Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*

Miao He[1], **Fabio Miyajima**[2,3], **Paul Roberts**[2,3], **Louise Ellison**[1], **Derek J. Pickard**[1], **Melissa J. Martin**[4], **Thomas R. Connor**[1], **Simon R. Harris**[1], **Derek Fairley**[5], **Kathleen B. Bamford**[6,7], **Stephanie D'Arc**[6,7], **Jon Brazier**[8], **Derek Brown**[9], **John E. Coia**[9], **Gill Douce**[9], **Dale Gerding**[10], **Hee Jung Kim**[11], **Tse Hsien Koh**[12], **Haru Kato**[13], **Mitsutoshi Senoh**[13], **Tom Louie**[14], **Stephen Michell**[15], **Emma Butt**[15], **Sharon J. Peacock**[1,16,17,18], **Nick M. Brown**[17,18], **Tom Riley**[19], **Glen Songer**[20], **Mark Wilcox**[21], **Munir Pirmohamed**[2,3], **Ed Kuijper**[22], **Peter Hawkey**[23], **Brendan W. Wren**[4], **Gordon Dougan**[1], **Julian Parkhill**[1,*], and **Trevor D. Lawley**[1,*]

[1]Wellcome Trust Sanger Institute, Hinxton, UK [2]Department of Molecular and Clinical Pharmacology, University of Liverpool, Liverpool, UK [3]University of Liverpool and Royal Liverpool and Broadgreen University Hospital NHS Trust, Liverpool, UK [4]Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, London, UK [5]Belfast Health and Social Trust, Belfast, Northern Ireland [6]Department of Infectious Diseases and Immunity, Imperial College London, London, UK [7]Department of Bacteriology, Imperial College Healthcare NHS Trust, London, UK [8]Anaerobic Reference Laboratory, Cardiff, UK [9]Scottish Salmonella, Shigella and Clostridium difficile Reference Laboratory, Glasgow, UK [10]Hines VA Hospital, Hines, IL USA [11]Yonsei University, Seoul, South Korea [12]Singapore General Hospital, Singapore [13]Department of Bacteriology, National Institute of Infectious Diseases, Tokyo, Japan [14]Department of Microbiology, Immunology & Infectious Diseases, University of Calgary, Calgary, Canada [15]College of Life and Environmental Sciences, University of Exeter, Exeter, UK [16]Department of Medicine, University of Cambridge, Cambridge, UK [17]Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK [18]Health Protection Agency, London, UK [19]School of Pathology and Laboratory Medicine, The University of Western Australia, Crawley, Australia [20]Department of Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, USA [21]Healthcare Associate Infection Research Group, University of Leeds, Leeds, UK [22]Department of Experimental Microbiology, Leiden University Medical Centre, Leiden, The Netherlands [23]School of Immunity and Infection, University of Birmingham, Birmingham, UK

Abstract

Epidemic *Clostridium difficile* (027/BI/NAP1) rapidly emerged in the past decade as the leading cause of antibiotic-associated diarrhea worldwide. However, the key moments in the evolutionary history leading to its emergence and subsequent patterns of global spread remain unknown. Here

we define the global population structure of *C. difficile* 027/BI/NAP1 based on whole-genome sequencing and phylogenetic analysis. We demonstrate that two distinct epidemic lineages, FQR1 and FQR2, not one as previously thought, emerged in North America within a relatively short period after acquiring the same fluoroquinolone resistance mutation and a highly-related conjugative transposon. The two epidemic lineages displayed distinct patterns of global spread, and the FQR2 lineage spread more widely leading to healthcare outbreaks in the UK, continental Europe and Australia. Our analysis identifies key genetic changes linked to the rapid trans-continental dissemination of epidemic *C. difficile* 027/BI/NAP1 and highlights the routes by which it spreads through the global healthcare system.

Clostridium difficile is the most common infectious cause of antibiotic-associated diarrhea and healthcare infection in the developed world[1,2]. Antibiotic treatment and hospitalization are major risk factors for *C. difficile* colonization leading to asymptomatic carriage, diarrhea, pseudomembranous colitis or death[3]. *C. difficile* pathogenesis is associated with the production of the enterotoxins TcdA and TcdB that are encoded within a pathogenicity locus (PaLoc)[4]. *C. difficile* produces highly resistant and infectious spores, which promote environmental transmission within the healthcare setting[5], and potentially facilitate spread over greater distances by those in the carrier state[6].

