, Scotland. Emerging Infect Dis **2009**; 15:817–818.

117. Metcalf DS, Costa MC, Dew WMV, Weese JS. Clostridium difficile in vegetables, Canada. Lett Appl Microbiol **2010**; 51:600–602.

118. Riley TV, Adams JE, O'Neill GL, Bowman RA. Gastrointestinal carriage of Clostridium difficile in cats and dogs attending veterinary clinics. Epidemiol Infect **1991**; 107:659–665.

119. O'Neill G, Adams JE, Bowman RA, Riley TV. A molecular characterization of Clostridium difficile isolates from humans, animals and their environments. Epidemiol Infect **1993**; 111:257–264.

120. Weese JS, Finley R, Reid-Smith RR, Janecko N, Rousseau J. Evaluation of Clostridium difficile in dogs and the household environment. Epidemiol Infect **2010**; 138:1100–1104.

121. Koene MGJ, Mevius D, Wagenaar JA, et al. Clostridium difficile in Dutch animals: their presence, characteristics and similarities with human isolates. Clin Microbiol Infect **2012**; 18:778–784.

122. Zidaric V, Beigot S, Lapajne S, Rupnik M. The occurrence and high diversity of Clostridium difficile genotypes in rivers. Anaerobe **2010**; 16:371–375.

123. Romano V, Pasquale V, Krovacek K, Mauri F, Demarta A, Dumontet S. Toxigenic Clostridium difficile PCR ribotypes from wastewater treatment plants in southern Switzerland. Appl Environ Microbiol **2012**; 78:6643–6646.

124. Pasquale V, Romano VJ, Rupnik M, et al. Isolation and characterization of Clostridium difficile from shellfish and marine environments. Folia Microbiol. (Praha) **2011**; 56:431–437.

125. Garner JS, The Hospital Infection Control Practices Advisory Committee. Guideline for isolation precautions in hospitals. Infect Control Hosp Epidemiol **1996**; 17:53–80.

126. Johnson S, Gerding DN, Olson MM, et al. Prospective, controlled study of vinyl glove use to interrupt Clostridium difficile nosocomial transmission. Am J Med **1990**; 88:137–140.

127. Wilcox MH, Fawley WN, Wigglesworth N, Parnell P, Verity P, Freeman J. Comparison of the effect of detergent versus hypochlorite cleaning on environmental contamination and incidence of Clostridium difficile infection. J Hosp Infect **2003**; 54:109–114.

128. Brooks SE, Veal RO, Kramer M, Dore L, Schupf N, Adachi M. Reduction in the incidence of Clostridium difficile-associated diarrhea in an acute care hospital and a skilled nursing facility following replacement of electronic thermometers with single-use disposables. Infect Control Hosp Epidemiol **1992**; 13:98–103.

129. Cartmill TD, Panigrahi H, Worsley MA, McCann DC, Nice CN, Keith E. Management and control of a large outbreak of diarrhoea due to Clostridium difficile. J Hosp Infect **1994**; 27:1–15.

130. Climo MW, Israel DS, Wong ES, Williams D, Coudron P, Markowitz SM. Hospital-wide restriction of clindamycin: effect on the incidence of Clostridium difficile-associated diarrhea and cost. Ann Intern Med **1998**; 128:989–995.

131. Wilcox MH, Freeman J, Fawley W, et al. Long-term surveillance of cefotaxime and piperacillin-tazobactam prescribing and incidence of Clostridium difficile diarrhoea. J Antimicrob Chemother **2004**; 54:168–172.

132. Thomas C, Riley TV. Restriction of third generation cephalosporin use reduces the incidence of Clostridium difficile-associated diarrhoea in hospitalised patients. Commun Dis Intell **2003**; 27 Suppl:S28–31.

133. Stone S, Kibbler C, How A, Balstrini A. Feedback is necessary in strategies to reduce hospital acquired infection. BMJ **2000**; 321:302–303.

134. Kufelnicka AM, Kirn TJ. Effective utilization of evolving methods for the laboratory diagnosis of Clostridium difficile infection. Clin Infect Dis **2011**; 52:1451–1457.

