Emergence and global spread of epidemic healthcare-associated Clostridium difficile

Miao He1, Fabio Miyajima2,3, Paul Roberts2,3, Louise Ellison1, Derek J Pickard1, Melissa J Martin4, Thomas R Connor1, Simon R Harris1, Derek Fairley5, Kathleen B Bamford6,7, Stephanie D’Arc6,7, Jon Brazier8, Derek Brown9, John E Coia9, Gill Douce9, Dale Gerding10, Hee Jung Kim11, Tse Hsien Koh12, Haru Kato13, Mitsutoshi Senoh13, Tom Louie14, Stephen Michell15, Emma Butt15, Sharon J Peacock16–18, Nick M Brown17,18, Tom Riley19, Glen Songer20, Mark Wilcox21, Munir Pirmohamed2,3, Ed Kuijper22, Peter Hawkey23, Brendan W Wren4, Gordon Dougan1, Julian Parkhill1 & Trevor D Lawley1

Epidemic C. difficile (027/BI/NAP1) has rapidly emerged in the past decade as the leading cause of antibiotic-associated diarrhea worldwide. However, the key events in evolutionary history leading to its emergence and the subsequent patterns of global spread remain unknown. Here, we define the global population structure of C. difficile 027/BI/NAP1 using whole-genome sequencing and phylogenetic analysis. We show 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–conferring mutation and a highly related conjugative transposon. The two epidemic lineages showed distinct patterns of global spread, and the FQR2 lineage spread more widely, leading to healthcare-associated outbreaks in the UK, continental Europe and Australia. Our analysis identifies key genetic changes linked to the rapid transcontinental dissemination of epidemic C. difficile 027/BI/NAP1 and highlights the routes by which it spreads through the global healthcare system.

C. difficile is the most common infectious cause of antibiotic-associated diarrhea and healthcare-related infection in the developed world1–2. Antibiotic treatment and hospitalization are major risk factors for C. difficile colonization, which leads to asymptomatic carriage, diarrhea, pseudomembranous colitis or death3. 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 setting5 and potentially facilitate spread over greater distances by those in the carrier state6.

The global emergence of C. difficile infection (CDI) in the past decade followed highly publicized C. difficile outbreaks in the United States7 and Canada8 that were associated with increased rates of disease recurrence and mortality7–10. The outbreaks were caused by a previously uncommon fluoroquinolone-resistant variant of C. difficile genotyped as 027/BI/NAP1 (refs. 7,8). Fluoroquinolone-resistant (FQR) C. difficile 027/BI/NAP1 is still the most common variant causing CDI throughout North America11. From 2004 to 2006, there were severe C. difficile 027/BI/NAP1 outbreaks in the UK. In subsequent years, FQR C. difficile 027/BI/NAP1 accounted for >40% of cases across the UK12 and is now commonly found in continental Europe13,14 and, more recently, in Australia15. 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 (ref. 16). We identified a total of 3,686 SNPs within the 3.8 Mb of non-repetitive core genome (representing 95%
of the genome). Of these SNPs, 3,150 (85%) 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 and Supplementary Note). These SNPs were removed, as they could mask the true phylogenetic signal, leaving 536 SNPs for downstream phylogenetic analysis. Other than these import events, homologous recombination has not had a major role in shaping the phylogeny of the global C. difficile 027/BI/NAP1 collection (Supplementary Fig. 2 and Supplementary Note).

We generated a maximum-likelihood phylogeny representing the C. difficile 027/BI/NAP1 global population structure (Fig. 1a). The phylogeny discriminates between >100 distinct genotypes within the global collection and shows limited geographic clustering, implying frequent long-range transmission among humans and, in a limited number of cases, two-way transmission between humans and animals 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 between $1.47 \times 10^{-7}$ and $5.33 \times 10^{-7}$ (95% confidence interval) substitutions per site per year, equivalent to 1–2 mutations per genome per year (Supplementary Fig. 4, Supplementary Table 2 and Supplementary Note). This rate is ~10 times slower than those observed in Streptococcus pneumoniae and Staphylococcus aureus over similar timescales, possibly owing to the fact that metabolically dormant spores do not accumulate mutations.

