Supplementary Appendix

Supplement to: Eyre DW, Cule ML, Wilson DJ, et al. Diverse sources of *C. difficile* infection identified on whole-genome sequencing. N Engl J Med 2013;369:1195-205. DOI: 10.1056/NEJMoa1216064

Supplementary Appendix

Diverse Sources of *C. difficile* Infection Identified on Whole-Genome Sequencing

David W Eyre, Madeleine L Cule, Daniel J Wilson, David Griffiths, Alison Vaughan, Lily O'Connor, Camilla LC Ip, Tanya Golubchik, Elizabeth M Batty, John M Finney, David H Wyllie, Xavier Didelot, Paolo Piazza, Rory Bowden, Kate E Dingle, Rosalind M Harding, Derrick W Crook, Mark H Wilcox, Tim EA Peto, A Sarah Walker

Supplementary Methods and Results are presented with the same order and sub-headings as the main text:

Supplementary Methods

Setting

Interventions in place at the start of the study (September 2007)	
Intervention	**Compliance monitoring**
Antibiotic policy: Cephalosporin and fluroquinolone use restricted. Community acquired pneumonia: amoxicillin / co-amoxiclav ± macrolide; urinary tract infection: nitrofurantoin / co-amoxiclav; cellultis: flucloxacillin. Cephalosporin use in penicillin allergy, fluroquinolone for severe beta-lactam allergy.	Pharmacy review of all inpatient prescriptions against hospital-wide antibiotic policy
C difficile isolation policy: All patients with suspected C difficile infection, including patients with diarrhea of unknown cause (≥3 unformed stools in 24 hours) isolated in side rooms with en suite bathroom or dedicated commode wherever possible. Hand decontamination with soap and water, aprons and gloves worn for contact. Continued in confirmed cases until 48 hours following return to normal bowel habit.	Monitored daily, Monday to Friday by infection control team review of all inpatients with diarrhea. Frequency of monitoring decreased to 3 times a week from 2008 to 2009, and once a week during 2010. Spot checks of patient notes undertaken to ensure infection control service aware of all patients with suspected *C difficile*, confirmed cases notified by laboratory to infection control team.
Cleaning: Actichor plus combined detergent and Sodium hypochlorite solution (called an "enhanced clean") used for once daily clean of every surface and wall in room with patient with suspected *C difficile*.	Daily monitoring, Monday to Friday of the number of additional enhanced cleans requested with the number of patients with suspected or confirmed active *C difficile* cases. Records from procurement monitored

	Monday to Friday. Frequency reduced over time as above.
C difficile testing: Three samples to be sent for C difficile testing from any patients with suspected infectious diarrhea (≥3 unformed stools in 24 hours)	Reviewed as part of infection control monitoring of these patients as above. Compliance with this policy was high with only 6% of samples testing positive.`[1]
Mandatory C difficile testing: Mandatory testing of all diarrheal samples sent from patients aged 65 years or older for *C. difficile*, regardless of whether test requested	Number of EIA tests increases from 400-500 per month before, to 800-1000 per month after introduction in May 2007
C difficile treatment: Empirical treatment of oral vancomycin 125mg qds, continued for 14 days if diagnosis confirmed	Initiation of treatment confirmed by infection control team for all confirmed cases. Pharmacy monitoring of duration of treatment.
Feedback: Failure to comply with infection control policy on isolation and cleaning escalated to ward and then senior management until compliance achieved.	

Study starts (September 2007)

Date of Change	**Intervention**	**Compliance Monitoring**
January 2009	***Care pathway:*** *C. difficile* care pathway added to policy Care checklist including communication of diagnosis to patient, family, and medical team; review of medication (antibiotics, acid suppression, laxatives); fluid balance monitoring (fluid and stool chart); nutritional	All confirmed *C difficile* cases reviewed by infection control team to ensure care pathway in place

	assessment; skin integrity assessment; sepsis monitoring; compliance with isolation as above	
September 2009	***C. difficile isolation policy:*** Isolation as previously, but now within 2hrs of suspicion Exception where patients requiring one-to-one monitoring or care and unsuitable for side room care	Monitored Monday to Friday by infection control service at introduction, then 3 times a week and then weekly as above
December 2009	***Sending of stool specimens:*** Reduction in number of stool samples tested. Sending one diarrheal specimen and if negative then send a second	Monitored weekly from January 2010, over testing reported back weekly to wards

Table S1 OUH Infection Control Practice. Interventions in place at the start of the study (September 2007), and changes during the study (through March 2011) are shown separately.