135. NHS Purchasing and Supply Agency. Clostridium difficile toxin detection assays. 2009. Available at: http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/documents/digitalasset/dh_127743.pdf. Accessed 27 October 2010.

136. Wilcox MH. Laboratory diagnosis of Clostridium difficile infection: in a state of transition or confusion or both? J Hosp Infect **2011**; 79:1–3.

137. Wilcox MH, Planche T. Summary of research underpinning the Department of Health's new *Clostridium difficile* testing guidance. 2011. Available at: http://www.hpa.org.uk/webc/HPAwebFile/HPAweb_C/1317132979562. Accessed 25 July 2012.

138. Wilcox MH, Planche T, Fang FC, Gilligan P. What is the current role of algorithmic approaches for diagnosis of Clostridium difficile infection? J Clin Microbiol **2010**; 48:4347–4353.

139. Fawley WN, Wilcox MH, Clostridium difficile Ribotyping Network for England and Northern Ireland. An enhanced DNA fingerprinting service to investigate potential Clostridium difficile infection case clusters sharing the same PCR ribotype. J Clin Microbiol **2011**; 49:4333–4337.

140. Roberts RJ, Murray K. Restriction endonuclease. Crit Rev Biochem Mol **1976**; 4:123–164.

141. Clabots CR, Johnson S, Bettin KM, et al. Development of a rapid and efficient restriction endonuclease analysis typing system for Clostridium difficile and correlation with other typing systems. J Clin Microbiol **1993**; 31:1870–1875.

142. Olive DM, Bean P. Principles and applications of methods for DNA-based typing of microbial organisms. J Clin Microbiol **1999**; 37:1661–1669.

143. Cartwright CP, Stock F, Beekmann SE, Williams EC, Gill VJ. PCR amplification of rRNA intergenic spacer regions as a method for epidemiologic typing of Clostridium difficile. J Clin Microbiol **1995**; 33:184–187.

144. Brazier JS. The epidemiology and typing of Clostridium difficile. J Antimicrob Chemother **1998**; 41 Suppl C:47–57.

145. Griffiths D, Fawley W, Kachrimanidou M, et al. Multilocus sequence typing of Clostridium difficile. J Clin Microbiol **2010**; 48:770–778.

146. Marsh JW, O'Leary MM, Shutt KA, et al. Multilocus variable-number tandem-repeat analysis for investigation of Clostridium difficile transmission in Hospitals. J Clin Microbiol **2006**; 44:2558–2566.

147. Killgore G, Thompson A, Johnson S, et al. Comparison of seven techniques for typing international epidemic strains of Clostridium difficile: restriction endonuclease analysis, pulsed-field gel electrophoresis, PCR-ribotyping, multilocus sequence typing, multilocus variable-number tandem-repeat analysis, amplified fragment length polymorphism, and surface layer protein A gene sequence typing. J Clin Microbiol **2008**; 46:431–437.

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

149. Jolley KA, Maiden MCJ. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics **2010**; 11:595.

150. Hunter PR, Gaston MA. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J Clin Microbiol **1988**; 26:2465–2466.

151. Tenover FC, Akerlund T, Gerding DN, et al. Comparison of strain typing results for Clostridium difficile isolates from North America. J Clin Microbiol **2011**; 49:1831–1837.

152. Carriço JA, Silva-Costa C, Melo-Cristino J, et al. Illustration of a common framework for relating multiple typing methods by application to macrolide-resistant Streptococcus pyogenes. J Clin Microbiol **2006**; 44:2524–2532.

153. Pinto FR, Melo-Cristino J, Ramirez M. A confidence interval for the wallace coefficient of concordance and its application to microbial typing methods. PLoS ONE **2008**; 3:e3696.

154. Garey KW, Sethi S, Yadav Y, DuPont HL. Meta-analysis to assess risk factors for recurrent Clostridium difficile infection. J Hosp Infect **2008**; 70:298–304.

155. Keighley MR, Burdon DW, Arabi Y, et al. Randomised controlled trial of vancomycin for pseudomembranous colitis and postoperative diarrhoea. BMJ **1978**; 2:1667–1669.

156. Johnson S, Homann SR, Bettin KM, et al. Treatment of asymptomatic Clostridium difficile carriers (fecal excretors) with vancomycin or metronidazole. A randomized, placebo-controlled trial. Ann Intern Med **1992**; 117:297–302.

