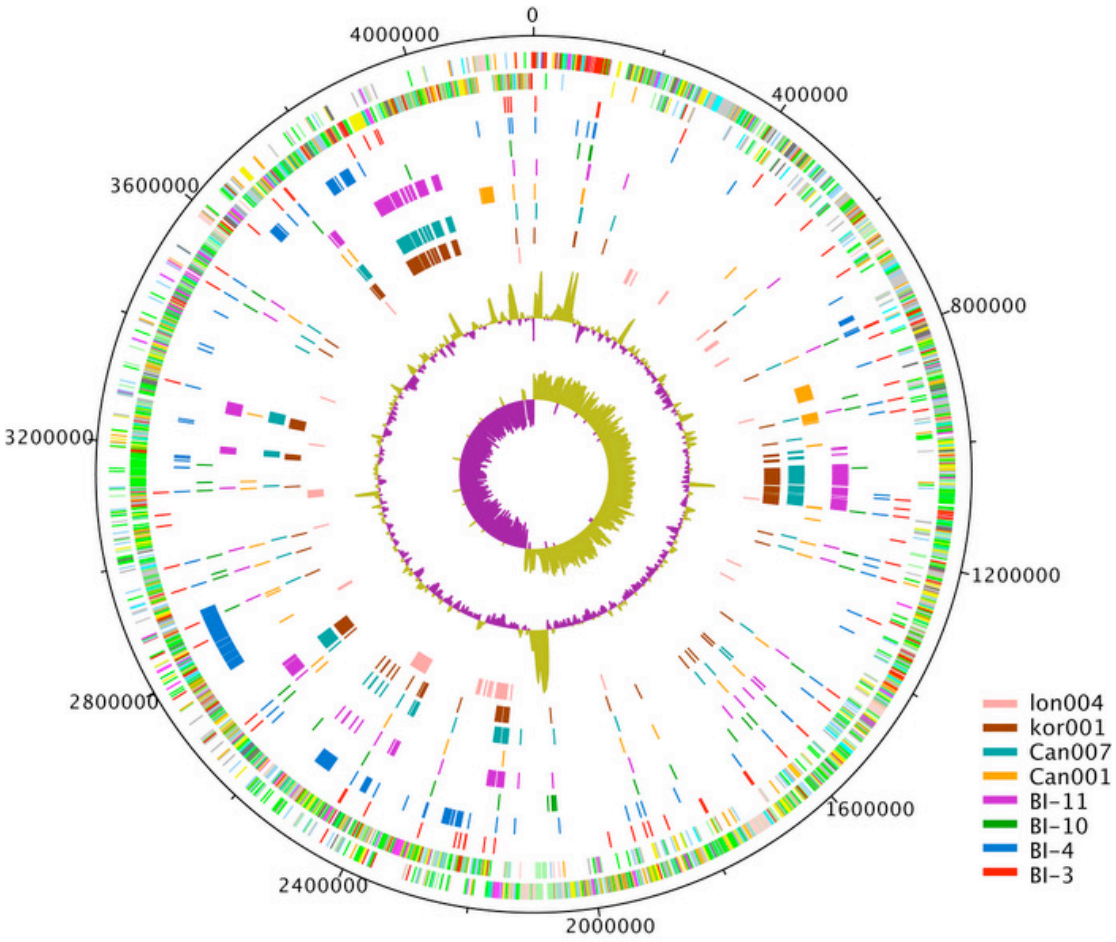
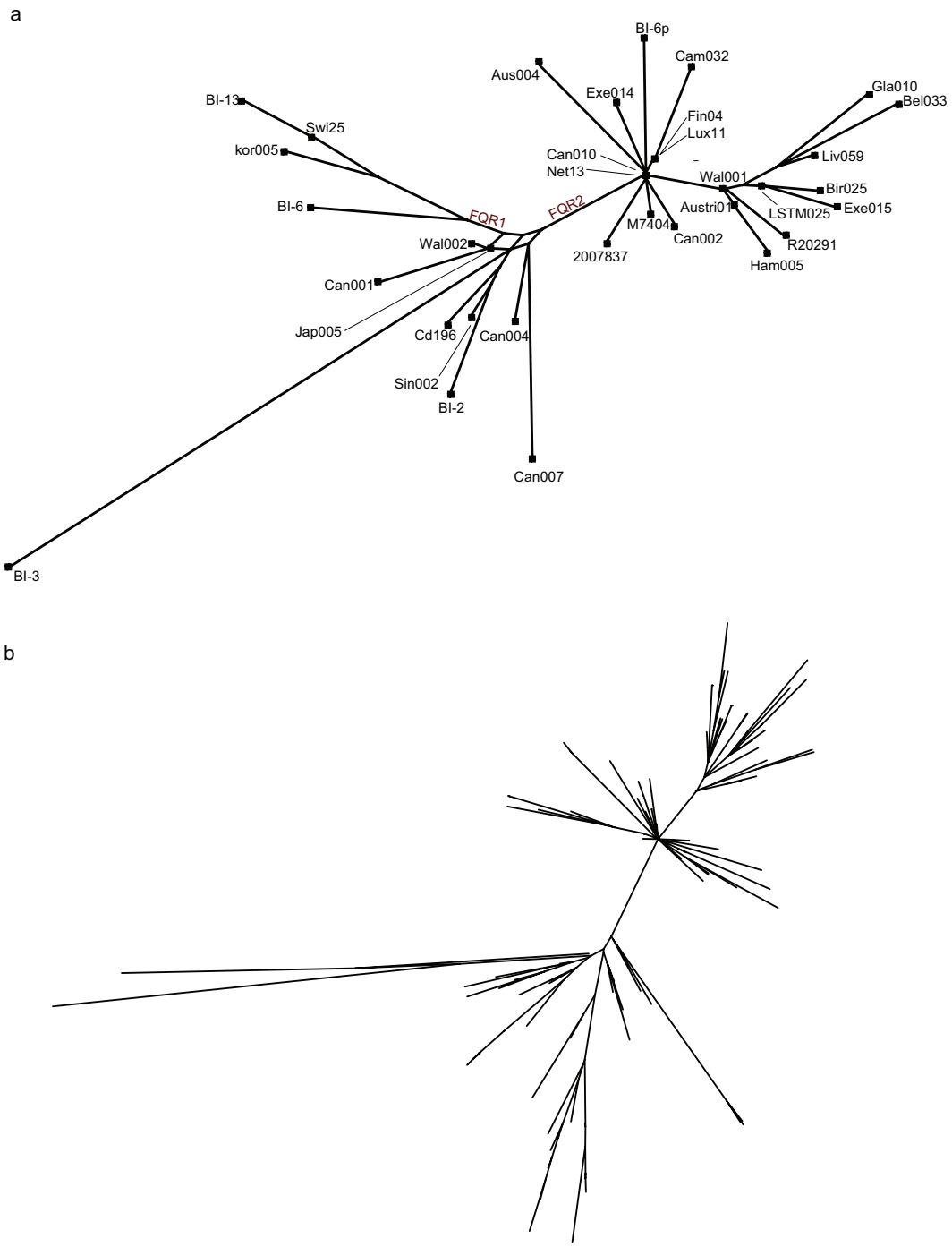
Supplementary Information

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

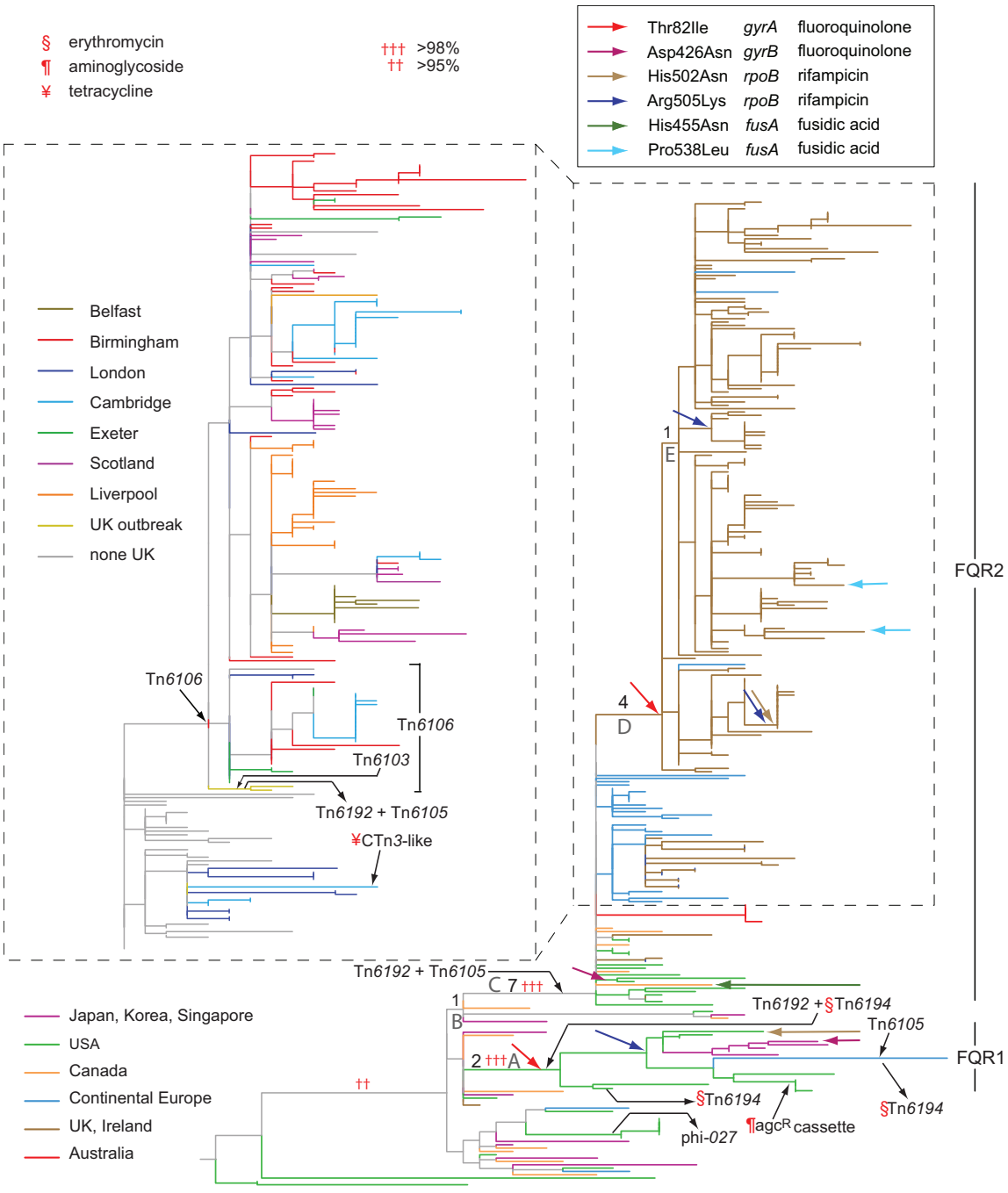
Miao He, Fabio Miyajima, Paul Roberts, Louise Ellison, Derek J. Pickard, Melissa J. Martin, Thomas R. Connor, Simon R. Harris, Derek Fairley, Kathleen B. Bamford, Stephanie D'Arc, Jon Brazier, Derek Brown, John E. Coia, Gill Douce, Dale Gerding, Hee Jung Kim, Tse Hsien Koh, Haru Kato, Mitsutoshi Senoh, Tom Louie, Stephen Michell, Emma Butt, Sharon J. Peacock, Nick M. Brown, Tom Riley, Glen Songer, Mark Wilcox, Munir Pirmohamed, Ed Kuijper, Peter Hawkey, Brendan W. Wren, Gordon Dougan, Julian Parkhill, Trevor D. Lawley



Supplementary Figure 1. Evidence of homologous recombination within eight *C. difficile* 027/BI/NAP1 isolates. Outer circle: Coding sequences of R20291 genome, shown on a pair of concentric rings representing both coding strands; two inner circles: G+C% content plot and GC deviation plot (>0% olive, <0% purple); in between: SNPs between R20291 and eight isolates (from outer to inner: BI-3, BI-4, BI-10, BI-11, Can001, Can007, kor001, lon004), colored according to legend.