Notably, the global phylogeny shows the presence of two genetically distinct lineages, which have independently acquired an identical mutation (encoding a p.Thr82Ile alteration) in the DNA gyrase subunit A gene gyrA by either mutation or recombination, leading to high-level fluoroquinolone resistance (Fig. 1). Although both lineages share this mutation, the independent emergence of both lineages is highly supported by maximum-likelihood and Bayesian methods (Supplementary Figs. 3 and 5), unequivocally demonstrating that fluoroquinolone resistance has arisen in two epidemic C. difficile 027/BI/NAP1 lineages in two separate events. We named these two lineages FQR1 and FQR2. Bayesian analysis estimated that FQR1 and FQR2 emerged recently, with the most recent common ancestors appearing in ~1993 and ~1994, respectively (median estimates of 95% highest posterior density intervals were 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, South Korea, Canada, the United States, the UK, Germany and France) and sampling times (1985–2009). To our knowledge, none of the isolates...
supplementary Fig. 5]. Country color indicates where fluoroquinolone-resistant and fluoroquinolone-sensitive C. difficile 027/BI/NAP1 isolates have been reported.\(^6\) The width of the arrow is approximately proportional to the number of descendents from each sublineage.

Inset, enlarged view of the transmission within Europe. (b) Inferred arrivals and transmissions of the FQR2 lineage into and within the UK based on phylogeographic analysis (Supplementary Fig. 5) and maximum-likelihood phylogeny (Fig. 1b).

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.

The FQR1 lineage contains epidemic isolates associated with healthcare-related outbreaks in the United States (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\(^22\) indicated that the FQR1 lineage originated in the United States (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 CD1 caused by FQR C. difficile 027/BI/NAP1 (refs. 9,23). Thus, FQR1 seems to represent an epidemic lineage that originally emerged in the northeastern United States and was subsequently transmitted to South Korea and Switzerland.

The FQR2 lineage contains the majority of epidemic isolates that show widespread geographic distribution (Fig. 1a). The most notable feature in the FQR2 phylogeny is a star-like topology in the early part of the lineage, which is 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 (ref. 24) and the other (with uncertain association with outbreak) from The Netherlands in 2006. Bayesian analysis suggested that the FQR2 lineage also originated in North America (59% probability of the United States and 33% probability of Canada). Notably, all but one Canadian isolate 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 sublineages of FQR2 associated with distinct geographic locations, suggesting rapid transcontinental dissemination from North America to continental Europe, the UK and Australia. Our analysis shows a single introduction 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 and Supplementary Fig. 5). Similarly, the descend-ants of the FQR2 lineage 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 from 2004 to 2006 (ref. 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 145 additional UK isolates from healthcare patients (Supplementary Table 3) in addition to the 43 UK isolates in our global collection. The maximum-likelihood phylogeny of these isolates (Fig. 1b) shows that long-range geographic transmission was frequent and extensive within the UK (Fig. 2b). This collection confirms that FQR2 C. difficile 027/BI/NAP1 probably 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 (Figs. 1b and 2b). Its 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 sublineage whose earliest representative in our collection is a 2002 Birmingham isolate. This dominant sublineage also underlies the Stoke Mandeville outbreak\(^27\) (Figs. 1b and 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\(^7,2,28\). However, we discovered only two 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 was no change within the entire PaLoc region between pre- and post-epidemic isolates of C. difficile 027/BI/NAP1 that could explain the emergence of the epidemic strains (Supplementary Note). We provide strong evidence that the acquisition of fluoroquinolone resistance in two distinct lineages is associated with the emergence of C. difficile 027/BI/NAP1. We next determined whether there were other genetic changes linked to the
acquisition of fluoroquinolone resistance that could underlie a presumed increase in fitness, contributing to the emergence of the FQR1 and FQR2 lineages. Only two and seven SNPs defined the branches leading to lineages FQR1 and FQR2, respectively. However, besides the gyrA mutation (Fig. 1), there were no SNPs shared by both lineages (Supplementary Table 4). Furthermore, there was little evidence that a substantial change in phenotype could result from any of the SNPs that define the FQR1 and FQR2 lineages (Supplementary Table 4 and Supplementary Note), except perhaps the gyrA mutation itself. Besides these SNPs and the gyrA mutation, we also identified ten nonsynonymous SNPs conflicting with the phylogeny (homoplasic; 1.9% of the total number) that are present mainly within the ten nonsynonymous SNPs differing throughout the FQR2 lineage (6192). Furthermore, there was little evidence that a substantial change in phenotype could result from any of the SNPs that define the FQR1 and FQR2 lineages (Supplementary Table 4 and Supplementary Note). However, these SNPs are only present within a small fraction of the FQR1 and FQR2 isolates, meaning that their presence cannot explain the emergence of C. difficile 027/B/NAP1.