Sample preparation and sequencing

DNA was extracted using a commercial kit (FastDNA, MP Biomedicals, California, USA; QIAamp, Qiagen, Hilden, Germany; QuickGene, Fujifilm, Tokyo, Japan), from a single colony sub-cultured onto a Columbia blood agar plate and incubated for 48 h. A combination of standard Illumina and adapted protocols was used to produce multiplexed paired-end libraries. Pools of 96 samples were sequenced at the Wellcome Trust Centre for Human Genetics, Oxford, UK, using sequencing-by-synthesis technology, on the Illumina Genome Analyzer II (GAII), GAIIx, and

HiSeq2000 platforms, generating 51, 101-108, 100 base paired-end reads respectively.

Sequence reads were analysed and assembled using a pipeline developed specifically for bacterial genomes. Each isolate was mapped using Stampy v1.0.11 (without Burrows-Wheeler Aligner pre-mapping, using an expected substitution rate of 0.01)[2] to the *C. difficile* 630 reference genome (Genbank: AM180355.1), CD630,[3] with the exception of Clade 2 (which includes ST1/ribotype-027/NAP1).[4] Clade 2 isolates were mapped to CD196 (Genbank: NC_013315.1) to allow the additional novel sequence in these samples to be compared.[5]

Base-pair calls were identified across all mapped non-repetitive core genome sites using SAMtools (version 0.1.12-10) mpileup with the extended base-alignment quality flag, after parameter tuning based on bacterial sequences. A consensus of ≥75% was required to support a call, and calls were required to be homozygous under a diploid model. Only calls supported by ≥5 reads, including one in each direction were accepted. Base-pair calls that were identified after quality filtering were used to identify single nucleotide variants (SNVs) between pairs of sequenced isolates and in phylogenetic comparisons.

To confirm the reproducible nature of the sequencing and pipeline 85 samples were sequenced multiple times. A total of 180 additional sequences were generated such that 38 samples were sequenced twice, 41 three times, and the remainder 4-22 times. Each replicate was compared to the sequence used in the manuscript (the most recently sequenced), only 2/180 sequences differed, each by only 1 SNV (1 SNV error per 90 genomes compared).

Microevolutionary rate

The rate of short-term evolution and within-host diversity was estimated from the SNV differences between *C. difficile* isolates (sharing the same ST) from the first and last faecal samples received from 145 patients who were tested on ≥2 occasions during the study, a median (IQR) [range] 51 (28-105) [0–561] days apart. As samples were selected conditional on the first and last samples sharing the same ST, re-infections with a new ST were excluded from estimates of evolutionary rates (however these samples are included in the main analysis as re-infections).[6]

In keeping with the method used in the main analysis, clade 2 samples were mapped to CD196, and samples from all other clades to CD630. Three patients were excluded with samples separated by 60, 60, and 374 days and isolates separated by 20, 823, and 20 SNVs, respectively. Each of these was considered more likely to represent a reinfection, as there was no evidence of a single recombination; furthermore, on both occasions where the isolates were separated by 20 SNVs, a sequence matching the later sample had previously been found in another patient.

A coalescent theory-based model[7] was fitted by maximum likelihood. The amount of diversity observed between a pair of samples obtained at times t_0 and

t_1 can be thought of as the sum of 2 Poisson processes, firstly the variation that arose in the time, *t*, between the samples being taken, and secondly the variation that has arose in the time, *u*, between the most recent common ancestor of the two samples and time t_0 (Figure S1).

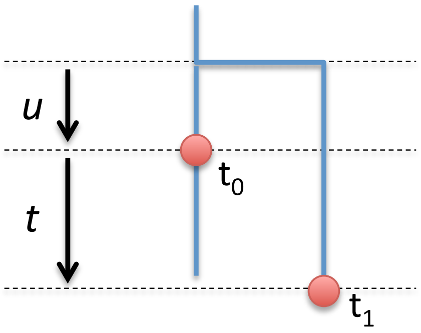

Figure S1. Coalescent-based model. t_0 denotes the time of the first sample, t_1 the time of the second sample. *t* is the time between the samples, and *u* the time between the first sample and the common ancestor of the samples.

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

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

Which can be written:

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

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

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

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

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

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

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

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

The data can be thought of as *n* independent observed pairs of samples separated by time t_i and with single nucleotide variants s_i between them. Therefore to calculate the likelihood of μ and N_e given the data, the product of the Pr(s_i) across all the *n* pairs of data was calculated:

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

Maximum likelihood values of μ and N_e were found by numerical optimization using R (http://www.r-project.org). 95% confidence intervals for parameter estimates were calculated by parametric bootstrap, using 1000 iterations.

Analysis

After accounting for repeated isolates of the same sequence type, ST, within a patient, 1250 infections were identified for sequencing. 1223/1250 (98%) were successfully sequenced, with mean 83.4% of the CD630 reference genome called after filtering for non-clade 2 samples, and 93.4% of the CD196 reference genome called for clade 2 samples. The remaining regions of the reference genome include repetitive regions that were explicitly masked, and several large mobile elements not present in many of the genomes mapped to CD630.