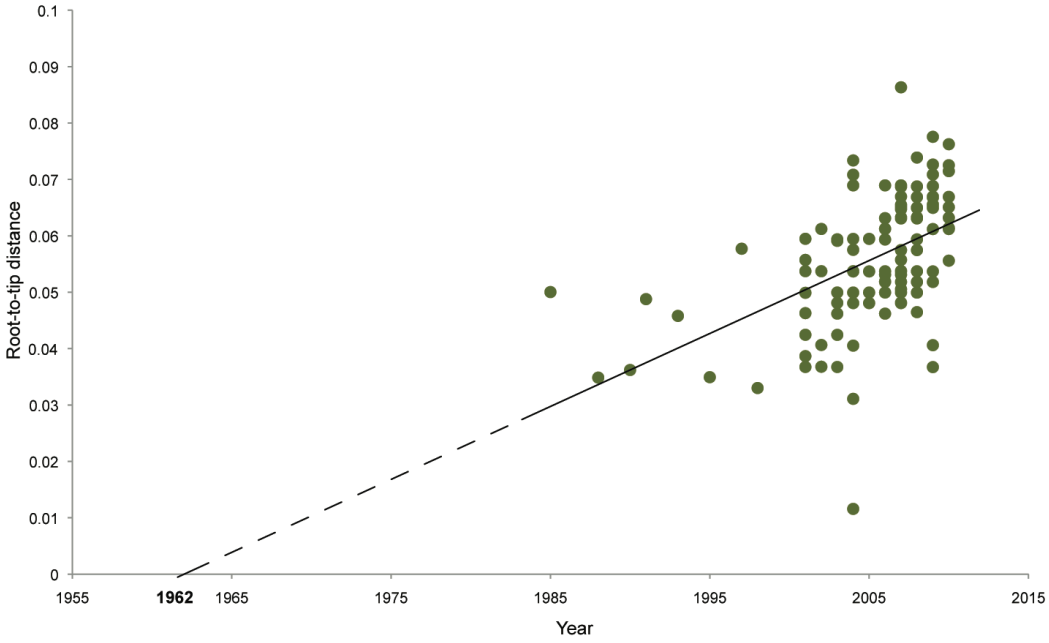


Supplementary Figure 2. Independent methods support robust global phylogeny of *C. difficile* 027/BI/NAP1. Phylogeny of *C. difficile* 027/BI/NAP1 isolates was inferred with split-decomposition algorithm (a) and neighbor-joining method (b). Numbers of isolates included are 34 (a) and 151 (b).

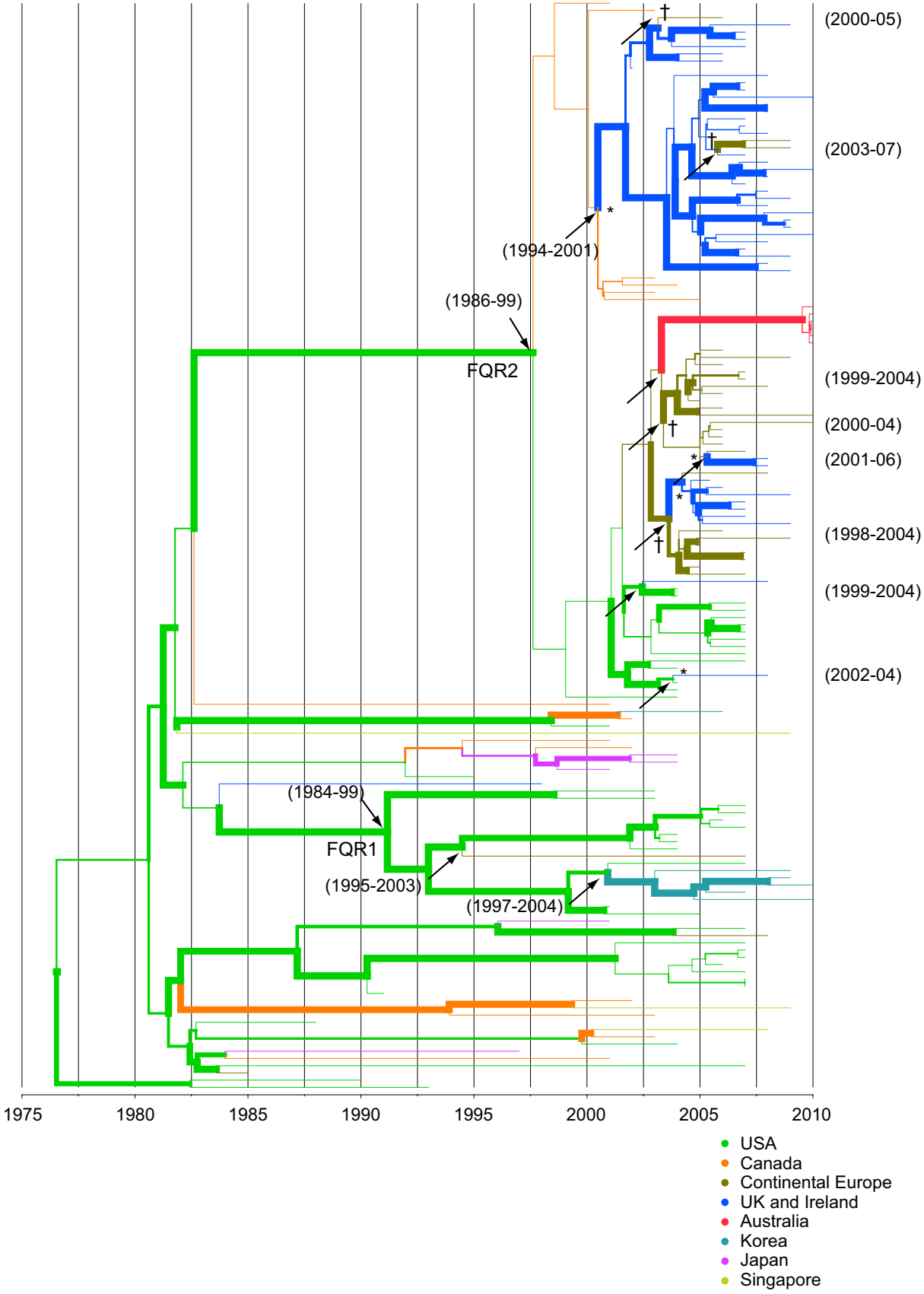


Supplementary Figure 3. Maximum-likelihood phylogeny of global and UK *C. difficile* 027/BI/NAP1 collection (n=296) based on core genome SNPs. Branches colors indicate the geographical sources of the isolates. An insert is shown to illustrate geographical sources of the UK isolates. Cross symbols give levels of support (in bootstrap values) for main branches. The numbers near the branches denote numbers of SNPs defining the main branches A-E. Arrows in colors indicate the isolates with

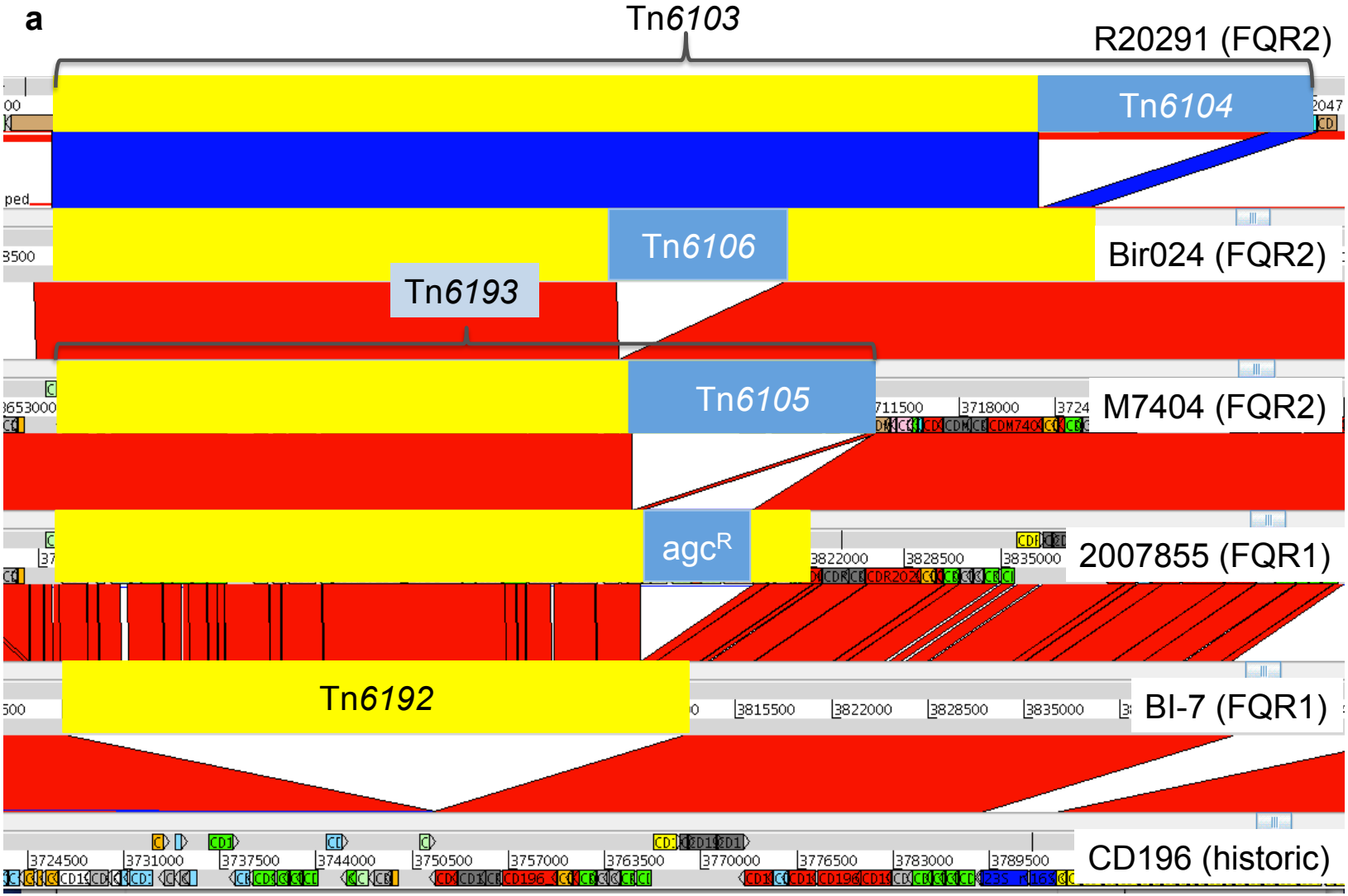
homoplasic SNPs associated with antibiotic resistance. Black arrows show insertion (towards the phylogeny) or deletion (away from the phylogeny) of selected mobile elements. The sub-lineage harboring Tn*6106* is depicted with a square bracket. Mobile elements carrying antibiotic resistance determinants are labeled with red symbols.