Horizontal gene transfer is a key mechanism driving C. difficile evolution29,30. We therefore assembled the C. difficile 027/B/NAP1 genomes and identified widespread acquisition of a range of mobile elements associated with erythromycin, chloramphenicol, tetraacycline and aminoglycoside resistance throughout the phylogeny (Supplementary Fig. 3 and Supplementary Note). Notably, we found a class of CTn5-like elements29 in almost all isolates in the FQR1 and FQR2 lineages, where they are inserted in the same genome location in the majority of isolates (Supplementary Fig. 6 and Supplementary Note). The FQR1 and FQR2 lineages both share a common CTn5-like element, which we named Tn6192, which contains genes encoding an ABC transport system, a two-component regulatory system 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 in length, Tn6105 (ref. 31), which contains 14 genes, 4 of which are predicted to be DNA-binding proteins or transcriptional regulators (Supplementary Table 5). This element has also acted as a hotspot for further integrations throughout the FQR2 lineage (Supplementary Fig. 6). The presence of Tn6192 is the only other shared genetic trait, aside from the fluoroquinolone resistance—conferring SNP, that differentiates the FQR1 and FQR2 lineages 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, Tn6192 may have had a role in the success of these lineages.

We show that the separate acquisitions of fluoroquinolone resistance and a conjugative transposon in two distinct lineages of C. difficile 027/B/NAP1 are the key genetic changes linked to its rapid emergence during the early 2000s7,8. Furthermore, 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/B/NAP1 in 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 2000s22, such that it is during this time that selective pressure for the acquisition and maintenance of fluoroquinolone resistance within healthcare settings would have been at its highest, explaining the near-simultaneous emergence of more than one clone of FQR C. difficile 027/B/NAP1. We also demonstrate the ease and rapidity with which FQR C. difficile 027/B/NAP1 has been transmitted internationally, highlighting the interconnectedness of the global healthcare system, which is 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 hospital33, and, although the low mutation rate of C. difficile might pose some limitations, we believe that this approach will have great potential for the study of C. difficile. Our analysis provides a genomic framework for understanding the population structure, geographic source, epidemiology and evolution of this highly transmissible healthcare-associated pathogen.

URLs. Path-O-Gen, http://tree.bio.ed.ac.uk/software/pathogen/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Illumina sequence reads for the C. difficile 027/B/NAP1 global and UK collections were deposited at the European Nucleotide Archive (full lists of accessions are given in Supplementary Tables 1 and 3).

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to members of the European Study Group of Clostridium difficile (ESGCD), a working group of ESCMID (European Society of Clinical Microbiology and Infectious Diseases), including F. Babut, T. Eckmanns, M.L. Lambert, F. Fitzpatrick, C. Wulf, H. Pittuch, P. Reichert, A.F. Widmer, F. Allerberger, D.W. Notermans, M. Delmée, R. Frei, O. Lytytkäinen, A. Ingebretsen and I.R. Poxton. We thank the Wellcome Trust Sanger Institute sequencing and informatics teams. This project was funded by the Wellcome Trust (grants 098051 and 086418), a Medical Research Council New Investigator Research Grant (T.D.L.; grant 93614) and the Scottish Infection Research Network.