157. Tannock GW, Munro K, Taylor C, et al. A new macrocyclic antibiotic, fidaxomicin (OPT-80), causes less alteration to the bowel microbiota of Clostridium difficile-infected patients than does vancomycin. Microbiology (Reading, Engl.) **2010**; 156:3354–3359.

158. Nelson RL, Kelsey P, Leeman H, et al. Antibiotic treatment for Clostridium difficile-associated diarrhea in adults. Cochrane Database Syst Rev **2011**; :CD004610.

159. Teasley DG, Gerding DN, Olson MM, et al. Prospective randomised trial of metronidazole versus vancomycin for Clostridium-difficile-associated diarrhoea and colitis. Lancet **1983**; 2:1043–1046.

160. Wenisch C, Parschalk B, Hasenhündl M, Hirschl AM, Graninger W. Comparison of vancomycin, teicoplanin, metronidazole, and fusidic acid for the treatment of Clostridium difficile-associated diarrhea. Clin Infect Dis **1996**; 22:813–818.

161. Zar FA, Bakkanagari SR, Moorthi KMLST, Davis MB. A comparison of vancomycin and metronidazole for the treatment of Clostridium difficile-associated diarrhea, stratified by disease severity. Clin Infect Dis **2007**; 45:302–307.

162. Bolton RP, Culshaw MA. Faecal metronidazole concentrations during oral and intravenous therapy for antibiotic associated colitis due to Clostridium difficile. Gut **1986**; 27:1169–1172.

163. Baines SD, O'Connor R, Freeman J, et al. Emergence of reduced susceptibility to metronidazole in Clostridium difficile. J Antimicrob Chemother **2008**; 62:1046–1052.

164. Peláez T, Alcalá L, Alonso R, et al. In vitro activity of ramoplanin against Clostridium difficile, including strains with reduced susceptibility to vancomycin or with resistance to metronidazole. Antimicrob Agents Chemother **2005**; 49:1157–1159.

165. Peláez T, Alcalá L, Alonso R, Rodríguez-Créixems M, García-Lechuz JM, Bouza E. Reassessment of Clostridium difficile susceptibility to metronidazole and vancomycin. Antimicrob Agents Chemother **2002**; 46:1647–1650.

166. Venugopal AA, Johnson S. Current State of Clostridium difficile Treatment Options. Clin Infect Dis **2012**; 55 Suppl 2:S71–6.

167. Louie TJ, Miller MA, Mullane KM, et al. Fidaxomicin versus vancomycin for Clostridium difficile infection. N Engl J Med **2011**; 364:422–431.

168. Crook DW, Walker AS, Kean Y, et al. Fidaxomicin Versus Vancomycin for Clostridium difficile Infection: Meta-analysis of Pivotal Randomized Controlled Trials. Clin Infect Dis **2012**; 55:S93–S103.

169. Guo B, Harstall C, Louie T, Veldhuyzen van Zanten S, Dieleman LA. Systematic review: faecal transplantation for the treatment of Clostridium difficile-associated disease. Aliment Pharmacol Ther **2012**; 35:865–875.

170. van Nood E, Vrieze A, Nieuwdorp M, et al. Duodenal Infusion of Donor Feces for Recurrent Clostridium difficile. N Engl J Med **2013**; 368:407–415.

171. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA **1977**; 74:5463–5467.

172. Smith LM, Sanders JZ, Kaiser RJ, et al. Fluorescence detection in automated DNA sequence analysis. Nature **1986**; 321:674–679.

173. Metzker ML. Sequencing technologies — the next generation. Nat Rev Genet **2009**; 11:31–46.

174. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science **1995**; 269:496–512.

175. The Sequence Explosion. Nature **2010**; 464:670–671. Available at: http://www.nature.com/news/2010/100331/pdf/464670a.pdf.

176. Nyrén P. The History of Pyrosequencing. Methods Mol Biol **2006**; 373:1–14.

177. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature **2008**; 456:53–59.

178. Loman NJ, Constantinidou C, Chan JZM, et al. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat Rev Microbiol **2012**; 10:599–606.