To analyze the genetic relationship between cases, sequences within 100 observed SNVs were grouped (for computational efficiency) and maximum likelihood trees estimated using PhyML[8] with a generalized time reversible substitution model. SNV distances between genomes for subsequent analyses were obtained from the estimated trees, or for larger distances from the observed SNVs (inter-100SNV groups).

Given the diversity observed within a host over time (see results), the probability of observing >10 SNVs through evolution during the 3.6-year study is

<0.001. We therefore considered cases >10 SNVs apart from any other as distinct subtypes to assess how many genetically-distinct clusters of cases occurred during the study. We also present subtypes >100 SNVs from any other case for comparison. The total number of distinct subtypes present in Oxfordshire was estimated using the number of distinct subtypes identified for a given sampling effort using models described by Lin[9,10] and Clench. [10,11]

The incidence of CDI caused by genetically-distinct isolates (>10SNVs from any previous case, 'introductions' to the symptomatic Oxfordshire population), was compared to the incidence of cases genetically-related to a previous case (≤2SNVs, possible secondary cases), to test the hypothesis that interventions aimed at reducing transmission might have preferentially affected genetically-related cases. Only cases from 01 April 2008, the final 3 years of the study, were analyzed; the first seven months of the study were denoted a run-in period to capture the increased rate of novel subtypes expected initially. Per annum rate ratios (change in incidence) were determined by Poisson regression. Heterogeneity p values were determined by stacked Poisson regression. During the study period sample submission rates also fell, but by a lesser extent than positive cases (per annum rate ratio 0.91 (95% CI 0.90-0.92), heterogeneity p<0.001 versus overall CDI incidence). Incidence is reported as monthly cases, where monthly cases is scaled to be 30 times the mean daily cases in the month, to allow for accurate comparison across months.

Epidemiological analysis

Approximate epidemiological relationships between genetically-related cases were classified as set out in the main Methods section. For genetically-related cases without hospital contact, we investigated whether patients with undiagnosed CDI/asymptomatic colonization could link the cases. For each such case we identified the most recent prior genetically-related case. We calculated the number of CDI-free hospital contacts the two cases shared between their diagnoses. The number of intermediate contacts between pairs of genetically distant cases (>10SNVs, randomly chosen, matched on the time between samples) was also calculated to estimate the background proportion of unrelated cases that shared intermediate patients.

Impact of recombination

We conducted a sensitivity analysis to assess the extent to which observed SNVs between samples could be accounted for by recombination rather than mutation. We took each of the groups of samples within 100 SNVs used in the PhyML analysis and obtained the pairwise SNV difference between all possible pairs of samples within the groups. For each pair of sequences, we estimated a recombination adjusted SNV difference.

To adjust the number of pairwise SNP differences (the *pairwise differences*) for the effect of homologous recombination, we used the model underlying ClonalFrame.[12] In the ClonalFrame model, mutation occurs at rate θ_S per nucleotide and recombination is initiated at rate ρ_S per nucleotide. When recombination occurs, the length of homologous DNA imported is assumed to follow a geometric distribution with mean δ nucleotides. The nucleotide divergence of the imported DNA is specified by the parameter ν.

We wished to estimate the number of pairwise differences attributable to mutation. To do so, we first estimated T, the length of the branch separating the two genomes in a two-taxa tree, measured in units of the expected number of point mutations per nucleotide. We then defined the adjusted pairwise difference to be TG, where G is the length of the genome.

To estimate T in a computationally efficient manner, we used the ClonalFrame model, which can be thought of as a hidden Markov model (HMM, see, e.g. [13]) when there are only two genomes. The hidden state of the HMM records whether each nucleotide was subject to recombination or not on the branch connecting the two genomes. Nucleotides unaffected by recombination are said to be *unimported* and nucleotides subject to recombination are said to be *imported*.[12] Based on the ClonalFrame model, we defined the following transition probability matrix for the hidden variable between adjacent sites, H_i and H_{i+1}:

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

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

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

We used the forward algorithm[13] to calculate the likelihood $\Pr\left(D \mid T, \rho_s/\theta_s, \delta, \nu\right)$. We estimated *T* using maximum likelihood, implemented as a unidirectional optimization routine, taking as the values of the auxiliary parameters $\left\{\rho_s/\theta_s, \delta, \nu\right\}$ the point estimates from the full ClonalFrame analyses, which have previously been performed separately for 15 *C. difficile* sequence types.[14] Where parameter estimates were not available for the sequence types of a pair of sequences, parameter values from the genetically closest sequence type with estimates available were used.