Table 1 Nonsynonymous homoplasic SNPs identified in the core genomes of C. difficile 027/B/NAP1 isolates

| Position | Region                  | SNP                     | Substitution   | Antibiotic                      |
|----------|-------------------------|-------------------------|----------------|---------------------------------|
| 5420     | DNA gyrase subunit B (gyrB) | c.1276G>A               | p.Asp426Asn   | Fluoroquinolone                 |
| 6310     | DNA gyrase subunit A (gyrA) | c.245C>T               | p.Thr82Ile    | Fluoroquinolone                 |
| 9542     | DNA-directed RNA polymerase β chain (rpoB) | c.1504C>A | p.His502Asn | Rifampicin                      |
| 9542     | DNA-directed RNA polymerase β chain (rpoB) | c.1514G>A | p.Arg505Iys | Rifampicin                      |
| 103867   | Translation elongation factor G (fusA) | c.1363C>A, c. 1363C>T | p.His455Asn, p.His455Tyr | Fusidic acid                    |
| 104117   | Translation elongation factor G (fusA) | c.1613C>T | p.Pro538Leu | Fusidic acid                    |
| 1800920  | Two-component response regulator | c.31G>A | p.Glu111ys | Fusidic acid                    |
| 1802086  | Two-component sensor histidine kinase | c.446C>T | p.Thr149Ile | Fusidic acid                    |
| 3170481  | S-layer precursor protein (slpA) | c.467G>A, c.467G>T | p.Pro156Leu, p.Pro156Gln | Fusidic acid                    |
| 3170482  | S-layer precursor protein (slpA) | c.466G>A, c.466G>T | p.Pro156Ser, p.Pro156Thr | Fusidic acid                    |
| 3938789  | Putative membrane protein | c.996A>C | p.Tyr332* | Fusidic acid                    |

Positions refer to those in the R20291 genome. The alleles listed relate to the forward strand.

*Antibiotic to which the substitution confers resistance.
We acknowledge funding from the National Institute for Health Research (NIHR) Biomedical Research Centre in Liverpool. Both E.M. and P.R. were supported by the Liverpool BRC (Biomedical Research Centre). M.P. is an NIHR Senior Investigator.

AUTHOR CONTRIBUTIONS
M.H. analyzed the data. T.D.L., M.H., G. Dougan, B.W.W. and J.P. were involved in the study design. E.M., P.R., L.E., D.J.P., M.I.M., D.F., K.B.B., S.D., J.B., D.B., J.E.C., G. Douce, D.G., H.J.K., H.K., M.S., T.L., S.M., E.B., S.J.P., N.M.B., T.R., G.S., M.W., M.P., E.K., P.H. and B.W.W. were involved in isolate collection and DNA extraction. T.R.C. contributed to Bayesian analysis. M.H., J.P., T.D.L., G. Dougan, T.R.C. and S.R.H. contributed to data interpretation. M.H., J.P., T.D.L. and G. Dougan wrote the manuscript.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

G. Dougan wrote the manuscript.
T.R.C. and S.R.H. contributed to data interpretation. M.H., J.P., T.D.L. and G. Dougan wrote the manuscript.