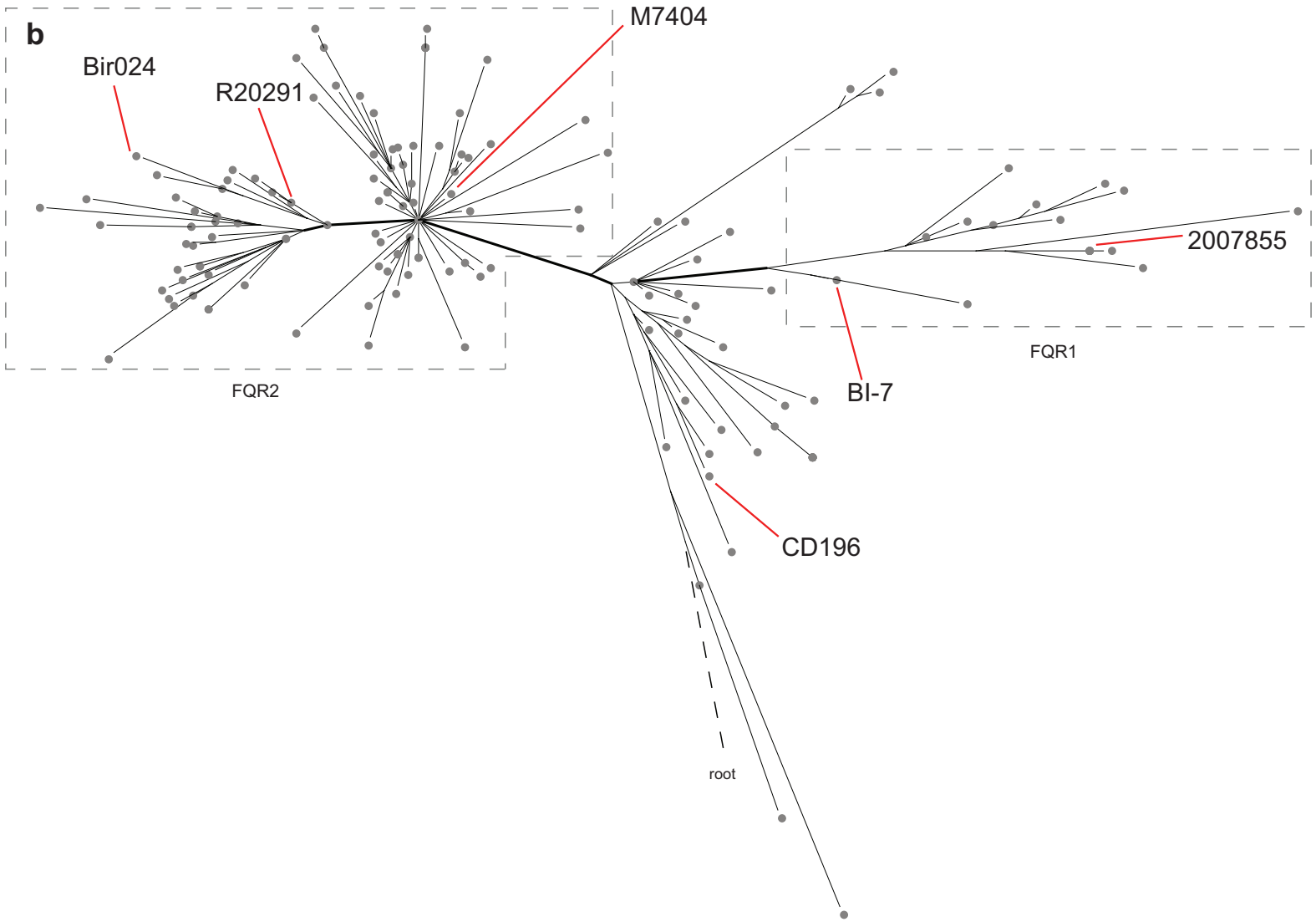


Supplementary Figure 4. Regression analysis between strain isolation dates (x-axis) and root-to-tip distance (y-axis). The analysis was conducted based on *C. difficile* 027/BI/NAP1 global collection (151 isolates). The point where the dashed line intersects with the x-axis gives the inferred date when the most recent common ancestor of *C. difficile* 027/BI/NAP1 emerged. R-square = 0.25. The weak correlation in this analysis is associated with the spore-forming lifestyle of *C. difficile*.

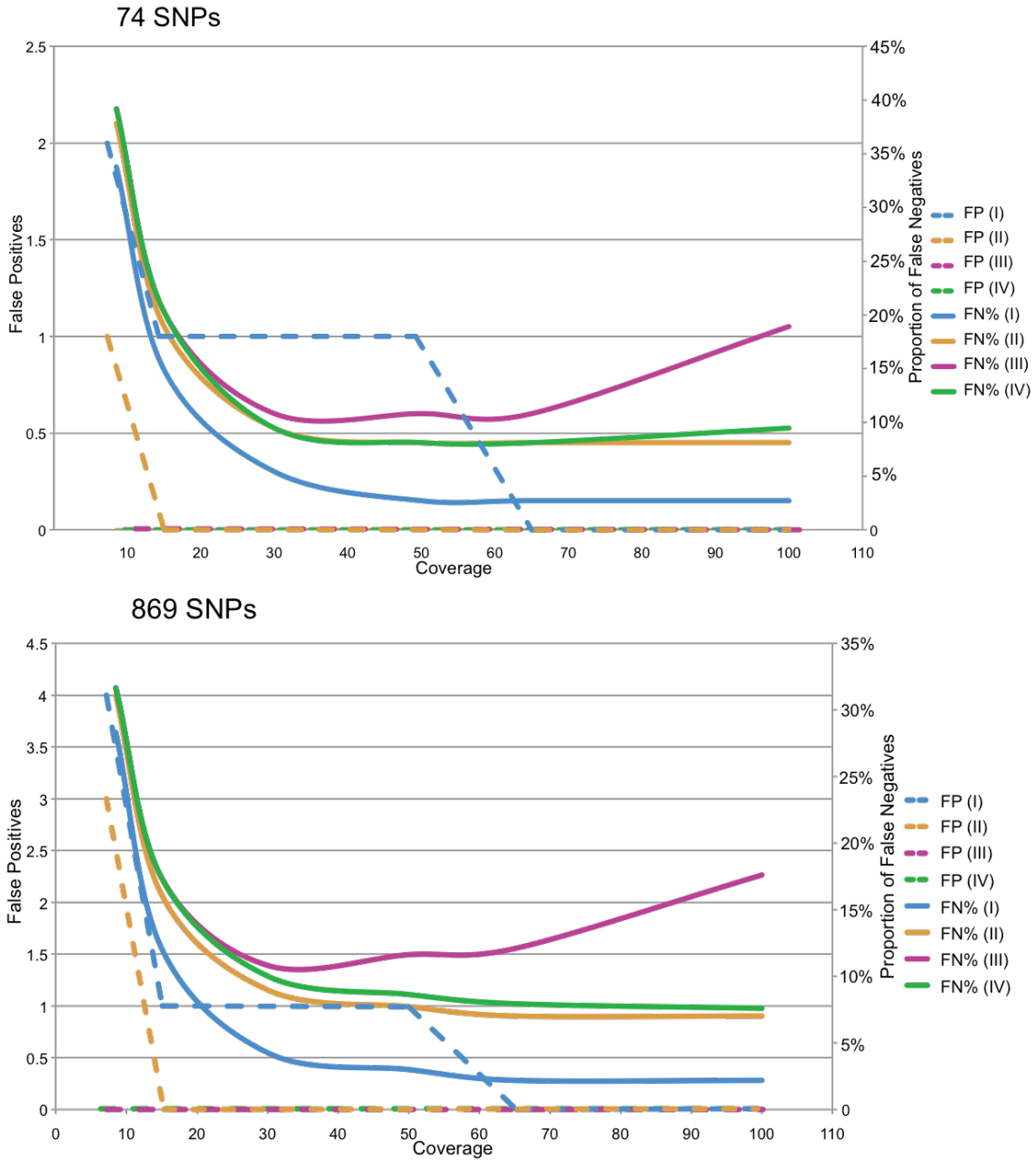

(2000-05)
(2003-07)
(1994-2001)
(1986-99)
FQR2
(1999-2004)
(2000-04)
(2001-06)
(1998-2004)
(1999-2004)
(2002-04)
(1984-99)
FQR1
(1995-2003)
(1997-2004)
1975
1980
1985
1990
1995
2000
2005
2010
USA
Canada
Continental Europe
UK and Ireland
Australia
Korea
Japan
Singapore

Supplementary Figure 5. Bayesian phylogeny of *C. difficile* 027/BI/NAP1 global collection with inferred geographic ancestry. Branches are colored according to actual or most probable origins of isolates. Line thickness of internal branches is proportional of the support (posterior probability) they receive. Arrow indicates inferred time to the most recent common ancestors (tMRCAs) of lineages or dates of major transmission events in 95% highest posterior density (HPD) intervals. Transmission events into the UK and continental Europe are labeled with asterisks and crosses respectively.


a
Tn6103
R20291 (FQR2)
Tn6104
Tn6106
Bir024 (FQR2)
Tn6193
Tn6105
M7404 (FQR2)
agcR
2007855 (FQR1)
Tn6192
BI-7 (FQR1)
CD196 (historic)

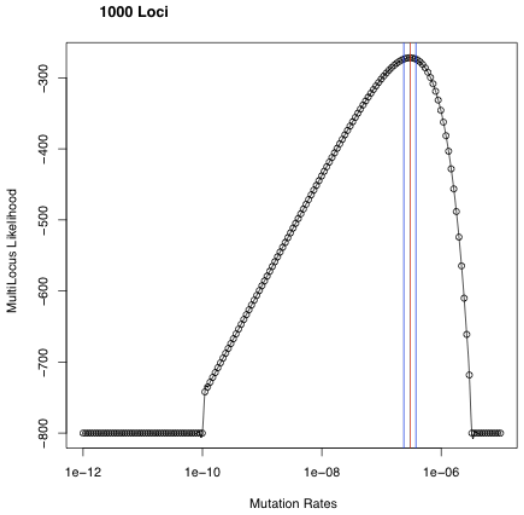

b
Bir024
R20291
M7404
2007855
FQR2
FQR1
BI-7
CD196
root

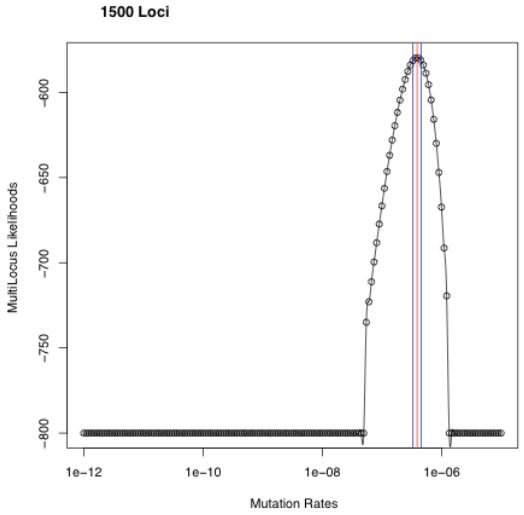
Supplementary Figure 6. Presence of five variants of CTn5-like element in the *C. difficile* 027/BI/NAP1 phylogeny. a) comparisons between CTn5-like elements in six representative *C. difficile* 027/BI/NAP1 genomes. The position and boundaries of each CTn5-like element or subsequent insertion are depicted by yellow or blue boxes, onto which the name of the element is labeled. The genome sequence is represented by a pair of thin grey rectangles (signifying both strands) and small colored boxes (coding sequences). The names of the isolates are given on the right, followed by brackets indicating lineage (FQR1, FQR2 or historic). Matching areas between sequences depict nucleotide similarity on the forward strand (red) or reverse strand (blue). b) positions of these six *C. difficile* 027/BI/NAP1 isolates (shown by red lines) in the same global phylogeny as Fig. 1a.

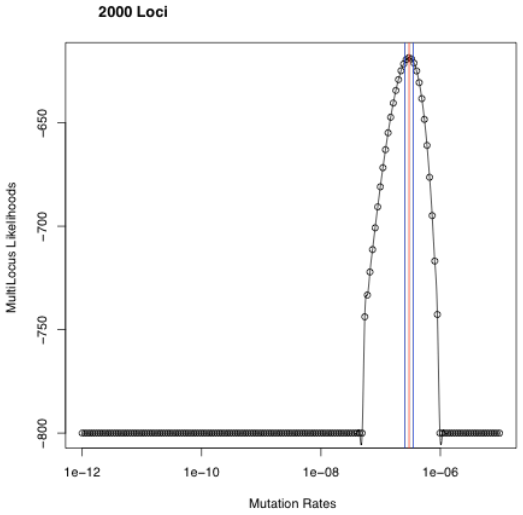


Supplementary Figure 7. Assessment of SNP detection accuracy. Numbers of false positive SNPs (dashed lines, left axis) and percentage of false negative SNPs (solid lines, right axis) are plotted in relation to sequencing data coverage (x-axis) and different SNP filtering and validation measures (colored lines). Scale 1 (top) and scale 2 (bottom) indicate two scenarios with different levels of divergence, as shown by the numbers of SNPs on top of each graph. SNP filtering and validation measures I – IV correspond to what stated in Supplementary Note.

