Adaptation of Host Transmission Cycle During *Clostridium difficile* Speciation

Nitin Kumar[1,*,§], Hilary P. Browne[1,*], Elisa Viciani[1], Samuel C. Forster[1,2,3], Simon Clare[4], Katherine Harcourt[4], Mark D. Stares[1], Gordon Dougan[4], Derek J. Fairley[5], Paul Roberts[6], Munir Pirmohamed[6], Martha RJ Clokie[7], Mie Birgitte Frid Jensen[8], Katherine R. Hargreaves[7], Margaret Ip[9], Lothar H. Wieler[10,11], Christian Seyboldt[12], Torbjörn Norén[13,14], Thomas V. Riley[15,16], Ed J. Kuijper[17], Brendan W. Wren[18], Trevor D. Lawley[1,§]

[1]Host-Microbiota Interactions Laboratory, Wellcome Sanger Institute, Hinxton, CB10 1SA, UK.
[2]Centre for Innate Immunity and Infectious Diseases, Hudson Institute of Medical Research, Clayton, Victoria, 3168, Australia.
[3]Department of Molecular and Translational Sciences, Monash University, Clayton, Victoria, 3800, Australia.
[4]Wellcome Sanger Institute, Hinxton, CB10 1SA, UK.
[5]Belfast Health and Social Care Trust, Belfast, Northern Ireland.
[6]University of Liverpool, Liverpool, UK.
[7]Department of Infection, Immunity and Inflammation, University of Leicester, Leicester, LE1 7RH, UK.
[8]Department of Clinical Microbiology, Slagelse Hospital, Ingemannsvej 18, 4200, Slagelse, Denmark.
[9]Department of Microbiology, Chinese University of Hong Kong, Shatin, Hong Kong.
[10]Institute of Microbiology and Epizootics, Department of Veterinary Medicine, Freie Universität Berlin, Berlin, Germany.
[11]Robert Koch Institute, Berlin, Germany.
[12]Institute of Bacterial Infections and Zoonoses, Federal Research Institute for Animal Health (Friedrich-Loeffler-Institut), Jena, Germany.
[13]Faculty of Medicine and Health, Örebro University, Örebro, Sweden.
[14]Department of Laboratory Medicine, Örebro University Hospital Örebro, Sweden
[15]Department of Microbiology, PathWest Laboratory Medicine, Queen Elizabeth II Medical Centre, Western Australia, Australia.
[16]School of Pathology & Laboratory Medicine, The University of Western Australia, Western Australia, Australia
[17]Section Experimental Bacteriology, Department of Medical Microbiology, Leiden University Medical Center, Leiden, Netherlands.
[18]Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, University of London, London, UK.

[*]These authors contributed equally to this work

[§]Corresponding authors

Trevor D. Lawley: Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK, CB10 1SA, Phone 01223 495 391, Fax 01223 495 239, Email: tl2@sanger.ac.uk

Nitin Kumar: Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK, CB10 1SA, Phone 01223 495 391, Fax 01223 495 239, Email: nk6@sanger.ac.uk

Bacterial speciation is a fundamental evolutionary process characterized by diverging genotypic and phenotypic properties. However, the selective forces impacting genetic adaptations and how they relate to the biological changes underpinning the formation of a new bacterial species remain poorly understood. Here we show that the spore-forming, healthcare-associated enteropathogen *Clostridium difficile* is actively undergoing speciation. Applying large-scale genomic analysis of 906 strains, we demonstrate that the ongoing speciation process is linked to positive selection on core genes in the newly forming species that are involved in sporulation and the metabolism of simple dietary sugars. Functional validation demonstrates the new *C. difficile* produce more resistant spores and show increased sporulation and host colonization capacity when glucose or fructose is available for metabolism. Thus, we report the formation of an emerging *C. difficil*e species, selected for metabolizing simple dietary sugars and producing high levels of resistant spores that is adapted for healthcare-mediated transmission.

The formation of a new bacterial species from its ancestor is characterized by genetic diversification and biological adaptation[1-4]. For decades, a polyphasic examination[5], relying on genotypic and phenotypic properties of a bacterium, has been used to define and discriminate a "species". The bacterial taxonomic classification framework has more recently used large-scale genome analysis to incorporate aspects of a bacterium's natural history, such as ecology[6], horizontal gene transfer[1], recombination[2] and phylogeny[3]. Although a more accurate definition of a bacterial species can be achieved with whole-genome-based approaches, we still lack a fundamental understanding of how selective forces impact adaptation of biological pathways and phenotypic changes leading to bacterial speciation. In this work, we describe the genome evolution and biological changes during the ongoing formation of a new *C. difficile* species that is highly specialized for human transmission in the modern healthcare system.

C. difficile is a strictly anaerobic, Gram-positive bacterial species that produces highly resistant, metabolically dormant spores capable of rapid transmission between mammalian hosts through environmental reservoirs[7]. Over the past four decades, *C. difficile* has emerged as the leading cause of antibiotic-associated diarrhea worldwide, with a large burden on the healthcare system[7,8]. To define the evolutionary history and genetic changes underpinning the emergence of *C. difficile* as a healthcare pathogen, we performed whole-genome sequence analysis of 906 strains isolated from humans (n = 761), animals (n = 116) and environmental sources (n = 29) with representatives from 33 countries and the largest proportion originating from the UK (n = 465) (Supplementary Fig. 1; Supplementary Table 1; Supplementary Table 2). This dataset is summarized visually here https://microreact.org/project/H1QidSp14. Our collection was designed to capture comprehensive *C. difficile* genetic diversity[9] and includes 13 high-quality and well-annotated reference genomes (Supplementary Table 2). Robust maximum likelihood phylogeny based on 1,322 concatenated single-copy core genes (Fig.