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

Supplementary Results

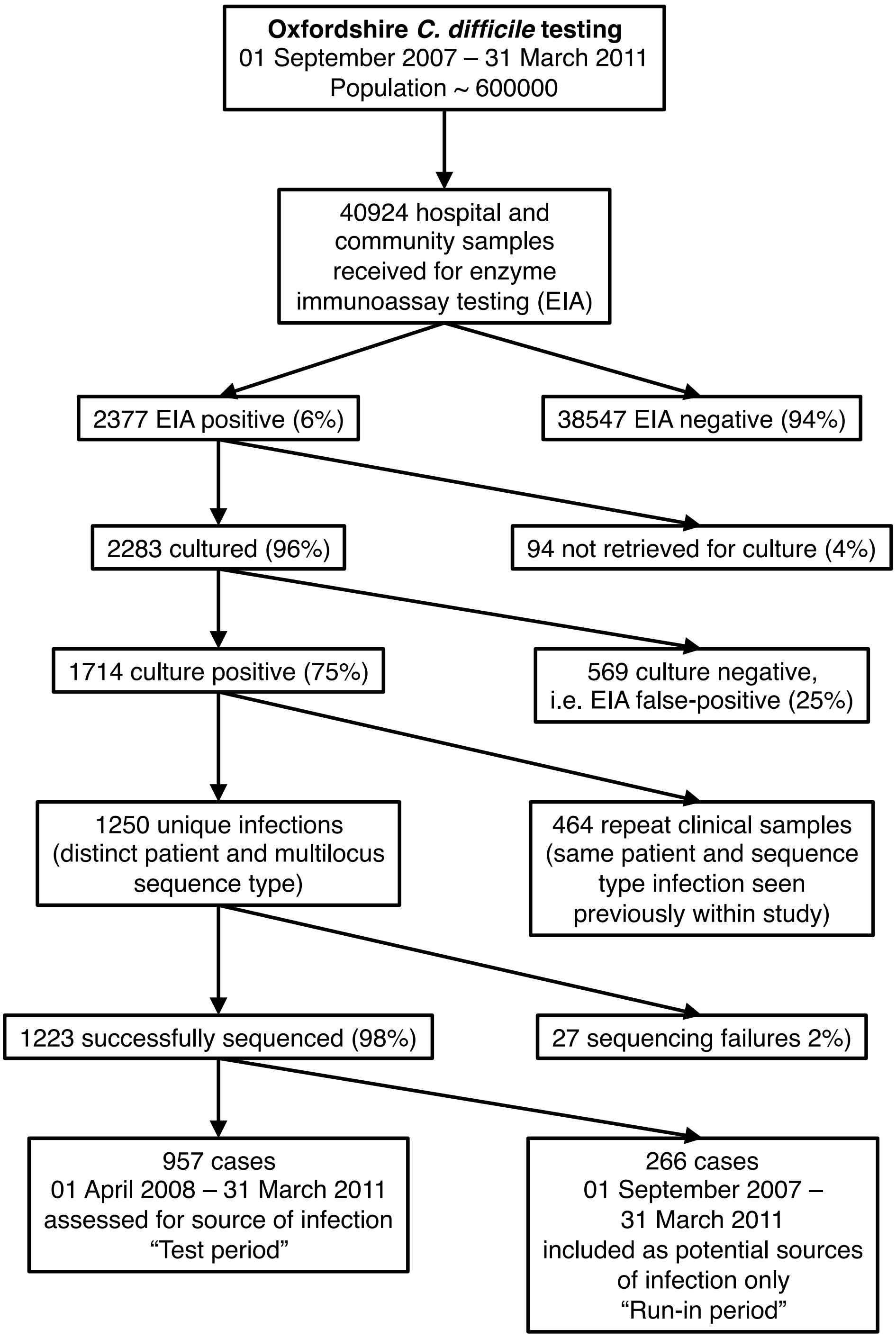
Oxfordshire *C. difficile* testing
01 September 2007 – 31 March 2011
Population ~ 600000
40924 hospital and community samples received for enzyme immunoassay testing (EIA)
2377 EIA positive (6%)
38547 EIA negative (94%)
2283 cultured (96%)
94 not retrieved for culture (4%)
1714 culture positive (75%)
569 culture negative, i.e. EIA false-positive (25%)
1250 unique infections (distinct patient and multilocus sequence type)
464 repeat clinical samples (same patient and sequence type infection seen previously within study)
1223 successfully sequenced (98%)
27 sequencing failures 2%)
957 cases 01 April 2008 – 31 March 2011 assessed for source of infection “Test period”
266 cases 01 September 2007 – 31 March 2011 included as potential sources of infection only “Run-in period”

Figure S2. Samples and patients, 01 September 2007 to 31 March 2011.