The global emergence of *C. difficile* infection (CDI) in the past decade followed highly-publicized *C. difficile* outbreaks in the USA[7] and Canada[8] that were associated with increased rates of disease recurrence and mortality[7-10]. The outbreaks were caused by a previously uncommon, fluoroquinolone resistant variant of *C. difficile* genotyped as 027/BI/NAP1[7,8]. Fluoroquinolone resistant (FQR) *C. difficile* 027/BI/NAP1 is still the most common variant causing CDI throughout North America[11]. During 2004-2006 there were severe *C. difficile* 027/BI/NAP1 outbreaks in the UK and in subsequent years FQR *C. difficile* 027/BI/NAP1 accounted for >40% of cases UK-wide[12], and is now commonly found in continental Europe[13,14] and more recently in Australia[15]. Although FQR *C. difficile* 027/BI/NAP1 is widespread in healthcare facilities worldwide the underlying reasons for its rapid emergence and the subsequent patterns of global spread remain unknown.

To address these questions, we sequenced the genomes of a global collection of *C. difficile* 027/BI/NAP1 (n=151) isolated primarily from hospital patients between 1985 and 2010 (Supplementary Table 1). Illumina reads were aligned to the genome of the *C. difficile* 027/BI/NAP1 strain R20291[16]. We identified a total of 3,686 single nucleotide polymorphisms (SNPs) within the 3.8 Mb non-repetitive, core genome (representing 95% of genome). Of these, 3,150 (85%) of the SNPs were clustered and private to 8 individual isolates, suggesting that these genomic regions were imported from outside the 027/BI/NAP1 lineage (Supplementary Fig. 1; Supplementary Note). These SNPs were removed as they could mask the true phylogenetic signal, leaving 536 SNPs for downstream phylogenetic analysis. Other than these events, homologous recombination has not played a major role in shaping the phylogeny of the global *C. difficile* 027/BI/NAP1 collection (Supplementary Fig. 2; Supplementary Note).

Figure 1a shows a maximum-likelihood phylogeny representing the *C. difficile* 027/BI/NAP1 global population structure. The phylogeny discriminates between >100 distinct genotypes within the global collection and shows limited geographical clustering, implying frequent, long-range transmission between humans, and in a limited number of cases two-way transmission between human and animal or food sources (Supplementary Note). The core genome of *C. difficile* 027/BI/NAP1 exhibits a relatively low-level of genetic diversity, consistent with its recent emergence, with very few SNPs defining the major branches in the phylogeny (Supplementary Fig. 3). Using three methods, we estimated the mutation rate of

C. difficile 027/BI/NAP1 to be 1.47×10^{-7}-5.33×10^{-7} (95% confidence intervals) substitutions per site per year, equivalent to 1-2 mutations per genome per year (Supplementary Fig. 4; Supplementary Table 2; Supplementary Note). Interestingly, this rate is ~10 times slower than *Streptococcus pneumoniae*[17] and *Staphylococcus aureus*[18] over similar timescales and may be due to the fact that metabolically dormant spores do not accumulate mutations[19].

Importantly, the global phylogeny demonstrates the presence of two genetically distinct lineages which have independently acquired an identical mutation (Thr82Ile) in DNA gyrase subunit A (*gyrA*), by either mutation or recombination, leading to high-level FQR[20] (Fig. 1). Although both lineages share this mutation, which is the most frequent in FQR isolates[20], both lineages are highly supported by maximum-likelihood and Bayesian methods (Supplementary Figs. 3, 5), unequivocally demonstrating that FQR has arisen in two epidemic *C. difficile* 027/BI/NAP1 lineages in two separate events. We named these lineages FQR1 and FQR2. Bayesian analysis estimates that FQR1 and FQR2 have emerged recently, with most recent common ancestors in ~1993 and ~1994, respectively (median estimates; 95% highest posterior density intervals are 1984-1999 for FQR1 and 1986-1999 for FQR2). Near the base of the *C. difficile* 027/BI/NAP1 phylogeny and outside of both FQR lineages are isolates from various global locations (Singapore, Japan, Korea, Canada, the USA, the UK, Germany and France) and sampling times (1985-2009). To our knowledge, none of the isolates in this part of the phylogeny are associated with major hospital outbreaks, suggesting that these isolates represent the pre-epidemic *C. difficile* 027/ BI/NAP1 genetic background from which FQR epidemic lineages emerged.

Lineage FQR1 contains epidemic isolates associated with healthcare outbreaks in the USA (Pennsylvania (2001), Oregon (2003), New Jersey (2004), Arizona (2006 and 2007) and Maryland (2007))[7] and isolates associated with sporadic infections in South Korea[21] and Switzerland between 2007 and 2010. Bayesian phylogeographic analysis[22] indicates that the FQR1 lineage originated from the USA (99% probability). The earliest isolate in the FQR1 lineage is from Pittsburgh, Pennsylvania in 2001, representing one of the earliest reports of an increase in CDI caused by FQR *C. difficile* 027/BI/NAP1[9,23]. Thus, FQR1 appears to represent an epidemic lineage that originally emerged in the North East of the USA and subsequently transmitted to South Korea and Switzerland.