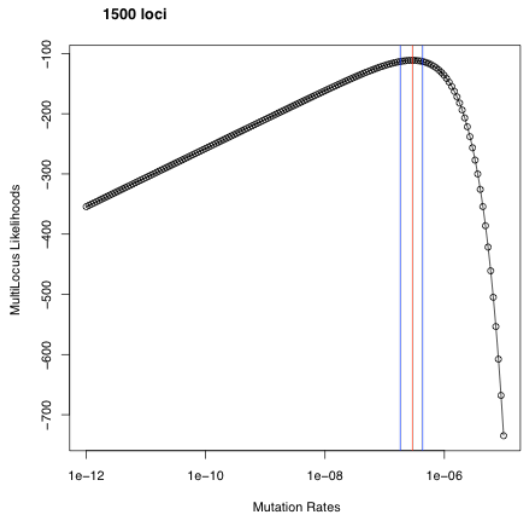
61

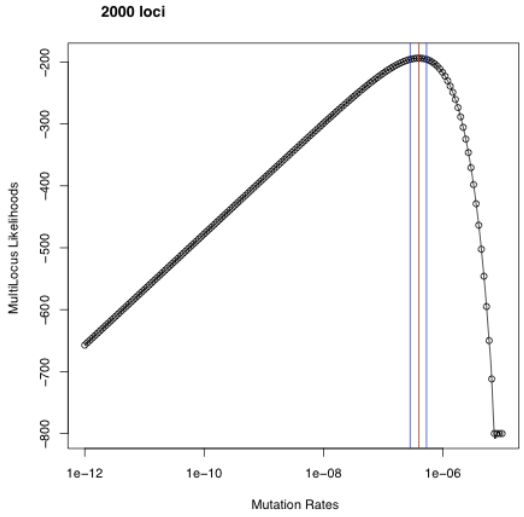

1000 Loci
MultiLocus Likelihood
−300
−400
−500
−600
−700
−800
1e−12
1e−10
1e−08
1e−06
Mutation Rates

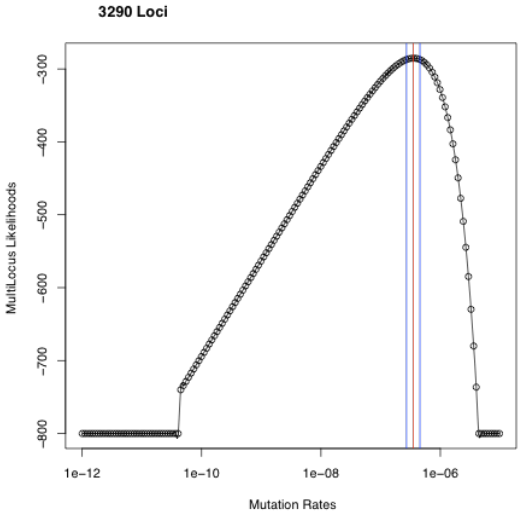

1500 Loci
MultiLocus Likelihoods
−600
−650
−700
−750
−800
1e−12
1e−10
1e−08
1e−06
Mutation Rates


2000 Loci
MultiLocus Likelihoods
−650
−700
−750
−800
1e−12
1e−10
1e−08
1e−06
Mutation Rates

18


1500 loci
MultiLocus Likelihoods
−100
−200
−300
−400
−500
−600
−700
1e−12
1e−10
1e−08
1e−06
Mutation Rates


2000 loci
MultiLocus Likelihoods
−200
−300
−400
−500
−600
−700
−800
1e−12
1e−10
1e−08
1e−06
Mutation Rates


3290 Loci
MultiLocus Likelihoods
−300
−400
−500
−600
−700
−800
1e−12
1e−10
1e−08
1e−06
Mutation Rates

Supplementary Figure 8. Likelihood surfaces with exact binomial computation[1] of FQR2 expansion mutation rate estimates. Mutation rates were estimated using two collections consisting of either 61 (upper graphs) or 18 (lower graphs) isolates in the star-like phylogeny in the FQR2 lineage. The number of coding sequences used in each calculation is labeled on top of each graph.

Supplementary Table 1. Details of 151 *C. difficile* 027/BI/NAP1 isolates forming the global collection. Pink and blue shaded areas denote isolates from FQR1 and FQR2 lineages respectively. The un-colored boxes represent isolates forming the background from which FQR lineages emerged. Non-human isolates are indicated by strain names followed by "^" (animal sources) "*" (food sources).

Strain Name	Isolation Year	Isolation Place	Other Name	Accession Number
Aus001	2010	Melbourne, Australia	ES231	ERS017078
Aus002	2010	Melbourne, Australia	ES232	ERS017091
Aus003	2010	Melbourne, Australia	ES233	ERS017092
Aus004	2010	Melbourne, Australia	ES235	ERS017093
Aus005	2010	Melbourne, Australia	ES236	ERS017094
Aus006	2010	Melbourne, Australia	ES238	ERS017095
BI-17	2004	Montreal, Canada	LSTM024	ERS017255
Can001	2001	Calgary, Canada	OB023	ERS017144
Can002	2001	Calgary, Canada	AO992	ERS017145
Can003	2001	Calgary, Canada	OB102	ERS017146
Can004	2001	Calgary, Canada	AO315	ERS017147
Can005	2002	Calgary, Canada	AO060	ERS017148
Can006	2002	Calgary, Canada	AO281	ERS017149
Can007	2002	Calgary, Canada	AO224	ERS017150
Can008	2003	Calgary, Canada	AO601	ERS017151
Can009	2003	Montreal, Canada	MTL047	ERS017275
Can010	2003	Montreal, Canada	MTL057	ERS017153
Can011	2003	Calgary, Canada	AO837	ERS017154
Can012	2003	Montreal, Canada	MTL077	ERS017155
M7404	2005	Montreal, Canada		ERS017219
Lei001	2006	Austria		ERS032926
Lei002	2005	Belgium		ERS032927
Lei003	2006	Belgium		ERS032928
Lei004	2009	Finland		ERS032929

Lei006	2007	France		ERS032930
Lei007	2006	Germany		ERS032931
Lei008	2008	Germany		ERS032932
Lei011	2007	Luxemburg		ERS032935
Lei012	2007	Luxemburg		ERS032936
Lei013	2006	Netherlands		ERS032938
Lei014	2006	Netherlands		ERS032939
Lei015	2006	Netherlands		ERS032940
Lei016	2006	Netherlands		ERS032941
Lei017	2005	Netherlands		ERS032942
Lei018	2006	Netherlands		ERS032943
Lei019	2006	Netherlands		ERS032944
Lei020	2007	Netherlands		ERS032945
Lei021	2007	Norway		ERS032946
Lei022	2009	Poland		ERS032947
Lei023	2010	Poland		ERS032948
Lei024	2008	Switzerland		ERS032949
Lei025	2007	Switzerland		ERS032950
Lei028	2008	Netherlands		ERS032953
Lei029	2009	Netherlands		ERS032954
Lei030	2010	Netherlands		ERS032955
Lei031	2007	Netherlands		ERS032956
Lei032	2005	Netherlands		ERS032957
CD196	1985	France		ERS017271/FN538970
Lei036	2009	Poland		ERS032961
Jap001	1997	Japan	JND-6-119	ERS032963
Jap002	2001	Japan	DJNS 5-23	ERS032964
Jap003	2001	Japan	JND 8-213	ERS032965
Jap004	2004	Japan	JND 8-064	ERS032966
Jap005	2004	Japan	JND 9-053	ERS032967
kor001	2006	Korea	Y11	ERS017225

kor002	2007	Korea	512	ERS017226
kor003	2009	Korea	Y93	ERS017227
kor004	2009	Korea	09-48	ERS017228
kor005	2010	Korea	Y175	ERS017229
kor006	2010	Korea	Y180	ERS017230
Sin001	2008	Singapore	DA3689	ERS012935
Sin002	2009	Singapore	DA1122	ERS012936
Sin003	2009	Singapore	DA3115	ERS012937
Lei010	2008	Ireland		ERS032934
Lei009	2009	Ireland		ERS032933
Bel028	2009	Belfast, Northern Ireland	CD090457	ERS012944
Bel032	2009	Belfast, Northern Ireland	CD090156	ERS012948
Bel033	2010	Belfast, Northern Ireland	CD100313	ERS012949
LSTM032	2008	Ayrshire, UK	BI-8	ERS017264
Bir012	2010	Birmingham, UK		ERS017109
Bir014	2008	Birmingham, UK		ERS017111
Bir024	2009	Birmingham, UK		ERS017122
Bir025	2007	Birmingham, UK		ERS017123
Cam010	2008	Cambridge, UK		ERS012970
Cam011	2008	Cambridge, UK		ERS012971
Cam019	2007	Cambridge, UK		ERS012980
Cam020	2007	Cambridge, UK		ERS012981
Cam032	2007	Cambridge, UK		ERS012993
Cam037	2010	Cambridge, UK		ERS012998
Exe003	2008	Exeter, UK		ERS017076
Exe012	2007	Exeter, UK		ERS017085
Exe013	2007	Exeter, UK		ERS017086
Exe014	2008	Exeter, UK		ERS017087
Exe015	2007	Exeter, UK		ERS017088
Gla001	2007	Glasgow, UK		ERS008986
Gla010	2008	Glasgow, UK		ERS008995

Gla012	2009	Glasgow, UK		ERS008998
Gla020	2009	Glasgow, UK		ERS009006
ham001	2006	London, UK		ERS012950
ham005	2007	London, UK		ERS012954
ham006	2007	London, UK		ERS012955
ham007	2007	London, UK		ERS012956
ham009	2009	London, UK		ERS012958
ham010	2008	London, UK		ERS012959
ham011	2008	London, UK		ERS012960
Liv059	2008	Liverpool, UK		ERS012883
Liv1	2008	Liverpool, UK		
Liv100	2010	Liverpool, UK		ERS012926
Liv180	2009	Liverpool, UK		ERS017194
Liv188	2007	Liverpool, UK		ERS017203
LSTM025	2007	Dundee, UK	20070031	ERS017257
LSTM028	2008	Edinburgh, UK	20080323	ERS017260
LSTM030	2008	Inverness, UK	20080684	ERS017262
LSTM35	2006	Nottingham, UK		ERS017267
LSTM36	2007	Maidstone, UK	DS902-06	ERS017268
R20291	2006	Stoke Mandeville, UK		ERS032962/FN545816
Wal001	2002	Birmingham, UK	R16760	ERS017096
Wal002	1998	Preston, UK	R12628	ERS017097
2004013	2004	Maine, USA		ERS004201
2004102	2004	New Jersey, USA		ERS004208
2004118	2004	Maine, USA		ERS004209
2004163	2004	Pennsylvania, USA		ERS004210
2006439	2006	Arizona, USA		ERS004207
2007140	2007	Maryland, USA		ERS004205
2007218	2007	USA		ERS004214
2007825	2007	USA		ERS004204
2007833	2007	Arizona, USA		ERS004212

2007837	2007	USA		ERS004206
2007850	2007	USA		ERS001492
2007855 ^	2007	USA		FN665654
2007221 *	2007	Arizona, USA		ERS004211
2007223 *	2007	Arizona, USA		ERS001491
BI-1	1988	Minneapolis, Minnesota, USA	LSTM021	ERS017252/FN668941
BI-10	2001	Pittsburgh, Pennsylvania, USA		ERS017143
BI-11	2001	Pittsburgh, Pennsylvania, USA		ERS003831
BI-12	2004	Camp Hill, Pennsylvania, USA	LSTM023	ERS017254
BI-13	2004	New Jersey, USA		ERS003836
BI-15	2004	New Jersey, USA		ERS003835
BI-2	1991	Tuscon, Arizona, USA		ERS017141
BI-3	1990	Minneapolis, Minnesota, USA		ERS017142
BI-4	1993	Minneapolis, Minnesota, USA		ERS003827
BI-5	1995	Albany, New York, USA		ERS003828
BI-6	2003	Portland, Oregon, USA		ERS003833
BI-6p	2004	Atlanta, Georgia, USA		ERS003829
BI-7	2003	Portland, Oregon, USA	LSTM022	ERS017253
BI-8	2004	Portland, Maine, USA		ERS003834
2004101	2004	Maine, USA	LSTM001	ERS017231
2005079	2005	Pennsylvania, USA	LSTM002	ERS017232
2006237 ^	2006	Arizona, USA	LSTM003	ERS017233
2007014	2007	Connecticut, USA	LSTM004	ERS017234
2007042 *	2007	Arizona, USA	LSTM005	ERS017235
2007043 *	2007	Arizona, USA	LSTM006	ERS017236
2007053	2007	Tennessee, USA	LSTM007	ERS017238
2007222 *	2007	Arizona, USA	LSTM009	ERS017240
2007235 *	2007	Arizona, USA	LSTM011	ERS017241
2007616 *	2007	Arizona, USA	LSTM012	ERS017242

2007828	2007	USA	LSTM013	ERS017243
2007832	2007	USA	LSTM014	ERS017244
2007839	2007	USA	LSTM016	ERS017247
2007850	2007	USA	LSTM019	ERS017250
2007851	2007	USA	LSTM020	ERS017251

Supplementary Table 2. Estimates of mutation rate of FQR2 star-like phylogeny with a rapid expansion model[1]. Lower and upper refer to the 95% lower and upper boundaries for the mutation estimates in the likelihood distribution (Supplementary Fig. 8).