***C. difficile* evolution and within-host diversity**

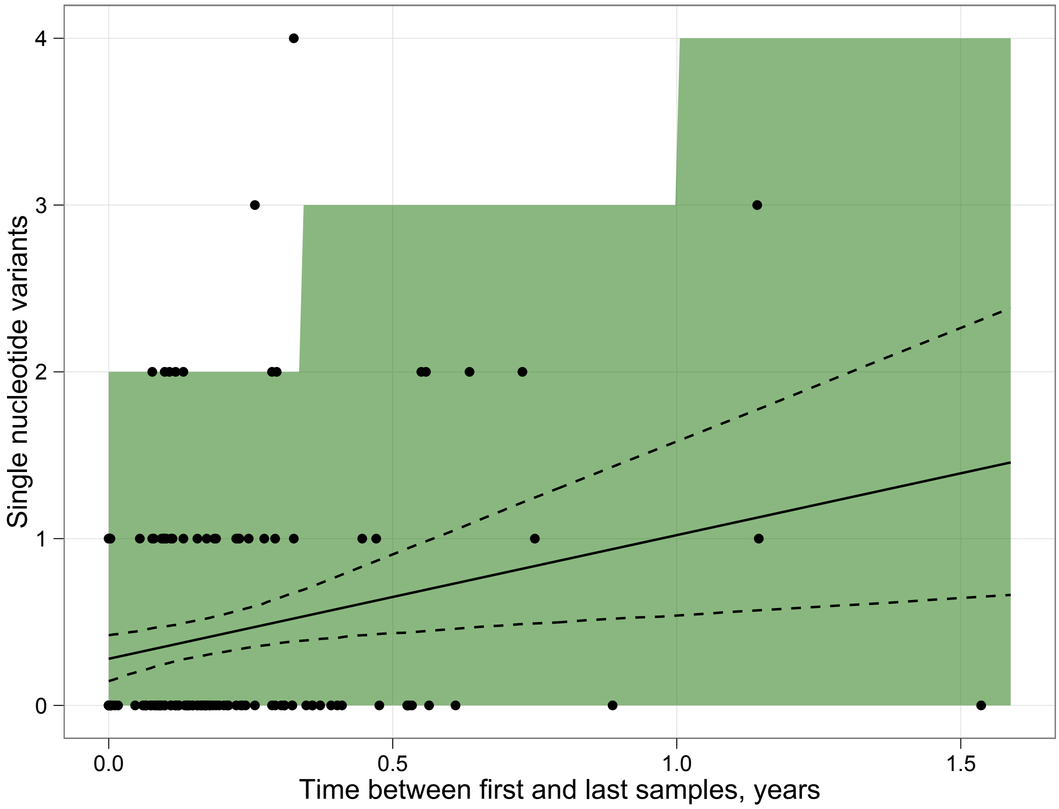

Figure S3. *C. difficile* microevolution in 145 serially sampled patients. The number of single nucleotide variants (SNVs) observed during 145 *C. difficile* infections is plotted against the time between the first and last sample sequenced from each patient. The line shows a maximum likelihood fitted coalescent theory-based model for the rate of evolution (95% confidence interval, dashed line). The slope of the line reflects the rate of evolution and the non-zero intercept the expected number of SNVs from within-host diversity. The green shaded area contains the 95% prediction interval from the model, i.e. the area within which 95% of observations are expected to lie after accounting for chance (assuming SNVs arise as a Poisson process). Although the evolutionary rate point estimate was lower in ST1/ribotype-027 versus non-ST1s, the difference did not reach statistical significance (p=0.23). The relatively limited within-host

diversity and low rate of evolution compared to the population diversity (Figure 1, Figure 3) underlies the discriminatory power of whole genome sequencing.

Impact of recombination

Across the groupings of samples within 100 SNVs, only 66/42805 pairs (0.15%) had >2 observed SNVs and ≤2 adjusted variants (i.e. variants due to mutation excluding recombination), suggesting relatively few very closely related sequences are erroneously described as more distant in the main analysis. Similarly, 27817/30057 (93%) of pairs >10 SNVs apart remained more than 10 variants different after adjustment. The total number of adjusted variants for each pair of sequences is shown in table S2 and figure S4.

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

Table S2. Comparison of SNV differences between pairs of samples, with and without adjustment for recombination.