The FQR2 lineage contains the majority of epidemic isolates that are widely spread geographically (Figure 1a). The most striking feature in FQR2 is a star-like topology in the early part of the lineage generally consistent with rapid population expansion from a common progenitor (Supplementary Fig. 5). Two isolates in our collection were found on the node at the base of the star-like topology, one associated with an outbreak in Montreal in 2003[24] and the other from the Netherlands in 2006. Bayesian analysis suggests the FQR2 lineage also originated in North America (with 59% probability from the USA and 33% probability from Canada). Interestingly, all but one Canadian isolates found in this lineage were from Montreal where FQR *C. difficile* 027/BI/NAP1 outbreaks were initially reported in Canada[8,10].

The Bayesian phylogeny[25] contains multiple highly-supported sub-lineages of FQR2 associated with distinct geographic locations suggesting rapid trans-continental dissemination from North America to continental Europe, the UK and Australia. Our analysis demonstrates a single introductory event of FQR2 into Australia and at least four separate introductions into continental Europe, including two trans-Atlantic transmission events (one of which reached the Netherlands) and two from the UK, giving rise to present-day isolates in Austria and Poland[26] (Fig. 2a; Supplementary Fig. 5). Similarly, the descendants of FQR2 were introduced into the UK on at least four occasions, including three

trans-Atlantic transmission events from North America and one from continental Europe, leading to a series of highly-publicized outbreaks in UK hospitals during 2004-2006[27].

To investigate the introduction and subsequent spread of *C. difficile* 027/BI/NAP1 within the UK we analyzed the genomes of a collection of a further 145 UK isolates from healthcare patients (Supplementary Table 3) (in addition to the 43 UK isolates in our global collection). A maximum-likelihood phylogeny (Fig. 1b) demonstrates that long-range geographical transmission was frequent and extensive within the UK (Fig. 2b). This collection confirms that FQR2 *C. difficile* 027/BI/NAP1 likely reached the UK on at least four separate occasions, arriving independently in Exeter, Ayrshire, and Birmingham from North America and arriving in Maidstone from continental Europe (Fig. 1b and Fig. 2b). The introduction into the Maidstone area likely triggered a large-scale hospital outbreak that affected London and Cambridge[27]. Despite multiple introductions, the majority of present-day UK isolates (including those from Liverpool, Belfast, Birmingham, London, Cambridge, Exeter and multiple sites in Scotland) are descendants of one sub-lineage whose earliest representative in our collection is a 2002 Birmingham isolate. This dominant sub-lineage also underlies the Stoke Mandeville outbreak[27] (Fig. 1b and Fig. 2b).

The emergence of epidemic *C. difficile* 027/BI/NAP1 has been proposed to be due to genetic changes in the *tcdC* gene of the PaLoc[4,7,8,28]. However, we discovered only 2 SNPs in the entire 19.6 kb PaLoc region in our 027/BI/NAP1 collection, both private to a single isolate. Consistent with a previous proposal based on two *C. difficile* 027/BI/NAP1 genomes[16], there is no change within the entire PaLoc region between pre- and post-epidemic isolates of *C. difficile* 027/BI/NAP1 that could explain the emergence (Supplementary Note). As described above, we provide strong evidence that the acquisition of FQR in two distinct lineages is associated with the emergence of *C. difficile* 027/BI/NAP1. We next determined if there are other genetic changes linked to the acquisition of FQR that could underlie a presumed increase of fitness contributing to the emergence of FQR1 and FQR2. Only 2 and 7 SNPs define the branches leading to lineages FQR1 and FQR2, respectively. However, besides the *gyrA* mutation (Figure 1) there are no SNPs shared by both lineages (Supplementary Table 4). Further, there is little evidence that a significant change in phenotype could result from any of the SNPs that define the FQR1 and FQR2 lineages (Supplementary Table 4; Supplementary Note), except perhaps the *gyrA* mutation itself. Besides these SNPs and the *gyrA* mutation we also identified 10 non-synonymous SNPs conflicting with the phylogeny (homoplasic, 1.9% of the total number) that are present mainly within FQR1 and FQR2. These 10 SNPs affect 8 genes associated with antimicrobial resistance or cell surface modification (Table 1; Supplementary Fig. 3; Supplementary Note). However, these SNPs are only present within a small fraction of the FQR1 and FQR2 isolates so their presence cannot explain the emergence of *C. difficile* 027/BI/NAP1.