NumStrain	NumLoci	Rate	Lower	Upper	NumSNPs
61	1000	3.02E-07	2.36E-07	3.76E-07	168
61	1500	3.84E-07	3.24E-07	4.50E-07	168
61	2000	2.99E-07	2.55E-07	3.51E-07	168
18	1500	2.91E-07	1.82E-07	4.33E-07	66
18	2000	3.96E-07	2.82E-07	5.33E-07	66
18	3290	3.55E-07	2.71E-07	4.55E-07	66

Supplementary Table 3. Details of 188 *C. difficile* 027/BI/NAP1 isolates forming the UK national collection.

Strain Name	Isolation Year	Isolation Place	Other Name	Accession Number
Bel023	2009	Belfast	CD090444	ERS012938
Bel024	2009	Belfast	CD090028	ERS012939
Bel026	2009	Belfast	CD090387	ERS012942
Bel027	2009	Belfast	CD090480	ERS012943
Bel028	2009	Belfast	CD090457	ERS012944
Bel029	2009	Belfast	CD090095	ERS012945
Bel030	2009	Belfast	CD090511	ERS012946
Bel031	2009	Belfast	CD090271	ERS012947
Bel032	2009	Belfast	CD090156	ERS012948
Bel033	2010	Belfast	CD100313	ERS012949
Bel025		Belfast	CD100056	ERS012940
Wal001	2002	Birmingham	R16760	ERS017096
Bir001	2007	Birmingham		ERS017079
Bir007	2007	Birmingham		ERS017104
Bir009	2007	Birmingham		ERS017106
Bir013	2007	Birmingham		ERS017110
Bir015	2007	Birmingham		ERS017112
Bir017	2007	Birmingham		ERS017115
Bir018	2007	Birmingham		ERS017116
Bir019	2007	Birmingham		ERS017117
Bir021	2007	Birmingham		ERS017119
Bir023	2007	Birmingham		ERS017121
Bir025	2007	Birmingham		ERS017123
Bir027	2007	Birmingham		ERS017125
Bir029	2007	Birmingham		ERS017128
Bir003	2008	Birmingham		ERS017099
Bir005	2008	Birmingham		ERS017102
Bir006	2008	Birmingham		ERS017103
Bir011	2008	Birmingham		ERS017108
Bir014	2008	Birmingham		ERS017111
Bir030	2008	Birmingham		ERS017129
Bir031	2008	Birmingham		ERS017130
Bir032	2008	Birmingham		ERS017131
Bir033	2008	Birmingham		ERS017132
Bir034	2008	Birmingham		ERS017133
Bir035	2008	Birmingham		ERS017134
Bir036	2008	Birmingham		ERS017135
Bir037	2008	Birmingham		ERS017136
Bir038	2008	Birmingham		ERS017137
Bir039	2008	Birmingham		ERS017138
Bir010	2009	Birmingham		ERS017107
Bir024	2009	Birmingham		ERS017122
Bir028	2009	Birmingham		ERS017127
Bir002	2010	Birmingham		ERS017098
Bir004	2010	Birmingham		ERS017101
Bir008	2010	Birmingham		ERS017105
Bir012	2010	Birmingham		ERS017109
Bir016	2010	Birmingham		ERS017114

Bir020	2010	Birmingham		ERS017118
Bir022	2010	Birmingham		ERS017120
Bir026	2010	Birmingham		ERS017124
Cam001	2008	Cambridge	L,08.7605514.Y	ERS012961
Cam002	2008	Cambridge	L,08.7605513.E	ERS012962
Cam003	2008	Cambridge	L,08.7612367.D	ERS012963
Cam004	2008	Cambridge	L,08.7604515.S	ERS012964
Cam005	2008	Cambridge	L,08.7604867.M	ERS012965
Cam006	2008	Cambridge	L,08.7604518.H	ERS012966
Cam007	2008	Cambridge	L,08.7604292.G	ERS012967
Cam008	2008	Cambridge	L,09.7720057.L	ERS012968
Cam009	2008	Cambridge	L,09.7720055.B	ERS012969
Cam010	2008	Cambridge	L,08.7612770.M	ERS012970
Cam011	2008	Cambridge	L,08.7612904.H	ERS012971
Cam012	2008	Cambridge	L,08.7612907.C	ERS012972
Cam013	2008	Cambridge	L,08.7612906.A	ERS012973
Cam014	2008	Cambridge	L,08.7612694.Z	ERS012974
Cam015	2008	Cambridge	L,08.7612621.Q	ERS012975
Cam016	2008	Cambridge	L,08.7603905.D	ERS012976
Cam017	2007	Cambridge	L,08.7597036.J	ERS012978
Cam018	2007	Cambridge	L,08.7597110.Z	ERS012979
Cam019	2007	Cambridge	L,07.7272911.Y	ERS012980
Cam020	2007	Cambridge	L,07.7271951.S	ERS012981
Cam021	2007	Cambridge	L,07.7272912.P	ERS012982
Cam022	2007	Cambridge	L,07.7272687.K	ERS012983
Cam023	2007	Cambridge	L,07.7272910.E	ERS012984
Cam024	2007	Cambridge	L,07.7272685.A	ERS012985
Cam025	2007	Cambridge	L,07.7272686.C	ERS012986
Cam026	2007	Cambridge	L,07.7272608.C	ERS012987
Cam027	2007	Cambridge	L,07.7272609.K	ERS012988
Cam028	2007	Cambridge	L,07.7272610.W	ERS012989
Cam029	2007	Cambridge	L,07.7272611.A	ERS012990
Cam030	2007	Cambridge	L,07.7272605.H	ERS012991
Cam031	2007	Cambridge	L,07.7272607.A	ERS012992
Cam032	2007	Cambridge	L,07.7270069.E	ERS012993
Cam033	2007	Cambridge	L,07.7271955.W	ERS012994
Cam034	2007	Cambridge	L,07.7270061.C	ERS012995
Cam035	2007	Cambridge	L,07.7270314.G	ERS012996
Cam036	2009	Cambridge		ERS012997
Cam037	2010	Cambridge		ERS012998
Cam038	2010	Cambridge		ERS012999
Cam039	2010	Cambridge		ERS013000
Exe001	2007	Exeter	FA0710129	ERS017074
Exe002	2007	Exeter	FA07011764	ERS017075
Exe004	2007	Exeter	FA07001994	ERS017080
Exe005	2007	Exeter	FA07003754	ERS017081
Exe006	2007	Exeter	FA07004642	ERS017082
Exe009	2007	Exeter	FA0700757	ERS017083
Exe010	2007	Exeter	FA07008773	ERS017077
Exe011	2007	Exeter	FA07004630	ERS017084
Exe012	2007	Exeter	FA07006324	ERS017085
Exe013	2007	Exeter	FA07008357	ERS017086
Exe015	2007	Exeter	FA07011040	ERS017088

Exe003	2008	Exeter	FA08003371	ERS017076
Exe014	2008	Exeter	FA08002245	ERS017087
Liv188	2007	Liverpool		ERS017203
Liv189	2007	Liverpool		ERS017204
Liv190	2007	Liverpool		ERS017205
Liv062	2008	Liverpool	98014	ERS012886
Liv083	2008	Liverpool	98038	ERS012908
Liv056	2008	Liverpool	98006	ERS012880
Liv081	2008	Liverpool	98331	ERS012906
Liv020	2009	Liverpool	98076	ERS017217
Liv053	2008	Liverpool	98034	ERS012877
Liv055	2008	Liverpool	98043	ERS012879
Liv057	2008	Liverpool	98007	ERS012881
Liv059	2008	Liverpool	98009	ERS012883
Liv061	2008	Liverpool	98010	ERS012885
Liv086	2009	Liverpool	98333	ERS012911
Liv096	2009	Liverpool	98350	ERS012922
Liv097	2009	Liverpool	98420	ERS012923
Liv1	2008	Liverpool	98011	ERS124182
Liv100	2010	Liverpool	108042	ERS012926
Liv101	2010	Liverpool	108056	ERS012927
Liv102	2010	Liverpool	108087	ERS012928
Liv103	2010	Liverpool	108059	ERS012929
Liv104	2010	Liverpool	108075	ERS012930
Liv10	2009	Liverpool	98237	ERS017274
Liv11	2008	Liverpool	98037	ERS124194
Liv127	2010	Liverpool	108121	ERS012931
Liv131	2010	Liverpool	108127	ERS012932
Liv132	2010	Liverpool	108061	ERS012933
Liv136	2010	Liverpool	108129	ERS012934
Liv15	2009	Liverpool	98276	ERS124201
Liv180	2009	Liverpool	98338	ERS017194
Liv182	2009	Liverpool	98273	ERS017196
Liv184	2009	Liverpool	98098	ERS017198
Liv186	2009	Liverpool	98152	ERS017201
Liv3	2009	Liverpool	98016	ERS017272
Liv41	2009	Liverpool	98580	ERS012869
CD679	2009	London		ERS012925
ham001	2006	London	31H	ERS012950
ham002	2007	London	133H	ERS012951
ham003	2009	London	332H	ERS012952
ham004	2007	London	97C	ERS012953
ham005	2007	London	44C	ERS012954
ham006	2007	London	53C	ERS012955
ham007	2007	London	101C	ERS012956
ham009	2009	London	217C	ERS012958
ham010	2008	London	211H	ERS012959
ham011	2008	London	164C	ERS012960
lon001		London	CD676 II	ERS013001
lon004		London	CD630	ERS013005
lon005		London	CD682	ERS013006
lon006		London	CD683	ERS013007
LSTM36		Maidstone	DS902-06	ERS017268

LSTM35	2006	Nottingham		ERS017267
Wal002	1998	Preston	R12628	ERS017097
LSTM032	2008	Ayrshire	BI-8	ERS017264
Gla001	2007	Glasgow		ERS008986
LSTM025	2007	Dundee	20070031	ERS017257
LSTM026	2008	Dumfries	20080107	ERS017258
LSTM027	2008	Ayrshire	20080195	ERS017259
LSTM028	2008	Edinburgh	20080323	ERS017260
LSTM029	2008	Dumbarton	20080533	ERS017261
Gla002	2008	Glasgow		ERS008987
Gla003	2008	Glasgow		ERS008988
Gla004	2008	Glasgow		ERS008989
Gla005	2008	Glasgow		ERS008990
Gla006	2008	Glasgow		ERS008991
Gla007	2008	Glasgow		ERS008992
Gla008	2008	Glasgow		ERS008993
Gla009	2008	Glasgow		ERS008994
Gla010	2008	Glasgow		ERS008995
LSTM030	2008	Inverness	20080684	ERS017262
Gla012	2009	Glasgow		ERS008998
Gla013	2009	Glasgow		ERS008999
Gla014	2009	Glasgow		ERS009000
Gla015	2009	Glasgow		ERS009001
Gla016	2009	Glasgow		ERS009002
Gla017	2009	Glasgow		ERS009003
Gla018	2009	Glasgow		ERS009004
Gla019	2009	Glasgow		ERS009005
Gla020	2009	Glasgow		ERS009006
Gla021	2009	Glasgow		ERS009007
Gla022	2009	Glasgow		ERS009008
Lei027	2009	Glasgow		ERS032952
R20291	2005	Stoke Mandeville		ERS032962/FN545816
LSTM031	2008	Glasgow	20080783	ERS017263
Lei026	2006	UK		ERS032951

Supplementary Table 4. Discriminatory SNPs and predicted functional impact within the *C. difficile* 027/BI/NAP1 lineage. Position and reference residue refer to those in R20291. Branches A-E correspond to the branches labeled in Supplementary Fig. 3. nonsyn – nonsynonymous.