Published online at http://www.nature.com/doifinder/10.1038/ng.2478.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Bartlett, J.G. Cladostium difficile: progress and challenges. Ann. NY Acad. Sci. 1213, 62–69 (2010).
2. Kelly, G.P. & LaMont, J.T. Cladostium difficile—more difficult than ever. N. Engl. J. Med. 359, 1932–1940 (2008).
3. Kuijper, E.J., Cognard, B. & Tul, P. Emergence of Cladostium difficile-associated disease in North America and Europe. Clin. Microbiol. Infect. 12 (suppl. 6), 2–18 (2006).
4. Wany, M. et al. Toxin production by an emerging strain of Cladostium difficile associated with outbreaks of severe disease in North America and Europe. Lancet 366, 1079–1084 (2005).
5. Rigs, M.M. et al. Asymptomatic carriers are a potential source for transmission of epidemic and nonepidemic Cladostium difficile strains among long-term care facility residents. Clin. Infect. Dis. 45, 992–998 (2007).
6. Clements, A.C., Magalhaes, R.J., Tatem, A.J., Parson, D.L. & Riley, T.V. Cladostium difficile PCR ribotype 027: assessing the risks of further worldwide spread. Lancet Infect. Dis. 10, 395–404 (2010).
7. McDonald, L.C. et al. An epidemic, toxin gene-variant strain of Cladostium difficile. N. Engl. J. Med. 353, 2433–2441 (2005).
8. Loo, V.G. et al. A predominantly clonal multi-institutional outbreak of Cladostium difficile-associated diarrhea with high morbidity and mortality. N. Engl. J. Med. 353, 2442–2449 (2005).
9. Muto, C.A. et al. A large outbreak of Cladostium difficile-associated disease with an unexpected proportion of deaths and colectomies at a teaching hospital following increased fluoroquinolone use. Infect. Control Hosp. Epidemiol. 26, 273–280 (2005).
10. Pêpin, J., Valiquette, L. & Cossette, B. Mortality attributable to nosocomial Cladostium difficile-associated disease during an epidemic caused by a hypervirulent strain in Quebec. CMAJ 173, 1037–1042 (2005).
11. O’Connor, J.R., Johnson, S. & Gerding, D.N. Cladostium difficile infection caused by the epidemic BI/NAP1/027 strain. Gastroenterology 136, 1913–1924 (2009).
12. Brazier, J.S. et al. Distribution and antimicrobial susceptibility patterns of Cladostium difficile PCR ribotypes in English hospitals, 2007–08. Euro Surveill. 13 pii: 19000 (2008).
13. Bauer, M.P. et al. Cladostium difficile infection in Europe: a hospital-based survey. Lancet 377, 63–73 (2011).
14. Kuijper, E.J. et al. Update of Cladostium difficile infection due to PCR ribotype 027 in Europe. 2008. Euro Surveill. 13 pii: 18942 (2008).
15. Richards, M. et al. Severe infection with Cladostium difficile PCR ribotype 027 acquired in Melbourne, Australia. Med. J. Aust. 194, 369–371 (2011).
16. Stabler, L. et al. Comparative genome and phenotypic analysis of Cladostium difficile 027 strains provides insight into the evolution of a hypervirulent bacterium. Genome Biol. 10, R102 (2009).
17. Croucher, N.J. et al. Rapid pneumococcal evolution in response to clinical interventions. Science 331, 430–434 (2011).
18. Harris, S.R. et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science 327, 469–474 (2010).
19. Pearson, T. et al. Phylogenetic discovery bias in Bacillus anthracis using single-nucleotide polymorphisms from whole-genome sequencing. Proc. Natl. Acad. Sci. USA 101, 13536–13541 (2004).
20. Spigaglia, P., Carattoli, A., Barbanti, F. & Mastrantonio, P. Detection of gyrA and gyrB mutations in Cladostium difficile isolates by real-time PCR. Mol. Cell. Probes 24, 61–67 (2010).
21. Kim, H. et al. Emergence of Cladostium difficile ribotype 027 in Korea. Korean J. Lab. Med. 31, 191–196 (2011).
22. Lemey, P., Rambaut, A., Drummond, A.J. & Suchard, M.A. Bayesian phylogenography finds its roots. PLOS Comput. Biol. 5, e1000520 (2009).
23. Dallai, R. et al. Fulminant Cladostium difficile: an underappreciated and increasing cause of death and complications. Ann. Surg. 235, 363–372 (2002).
24. MacCannell, D.R. et al. Molecular analysis of Cladostium difficile PCR ribotype 027 isolates from Eastern and Western Canada. J. Clin. Microbiol. 44, 2147–2152 (2006).
25. Drummond, A.J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 214 (2007).
26. Nytk, O., Pituch, H., Malejekova, J., Obuch-Woszczytsnki, P. & Kuijper, E.J. Cladostium difficile PCR ribotype 176 in the Czech Republic and Poland. Lancet 377, 1407 (2011).
27. Brazier, J.S. Cladostium difficile: the anaerobe that made the grade. Anaerobe 18, 197–199 (2012).
28. Carter, G.P., Awad, M.M., Kelly, M.L., Rodd, J.L. & Lyra, D. TcdB or not TcdB: a tale of two Cladostium difficile toxins. Future Microbiol. 6, 121–123 (2011).
29. Sebahia, M. et al. The multidrug-resistant human pathogen Cladostium difficile has a highly mobile, mosaic genome. Nat. Genet. 38, 779–786 (2006).
30. He, M. et al. Evolutionary dynamics of Cladostium difficile over short and long time scales. Proc. Natl. Acad. Sci. USA 107, 7527–7532 (2010).
31. Brouwer, M.S.M., Warburton, P.J., Roberts, A.P., Mullany, P. & Allan, E. Genetic organisation, mobility and predicted functions of genes on integrated, mobile genetic elements in sequenced strains of Cladostium difficile. PLoS ONE 6, e23014 (2011).
32. Linder, J.A., Huang, E.S., Steinman, M.A., Gonzales, R. & Stafford, R.S. Fluoroquinolone prescribing in the United States: 1995 to 2002. Am. J. Med. 118, 259–268 (2005).
33. Köser, C.U. et al. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. N. Engl. J. Med. 366, 2267–2275 (2012).
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 2 previous studies. New genome sequencing data were generated for six of these isolates. Details on the isolates are summarized in Supplementary Tables 1 and 3. Genomic DNA was extracted as previously described.