1a; Supplementary Table 3) illustrates the existence of four major phylogenetic groups within this collection. Bayesian analysis of population structure (BAPS) using concatenated alignment of 1,322 single-copy core genes corroborated the presence of the four distinct phylogenetic groupings (PGs 1-4) (Fig. 1a) that each harbor strains from different geographical locations, hosts and environmental sources which indicates signals of sympatric speciation. Each phylogenetic group also harbors distinct clinically relevant ribotypes (RT): PG1 (RT001, 002, 014); PG2 (RT027 and 244); PG3 (RT023 and 017); PG4 (RT078, 045 and 033).

The phylogeny was rooted using closely related species (*C. bartlettii, C. hiranonis*, *C. ghonii* and *C. sordellii*) as outgroups (Fig. 1a). This analysis indicated that three phylogenetic groups (PG1, 2 and 3) of *C. difficile* descended from the most diverse phylogenetic group (PG4). This was also supported by the frequency of single-nucleotide polymorphism (SNP) differences in pairwise comparisons between strains of PG4 and each of the other PGs versus the level of pairwise SNP differences between comparisons of PGs 1, 2 and 3 to each other (Supplementary Fig. 2). Interestingly, bacteria from PG4 display distinct colony morphologies compared to bacteria from PG 1, 2 and 3 when grown on nutrient agar plates (Supplementary Fig. 3), suggesting a link between *C. difficile* colony phenotype and genotype that distinguishes PG 1, 2 and 3 from PG4.

Our previous genomic study using 30 *C. difficile* genomes indicated an ancient, genetically diverse species that likely emerged 1 to 85 million years ago[10]. Testing this estimate using our larger dataset indicated the species emerged approximately 13.5 million years (12.7-14.3 million) ago. Using the same BEAST[11] analysis on our substantially expanded collection, we estimate the most recent common ancestor (MRCA) of PG4 (using RT078 lineage) arose approximately 385,000 (297,137-582,886) years ago. In contrast, the MRCA of the PG1, 2 and 3 groups (using RT027 lineage) arose approximately 76,000

(40,220-214,555) years ago. Bayesian skyline analysis reveals a population expansion of PG1, 2 and 3 groups (using RT027 lineage) around 1595 A.D., which occurred shortly before the emergence of the modern healthcare system in the 18th century (Supplementary Fig. 4). Combined, these observations suggest that PG4 emerged prior to the other PGs and that the PG1, 2 and 3 population structure started to expand just prior to the implementation of the modern healthcare system[12]. We therefore refer to PG1, 2 and 3 groups as *C. difficile* "clade A" and PG4 as *C. difficile* "clade B".

To investigate genomic relatedness, we next performed pairwise Average Nucleotide Identity (ANI) analysis (Fig. 1b). This analysis revealed high nucleotide identity (ANI > 95%) between PGs 1, 2 and 3 indicating that bacteria from these groups belong to the same species; however, ANI between PG4 and any other PG was either less than the species threshold (ANI > 95%) or on the borderline of the species threshold (94.04%-96.25%) (Fig. 1b). To detect recombination events, FastGEAR analysis[13] was performed on whole-genome sequences of 906 strains. While analysis identified increased recombination within *C. difficile* clade A (PG1-PG2: 1-102, PG1-PG3: 1-214, PG2-PG3: 1-96) (Supplementary Fig. 5) a restricted number of recombination events between *C. difficile* clade A and clade B was observed (PG1-PG4: 1-20, PG2-PG4: 1-25, PG3-PG4: 1-46). This analysis strongly indicates the presence of recombination barriers in the core genome that further distinguishes the two *C. difficile* clades and could encourage sympatric speciation. Furthermore, accessory genome functional analysis also shows a clear separation between clade A and clade B (Supplementary Fig. 6; Supplementary Table 4-5). We also observe a higher number of pseudogenes in clade A compared to clade B (Supplementary Fig. 7; Supplementary Table 6-11). Taken together, these results indicate different selection pressures on the genomes of *C. difficile* clades A and B.

In addition to reduced rates of recombination events, advantageous genetic variants in a population driven by positive selective pressures, termed positive selection, are also a marker of speciation[6]. We determined the Ka/Ks ratios and identified 172 core genes in clade A and 93 core genes in clade B that were positively selected (Ka/Ks >1) (Fig. 2a; Supplementary Table 12-13). Functional annotation and enrichment analysis identified positively selected genes involved in carbohydrate and amino acid metabolism, sugar phosphotransferase system (PTS) and spore coat architecture and spore assembly in clade A (Fig. 2b). In contrast, the sulphur relay system was the only enriched functional category in positively selected genes from clade B. Notably, 26% (45 in total) of the positively selected genes in *C. difficile* clade A produce proteins that are either directly involved in spore production, are present in the mature spore proteome[14] or are regulated by Spo0A[15] or its sporulation-specific sigma factors[16] (Fig. 2c). In contrast, no positively selected genes are directly involved in spore production in *C. difficile* clade B; however, 22.5% (21 genes in total) are either present in the mature spore proteome or are regulated by Spo0A or its sporulation specific sigma factors (Supplementary Fig. 8). The lack of overlap between sporulation-associated positively selected genes in the two lineages suggests a divergence of spore-mediated transmission functions. In addition, these results suggest functions important for host-to-host transmission have evolved in *C. difficile* clade A.