Position	Branch	Type	Product	Reference residue	Alternative residue	Amino acid change	Predicted impact
2656051	B	Nonsyn	quinolinate synthetase A	L	I	conservative	None
6310	A, C	Nonsyn	DNA gyrase subunit A	I	T		Fluoroquinolone resistance
118571	C	Nonsyn	putative ribosomal protein	D	N	conservative	None
1239212	C	Nonsyn	conserved hypothetical protein, DUF_177 family	T	N	conservative	None
1466990	C	Nonsyn	Probable cation transporter, Nramp homolog	D	A	non-conservative	Possible transmembrane helix disruption
2938388	C	Nonsyn	hypothetical protein (Small, very Histidine-rich protein, unique to *Clostridia*)	F	L	non-conservative	Unknown
3118366	C	synonymous	phosphoenolpyruvate-protein phosphotransferase	P	P	none	None
3507157	C	Intergenic				none	200 bp upstream of conserved hypothetical protein
120932	D	synonymous	DNA-directed RNA polymerase alpha chain	I	I		
2304160	D	Nonsyn	conserved hypothetical protein, DUF162 family	E	G	non-conservative	Unknown
2983263	D	Nonsyn	UDP-N-acetylmuramoylalanine--D-glutamate ligase	T	I	non-conservative	Unknown; not within any Pfam domain
3538081	D	Nonsyn	PTS system, IIabc component	V	I	conservative	None
886105	E	Nonsyn	penicillin-binding protein	T	I	non-conservative	Transpeptidase, could potentially result in penicillin resistance
2575047	A	Nonsyn	hypothetical protein	V	I	conservative	Unknown
2948945	A	Intergenic				none	60 bp upstream of a putative membrane protein

Supplementary Table 5. Coding sequences carried by CTn5-like elements Tn*6192* and Tn*6105* in *C. difficile* 027/BI/NAP1 FQR1 and FQR2 lineages. Coding sequence corresponds to that in isolate R20291.

Coding sequence	Product	Transposon
CDR20291_1789	putative conjugative transposon membrane protein	Tn*6192*
CDR20291_1790	putative conjugal transfer protein	Tn*6192*
CDR20291_1791	conserved hypothetical protein	Tn*6192*
CDR20291_1792	putative cell surface protein	Tn*6192*
CDR20291_1793	putative DNA topoisomerase	Tn*6192*
CDR20291_1794	putative uncharacterized protein	Tn*6192*
CDR20291_1795	putative helicase	Tn*6192*
CDR20291_1796	putative uncharacterized protein	Tn*6192*
CDR20291_1797	putative conjugative transposon DNA recombination protein	Tn*6192*
CDR20291_1798	putative conjugative transposon conserved hypothetical protein	Tn*6192*
CDR20291_1799	putative uncharacterized protein	Tn*6192*
CDR20291_1800	putative conjugative transposon mobilization protein	Tn*6192*
CDR20291_1801	putative exported protein	Tn*6192*
CDR20291_1802	putative iron-sulfur-binding membrane protein	Tn*6192*
CDR20291_1803	ABC transporter, permease protein	Tn*6192*
CDR20291_1804	abc transporter, atp-binding protein	Tn*6192*
CDR20291_1805	ABC transporter, permease protein	Tn*6192*
CDR20291_1806	two-component system, response regulator	Tn*6192*
CDR20291_1807	two-component system, sensor protein	Tn*6192*
CDR20291_1808	probable transcription regulator	Tn*6192*
CDR20291_1809	site-specific recombinase	Tn*6192*
CDR20291_1765	conserved hypothetical protein	Tn*6105*
CDR20291_1766	probable transcription regulator (contains HtH domain)	Tn*6105*
CDR20291_1767	conserved hypothetical protein	Tn*6105*
CDR20291_1767A	conserved hypothetical protein	Tn*6105*

CDR20291_1768	putative plasmid mobilisation protein	Tn*6105*
CDR20291_1769	putative DNA-binding protein (contains zinc-finger domain)	Tn*6105*
CDR20291_1770	putative P-loop NTPase	Tn*6105*
CDR20291_1771	site-specific recombinase	Tn*6105*
CDR20291_1771A	hypothetical protein	Tn*6105*
CDR20291_1772	site-specific recombinase	Tn*6105*
CDR20291_1773	probable regulator (contains HtH domain from sigma 70)	Tn*6105*
CDR20291_1774	radical SAM enzyme	Tn*6105*
CDR20291_1775	two-component system regulatory protein	Tn*6105*
CDR20291_1775A	conserved hypothetical protein	Tn*6105*

Supplementary Table 6. Mobile elements carried by *C. difficile* 027/BI/NAP1 isolates. The presence of an element in a strain is depicted by a tick in the corresponding box.

Strain Name	Other Name	Lineage	CTn1-like	CTn*5*-like						phi-027
				Tn*6192*	Tn*6105*	Tn*6106*	Tn*6104*	agc^R	Tn*6194*	
2004102		FQR1	√	√				√	√	√
2005079	LSTM002	FQR1	√	√					√	√
2006439		FQR1	√	√				√	√	√
2007042	LSTM005	FQR1	√	√				√	√	√
2007043	LSTM006	FQR1	√	√				√	√	√
2007140		FQR1	√	√					√	√
2007855		FQR1	√	√				√	√	√
BI-10		FQR1	√	√					√	√
BI-13		FQR1	√	√				√	√	√
BI-15		FQR1	√	√				√	√	√
BI-6		FQR1	√	√					√	√
BI-7	LSTM022	FQR1	√	√						√
kor002	512	FQR1	√	√					√	√
kor003	Y93	FQR1	√	√					√	√
kor004	09-48	FQR1	√	√					√	√
kor005	Y175	FQR1	√	√					√	√
kor006	Y180	FQR1	√	√					√	√
Lei025		FQR1	√	√	√					√
2004013		FQR2	√	√	√					√
2004101	LSTM001	FQR2	√	√	√					√
2004118		FQR2	√	√	√					√
2004163		FQR2	√	√	√					√
2007014	LSTM004	FQR2	√	√	√					√
2007053	LSTM007	FQR2	√	√	√					√
2007825		FQR2	√	√	√					√

2007833		FQR2	√	√	√					√
2007837		FQR2	√	√	√					√
2007839	LSTM016	FQR2	√	√	√					√
2007850		FQR2	√	√	√					√
2007850	LSTM019	FQR2	√	√	√					√
2007851	LSTM020	FQR2	√	√	√					√
Aus001	ES231	FQR2	√	√	√					√
Aus002	ES232	FQR2	√	√	√					√
Aus003	ES233	FQR2	√	√	√					√
Aus004	ES235	FQR2	√	√	√					√
Aus005	ES236	FQR2	√	√	√					√
Aus006	ES238	FQR2	√	√	√					√
Bel023	CD090444	FQR2	√	√	√					√
Bel024	CD090028	FQR2	√	√	√					√
Bel025	CD100056	FQR2	√	√	√					√
Bel026	CD090387	FQR2	√	√	√					√
Bel027	CD090480	FQR2	√	√	√					√
Bel028	CD090457	FQR2	√	√	√					√
Bel029	CD090095	FQR2	√	√	√					√
Bel030	CD090511	FQR2	√	√	√					√
Bel031	CD090271	FQR2	√	√	√					√
Bel032	CD090156	FQR2	√	√	√					√
Bel033	CD100313	FQR2	√	√	√					√
BI-12	LSTM023	FQR2	√	√	√					√
BI-17	LSTM024	FQR2	√	√	√					√
BI-6p		FQR2	√	√	√					√
Bir001		FQR2	√	√	√					√
Bir002		FQR2	√	√	√					√
Bir003		FQR2	√	√	√					√
Bir004		FQR2	√	√	√					√
Bir005		FQR2	√	√	√					√

Bir006		FQR2	√	√	√					√
Bir007		FQR2	√	√	√					√
Bir008		FQR2	√	√	√					√
Bir009		FQR2	√	√	√					√
Bir010		FQR2	√	√	√					√
Bir011		FQR2	√	√	√					√
Bir012		FQR2	√	√	√					√
Bir013		FQR2	√	√	√					√
Bir014		FQR2	√	√	√					√
Bir015		FQR2	√	√	√					√
Bir016		FQR2	√	√	√					√
Bir017		FQR2	√	√	√					√
Bir018		FQR2	√	√	√					√
Bir019		FQR2	√	√	√					√
Bir020		FQR2	√	√	√					√
Bir021		FQR2	√	√	√					√
Bir022		FQR2	√	√	√					√
Bir023		FQR2	√	√	√	√				√
Bir024		FQR2	√	√	√	√				√
Bir025		FQR2	√	√	√					√
Bir026		FQR2	√	√	√					√
Bir027		FQR2	√	√	√					√
Bir028		FQR2	√	√	√					√
Bir029		FQR2	√	√	√	√				√
Bir030		FQR2	√	√	√					√
Bir031		FQR2	√	√	√	√				√
Bir032		FQR2	√	√	√	√				√
Bir033		FQR2	√	√	√	√				√
Bir034		FQR2	√	√	√	√				√
Bir035		FQR2	√	√	√					√
Bir036		FQR2	√	√	√					√

Bir037		FQR2	√	√	√					√
Bir038		FQR2	√	√	√					√
Bir039		FQR2	√	√	√					√
Cam001	L,08.7605514.Y	FQR2	√	√	√	√				√
Cam002	L,08.7605513.E	FQR2	√	√	√					√
Cam003	L,08.7612367.D	FQR2	√	√	√					√
Cam004	L,08.7604515.S	FQR2	√	√	√					√
Cam005	L,08.7604867.M	FQR2	√	√	√	√				√
Cam006	L,08.7604518.H	FQR2	√	√	√	√				√
Cam007	L,08.7604292.G	FQR2	√	√	√					√
Cam008	L,09.7720057.L	FQR2	√	√	√					√
Cam009	L,09.7720055.B	FQR2	√							√
Cam010	L,08.7612770.M	FQR2	√	√	√					√
Cam011	L,08.7612904.H	FQR2	√	√	√					√
Cam012	L,08.7612907.C	FQR2	√	√	√					√
Cam013	L,08.7612906.A	FQR2	√	√	√					√
Cam014	L,08.7612694.Z	FQR2	√	√	√					√
Cam015	L,08.7612621.Q	FQR2	√	√	√					√
Cam016	L,08.7603905.D	FQR2	√	√	√	√				√
Cam017	L,08.7597036.J	FQR2	√	√	√	√				√
Cam018	L,08.7597110.Z	FQR2	√	√	√	√				√
Cam019	L,07.7272911.Y	FQR2	√	√	√	√				√
Cam020	L,07.7271951.S	FQR2	√	√	√					√
Cam021	L,07.7272912.P	FQR2	√	√	√	√				√
Cam022	L,07.7272687.K	FQR2	√	√	√	√				√
Cam023	L,07.7272910.E	FQR2	√	√	√	√				√
Cam024	L,07.7272685.A	FQR2	√	√	√	√				√
Cam025	L,07.7272686.C	FQR2	√	√	√					√
Cam026	L,07.7272608.C	FQR2	√	√	√	√				√
Cam027	L,07.7272609.K	FQR2	√	√	√	√				√
Cam028	L,07.7272610.W	FQR2	√	√	√					√