179. Lunter G, Goodson M. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res **2011**; 21:936–939.

180. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics **2009**; 25:2078–2079.

181. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res **2008**; 18:821–829.

182. Sebaihia M, Wren BW, Mullany P, et al. The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome. Nat Genet **2006**; 38:779–786.

183. Wüst J, Sullivan NM, Hardegger U, Wilkins TD. Investigation of an outbreak of antibiotic-associated colitis by various typing methods. J Clin Microbiol **1982**; 16:1096–1101.

184. Stabler RA, He M, Dawson L, et al. Comparative genome and phenotypic analysis of Clostridium difficile 027 strains provides insight into the evolution of a hypervirulent bacterium. Genome Biol **2009**; 10:R102.

185. He M, Sebaihia M, Lawley TD, et al. Evolutionary dynamics of Clostridium difficile over short and long time scales. Proc Natl Acad Sci USA **2010**; 107:7527–7532.

186. Planche T, Wilcox M. Reference assays for Clostridium difficile infection: one or two gold standards? J Clin Pathol **2011**; 64:1–5.

187. Planche T, Aghaizu A, Holliman R, et al. Diagnosis of Clostridium difficile infection by toxin detection kits: a systematic review. Lancet Infect Dis **2008**; 8:777–784.

188. Grundmann H, Bärwolff S, Tami A, et al. How many infections are caused by patient-to-patient transmission in intensive care units? Crit Care Med **2005**; 33:946–951.

189. Rodriguez-Palacios A, Staempfli HR, Duffield T, Weese JS. Clostridium difficile in retail ground meat, Canada. Emerging Infect Dis **2007**; 13:485–487.

190. Gould LH, Limbago B. Clostridium difficile in food and domestic animals: a new foodborne pathogen? Clin Infect Dis **2010**; 51:577–582.

191. Goldenberg SD, French GL. Diagnostic testing for Clostridium difficile: a comprehensive survey of laboratories in England. J Hosp Infect **2011**; 79:4–7.

192. Eastwood K, Else P, Charlett A, Wilcox M. Comparison of nine commercially available Clostridium difficile toxin detection assays, a real-time PCR assay for C. difficile tcdB, and a glutamate dehydrogenase detection assay to cytotoxin testing and cytotoxigenic culture methods. J Clin Microbiol **2009**; 47:3211–3217.

193. Broukhanski G, Simor A, Pillai DR. Defining criteria to interpret multilocus variable-number tandem repeat analysis to aid Clostridium difficile outbreak investigation. J Med Microbiol **2011**; 60:1095–1100.

194. Tanner HE, Hardy KJ, Hawkey PM. Coexistence of multiple multilocus variable-number tandem-repeat analysis subtypes of Clostridium difficile PCR ribotype 027 strains within fecal specimens. J Clin Microbiol **2010**; 48:985–987.

195. van den Berg RJ, Ameen HA, Furusawa T, Claas EC, van der Vorm ER, Kuijper EJ. Coexistence of multiple PCR-ribotype strains of Clostridium difficile in faecal samples limits epidemiological studies. J Med Microbiol **2005**; 54:173–179.

196. Wroblewski D, Hannett GE, Bopp DJ, et al. Rapid molecular characterization of Clostridium difficile and assessment of populations of C. difficile in stool specimens. J Clin Microbiol **2009**; 47:2142–2148.

197. Eyre DW, Walker AS, Griffiths D, et al. Clostridium difficile Mixed Infection and Reinfection. J Clin Microbiol **2012**; 50:142–144.

198. Eyre DW, Golubchik T, Gordon NC, et al. A pilot study of rapid benchtop sequencing of Staphylococcus aureus and Clostridium difficile for outbreak detection and surveillance. BMJ Open **2012**; 2:e001124.

199. Dingle KE, Griffiths D, Didelot X, et al. Clinical Clostridium difficile: clonality and pathogenicity locus diversity. PLoS ONE **2011**; 6:e19993.

200. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **1997**; 25:3389–3402.

201. Rodrigo AG, Felsenstein J. Coalescent approaches to HIV population genetics. In: Crandall KA, ed. Evolution of HIV. Baltimore, MD: Johns Hopkins University Press, 1999: 233–272.