We found 20 positively selected genes (Supplementary Table 12) in clade A whose products are components of the mature spore[14,15] and could contribute to environmental survival[17]. As an example, *sodA* (superoxide dismutase A), a gene associated with spore coat assembly, has three-point mutations which are present in all clade A genomes but absent in clade B genomes (Supplementary Fig. 9). Spores derived from diverse *C. difficile* clades have a wide variation in resistance to microbiocidal free radicals from gas plasma[18]. To investigate if the phenotypic resistance properties of spores from the new lineage have evolved, we exposed spores from both clades to hydrogen peroxide, a commonly used healthcare environmental disinfectant[17]. Spores derived from clade A were more resistant to 3% (P = 0.0317) and 10% hydrogen peroxide (P = 0.0317) when compared to spores from clade B, although there was no difference in survival at 30% peroxide likely due to the overpowering bactericidal effect at this concentration (P = 0.1667) (Fig. 3a).

The master regulator of *C. difficile* sporulation, *Spo0A*, is under positive selection in *C. difficile* clade A only. *Spo0A* also controls other host colonization factors, such as flagella, and carbohydrate metabolism, potentially serving to mediate cellular processes to direct energy to spore production and host colonization to facilitate host-to-host transmission[15]. Interestingly, the clade A genomes contain genes under positive selection that are involved in fructose metabolism (*fruABC* and *fruK*), glycolysis (*pgk* and *pyk*), sorbitol (CD630_24170) and ribulose metabolism (*rep1*), and conversion of pyruvate to lactate (*ldh*). To further explore the link between sporulation and carbohydrate metabolism in clade A, we analyzed positively selected genes using KEGG pathways[19] and manual curation. Manual curation of key enriched pathways across the 172 positively selected core genes in *C. difficile* clade A identified a complete fructose-specific PTS pathway and identified four genes (30%, 4/13) involved in anaerobic glycolysis during glucose metabolism (Supplementary Fig. 10). Other genes associated with enriched PTS pathways include genes used for the cellular uptake and metabolism of mannitol, cellobiose, glucitol/sorbitol, galactitol, mannose and ascorbate. Furthermore, comparative analysis of carbohydrate active enzymes (CAZymes)[20] identified a clear separation of CAZymes between *C. difficile* clade A and clade B (Supplementary Fig. 11; Supplementary Table 14). Combined, these observations suggest a divergence of functions between two *C. difficile* clades linked to metabolism of a broad range of simple dietary sugars [21].

The simple sugars glucose and fructose are increasingly used in diets within Western societies[21]. Interestingly, trehalose, a disaccharide of glucose, used as a food additive has impacted the emergence of some human virulent *C. difficile* variants[22]. Based on our genomic analysis, we hypothesized that dietary glucose or fructose could differentially impact host colonization by spores from *C. difficile* clade A or clade B. We therefore supplemented the drinking water of mice with either glucose, fructose or ribose and challenged with clade A or clade B strains. Ribose metabolic genes were not under positive selection so this sugar was included as a control. Mice challenged with clade A spores exhibited increased bacterial load when exposed to dietary glucose (P = 0.048) or fructose (P = 0.0045) compared to clade B (Fig. 3b). No difference in bacterial load was observed between *C. difficile* clade A and clade B without supplemented sugars or when supplemented with ribose (P = 0.2709) (Fig. 3b).

The infectivity and transmission of *C. difficile* within healthcare settings is facilitated by environmental spore density[23,24]. To determine the impact of simple sugar availability on spore production rates we assessed the ability of the two lineages to form spores in basal defined medium (BDM) alone or supplemented with either glucose, fructose or ribose. While no difference was observed on the ribose control (P = 0.3095), *C. difficile* clade A strains exhibited increased spore production on glucose (P = 0.0317) or fructose (P = 0.0317) (Fig. 3c). These results provide experimental validation and, together with our genomic predictions, suggest that enhanced host colonization and onward spore-mediated transmission with the consumption of simple dietary sugars is a feature of *C. difficile* clade A but not clade B.

The rapid recent emergence of *C. difficile* as a significant healthcare pathogen has mainly been attributed to the genomic acquisition of antibiotic resistance and carbohydrate metabolic functions on mobile elements via horizontal gene transfer[22,25]. Here we show that these recent genomic adaptations have occurred in established, distinct evolutionary lineages each with core genomes expressing unique, pre-existing transmission properties. We reveal the ongoing formation of a new species with biological and phenotypic properties consistent with a transmission cycle that was primed for human transmission in the modern healthcare system (Fig. 3d). Indeed, different transmission dynamics and host epidemiology have also been reported for *C. difficile* clade A (027 lineage[26] and 017 lineage[27]) endemic in healthcare systems in different parts of the world, and the 078 lineage that likely enters the human population from livestock[28-30]. Further, broad epidemiological screens of *C. difficile* present in the healthcare system often highlight high abundances of *C. difficil*e clade A as they represent 68.5% (USA), 74% (Europe) and 100% (Mainland China) of the infecting strains[7,8,31,32]. Thus, we report a link between *C. difficile* clade A speciation, adapted biological pathways and epidemiological patterns. In summary, our study elucidates how bacterial speciation may prime lineages to emerge and transmit in a process accelerated by modern human diet, the acquisition of antibiotic resistance or healthcare regimes.

Acknowledgements

This work was supported by the Wellcome Trust [098051]; the United Kingdom Medical Research Council [PF451 and MR/K000511/1] and the Australian National Health and Medical Research Council [1091097 to SF] and the Victorian government. The authors thank Scott Weese, Fabio Miyajima, Glen Songer, Thomas Louie, Julian Rood, and Nicholas M. Brown for *C. difficile* strains. The authors thank Anne Neville, Daniel Knight and Bastian Hornung for critical reading and comments. The authors would also like to acknowledge the support of the Wellcome Sanger Institute Pathogen Informatics Team.