Cam029	L,07.7272611.A	FQR2	√	√	√					√
Cam030	L,07.7272605.H	FQR2	√	√	√					√
Cam031	L,07.7272607.A	FQR2	√	√	√					√
Cam032	L,07.7270069.E	FQR2	√	√	√					√
Cam033	L,07.7271955.W	FQR2	√	√	√					√
Cam034	L,07.7270061.C	FQR2	√	√	√					√
Cam035	L,07.7270314.G	FQR2	√	√	√					√
Cam036		FQR2	√	√	√					√
Cam037		FQR2	√	√	√					√
Cam038		FQR2	√	√	√					√
Cam039		FQR2	√	√	√					√
Can002	AO992	FQR2	√	√	√					√
Can009	MTL047	FQR2	√	√	√					√
Can010	MTL057	FQR2	√	√	√					√
Can012	MTL077	FQR2	√	√	√					√
CD679		FQR2	√	√	√					√
Exe001	FA0710129	FQR2	√	√	√					√
Exe002	FA07011764	FQR2	√	√	√					√
Exe003	FA08003371	FQR2	√	√	√					√
Exe004	FA07001994	FQR2	√	√	√	√				√
Exe005	FA07003754	FQR2	√	√	√	√				√
Exe006	FA07004642	FQR2	√	√	√	√				√
Exe009	FA0700757	FQR2	√	√	√	√				√
Exe010	FA07008773	FQR2	√	√	√	√				√
Exe011	FA07004630	FQR2	√	√	√	√				√
Exe012	FA07006324	FQR2	√	√	√	√				√
Exe013	FA07008357	FQR2	√	√	√					√
Exe014	FA08002245	FQR2	√	√	√					√
Exe015	FA07011040	FQR2	√	√	√					√
Gla001		FQR2	√	√	√					√
Gla002		FQR2	√	√	√					√

Gla003		FQR2	√	√	√					√
Gla004		FQR2	√	√	√					√
Gla005		FQR2	√	√	√					√
Gla006		FQR2	√	√	√					√
Gla007		FQR2	√	√	√					√
Gla008		FQR2	√	√	√					√
Gla009		FQR2	√	√	√					√
Gla010		FQR2	√	√	√					√
Gla012		FQR2	√	√	√					√
Gla013		FQR2	√	√	√					√
Gla014		FQR2	√	√	√					√
Gla015		FQR2	√	√	√					√
Gla016		FQR2	√	√	√					√
Gla017		FQR2	√	√	√					√
Gla018		FQR2	√	√	√					√
Gla019		FQR2	√	√	√					√
Gla020		FQR2	√	√	√					√
Gla021		FQR2	√	√	√					√
Gla022		FQR2	√	√	√					√
ham001	31H	FQR2	√	√	√					√
ham002	133H	FQR2	√	√	√					√
ham003	332H	FQR2	√	√	√					√
ham004	97C	FQR2	√	√	√	√				√
ham005	44C	FQR2	√	√	√	√				√
ham006	53C	FQR2	√	√	√					√
ham007	101C	FQR2	√	√	√					√
ham009	217C	FQR2	√	√	√					√
ham010	211H	FQR2	√	√	√					√
ham011	164C	FQR2	√	√	√					√
Lei001		FQR2	√	√	√	√				√
Lei002		FQR2	√	√	√					√

Lei003		FQR2	√	√	√					√
Lei004		FQR2	√	√	√					√
Lei006		FQR2	√	√	√					√
Lei007		FQR2	√	√	√					√
Lei009		FQR2	√	√	√					√
Lei010		FQR2	√	√	√					√
Lei011		FQR2	√	√	√					√
Lei012		FQR2	√	√	√					√
Lei013		FQR2	√	√	√					√
Lei014		FQR2	√	√	√					√
Lei015		FQR2	√	√	√					√
Lei016		FQR2	√	√	√					√
Lei017		FQR2	√	√	√					√
Lei018		FQR2	√	√	√					√
Lei019		FQR2	√	√	√					√
Lei020		FQR2	√	√	√					√
Lei021		FQR2	√	√	√					√
Lei022		FQR2	√	√	√					√
Lei023		FQR2	√	√	√					√
Lei024		FQR2	√	√	√					√
Lei026		FQR2	√	√	√	√				√
Lei027		FQR2	√	√	√					√
Lei028		FQR2	√	√	√					√
Lei029		FQR2	√	√	√					√
Lei030		FQR2	√	√	√					√
Lei031		FQR2	√	√	√					√
Lei032		FQR2	√	√	√					√
Lei036		FQR2	√	√	√					√
Liv020	98076	FQR2	√	√	√					√
Liv053	98034	FQR2	√	√	√					√
Liv055	98043	FQR2	√	√	√					√

Liv056	98006	FQR2	√	√	√					√
Liv057	98007	FQR2	√	√	√					√
Liv059	98009	FQR2	√	√	√					√
Liv061	98010	FQR2	√	√	√					√
Liv062	98014	FQR2	√	√	√					√
Liv081	98331	FQR2	√	√	√					√
Liv083	98038	FQR2	√	√	√					√
Liv086	98333	FQR2	√	√	√					√
Liv096	98350	FQR2	√	√	√					√
Liv097	98420	FQR2	√	√	√					√
Liv1	98011	FQR2	√	√	√					√
Liv10	98237	FQR2	√	√	√					√
Liv100	108042	FQR2	√	√	√					√
Liv101	108056	FQR2	√	√	√					√
Liv102	108087	FQR2	√	√	√					√
Liv103	108059	FQR2	√	√	√					√
Liv104	108075	FQR2	√	√	√					√
Liv11	98037	FQR2	√	√	√					√
Liv127	108121	FQR2	√	√	√					√
Liv131	108127	FQR2	√	√	√					√
Liv132	108061	FQR2	√	√	√					√
Liv136	108129	FQR2	√	√	√					√
Liv15	98276	FQR2	√	√	√					√
Liv180	98338	FQR2	√	√	√					√
Liv182	98273	FQR2	√	√	√					√
Liv184	98098	FQR2	√	√	√					√
Liv186	98152	FQR2	√	√	√					√
Liv188		FQR2	√	√	√					√
Liv189		FQR2	√	√	√					√
Liv190		FQR2	√	√	√					√
Liv3	98016	FQR2	√	√	√					√

Liv41	98580	FQR2	√	√	√					√
lon001	CD676 II	FQR2	√	√	√					√
lon004	CD630	FQR2	√	√	√					√
lon005	CD682	FQR2	√	√	√					√
lon006	CD683	FQR2	√	√	√					√
LSTM025	20070031	FQR2	√	√	√					√
LSTM026	20080107	FQR2	√	√	√					√
LSTM027	20080195	FQR2	√	√	√					√
LSTM028	20080323	FQR2	√	√	√					√
LSTM029	20080533	FQR2	√	√	√					√
LSTM030	20080684	FQR2	√	√	√					√
LSTM031	20080783	FQR2	√	√	√					√
LSTM032	BI-8	FQR2	√	√	√					√
LSTM35		FQR2	√	√	√	√	√			√
LSTM36	DS902-06	FQR2	√	√	√					√
M7404		FQR2	√	√	√					√
R20291		FQR2	√	√	√	√	√			√
Wal001	R16760	FQR2	√	√	√	√				√
2006237	LSTM003	Background	√							
2007218		Background	√							
2007221		Background	√							
2007222	LSTM009	Background	√							
2007223		Background	√							
2007235	LSTM011	Background	√							
2007616	LSTM012	Background	√							
2007828	LSTM013	Background	√							√
2007832	LSTM014	Background	√							√
BI-1	LSTM021	Background	√							√
BI-11		Background	√							√
BI-2		Background	√							
BI-3		Background	√							√

BI-4		Background	√							√
BI-5		Background	√							√
BI-8		Background	√							√
Can001	OB023	Background	√							√
Can003	OB102	Background	√							√
Can004	AO315	Background	√							√
Can005	AO060	Background	√							√
Can006	AO281	Background	√							√
Can007	AO224	Background	√							√
Can008	AO601	Background	√							√
Can011	AO837	Background	√							√
CD196		Background	√							√
Jap001	JND-6-119	Background	√							√
Jap002	DJNS 5-23	Background	√							√
Jap003	JND 8-213	Background	√							√
Jap004	JND 8-064	Background	√							√
Jap005	JND 9-053	Background	√							√
kor001	Y11	Background	√							√
Lei008		Background	√							√
Sin001	DA3689	Background	√							√
Sin002	DA1122	Background	√							√
Sin003	DA3115	Background	√							√
Wal002	R12628	Background	√							√

Supplementary Note

Validation of SNP detection

A simulation approach was used to assess the accuracy of short reads mapping and variant detection. A pseudo-sequence was made by introducing artificial variants (single base substitutions, insertions and deletions) into the genome of R20291. The software INDELible[2] was used for this purpose. Default options were implemented, including a JC model[3] and insertion and deletion rates both of 0.1 relative to substitution rate. Two pseudo-genomes (named "scale 1" and "scale 2" for simplicity) with different levels of divergence were created. Scale 1 genome differs from R20291 by 74 SNPs; scale 2 genome differs from R20291 by 869 SNPs. Paired-end Illumina reads generated for R20291 were aligned to pseudo references. This step was performed multiple times, each time with Illumina reads of a different coverage. The range of the tested data coverage is 8.5-fold to 100-fold. For variant detection, four different SNP filtering and validation measures were tested: (I) use default settings in BWA[4] and specify a coverage cut off of >5-fold and < three times the average coverage; (II) by excluding SNPs within repetitive regions following the measure stated in (I); (III) validate SNP alleles by checking at all variant positions in all sequencing reads following the measures stated in (II) and only consider a SNP allele true if it is supported by all sequencing reads; and (IV) validate SNP alleles by checking at all variant positions in all sequencing reads following the measures stated in (II) and only consider a SNP allele true if it is supported by all sequencing reads and the depth for this position is no less than 40, or if it is supported by > 92.5% of the reads and the depth for this position is > 40. The numbers of false positives and false negatives were calculated in each case. The results were plotted as

Supplementary Fig. 7.

The number of false positive SNPs decreases as data coverage increases, and a minimum of 15-fold coverage is necessary to achieve a result of no false positive SNPs using measures (II), (III) and (IV). The proportion of false negative SNPs also decreases as data coverage increases, except for SNP validation measure (III). The overly stringent validation criterion of (III) rejects more SNPs when data coverage is higher, as this method only considers a SNP allele correct if it is supported by all sequencing reads. A significant number of true SNPs were therefore missed due to a few sequencing errors in abundant reads covering the variant position, despite the majority of the reads indicating the correct allele. Method (IV) is an improvement with respect to this situation, as shown by a false negative rate comparable to method (II), which does not include a SNP validation step. After comparing four SNP validation methods, method (IV) was selected for analyzing the actual sequencing data of *C. difficile* 027/BI/NAP1/027 isolates. This method allows for no false positive SNPs and a false negative rate of 7%-10% when data coverage is above 30-fold, depending on the similarity between subject sequence and reference sequence.

Estimates of mutation rate

We used a full maximum-likelihood model that assumes after an introduction event population expansion is strong enough to result in perfect star genealogies[1]. Almost all SNPs discovered in the early part of FQR2 lineage are private to individual isolates resulting in a star-like topology (Fig. 1a), which supports this assumption. In this calculation, we assumed a population expansion event that started in 2003 reflecting the earliest date of the two isolates found on the node at the base of the star-

like topology. We first conducted the analysis using 61 isolates that are present in the entire star-like topology in the early part of FQR2 lineage. Because this star-like topology still contains bifurcated sub-lineages, which can be considered as a deviation from the expansion model, we selected a subset consisting of 18 isolates that form a perfect star genealogy and performed the analysis again. For both datasets, we randomly selected different numbers of loci or used all loci (Supplementary Table 2) from a total of 3,290 coding sequences spanning 3.34 Mb from the non-repetitive genome. Comparable results were obtained for both data sets (Supplementary Table 2).

We also used two other methods to estimate the mutation rate of *C. difficile* 027/BI/NAP1. Our estimate is 1.83×10^{-7} substitutions per site per year by Path-O-Gen (http://tree.bio.ed.ac.uk/software/pathogen/) (Supplementary Fig. 4) and 1.88×10^{-7} substitutions per site per year by BEAST (95% HPD interval is $1.47\times10^{-7} \sim 2.32\times10^{-7}$). Although these estimates are comparable, there is overall weak correlation in the linear regression analysis (Supplementary Fig. 4). We believe the the spore forming lifestyle of *C. difficile* underlies this weak correlation, as the length of time spent in vegetative form is influenced by stochastic environmental factors, such as host and transmission dynamics, and therefore varies between individual lineages, but the estimates should reflect an average mutation rate.