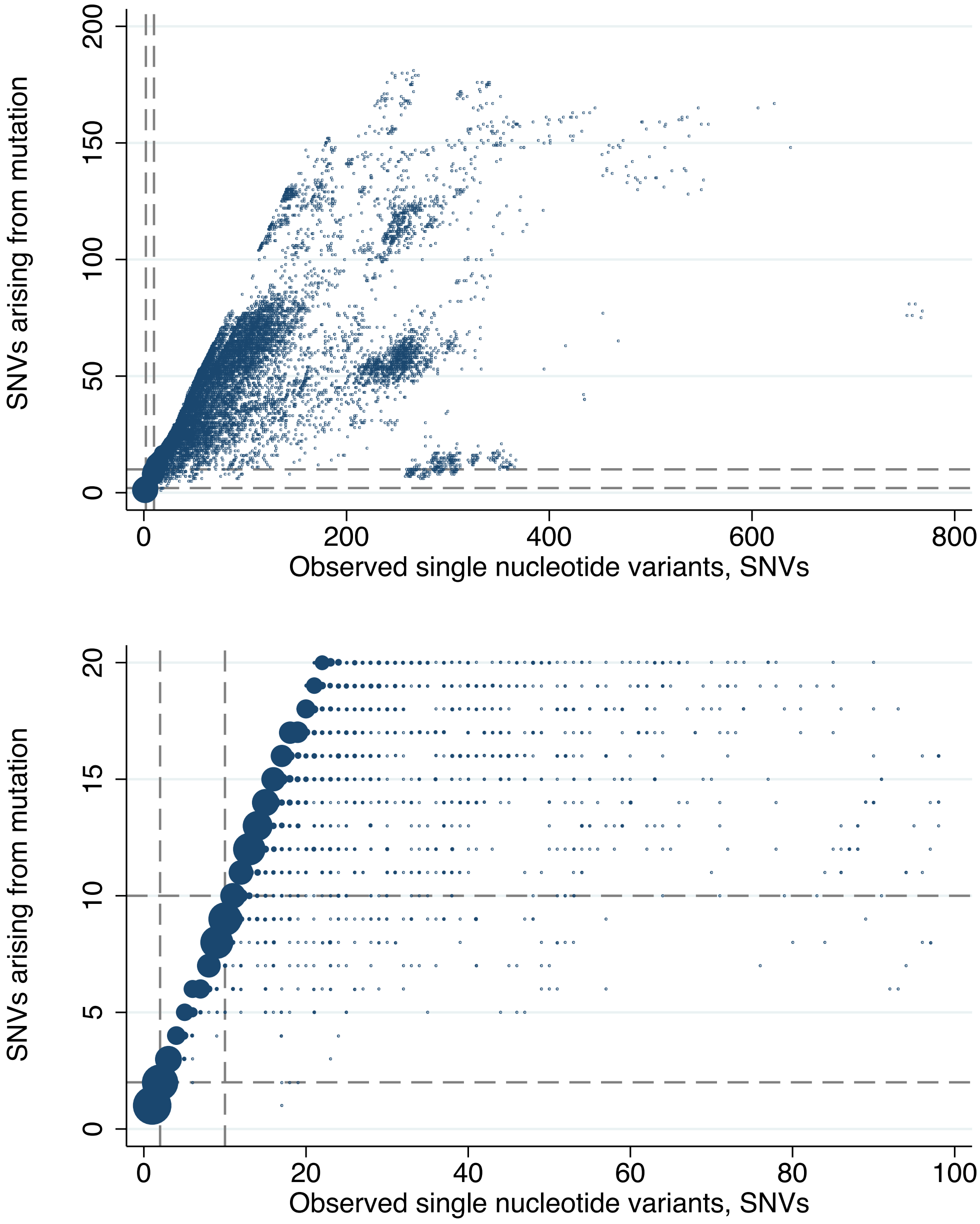

Figure S4. Comparison of SNV differences between pairs of samples, with and without adjustment for recombination. The number of variants due to mutation after adjusting for recombination is plotted against the number of observed SNVs between samples. Markers are weighted by the number of pairs of samples with each value. Panel A shows all pairs, for ease of visualization excluding 3 pairs with >1000 observed SNVs and recombination adjusted

variants of 46, 47, 49. Panel B shows the same plot in more detail for lower observed and adjusted SNV values.

Genetic diversity within Oxfordshire cases

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

Table S3. Sensitivity Analysis: Epidemiological relationships between each of 957 CDI cases, 01 April 2008 to 31 March 2011, and the most genetically similar previous case. For each case the nature of the closest epidemiological links with all previous cases within a given SNV limit is shown. The duration of infectious/incubation/ward contamination periods were unlimited in this

sensitivity analysis. Cases sharing time and space on the same hospital ward are denoted “ward-contact”. Cases sharing a common ward, between discharge of the first case (or end of infectivity of the first case) and admission of the second, and no other form of shared hospital exposure are denoted “ward contamination only”. Cases present at the same time in the same hospital, but not sharing the same ward, are denoted “shared specialty only” if they shared a common specialty, and “other hospital-wide only” if not. Where patients have a mix of “ward contamination contact” and “hospital-wide contact”, but without a direct “ward-contact” this is denoted “mixed spore and hospital-wide”. Community links are made when two cases share the same general practice or where two cases share the same home postcode district. Cases without any form of shared hospital exposure or community contact as defined above are denoted “no known epidemiological link”.