Author Contributions

N.K. and T.D.L. conceived and managed the study. N.K., S.C.F., E.V., H.P.B. and T.D.L. wrote the manuscript. D.J.F., P.R., M.P., M.RJ.C., M.B.F.J., K.R.H., M.I., L.H.W., C.S., T.N., G.D., T.V.R., E.J.K., B.W.W. provided critical input and contributed to the editing of the manuscript. N.K. performed the computational analysis. H.P.B. performed genome annotation of reference genomes. D.J.F., P.R., M.P., M.RJ.C., M.B.F.J., K.R.H., M.I., L.H.W., C.S., T.N. provided *C. difficile* strains. E.V., H.P.B., S.C.F. and T.D.L. designed *in vitro* and *in vivo* experiments. H.P.B., E.V. and M.S. performed *in vitro* experiments. E.V., M.D.S., S.C. and K.H. performed *in vivo* experiments.

Conflict of interests

The authors declare no competing financial interests.

References:

1. Lawrence, J.G. & Retchless, A.C. The interplay of homologous recombination and horizontal gene transfer in bacterial speciation. *Methods Mol Biol* **532**, 29-53 (2009).
2. Fraser, C., Alm, E.J., Polz, M.F., Spratt, B.G. & Hanage, W.P. The bacterial species challenge: making sense of genetic and ecological diversity. *Science* **323**, 741-6 (2009).
3. Staley, J.T. The bacterial species dilemma and the genomic-phylogenetic species concept. *Philos Trans R Soc Lond B Biol Sci* **361**, 1899-909 (2006).
4. Moeller, A.H. *et al.* Cospeciation of gut microbiota with hominids. *Science* **353**, 380-382 (2016).
5. Vandamme, P. *et al.* Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiol Rev* **60**, 407-38 (1996).
6. Cohan, F.M. & Perry, E.B. A systematics for discovering the fundamental units of bacterial diversity. *Curr Biol* **17**, R373-86 (2007).
7. Martin, J.S., Monaghan, T.M. & Wilcox, M.H. Clostridium difficile infection: epidemiology, diagnosis and understanding transmission. *Nat Rev Gastroenterol Hepatol* **13**, 206-16 (2016).
8. Lessa, F.C., Winston, L.G., McDonald, L.C. & Emerging Infections Program, C.d.S.T. Burden of Clostridium difficile infection in the United States. *N Engl J Med* **372**, 2369-70 (2015).
9. Stabler, R.A. *et al.* Macro and micro diversity of Clostridium difficile isolates from diverse sources and geographical locations. *PLoS One* **7**, e31559 (2012).
10. He, M. *et al.* Evolutionary dynamics of Clostridium difficile over short and long time scales. *Proc Natl Acad Sci U S A* **107**, 7527-32 (2010).
11. Drummond, A.J., Suchard, M.A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* **29**, 1969-73 (2012).
12. Jackson, M. & Spray, E.C. Health and Medicine in the Enlightenment. (Oxford University Press, 2012).
13. Mostowy, R. *et al.* Efficient Inference of Recent and Ancestral Recombination within Bacterial Populations. *Mol Biol Evol* **34**, 1167-1182 (2017).
14. Lawley, T.D. *et al.* Proteomic and genomic characterization of highly infectious Clostridium difficile 630 spores. *J Bacteriol* **191**, 5377-86 (2009).
15. Pettit, L.J. *et al.* Functional genomics reveals that Clostridium difficile Spo0A coordinates sporulation, virulence and metabolism. *BMC Genomics* **15**, 160 (2014).
16. Fimlaid, K.A. *et al.* Global analysis of the sporulation pathway of Clostridium difficile. *PLoS Genet* **9**, e1003660 (2013).

17. Lawley, T.D. *et al.* Use of purified Clostridium difficile spores to facilitate evaluation of health care disinfection regimens. *Appl Environ Microbiol* **76**, 6895-900 (2010).

18. Connor, M. *et al.* Evolutionary clade affects resistance of Clostridium difficile spores to Cold Atmospheric Plasma. *Sci Rep* **7**, 41814 (2017).

19. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457-62 (2016).

20. Cantarel, B.L. *et al.* The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res* **37**, D233-8 (2009).

21. Lustig, R.H., Schmidt, L.A. & Brindis, C.D. Public health: The toxic truth about sugar. *Nature* **482**, 27-9 (2012).

22. Collins, J. *et al.* Dietary trehalose enhances virulence of epidemic Clostridium difficile. *Nature* (2018).

23. Browne, H.P. *et al.* Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature* **533**, 543-546 (2016).

24. Merrigan, M. *et al.* Human hypervirulent Clostridium difficile strains exhibit increased sporulation as well as robust toxin production. *J Bacteriol* **192**, 4904-11 (2010).

25. Sebaihia, M. *et al.* The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome. *Nat Genet* **38**, 779-86 (2006).

26. He, M. *et al.* Emergence and global spread of epidemic healthcare-associated Clostridium difficile. *Nat Genet* **45**, 109-13 (2013).

27. Cairns, M.D. *et al.* Comparative Genome Analysis and Global Phylogeny of the Toxin Variant Clostridium difficile PCR Ribotype 017 Reveals the Evolution of Two Independent Sublineages. *J Clin Microbiol* **55**, 865-876 (2017).

28. Dingle, K.E. *et al.* A Role for Tetracycline Selection in Recent Evolution of Agriculture-Associated Clostridium difficile PCR Ribotype 078. *MBio* **10**(2019).

29. Knetsch, C.W. *et al.* Zoonotic Transfer of Clostridium difficile Harboring Antimicrobial Resistance between Farm Animals and Humans. *J Clin Microbiol* **56**(2018).