Horizontal gene transfer is a key mechanism driving *C. difficile* evolution[29,30]. We therefore assembled the *C. difficile* 027/BI/NAP1 genomes and identified widespread acquisition of a range of mobile elements associated with erythromycin, chloramphenicol, tetracycline and aminoglycoside resistance throughout the phylogeny (Supplementary Fig. 3; Supplementary Note). Importantly, we found a class of CTn*5*-like elements[29] in almost all isolates in FQR1 and FQR2 lineages, where they are inserted in the same genome location in the majority of isolates (Supplementary Fig. 6; Supplementary Note). The FQR1 and FQR2 lineages both share a common CTn*5*-like element, which we named Tn*6192*, that contains genes encoding an ABC transport system, a two-component transcriptional regulator and a predicted DNA-binding protein (Supplementary Table 5). The version found in the FQR2 lineage differs from that in the FQR1 lineage by harboring a contiguous insertion of 15.7 kb, Tn*6105*[31], which contains 14 genes, four of which are predicted to be DNA-binding proteins or

transcriptional regulators (Supplementary Table 5). This element has also acted as a hot-spot for further integrations throughout the FQR2 lineage (Supplementary Fig. 6). The presence of Tn*6192* is the only other significant genetic trait, aside from the fluoroquinolone resistance SNP, that differentiates FQR1 and FQR2 from the pre-epidemic isolates (Supplementary Table 6). It is unknown whether the genes carried by this element have any phenotypic effect on the core genome. However, Tn*6192* may have played a role in the success of these lineages.

We demonstrate that the separate acquisitions of FQR and a conjugative transposon in two distinct lineages of *C. difficile* 027/BI/NAP1 are the key genetic changes linked to its rapid emergence during the early 2000s[7,8]. Further, our data suggest that the acquisition of resistance to commonly used antibiotics is a major feature of the continued evolution and persistence of *C. difficile* 027/BI/NAP1 in the healthcare settings. It is notable that fluoroquinolone antibiotics were one of the most commonly prescribed antibiotic classes in North America during the late 1990s and early 2000s[32], so it is during this time that the selective pressure for the acquisition and maintenance of FQR within healthcare settings would have been at its highest, explaining the near simultaneous emergence of more than one clone of FQR *C. difficile* 027/BI/NAP1. We also demonstrate the ease and rapidity with which FQR *C. difficile* 027/BI/NAP1 has transmitted internationally highlighting the interconnectedness of the global healthcare system facilitated by rapid human travel. Whole-genome sequencing and phylogenetic analysis has been used successfully for clinical transmission studies of methicillin resistant *S. aureus* within a hospital[33] and although the low mutation rate may pose some limitations we believe this approach will have great potential for *C. difficile*. Our analysis provides a genomic framework for understanding the population structure, geographical source, epidemiology and evolution of this highly transmissible healthcare pathogen.

Online Methods

C. difficile strain collection

The *C. difficile* isolates chosen for this study were characterized as PCR-ribotype 027 or 176, REA type BI or PFGE type NAP1 and include 151 isolates from a global collection and 188 isolates from the UK, with an overlap of 43 isolates between the collections. The global collection includes 25 isolates from two previous studies[16,30]. New genome sequencing data was generated for 6 of these isolates. Supplementary Table 1 and Supplementary Table 3 summarize the isolate details. Genomic DNA was extracted as previously described[30].

Sequencing, mapping, and SNP detection

Paired-end multiplex libraries were created as described[18], followed by sequencing on Illumina GAIIx and HiSeq2000 platforms. The read length was 54 bp for samples Liv1-Liv21, 108 bp for samples Gla001-Gla022, and 76 bp for the rest. All isolates were sequenced to an average coverage of 110-fold across the isolates. Sequencing reads were aligned with BWA[34] against the genome sequence of the ribotype 027 reference strain R20291[16]. SNPs were identified with SAMtools[35]. A coverage cut-off of >5-fold and <three times the average coverage was set for each individual isolate during SNP detection. Repetitive regions in the reference genome sequence were characterized using REPuter[36] and the repeat finder functions in the MUMmer package[37]. The boundaries of repetitive regions were extended to include the mobile elements in R20291. SNPs falling within these repetitive regions were excluded. To confirm the alleles for each variant position, SNPs were checked at each position in all sequencing reads in all isolates. An allele was only considered to be valid if supported by all reads (if 5<coverage<=40) or >92.5% of the total reads (if coverage >40) covering the position; otherwise it was treated as missing data. Our

method allows no false positive SNPs and a false negative rate of ~8% (Supplementary Fig. 7; Supplementary Note).