Epidemiologically unexplained genetically-related cases

		0 SNVs	0-2 SNVs	0-10 SNVs
Possible explanations for cases without a clear epidemiological link	**No known epidemiological link** *within hospital exposure limits in Table 1*	**73**	**120**	**242**
	Samples obtained within ≤7 days *Possible point source / Laboratory contamination*	7 (10%)	11 (9%)	18 (7%)
	EIA negative third party contact in common *(Samples >7 days apart)*	19 (26%)	32 (27%)	43 (18%)
	Other inpatient contact in common *(No EIA negative contact, samples >7 days apart)*	3 (4%)	4 (3%)	7 (3%)
	Ward contamination if assumed to persist infinitely *(and none of above)*	9 (12%)	13 (11%)	44 (18%)
	Hospital contact with another symptomatic case if CDI assumed to be indefinitely infectious, and incubate for an unlimited period *(and none of the above)*	6 (8%)	12 (10%)	24 (10%)
	Genetically related cases remaining without any plausible epidemiological link found	**29 (40%)**	**48 (40%)**	**106 (44%)**

Table S4. Potential relationships between each new CDI, 01 April 2008 to 31 March 2011 and the most genetically similar previous cases where no clear hospital/community link found.

Temporal patterns in diversity

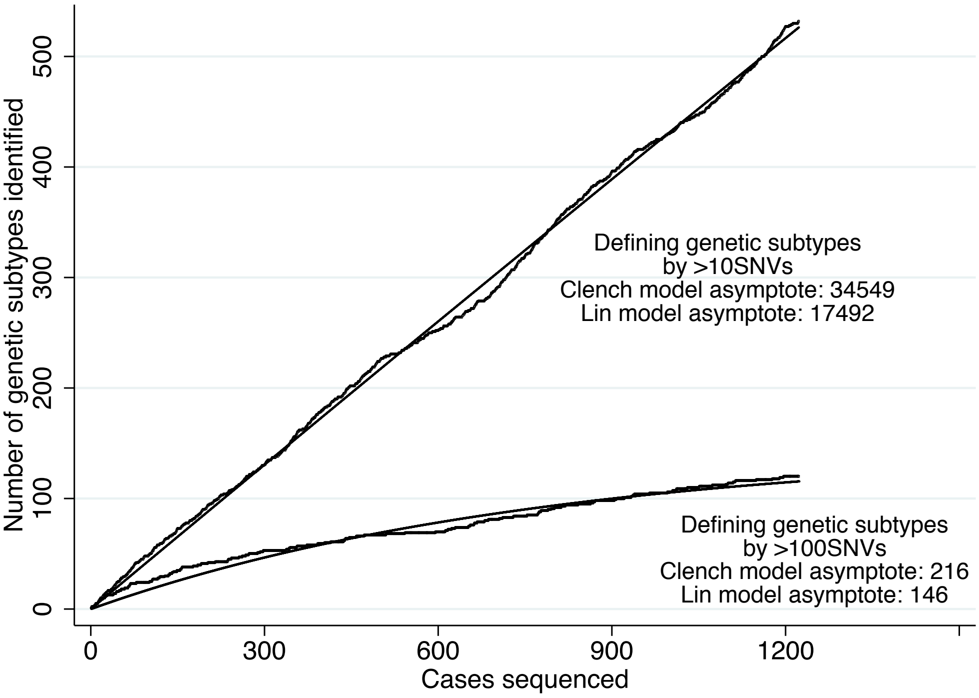

Figure S5. Timing and size of *C. difficile* genetic clusters, total population diversity models. The number of distinct genetic types of *C. difficile* identified for a given sampling effort over calendar time is plotted for 2 different SNV thresholds. Two models estimating the total population diversity are plotted shown in bold. (See supplementary materials for details of models).