30. Knight, D.R., Squire, M.M. & Riley, T.V. Nationwide surveillance study of Clostridium difficile in Australian neonatal pigs shows high prevalence and heterogeneity of PCR ribotypes. *Appl Environ Microbiol* **81**, 119-23 (2015).

31. Bauer, M.P. *et al.* Clostridium difficile infection in Europe: a hospital-based survey. *Lancet* **377**, 63-73 (2011).

32. Tang, C. *et al.* The incidence and drug resistance of Clostridium difficile infection in Mainland China: a systematic review and meta-analysis. *Sci Rep* **6**, 37865 (2016).

33. Argimon, S. *et al.* Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb Genom* **2**, e000093 (2016).

34. Croucher, N.J. *et al.* Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**, 430-4 (2011).

35. Harris, S.R. *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**, 469-74 (2010).

36. Quail, M.A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nat Methods* **5**, 1005-10 (2008).

37. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-9 (2008).

38. Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578-9 (2011).

39. Boetzer, M. & Pirovano, W. Toward almost closed genomes with GapFiller. *Genome Biol* **13**, R56 (2012).

40. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-9 (2014).

41. Chain, P.S. *et al.* Genomics. Genome project standards in a new era of sequencing. *Science* **326**, 236-7 (2009).

42. Page, A.J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691-3 (2015).

43. Katoh, K. & Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-80 (2013).

44. Croucher, N.J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* **43**, e15 (2015).

45. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-3 (2014).

46. Milne, I. *et al.* TOPALi v2: a rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops. *Bioinformatics* **25**, 126-7 (2009).

47. Ondov, B.D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**, 132 (2016).

48. Popescu, A.A., Huber, K.T. & Paradis, E. ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. *Bioinformatics* **28**, 1536-7 (2012).

49. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* **39**, W475-8 (2011).

50. Delcher, A.L., Phillippy, A., Carlton, J. & Salzberg, S.L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* **30**, 2478-83 (2002).

51. Cheng, L., Connor, T.R., Siren, J., Aanensen, D.M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol Biol Evol* **30**, 1224-8 (2013).

52. Tatusov, R.L., Galperin, M.Y., Natale, D.A. & Koonin, E.V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**, 33-6 (2000).

53. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet* **11**, 94 (2010).

54. Jombart, T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403-5 (2008).

55. Yin, Y. *et al.* dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res* **40**, W445-51 (2012).

56. Riley, M. Functions of the gene products of Escherichia coli. *Microbiol Rev* **57**, 862-952 (1993).

57. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J Mol Biol* **428**, 726-731 (2016).

58. Finn, R.D. *et al.* Pfam: the protein families database. *Nucleic Acids Res* **42**, D222-30 (2014).

59. Lerat, E. & Ochman, H. Recognizing the pseudogenes in bacterial genomes. *Nucleic Acids Res* **33**, 3125-32 (2005).

60. Rambaut, A., Drummond, A.J., Xie, D., Baele, G. & Suchard, M.A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst Biol* **67**, 901-904 (2018).

61. Karasawa, T., Ikoma, S., Yamakawa, K. & Nakamura, S. A defined growth medium for Clostridium difficile. *Microbiology* **141 (Pt 2)**, 371-5 (1995).

62. Duncan, S.H., Hold, G.L., Harmsen, H.J., Stewart, C.S. & Flint, H.J. Growth requirements and fermentation products of Fusobacterium prausnitzii, and a proposal to reclassify it as Faecalibacterium prausnitzii gen. nov., comb. nov. *Int J Syst Evol Microbiol* **52**, 2141-6 (2002).

Figure legends:

Figure 1. Phylogeny and population structure of *Clostridium difficile*. (a) Maximum likelihood tree of 906 *C. difficile* strains constructed from the core genome alignment, excluding recombination events. Collapsed clades as triangles represent four Phylogenetic groups (PG1-4) identified by Bayesian analysis of population structure (BAPS). Number in parentheses indicates number of strains. Key PCR ribotypes in each PG are shown. Bootstrap values of key branches are shown next to the branches. Dates indicate estimated emergence of *C. difficile* species-13.5 million (range 12.7-14.3) years ago, PG4- 385,000 (range 297,137-582,886) years ago and PG1-3- 76,000 (range 40,220-214,555) years ago. *C. bartlettii*, *C. hiranonis*, *C. ghonii* and *C. sordellii* were used as outgroups to root the tree. Scale bar indicates number of substitutions per site. (b) Distribution pattern of average nucleotide identity (ANI) for 906 *C. difficile* strains. Pairwise ANI calculations between different PGs are shown. Dotted red line indicates bacterial species cut-off.

Figure 2. Adaptation of sporulation and metabolic genes in *Clostridium difficile* clade A. Positive selection analysis of *C. difficile* clade A and B based on 1,322 core genes. (a) Distribution of Ka/Ks ratio for the positively selected genes in *C. difficile* clade A (n = 172 genes) and clade B (n = 93 genes) is shown. Error bars are standard error of the mean (SEM).

(b) Enriched functions in the positively selected genes of *C. difficile* clade A (n = 172 genes) and clade B (n = 93 genes) are shown. Y–axis represents number of positive selected genes in each enriched function. All are statistically significant (sugar phosphotransferase system (q = 0.00167), fructose and mannose metabolism (q = 0.001173), sporulation, differentiation and germination (q = 0.0165), cysteine and methionine metabolism (q = 0.00279), sulphur relay system (q = 0.00791)). One-sided Fisher's exact test with *P* value adjusted by Hochberg method. (c) Positively selected sporulation-associated genes in *C. difficile* clade A are shown in blue. Of the 172 genes under positive selection, 26% (45 in total) are either involved in spore production (sporulation stages I, III, IV and V), their proteins are present in the mature spore proteome or they are regulated by Spo0A or its sporulation specific sigma factors.