Phylogenetic analysis

An appropriate evolutionary model (simple GTR) was determined using jModelTest 0.1.1[38]. Phylogenetic relationships were inferred using three methods: 1), split-decomposition and neighbor-net methods in SplitsTree4[39]; 2), the program PHYML[40], and 3), the program BEAST[25]. In the first two cases, a simple GTR model was used. Neighbor-joining tree (Supplementary Fig. 2b) were also constructed with PHYML, and the results were compared. Two non-027/BI/NAP1 *C. difficile* isolates 630[29] and CF5[30] were used to root the global phylogeny.

In the BEAST analysis, three clock models (strict, relaxed lognormal and relaxed exponential) and two population models (constant and skyline) were tested initially. The relaxed exponential clock model in combination with skyline population model was determined as more suitable based on Bayes Factor calculations[41] and was used for later BEAST runs. The program was specified to estimate tMRCA (time to the most recent common ancestor) of taxon groupings. All other parameters were set to default. These analyses were carried out with a chain length of 400,000,000 states, and re-sampling every 20,000 states. The phylogeographic history was also inferred with BEAST using a Bayesian method as described[22]. The ESS values were >200 for all parameters.

Accessory genome

For each genome, the unaligned sequencing reads were assembled using Velvet[42]. To assess whether the resulting contigs were unique, each contig with a length >1kb was searched using BLASTN against the current pan-genome, which was made by concatenating the draft genome sequence of M7404 and already determined unique contigs. Any unique contigs were added to the pan-genome. If the match was of >80% identity and covered >40% of the contig length, this contig was not considered to be unique, and was not added to the current pan-genome. The resulting unique contigs were individually searched against the NCBI bacterial genome database to check for contamination. The filtered unique contigs were added to the genome sequence of M7404 to create a pan-genome. Finally, the sequencing reads from each strain were aligned against the constructed pan-genome to assess for the presence and absence of genomic regions in each isolate. The raw Illumina data were also assembled *de novo*. Draft genome sequences were produced for isolates by ordering contigs against a suitable reference sequence. Pairwise genome comparisons were made in Artemis Comparison Tool[43]. Key mobile elements and their nomenclature are described in Supplementary Note.

Identification of homoplasic characters and homologous recombination

Homoplasic SNPs were identified by examining the SNP allele pattern across all isolates in relation to the phylogenetic tree. A SNP was considered homoplasic if the allele pattern did not agree with the tree topology. Genomic regions affected by homologous recombination were identified by a) clusters of SNPs within 2,000 bp windows and b) an iterative method to eliminate recombination sites as described[17] The identified homologous recombination blocks were excluded from phylogenetic and population genetic analysis.

Mutation rate

The apparent mutation rate was estimated using three methods: 1), A full maximum-likelihood model assuming a rapid expansion which results in perfect star genealogies, implemented in an R script[44] (Supplementary Note; Supplementary Fig. 8); 2), the program

BEAST[25]; and 3), the program Path-O-Gen v1.3 (See URL). A maximum-likelihood tree was used in this analysis. BEAST analyses were carried out as stated above. The dataset used for 2) and 3) is the final SNP alignment of the global collection. A final mutation rate was determined by combining the single estimate from method 3) and 95% confidence intervals from methods 1) and 2)[45].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

REFERENCES

1. Bartlett JG. Clostridium difficile: progress and challenges. Ann N Y Acad Sci. 2010; 1213:62–69. [PubMed: 21175676]

2. Kelly CP, LaMont JT. Clostridium difficile--more difficult than ever. N Engl J Med. 2008; 359:1932–1940. [PubMed: 18971494]

3. Kuijper EJ, Coignard B, Tull P. Emergence of Clostridium difficile-associated disease in North America and Europe. Clin Microbiol Infect. 2006; 12(Suppl 6):2–18. [PubMed: 16965399]

4. Warny M, et al. Toxin production by an emerging strain of Clostridium difficile associated with outbreaks of severe disease in North America and Europe. Lancet. 2005; 366:1079–1084. [PubMed: 16182895]

5. Riggs MM, et al. 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. [PubMed: 17879913]

6. Clements AC, Magalhaes RJ, Tatem AJ, Paterson DL, Riley TV. Clostridium difficile PCR ribotype 027: assessing the risks of further worldwide spread. Lancet Infect Dis. 2010; 10:395–404. [PubMed: 20510280]

7. McDonald LC, et al. An epidemic, toxin gene-variant strain of Clostridium difficile. N Engl J Med. 2005; 353:2433–2441. [PubMed: 16322603]

8. Loo VG, 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. [PubMed: 16322602]

9. Muto CA, et al. A large outbreak of Clostridium difficile-associated disease with an unexpected proportion of deaths and colectomies at a teaching hospital following increased fluoroquinolone use. Infect Control Hosp Epidemiol. 2005; 26:273–280. [PubMed: 15796280]