202. Burnham KP, Anderson DR. Model selection and multimodel inference. 2nd ed. New York, NY: Springer, 2002.

203. Orton RJ, Wright CF, Morelli MJ, et al. Observing micro-evolutionary processes of viral populations at multiple scales. Phil Trans R Soc B **2013**; 368:20120203.

204. Ford CB, Lin PL, Chase MR, et al. Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. Nat Genet **2011**; 43:482–486.

205. Baines SD, O'Connor R, Saxton K, Freeman J, Wilcox MH. Activity of vancomycin against epidemic Clostridium difficile strains in a human gut model. J Antimicrob Chemother **2009**; 63:520–525.

206. Best EL, Freeman J, Wilcox MH. Models for the study of Clostridium difficile infection. Gut Microbes **2012**; 3:145–167.

207. Chilton CH, Freeman J, Crowther GS, Todhunter SL, Nicholson S, Wilcox MH. Co-amoxiclav induces proliferation and cytotoxin production of Clostridium difficile ribotype 027 in a human gut model. J Antimicrob Chemother **2012**; 67:951–954.

208. Baines SD, Crowther GS, Todhunter SL, Freeman J, Wilcox MH. *In vitro* activity of cadazolid (act-179811) against *Clostridium difficile* and in an *in vitro* gut model of *C. difficile* infection [Abstract B-662]. In: Program and abstracts of 52nd Interscience Conference on Antimicrobial Agents and Chemotherapy (San Francisco). 2012.

209. Baldoni D, Gutierrez M, Dingemanse J, Timmer W. Cadazolid, a novel antibiotic with potent activity against *Clostridium difficile*: safety, tolerability, and pharmacokinetics in healthy subjects following single and multiple oral doses [Abstract A-1273]. In: Program and abstracts of 52nd Interscience Conference on Antimicrobial Agents and Chemotherapy (San Francisco). 2012.

210. Bartlett JG. Clostridium difficile infection: historic review. Anaerobe **2009**; 15:227–229.

211. Walker AS, Eyre DW, Wyllie DH, et al. Characterisation of Clostridium difficile hospital ward-based transmission using extensive epidemiological data and molecular typing. PLoS Med **2012**; 9:e1001172:1–12.

212. Rohde H, Qin J, Cui Y, et al. Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4. N Engl J Med **2011**; 365:718–724.

213. Rasko DA, Webster DR, Sahl JW, et al. Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany. N Engl J Med **2011**; 365:709–717.

214. Mellmann A, Harmsen D, Cummings CA, et al. Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. PLoS ONE **2011**; 6:e22751.

215. Chin C-S, Sorenson J, Harris JB, et al. The origin of the Haitian cholera outbreak strain. N Engl J Med **2011**; 364:33–42.

216. Köser CU, Holden MTG, Ellington MJ, et al. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. N Engl J Med **2012**; 366:2267–2275.

217. Young BC, Golubchik T, Batty EM, et al. Evolutionary dynamics of Staphylococcus aureus during progression from carriage to disease. Proc Natl Acad Sci USA **2012**; 109:4550–4555.

218. Royston P, Sauerbrei W. Multivariable Model-building: A Pragmatic Approach to Regression Analysis Based on fractional Polynomials for Modeling Continuous Variables. Chichester, UK: Wiley, 2008.

219. Lambert M-L, Hasker E, Van Deun A, Roberfroid D, Boelaert M, Van der Stuyft P. Recurrence in tuberculosis: relapse or reinfection? Lancet Infect Dis **2003**; 3:282–287.

220. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet **2012**; 13:601–612.

221. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res **2010**; 20:1297–1303.

222. Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinformatics **2011**; 12:119.

223. Prosperi MCF, Salemi M. QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. Bioinformatics **2011**; 28:132–133.

224. Cule ML, Bowden R, Eyre DW, et al. The biology and epidemiology of *Clostridium difficile* in Oxfordshire hospitals 2007-2010. In: Program and abstracts of ID Week (San Diego). 2012.

225. Didelot X, Eyre DW, Cule ML, et al. Microevolutionary analysis of Clostridium difficile genomes to investigate transmission. Genome Biol **2012**; 13:R118.

226. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **2009**; 25:1754–1760.

227. Clostridium difficile MLST Home Page. Available at: http://pubmlst.org/cdifficile/. Accessed 30 December 2011.

228. Maidak BL, Cole JR, Lilburn TG, et al. The RDP-II (Ribosomal Database Project). Nucleic Acids Res **2001**; 29:173–174.

229. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol **2003**; 52:696–704.

230. Muto CA. Asymptomatic Clostridium difficile colonization: is this the tip of another iceberg? Clin Infect Dis **2007**; 45:999–1000.

231. Soberón J, Llorente J. The use of species accumulation functions for the prediction of species richness. Conserv Biol **1993**; 7:480–488.

232. Thompson GG, Withers PC, Pianka ER, Thompson SA. Assessing biodiversity with species accumulation curves; inventories of small reptiles by pit-trapping in Western Australia. Austral Ecology **2003**; 28:361–383.

233. Clench H. How to make regional list of butterflies: Some thoughts. J Lepid Soc **1979**; 33:216–231.

234. Didelot X, Falush D. Inference of bacterial microevolution using multilocus sequence data. Genetics **2007**; 175:1251–1266.

235. Durbin R, Eddy SR, Krogh A, Mitchison G. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge: Cambridge University Press, 1998.

236. Health Protection Agency. Results from the mandatory *Clostridium difficile* reporting scheme. 2012. Available at: http://www.hpa.org.uk/web/HPAweb&HPAwebStandard/HPAweb_C/1195733750761. Accessed 2 November 2012.

237. Bartsch SM, Curry SR, Harrison LH, Lee BY. The potential economic value of screening hospital admissions for Clostridium difficile. Eur J Clin Microbiol Infect Dis **2012**; 31:3163–3171.

238. Department of Health. Updated guidance on the diagnosis and reporting of Clostridium difficile. 2012. Available at:

http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_132927. Accessed 2 November 2012.

239. Wilcox MH, Shetty N, Fawley WN, et al. Changing Epidemiology of Clostridium difficile Infection Following the Introduction of a National Ribotyping-Based Surveillance Scheme in England. Clin Infect Dis **2012**; 55:1056–1063.

240. Wyllie DH, Walker AS, Miller R, et al. Decline of meticillin-resistant Staphylococcus aureus in Oxfordshire hospitals is strain-specific and preceded infection-control intensification. BMJ Open **2011**; 1:e000160.

241. Delmée M, Vandercam B, Avesani V, Michaux JL. Epidemiology and prevention of Clostridium difficile infections in a leukemia unit. Eur J Clin Microbiol **1987**; 6:623–627.

242. Clabots CR, Gerding SJ, Olson MM, Peterson LR, Gerding DN. Detection of asymptomatic Clostridium difficile carriage by an alcohol shock procedure. J Clin Microbiol **1989**; 27:2386–2387.

243. Foglia G, Shah S, Luxemburger C, Pietrobon PJF. Clostridium difficile: development of a novel candidate vaccine. Vaccine **2012**; 30:4307–4309.

244. Steele J, Mukherjee J, Parry N, Tzipori S. Antibody against TcdB, but not TcdA, prevents development of gastrointestinal and systemic Clostridium difficile disease. J Infect Dis **2013**; 207:323–330.

245. Louie TJ, Cannon K, Byrne B, et al. Fidaxomicin Preserves the Intestinal Microbiome During and After Treatment of Clostridium difficile Infection (CDI)

and Reduces Both Toxin Reexpression and Recurrence of CDI. Clin Infect Dis **2012**; 55 Suppl 2:S132–42.

246. Mutters NT, Brooke RJ, Frank U, Heeg K. Low risk of apparent transmission of vancomycin-resistant Enterococci from bacteraemic patients to hospitalized contacts. Am J Infect Control **2013**;

247. Harris SR, Cartwright EJ, Török ME, et al. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. Lancet Infect Dis **2013**; 13:130–136.

248. Optimer Pharmaceuticals. Anti-infective drugs advisory committee meeting briefing document. 2011. Available at: http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/Anti-InfectiveDrugsAdvisoryCommittee/UCM249354.pdf. Accessed 10 March 2013.