Figure 3. Bacterial speciation is linked to increased host adaptation and transmission ability. (a) Spores of *C. difficile* clade A are more resistant to widely used hydrogen peroxide disinfectant. Spores of *C. difficile* clade A and clade B (n = 5 different ribotypes for both lineages) were exposed to hydrogen peroxide for 5 minutes, washed and plated. Recovered CFUs representing surviving germinated spores were counted and presented as a percentage of spores exposed to PBS. Mean and range of 3 independent experiments is presented, Mann-Whitney unpaired two-tailed test. (b) Intestinal colonization of clade A strains is increased in the presence of simple sugars compared to clade B strains. Comparison of vegetative cell loads between *C. difficile* clade A (n = 1, RT027) and clade B (n = 1, RT078) strains in mice whose diet was supplemented with different sugars before challenging with spores. CFUs from fecal samples cultured 16 hours after *C. difficile* challenge are presented. Mean values of 5 mice are presented from 1 representative experiment which was repeated once with similar results, standard error of the mean (SEM), unpaired two-tailed *t* test. (c) Clade A strains produce more spores in the presence of simple sugars. *C. difficile* clade A and clade B

(n = 5 different ribotypes for both lineages) strains were grown on basal defined media in the presence or absence of different sugars, vegetative cells were killed by ethanol exposure and the number of CFUs representing germinated spores were counted. The percentage of spores recovered in the presence of sugars compared to BDM alone is presented. Mean and range of 3 independent experiments is presented, Mann-Whitney unpaired two-tailed test. (d) Overview of adaptations in key aspects of the *C. difficile* clade A transmission cycle in human population.

Online Methods

Collection of *C. difficile* strains

Laboratories worldwide were asked to send a diverse representation of their *C. difficile* collections to the Wellcome Sanger Institute (WSI). After receiving all shipped samples the DNA extraction was performed batch-wise using the same protocol and reagents to minimize bias. Phenol-Chloroform was the preferred method for extraction since it provides high DNA yield and intact chromosomal DNA.

The genomes of 382 strains designated as *C. difficile*, by PCR ribotyping were sequenced and combined with our previous collection of 506 *C. difficile* strains, 13 high quality *C. difficile* reference strains and 5 publicly available *C. difficile* RT 244 strains making a total of 906 strains analyzed in this study. This genome collection includes strains from humans (n = 761), animals (n =116) and the environment (n = 29) that were collected from diverse geographic locations (UK; n = 465, Europe; n = 230, N-America; n = 111, Australia; n = 62, Asia; n = 38). Details of all strains are listed in Supplementary Table 1 and Supplementary Table 2, including the European Nucleotide Archive (ENA) sample accession numbers. Metadata of this *C. difficile* collection have been made freely publicly available through Microreact[33] (https://microreact.org/project/H1QidSp14).

Bacterial culture and genomic DNA preparation

C. difficile strains were cultured on blood agar plates for 48 hours, inoculated into brain–heart infusion broth supplemented with yeast extract and cysteine and grown overnight (16 hours) anaerobically at 37 °C. Cells were pelleted, washed with PBS, and genomic DNA preparation was performed using a phenol–chloroform extraction as previously described[34]. All culturing of *C. difficile* took place in anaerobic conditions (10% CO_2, 10% H_2, 80% N_2) in a Whitley DG250 workstation at 37 °C. All reagents and media were reduced for 24 hours in anaerobic conditions before use.

DNA sequencing, assembly and annotation

Paired-end multiplex libraries were prepared and sequenced using Illumina Hi-Seq platform with fragment size of 200-300 bp and a read length of 100 bp, as previously described[35,36]. An in-house pipeline developed at the WSI (https://github.com/sanger-pathogens/Bio-AutomatedAnnotation) was used for bacterial assembly and annotation. It consisted of *de novo* assembly for each sequenced genome using Velvet v1.2.10[37], SSPACE v2.0[38] and GapFiller v1.1[39] followed by annotation using Prokka v1.5-1[40]. For the 13 high-quality reference genomes, strains Liv024, TL178, TL176, TL174, CD305 and Liv022 were sequenced using 454 and Illumina sequencing platforms, BI-9 and M68 were sequenced using 454 and capillary sequencing technologies with the assembled data for these 8 strains been improved to an 'Improved High Quality Draft' genome standard[41]. Optical maps using the Argus Optical Mapping system were also generated for Liv024, TL178, TL176, TL174, CD305 and Liv022. The remaining strains are all contiguous and were all sequenced using 454 and capillary sequencing technologies except for R20291 which also had Illumina data incorporated and 630 which was sequenced using capillary sequence data alone.

Phylogenetic analysis, Pairwise SNP distances analysis and Average Nucleotide Identity analysis

The phylogenetic analysis was conducted by extracting nucleotide sequence of 1,322 single copy core gene from each *C. difficile* genome using Roary[42]. The nucleotide sequences were concatenated and aligned with MAFFT v7.20[43]. Gubbins[44] was used to mask recombination from concatenated alignment of these core genes and a maximum-likelihood tree was constructed using RAxML v8.2.8[45] with the best-fit model of nucleotide substitution (GTRGAMMA) calculated from ModelTest embedded in TOPALi v2.5[46] and 500 bootstrap replicates. The phylogeny was rooted using a distance-based tree generated using Mash v2.0[47], R package APE[48] and genome assemblies of closely related species (*C. bartlettii, C. hiranonis*, *C. ghonii* and *C. sordellii*). All phylogenetic trees were visualized in iTOL[49]. Genomes of closely related *C. difficile* were downloaded from NCBI. Pairwise SNP distances analysis was performed on concatenated alignment of 1,322 single-copy core genes using SNP-Dist (https://github.com/tseemann/snp-dists). Average nucleotide analysis (ANI) was calculated by performing pairwise comparison of genome assemblies using MUMmer[50].