10. Pepin J, Valiquette L, Cossette B. Mortality attributable to nosocomial Clostridium difficile-associated disease during an epidemic caused by a hypervirulent strain in Quebec. CMAJ. 2005; 173:1037–1042. [PubMed: 16179431]

11. O'Connor JR, Johnson S, Gerding DN. Clostridium difficile infection caused by the epidemic BI/NAP1/027 strain. Gastroenterology. 2009; 136:1913–1924. [PubMed: 19457419]

12. Brazier JS, et al. Distribution and antimicrobial susceptibility patterns of Clostridium difficile PCR ribotypes in English hospitals, 2007-08. Euro Surveill. 2008:13.

13. Bauer MP, et al. Clostridium difficile infection in Europe: a hospital-based survey. Lancet. 2011; 377:63–73. [PubMed: 21084111]

14. Kuijper EJ, et al. Update of Clostridium difficile infection due to PCR ribotype 027 in Europe, 2008. Euro Surveill. 2008:13.

15. Richards M, et al. Severe infection with Clostridium difficile PCR ribotype 027 acquired in Melbourne, Australia. Med J Aust. 2011; 194:369–371. [PubMed: 21470090]

16. Stabler RA, 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. [PubMed: 19781061]

17. Croucher NJ, et al. Rapid pneumococcal evolution in response to clinical interventions. Science. 2011; 331:430–434. [PubMed: 21273480]

18. Harris SR, et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science. 2010; 327:469–474. [PubMed: 20093474]

19. Pearson T, et al. Phylogenetic discovery bias in Bacillus anthracis using single-nucleotide polymorphisms from whole-genome sequencing. Proc Natl Acad Sci U S A. 2004; 101:13536–13541. [PubMed: 15347815]

20. Spigaglia P, Carattoli A, Barbanti F, Mastrantonio P. Detection of gyrA and gyrB mutations in Clostridium difficile isolates by real-time PCR. Mol Cell Probes. 2010; 24:61–67. [PubMed: 19837154]

21. Kim H, et al. Emergence of Clostridium difficile Ribotype 027 in Korea. Korean J Lab Med. 2011; 31:191–196. [PubMed: 21779194]

22. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. PLoS Comput Biol. 2009; 5:e1000520. [PubMed: 19779555]

23. Dallal RM, et al. Fulminant Clostridium difficile: an underappreciated and increasing cause of death and complications. Ann Surg. 2002; 235:363–372. [PubMed: 11882758]

24. MacCannell DR, et al. Molecular analysis of Clostridium difficile PCR ribotype 027 isolates from Eastern and Western Canada. J Clin Microbiol. 2006; 44:2147–2152. [PubMed: 16757612]

25. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 2007; 7:214. [PubMed: 17996036]

26. Nyc O, Pituch H, Matejkova J, Obuch-Woszczatynski P, Kuijper EJ. Clostridium difficile PCR ribotype 176 in the Czech Republic and Poland. Lancet. 2011; 377:1407. [PubMed: 21515161]

27. Brazier JS. Clostridium difficile: The anaerobe that made the grade. Anaerobe. 2012

28. Carter GP, Awad MM, Kelly ML, Rood JI, Lyras D. TcdB or not TcdB: a tale of two Clostridium difficile toxins. Future Microbiol. 2011; 6:121–123. [PubMed: 21366410]

29. Sebaihia M, et al. The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome. Nat Genet. 2006; 38:779–786. [PubMed: 16804543]

30. He M, et al. Evolutionary dynamics of Clostridium difficile over short and long time scales. Proc Natl Acad Sci U S A. 2010; 107:7527–7532. [PubMed: 20368420]

31. Brouwer MSM, Warburton PJ, Roberts AP, Mullany P, Allan E. Genetic Organisation, Mobility and Predicted Functions of Genes on Integrated, Mobile Genetic Elements in Sequenced Strains of Clostridium difficile. PLoS ONE. 2011; 6:e23014. [PubMed: 21876735]

32. Linder JA, Huang ES, Steinman MA, Gonzales R, Stafford RS. Fluoroquinolone prescribing in the United States: 1995 to 2002. Am J Med. 2005; 118:259–268. [PubMed: 15745724]

33. Koser CU, et al. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. N Engl J Med. 2012; 366:2267–2275. [PubMed: 22693998]

34. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]

35. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–2079. [PubMed: 19505943]

36. Kurtz S, et al. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 2001; 29:4633–4642. [PubMed: 11713313]

37. Kurtz S, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004; 5:R12. [PubMed: 14759262]

38. Posada D. jModelTest: phylogenetic model averaging. Molecular Biology and Evolution. 2008; 25:1253. [PubMed: 18397919]

39. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 2006; 23:254–267. [PubMed: 16221896]

40. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 2003; 52:696–704. [PubMed: 14530136]

41. Suchard MA, Weiss RE, Sinsheimer JS. Bayesian selection of continuous-time Markov chain evolutionary models. Mol Biol Evol. 2001; 18:1001–1013. [PubMed: 11371589]

42. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18:821–829. [PubMed: 18349386]

43. Carver TJ, et al. ACT: the Artemis Comparison Tool. Bioinformatics. 2005; 21:3422–3423. [PubMed: 15976072]

44. Morelli G, et al. Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity. Nat Genet. 2010

45. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol. 2005; 22:1185–1192. [PubMed: 15703244]
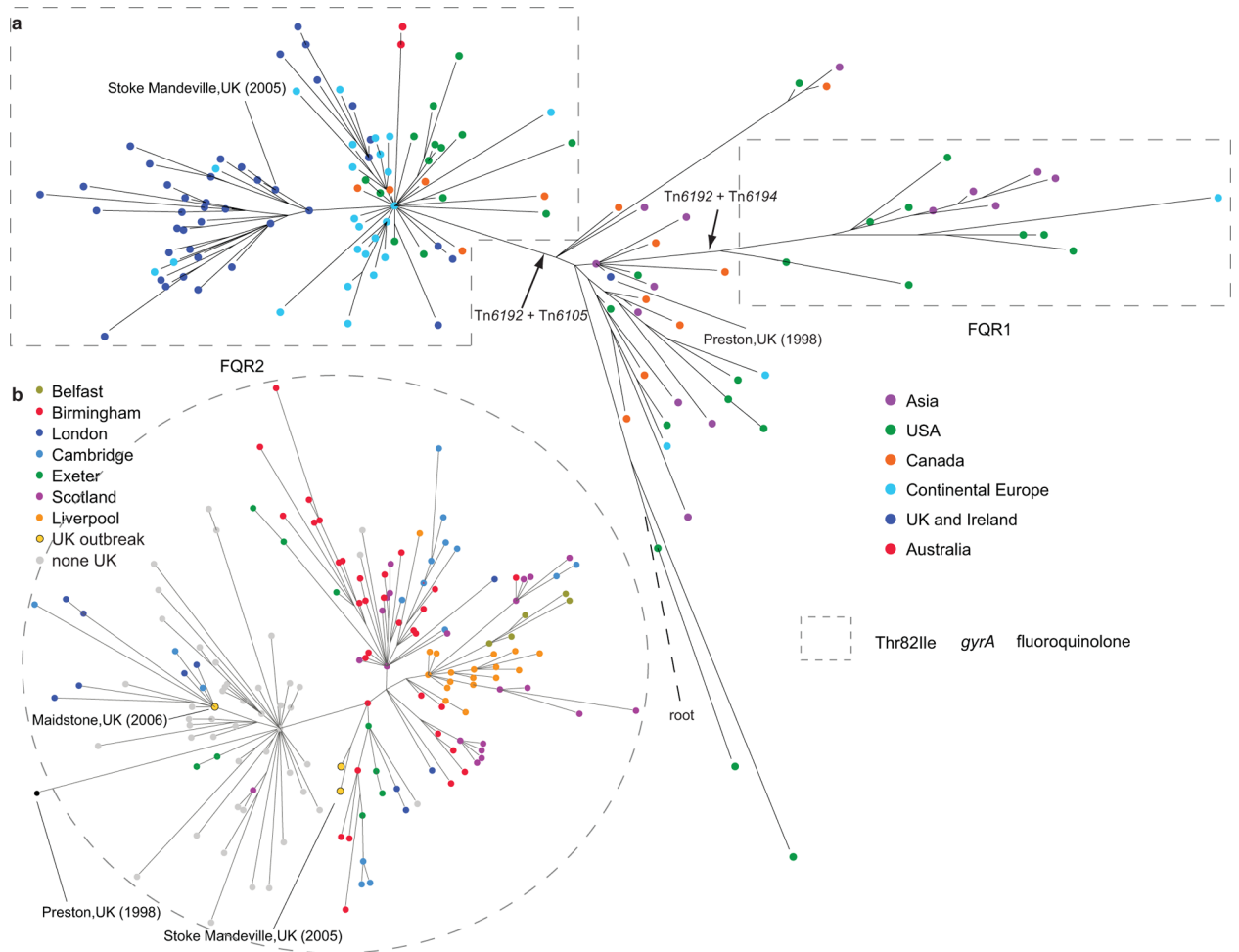
Figure 1.
Phylogeny of *C. difficile* 027/BI/NAP1 based on core genome SNPs. a, Global phylogeny
(n=151 isolates). Colored nodes indicate the geographical source of the isolates. The
position of the inferred root is indicated by a dashed line. b, Phylogeny of UK isolates
(n=188) colored according to geographical source of isolates (gray - non-UK isolates).
Dashed-line circled areas indicate the isolates with Thr82Ile substitution in *gyrA* associated
with fluoroquinolone resistance. Black arrows show insertion of selected mobile elements.
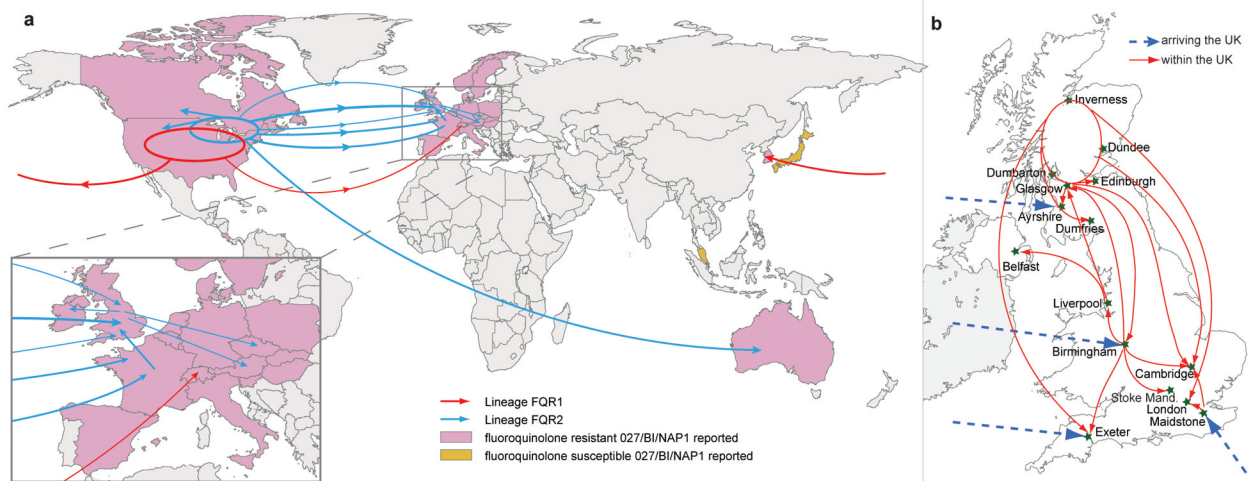
Figure 2.
Transmission events inferred for epidemic *C. difficile* 027/BI/NAP1. a, Global spread of lineages FQR1 (red arrows) and FQR2 (blue arrows) inferred from phylogeographic analysis (Supplementary Fig. 5). Global map demonstrating the countries where fluoroquinolone-resistant (pink) and fluoroquinolone-sensitive (yellow) *C. difficile* 027/BI/NAP1 has been reported[6]. Width of the arrow is approximately proportional to the number of descendents from each sub-lineage. b, Inferred arrivals (blue dashed arrows) and transmissions (red arrows) of FQR2 into and within the UK based on phylogeographic analysis (Supplementary Fig. 5) and maximum likelihood phylogeny (Fig. 1b).