Sequencing, mapping and SNP detection. Paired-end multiplex libraries were created as described, and sequencing was performed on the Illumina Genome Analyzer IIX and HiSeq 2000 platforms. The read length was 54 bp for samples Liv1–Liv21, 108 bp for samples Glu01–Glu022 and 76 bp for the remaining samples. All isolates were sequenced to an average coverage of 110-fold across the isolates. Sequencing reads were aligned with the Burrows-Wheeler Aligner (BWA) against the genome sequence of the ribotype 027 reference strain R20291 (ref. 16). SNPs were identified with SAMTOOLS. A coverage cutoff of >5-fold and <3 times the average coverage was set for each individual isolate during SNP detection. Repetitive regions in the reference genome sequence were characterized using REPuter and the repeat finder functions in the MUMmer package. The boundaries of repetitive regions were extended to include the mobile elements in R20291. SNPs falling within these repetitive regions were excluded from analysis. To confirm the alleles at each variant position, SNPs were examined at each position in all sequencing reads in all isolates. An allele was only considered to be valid if supported by all reads (with 5 <coverage ≤40) or by >92.5% of the total reads (with coverage >40) covering the position; otherwise it was treated as having missing data. These parameters ensure no false positive SNPs and a false negative rate of ~8% (Supplementary Fig. 7 and Supplementary Note).

Phylogenetic analysis. An appropriate evolutionary model (simple GTR (generalized time reversible)) was determined using jModelTest 0.1.1 (ref. 38). Phylogenetic relationships were inferred using three methods: (i) the split-decomposition and neighbor-net methods in SplitsTree4 (ref. 39); (ii) the PHYML program; and (iii) the BEAST program. In the first two cases, a simple GTR model was used. Neighbor-joining trees (Supplementary Fig. 2b) were also constructed with PHYML, and the results were compared. Two non-027/BI/NAP1 C. difficile isolates, 630 (ref. 29) and CF5 (ref. 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 the skyline population model was determined to be more suitable on the basis of Bayes factor calculations and was used for later BEAST runs. The program was specified to estimate the 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 with resampling every 20,000 states. Phylogeographic history was also inferred with BEAST using a Bayesian method as described. The ESS (effective sample size) values were >200 for all parameters.

Accessory genome. For each genome, the unaligned sequencing reads were assembled using Velvet. To assess whether the resulting contigs were unique, each contig with length of >1 kb was searched using BLASTN against the current pan-genome, which was made by concatenating the draft genome sequences of M7404 and already determined unique contigs. Any unique contigs were added to the pan-genome. If a resulting match had >80% identity and covered >40% of the contig length, the corresponding 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. 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 the presence and absence of genomic regions in each isolate. 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 the Artemis Comparison Tool. Key mobile elements and their nomenclature are described in the 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 (i) clusters of SNPs within 2,000-bp windows and (ii) an iterative method to eliminate recombination sites as described. The identified homologous recombination blocks were excluded from phylogenetic and population genetic analysis.

Mutation rate estimates. The apparent mutation rate was estimated using three methods: (i) a full maximum-likelihood model assuming rapid expansion, which results in perfect star genealogies, implemented in an R script (Supplementary Fig. 8 and Supplementary Note); (ii) the BEAST program; and (iii) the Path-O-Gen v1.3 program (see URLs). A maximum-likelihood tree was used in this analysis. BEAST analyses were carried out as stated. The data set used for the second and third methods was the final SNP alignment of the global collection. A final mutation rate was determined by combining the single estimate from the third method and the 95% confidence intervals from the first and second methods.