Population structure and recombination analysis

Population structure based on concatenated alignment of 1,322 single-copy core genes of *C. difficile* was inferred using the HierBAPS[51] with one clustering layers and 5, 10 and 20 expected numbers of clusters (k) as input parameters. Recombination events across the whole-genome sequences were detected by mapping genomes against a reference genome (NCTC 13366; RT027) and using FastGear[13] with default parameters.

Functional genomic analysis

To explore accessory genome and identify protein domains in a genome, we performed RPS-BLAST using COG database (accessed February 2019)[52]. All protein domains were classified in different functional categories using the COG database[52] and were used to perform Discriminant Analysis of Principle Components (DAPC)[53] implemented in the R package Adegenet v2.0.1[54]. Domain and functional enrichment analysis was calculated using one-sided Fisher's exact test with *P* value adjusted by Hochberg method in R v3.2.2.

Carbohydrate active enzymes (CAZymes) in a genome were identified using dbCAN v5.0[55] (HMM database of carbohydrate active enzyme annotation). Best hits include hits with E-value < 1×10^{-5} if alignment > 80 aa and hits with E-value < 1×10^{-3} if alignment < 80 aa, and alignment coverage > 0.3. Best hits were used to perform Discriminant Analysis of Principle Components (DAPC)[53] implemented in the R package Adegenet v2.0.1[54].

Functional annotation of positively selected genes was carried out using the Riley classification system[56], KEGG Orthology[57] and Pfam functional families[58].

Analysis of selective pressures.

The aligned nucleotide sequences of each 1,322 single copy core genes were extracted from Roary's output. The ratio between the number of non-synonymous mutations (Ka) and the number of synonymous mutations (Ks) was calculated for the whole alignment and for the respective subsets of strains belonging to the PG1, 2, 3 as a group and PG4. The Ka/Ks ratio for each gene alignment was calculated with SeqinR v3.1. A Ka/Ks > 1 was considered as the threshold for identifying genes under positive selection.

Pseudogenes analysis

Nucleotide annotations of genes within a genome within each phylogenetic group were mapped against the protein sequences of the reference genome for its phylogenetic group (PG1: NCTC 13307(RT012), PG2: SRR2751302 (RT244), PG3: NCTC 14169 (RT017), PG4: NCTC 14173 (RT078)) using TBLASTN as previously described[59]. Pseudogenes were called based on following criteria: genes with E value > 1×10^{-30} and sequence identity < 99% and which are absent in 90% members of a phylogenetics group. Genes in the reference genomes annotated as a pseudogene were also included in addition to genes in query genomes.

Analysis of estimating dates

The aligned nucleotide sequences of each 222 core genes of *C. difficile* which are under neutral selection (Ka/Ks = 1) were extracted from Roary's output. Gubbins[44] was used to mask recombination from concatenated alignment of these core genes and used as an input for Bayesian Evolutionary Analysis Sampling Trees (BEAST) software package v2.4.1[11]. In BEAST, the MCMC chain was run for 50 million generations, sampling every 1,000 states using the strict clock model (2.50×10^{-9} - 1.50×10^{-8} per site per year)[10] and HKY four discrete gamma substitution model, each run in triplicate. Convergence of parameters were verified with Tracer v1.5[60] by inspecting the Effective Sample Sizes ($\text{ESS} > 200$). LogCombiner was used to remove 10% of the MCMC steps discarded as burn-ins and combine triplicates. The resulting file was used to infer the time of divergence from the most recent common ancestor for *C. difficile*, *C. difficile* clade A and clade B. The Bayesian skyline plot was generated with Tracer v1.5[60].

***C. difficile* growth *in vitro* on selected carbon sources**

Basal defined medium (BDM)[61] was used as the minimal medium to which selected carbon sources (2 g/l of glucose, fructose or ribose from Sigma-Aldrich) were added. *C. difficile* strains were grown on CCEY agar (Bioconnections) for two days; 125-ml Erlenmeyer flasks containing 10 ml of BDM with or without carbon sources were inoculated with *C. difficile* strains and incubated in anaerobic conditions at 37 ºC shaking at 180 rpm. After 48 hours, spores were enumerated by centrifuging the culture to a pellet, carefully decanting the BDM and re-suspending in 70% ethanol for 4 hours to kill vegetative cells. Following ethanol shock, spores were washed twice in PBS and plated in a serial dilution on YCFA media[62] supplemented with 0.1% sodium taurocholate. Colony forming units (representing germinated spores) were counted 24 hours later. The experiment was performed independently 3 times for each strain. Clade A strains used were TL178 (RT002/ PG1), TL174 (RT015/ PG1), R20291 (RT027/ PG2), CF5 (RT017/ PG3) and CD305 (RT023/ PG3). Clade B strains used were MON024 (RT033), CDM120 (RT078), WA12 (RT291), WA13 (RT228) and MON013 (RT127). Data were presented using GraphPad Prism v7.03.