Table 1

Non-synonymous homoplasic SNPs identified in the core genomes of *C. difficile* 027/BI/NAP1 isolates. The positions refer to those in the R20291 genome. The alleles listed relate to the forward strand.

Position	Region	SNP	Substitution	Antibiotic
5420	DNA gyrase subunit B *(gyrB)*	1276G>A	Asp426Asn	Fluoroquinolone
6310	DNA gyrase subunit A *(gyrA)*	245C>T	Thr82Ile	Fluoroquinolone
95412	DNA-directed RNA polymerase beta chain *(rpoB)*	1504C>A	His502Asn	Rifampicin
95422	*rpoB*	1514G>A	Arg505Lys	Rifampicin
103867	translation elongation factor G *(fusA)*	1363C>A/1363C>T	His455Asn/His455Tyr	Fusidic acid
104117	*fusA*	1613C>T	Pro538Leu	Fusidic acid
1800920	two-component response regulator	31G>A	Glu11Lys	
1802086	two-component sensor histidine kinase	446C>T	Thr149Ile	
3170481	S-layer precursor protein *(slpA)*	467G>A/467G>T	Pro156Leu/Pro156Gln	
3170482	*slpA*	466G>A/466G>T	Pro156Ser/Pro156Thr	
3938789	putative membrane protein	996A>C	Tyr332*	