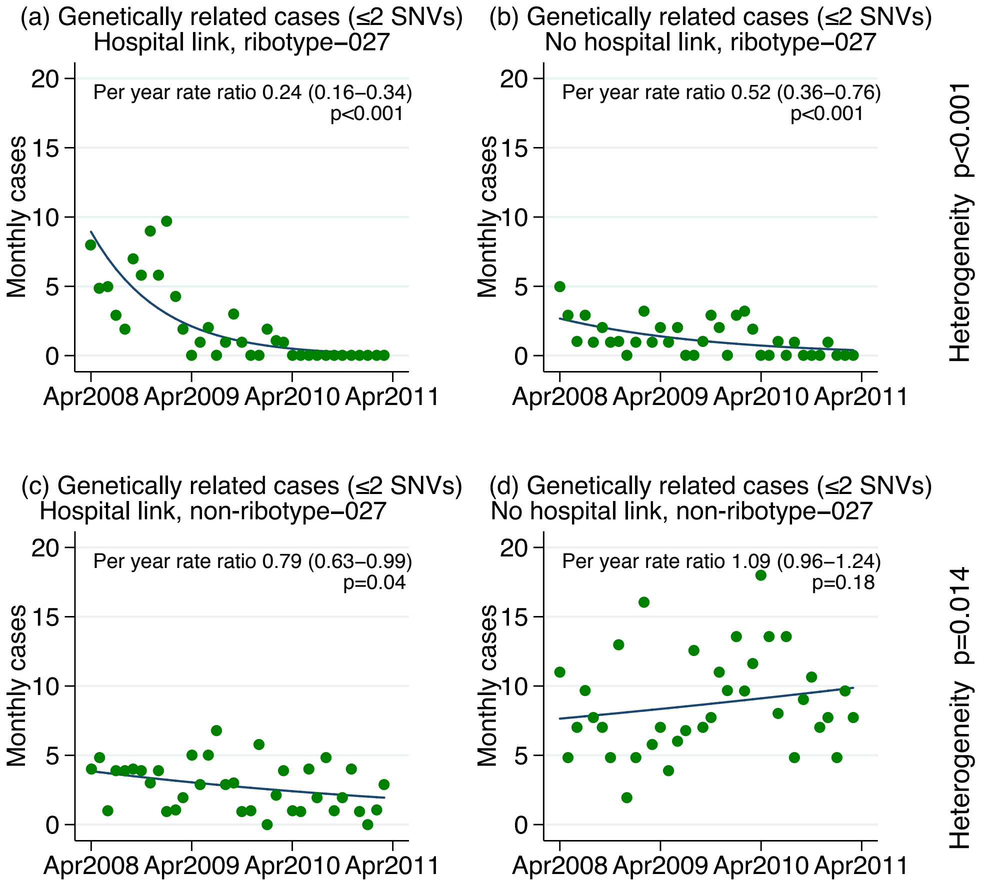

Figure S6. Secondary cases in the Oxfordshire symptomatic *C. difficile* population by ribotype-027 / non-ribotype-027. CDI cases caused by an isolate that was >10 SNVs apart from a previous case since 01 September 2007 (novel genetic subtypes) were denoted "introductions" to the symptomatic Oxfordshire population, and cases ≤2 SNVs as possible "secondary cases". Only cases from 01 April 2008 are shown as the first seven months of the study were denoted a run-in period to capture the increased rate of novel subtypes expected initially; thus, temporal patterns are analyzed over the final 3 years of the study. The study population during the study was 600000, therefore rates of 20 cases per month correspond to 3.3 per 100000 population per month. Per annum rate

ratio (change in incidence), determined by Poisson regression. Heterogeneity p values were determined by stacked Poisson regression. Plotted monthly cases are scaled 30 times the mean daily cases in the month, to allow for accurate comparison across months.

Data sharing

The sequences reported in this paper have been deposited in the European Nucleotide Archive Sequence Read Archive under study accession number PRJEB4556 and are available at http://www.ebi.ac.uk/ena/data/view/PRJEB4556.

References

1. Walker AS, Eyre DW, Wyllie DH, et al. Characterisation of Clostridium difficile hospital ward-based transmission using extensive epidemiological data and molecular typing. PLoS Med 2012;9(2):e1001172:1–12.

2. Lunter G, Goodson M. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 2011;21(6):936–9.

3. Sebaihia M, Wren BW, Mullany P, et al. The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome. Nat Genet 2006;38(7):779–86.

4. Dingle KE, Griffiths D, Didelot X, et al. Clinical Clostridium difficile: clonality and pathogenicity locus diversity. PLoS ONE 2011;6(5):e19993.

5. Stabler RA, He M, Dawson L, et al. Comparative genome and phenotypic analysis of Clostridium difficile 027 strains provides insight into the evolution of a hypervirulent bacterium. Genome Biol 2009;10(9):R102.

6. Eyre DW, Walker AS, Griffiths D, et al. Clostridium difficile Mixed Infection and Reinfection. J Clin Microbiol 2012;50(1):142–4.

7. Rodrigo AG, Felsenstein J. Coalescent approaches to HIV population genetics. In: Crandall KA, editor. Evolution of HIV. Baltimore, MD: Johns Hopkins University Press; 1999. p. 233–72.

8. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 2003;52(5):696–704.

9. Soberón J, Llorente J. The use of species accumulation functions for the prediction of species richness. Conserv Biol 1993;7:480–8.

10. Thompson GG, Withers PC, Pianka ER, Thompson SA. Assessing biodiversity with species accumulation curves; inventories of small reptiles by pit-trapping in Western Australia. Austral Ecology 2003;28(4):361–83.

11. Clench H. How to make regional list of butterflies: Some thoughts. J Lepid Soc 1979;33:216–31.

12. Didelot X, Falush D. Inference of bacterial microevolution using multilocus sequence data. Genetics 2007;175(3):1251–66.

13. Durbin R, Eddy SR, Krogh A, Mitchison G. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge: Cambridge University Press; 1998.

14. Didelot X, Eyre DW, Cule ML, et al. Microevolutionary analysis of Clostridium difficile genomes to investigate transmission. Genome Biol 2012;13(12):R118.