C. difficile spore resistance to disinfectant

Spores were prepared by adapting the previous protocol[18]. In brief, *C. difficile* strains were streaked on CCEY media, the cells were harvested from the plates 48 hours later and subjecting to exposure in 70% ethanol for 4 hours to kill vegetative cells. The solution was then centrifuged, ethanol was decanted and the spores were washed once in 5 ml sterile saline (0.9% w/v) solution before being suspended in 5 ml of saline (0.9% w/v) with Tween20 (0.05% v/v). 300 µl spore suspensions (at a concentration of approximately 10^6 spores) were exposed to 300 µl of 3%, 10% and 30% hydrogen peroxide (Fisher Scientific UK Limited) solutions for 5 minutes in addition to 300 µl PBS. The suspensions were then centrifuged, hydrogen peroxide or PBS was decanted and the spores were washed twice with PBS. Washed spores were plated on YCFA media with 0.1% sodium taurocholate to stimulate spore germination and colony forming units were counted 24 hours later. The experiment was performed independently 3 times for each strain. Clade A strains used were TL178 (RT002/ PG1), TL174 (RT015/ PG1), R20291 (RT027/ PG2), CF5 (RT017/ PG3) and CD305 (RT023/ PG3). Clade B strains used were MON024 (RT033), CDM120 (RT078), WA12 (RT291), WA13 (RT228) and MON013 (RT127). Data were presented using GraphPad Prism v7.03.

In vivo C. difficile colonization experiment

Five female 8-week-old C57BL/6 mice were given 250 mg/l clindamycin (Apollo Scientific) in drinking water. After 5 days, clindamycin treatment was interrupted and 100 mM of glucose, fructose or ribose was added to mouse drinking water for the rest of the experiment; no sugars were given to control mice. After 3 days, mice were infected orally with 6 × 10^3 spore/mouse of *C. difficile* R20291 (RT027) or M120 (RT078) strain. Fecal samples were collected from all mice before infection to check for pre-existing *C. difficile* contamination. Spore suspensions were prepared as described above[18]. After 16 hours, fecal samples were collected from all mice to determine viable *C. difficile* cell counts by serial dilution and plating on CCEY agar supplemented with 0.1% sodium taurocholate. The mean values of 5 mice are presented from 1 representative experiment which was repeated once with similar results. Data were presented using GraphPad Prism version 7.03. Ethical approval for mouse experiments was obtained from the Wellcome Sanger Institute.

Reporting Summary

Further information on research design is available in the Life Sciences Reporting Summary linked to this article.

Data Availability

Genomes have been deposited in the European Nucleotide Archive. Accession codes are listed in Supplementary Table 1. The 13 *C. difficile* reference isolates (Supplementary Table 2) are publicly available from the National Collection of Type Cultures (NCTC) and the annotation of these genomes are available from the Host-Microbiota Interactions Lab (HMIL; **www.lawleylab.com**), Wellcome Sanger Institute.

Code Availability

No custom code was used.

Methods-only References

33. Argimon, S. *et al.* Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb Genom* **2**, e000093 (2016).
34. Croucher, N.J. *et al.* Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**, 430-4 (2011).
35. Harris, S.R. *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**, 469-74 (2010).
36. Quail, M.A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nat Methods* **5**, 1005-10 (2008).
37. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-9 (2008).
38. Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578-9 (2011).
39. Boetzer, M. & Pirovano, W. Toward almost closed genomes with GapFiller. *Genome Biol* **13**, R56 (2012).
40. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-9 (2014).
41. Chain, P.S. *et al.* Genomics. Genome project standards in a new era of sequencing. *Science* **326**, 236-7 (2009).
42. Page, A.J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691-3 (2015).
43. Katoh, K. & Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-80 (2013).
44. Croucher, N.J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* **43**, e15 (2015).
45. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-3 (2014).

46. Milne, I. *et al.* TOPALi v2: a rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops. *Bioinformatics* **25**, 126-7 (2009).

47. Ondov, B.D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**, 132 (2016).

48. Popescu, A.A., Huber, K.T. & Paradis, E. ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. *Bioinformatics* **28**, 1536-7 (2012).

49. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* **39**, W475-8 (2011).

50. Delcher, A.L., Phillippy, A., Carlton, J. & Salzberg, S.L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* **30**, 2478-83 (2002).

51. Cheng, L., Connor, T.R., Siren, J., Aanensen, D.M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol Biol Evol* **30**, 1224-8 (2013).

52. Tatusov, R.L., Galperin, M.Y., Natale, D.A. & Koonin, E.V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**, 33-6 (2000).

53. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet* **11**, 94 (2010).

54. Jombart, T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403-5 (2008).

55. Yin, Y. *et al.* dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res* **40**, W445-51 (2012).

56. Riley, M. Functions of the gene products of Escherichia coli. *Microbiol Rev* **57**, 862-952 (1993).

57. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J Mol Biol* **428**, 726-731 (2016).

58. Finn, R.D. *et al.* Pfam: the protein families database. *Nucleic Acids Res* **42**, D222-30 (2014).

59. Lerat, E. & Ochman, H. Recognizing the pseudogenes in bacterial genomes. *Nucleic Acids Res* **33**, 3125-32 (2005).

60. Rambaut, A., Drummond, A.J., Xie, D., Baele, G. & Suchard, M.A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst Biol* **67**, 901-904 (2018).

61. Karasawa, T., Ikoma, S., Yamakawa, K. & Nakamura, S. A defined growth medium for Clostridium difficile. *Microbiology* **141 (Pt 2)**, 371-5 (1995).

62. Duncan, S.H., Hold, G.L., Harmsen, H.J., Stewart, C.S. & Flint, H.J. Growth requirements and fermentation products of Fusobacterium prausnitzii, and a proposal to reclassify it as Faecalibacterium prausnitzii gen. nov., comb. nov. *Int J Syst Evol Microbiol* **52**, 2141-6 (2002).