Imports from outside of the *C. difficile* 027/BI/NAP1 lineage

The largest homologous recombination blocks are found in isolates BI-4 (123kb), BI-11, kor001 and Can007 (134kb and 147kb; these are almost identical between these strains, implying a recombination event in their common ancestor). It is possible that

chromosomal mobilization mediated by integrated mobile elements is the mechanism for these large replacements. Hfr-type chromosomal mobilization from multiple sites in the genome was previously suggested for *S. agalactiae*[5]. A putative phage element was found adjacent to the 147 kb blocks in BI-11, kor001 and Can007.

Donors of the recombination blocks

We have investigated potential donors of the large recombination blocks in the *C. difficile* 027/BI/NAP1 phylogeny by searching relevant regions in our data set against the NCBI nucleotide collection and whole-genome shot gun databases using BLASTn. This was carried out using an automated pipeline and additional manual checks. The top hits (99%~100% similarity) are from a number of ribotype 027 genomes of various geographical sources (including Canada and France). However, the degree of divergence between these sequences and our regions of interest, which is still very low, is similar to the degree of divergence between these regions and other 027/BI/NAP1 isolates in our collection. For example, 235 SNPs are present in a recombination block of 141 kb in isolate kor001 when compared to the reference R20291; this is still 99.8% similarity. We did not identify any sequence that matches our regions of interest perfectly. When these regions were searched against our unpublished *C. difficile* genomes of ribotypes 001, 002, 014, 015, 023 and 106, BLAST search results reported much lower sequence similarity (50~94 in different comparisons). We therefore concluded that potential donors of these regions are likely to be isolates with genomes very similar to ribotype 027 ones.

Lack of homologous recombination within the *C. difficile* 027/BI/NAP1 lineage

Split-decomposition method was used to assess the level of recombination between 027/BI/NAP1 isolates. We constructed a split-decomposition network (Supplementary Fig. 2a) based on 176 SNPs discovered between 34 isolates within the *C. difficile* 027/BI/NAP1 global collection. No site with missing allele information or gap was included. We used 34 taxa instead of the entire *C. difficile* 027/BI/NAP1 global collection because split-decomposition method loses resolution when the number of taxa is large, which is a limitation intrinsic to the algorithm[6]. The fit value for this network is 100.0, suggesting a very good representation of the data[6]. This network is very tree-like, while also agrees with the topologies of the maximum likelihood (Fig. 1a) and neighbor-joining phylogenies (Supplementary Fig. 2b). The unresolved splits at the base of two FQR lineages (Supplementary Fig. 2a) are caused by the single *gyrA* mutation, which has arisen separately in the two FQR lineages. We consider these as sufficient evidence that homologous recombination has not played a major role in shaping the phylogeny of the *C. difficile* 027/BI/NAP1 global collection.

Phylogeny of non-human *C. difficile* 027 isolates in the global collection

All 9 non-human isolates (2 from animals and 7 from food sources) in our collection were found in the FQR1 lineage or the non-epidemic part of the phylogeny. The derived phylogeny suggests that *C. difficile* 027/BI/NAP1 has transmitted between human and non-human sources in both directions. For example, multiple isolates from food sources or animals in Arizona were derived from a historical Arizona human isolate (BI-2 from Tucson, 1991). In a sub-lineage of FQR1, a number of isolates from food sources or animals were found in the exact same position in the phylogeny as human isolates, suggesting an identical genotype; the tip of this sub-lineage being a human isolate from New Jersey (BI-13). These data suggests *C. difficile*

027/BI/NAP1 transmits through the food chain, and human *C. difficile* could contaminate the environment. However, a more comprehensive strain collection would be needed to confirm this.

Genetic variation in PaLoc region and emergence of epidemic *C. difficile* 027/BI/NAP1

Although it has been proposed that the emergence of *C. difficlie* 027/BI/NAP1 may be due to increased toxin production in these isolates[7-9], conflicting studies have shown that there was no significant difference in toxin production[10,11]. It was also proposed that the increased toxin production was linked to an 18-bp deletion within the *tcdC* gene of the PaLoc[7,8,9A]. *tcdC* is the negative regulator of toxins A and B (encoded by *tcdA* and *tcdB*) and deletion within this gene is likely to up-regulate toxin production. However, it was later shown that a single base deletion instead of the 18-bp deletion in *tcdC* was associated with increased toxin production[12].

Perhaps remarkably, only two SNPs were found in the entire 19.6 kb PaLoc region within our *C. difficile* 027/BI/NAP1 collection. One SNP results in a premature stop codon in *tcdB* in isolate 2007825, which could lead to a truncated TcdB that lacks 203 amino acid residues at its C-terminus. Another SNP leads to a residue change (Ser419Ala) in *tcdA* gene in isolate BI-7. Both SNPs are only private to a single isolate. No change in the *tcdC* gene was found. Thus, it is unlikely the genetic changes in PaLoc have had a large functional impact within the 027/BI/NAP1 lineage.

Discriminatory SNPs with potential notable functional impact

Among the discriminatory SNPs discovered (Supplementary Table 4) is an amino acid change (A240D) in a probable transporter. This amino acid change is within the transmembrane helix domain of the protein, which belongs to the PFAM Nramp (PF01566) family (natural resistance-associated macrophage protein family)[13]. This family of proteins normally acts as cation transporters and have been shown to be involved on both 'sides' in interactions between intracellular microbial pathogens and their hosts[14,15]. However, there is no evidence that this protein could be involved in host interactions in an extracellular pathogen such as *C. difficile*, and the nature of the amino acid change would suggest impaired protein function in isolates belonging to FQR2 lineage.

Non-synonymous homoplasic SNPs

Our findings include previously known amino acid substitutions (Asp426Asn in *gyrB*, His502Asn and Arg505Lys in *rpoB*)[16,17] as well as novel mutations in *fusA*, including a non-biallelic one. Although resistance to both rifampicin and fusidic acid occurred only in the fluoroquinolone-resistant lineages, there is no evidence for a specific multidrug-resistant strain or lineage, as none of the isolates are resistant to all three antibiotics. The earliest isolates in our collection that developed resistance to rifampicin and fusidic acid are from the USA (2004) and Canada (2003) respectively; while resistance to fluoroquinolones is likely to have developed earlier. Beyond drug resistance, two non-biallelic homoplasic SNPs were found to affect codon 156 (Proline) in *slpA,* which encodes S-layer precursor protein. This single protein undergoes post-translational cleavage and forms two proteins which are major components of surface layer (S-layer)[18]. S-layer acts as an important adhesin promoting interactions between host cells and *C. difficile* bacterium. The finding

seems to suggest strong selection pressure associated with cell-surface modification acting on the gene and the codon in particular.

Mobile elements in the *C. difficile* 027/BI/NAP1 lineage

We found five different versions of CTn5-like elements (Supplementary Fig. 6) in the *C. difficile* 027/BI/NAP1 lineage. These elements, which are highly similar to conjugative transposon 5 (CTn5) in *C. difficile* 630[19], are present in all isolates in FQR1 and FQR2 lineages, except isolate Cam009 (FQR2), but absent from all isolates outside of either FQR lineage. The simplest version of these elements is found in the majority of FQR1 isolates including BI-7. We named this element Tn*6192*. Tn*6192* is highly similar to CTn5 over the entire sequence. It is particularly conserved (99% nucleotide similarity) in the region encoding ABC transporters and two-component system, but less conserved (82% nucleotide similarity) in its conjugation module when compared to CTn5 in isolate 630. The other four versions of CTn5-like elements in the 027/BI/NAP1 lineage are formed by secondary or tertiary insertions into the Tn*6192* backbone; these inserted regions include a 7.5-kb aminoglycoside resistance cassette[20], three mobile elements named Tn*6104*, Tn*6105*, and Tn*6106* (Supplementary Fig. 6). Tn*6105* is found in all FQR2 isolates (except Cam009) and isolate Lei025 (FQR1) (Fig. 1a and Supplementary Fig. 3), while Tn*6106* is exclusively carried by a sub-lineage within the FQR2. Tn*6104* – Tn*6106* have been previously described in R20291[21]. We assigned the name Tn*6193* to the version of combined Tn*6192* and Tn*6105* (Supplementary Fig. 6). Interestingly, except in R20291 and LSTM035 all CTn5-like elements were found at the same location within the genome, suggesting a hot-spot for integration. In R20291 and LSTM035 it is found in the reverse strand at a different location, implying deletion of

Tn*6193* and acquisition of Tn*6103*, which uniquely contains Tn*6104*. The complete list of isolates carrying different versions of CTn5-like elements is given as Supplementary Table 6. Beyond CTn5-like elements, all 027/BI/NAP1 isolates in our collection harbor a CTn1-like element carrying chloramphenicol resistance determinant (Supplementary Table 6). FQR1 isolates except Lei025 and BI-7 harbor an element we named Tn*6194* (previously called CTnCD11[20]) that contains *ermB*, a gene conferring resistance to erythromycin. We also found a transposon highly similar to CTn3 (630) in isolate Cam036. This CTn3-like element carries the tetracycline-resistance genes *tetM* and *tetL*, while only *tetM* is found in CTn3 in 630.

REFERENCES

1 Morelli, G. *et al.* Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity. *Nat Genet*, (2010).

2 Fletcher, W. & Yang, Z. INDELible: a flexible simulator of biological sequence evolution. *Mol Biol Evol* **26**, 1879-1888, (2009).

3 Jukes, T. H. & Cantor, C. R. Evolution of protein molecules. *Mammalian protein metabolism* **3**, 21ñ132, (1969).

4 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, (2009).

5 Brochet, M. *et al.* Shaping a bacterial genome by large chromosomal replacements, the evolutionary history of Streptococcus agalactiae. *Proc Natl Acad Sci U S A* **105**, 15961-15966, (2008).

6 Huson, D. H. What if I don't have a tree? Split decomposition and related models. *Curr Protoc Bioinformatics* **Chapter 6**, Unit 6 7, (2003).

7 Loo, V. G. *et al.* A predominantly clonal multi-institutional outbreak of Clostridium difficile-associated diarrhea with high morbidity and mortality. *N Engl J Med* **353**, 2442-2449, (2005).

8 McDonald, L. C. *et al.* An epidemic, toxin gene-variant strain of Clostridium difficile. *N Engl J Med* **353**, 2433-2441, (2005).

9 Warny, M. *et al.* Toxin production by an emerging strain of Clostridium difficile associated with outbreaks of severe disease in North America and Europe. *Lancet* **366**, 1079-1084, (2005).

10 Akerlund, T. *et al.* Increased sporulation rate of epidemic Clostridium difficile Type 027/NAP1. *J Clin Microbiol* **46**, 1530-1533, (2008).

11 Merrigan, M. *et al.* Human hypervirulent Clostridium difficile strains exhibit increased sporulation as well as robust toxin production. *J Bacteriol* **192**, 4904-4911, (2010).

12 Matamouros, S., England, P. & Dupuy, B. Clostridium difficile toxin expression is inhibited by the novel regulator TcdC. *Mol Microbiol* **64**, 1274-1288, (2007).

13 Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res* **36**, D281-288, (2008).

14 Govoni, G. & Gros, P. Macrophage NRAMP1 and its role in resistance to microbial infections. *Inflamm Res* **47**, 277-284, (1998).

15 Pinner, E., Gruenheid, S., Raymond, M. & Gros, P. Functional complementation of the yeast divalent cation transporter family SMF by NRAMP2, a member of the mammalian natural resistance-associated macrophage protein family. *J Biol Chem* **272**, 28933-28938, (1997).

16 O'Connor, J. R. *et al.* Rifampin and rifaximin resistance in clinical isolates of Clostridium difficile. *Antimicrob Agents Chemother* **52**, 2813-2817, (2008).

17 Spigaglia, P. *et al.* Fluoroquinolone resistance in Clostridium difficile isolates from a prospective study of C. difficile infections in Europe. *Journal of medical microbiology* **57**, 784-789, (2008).

18 Calabi, E. *et al.* Molecular characterization of the surface layer proteins from Clostridium difficile. *Molecular Microbiology* **40**, 1187-1199, (2001).

19 Sebaihia, M. *et al.* The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome. *Nat Genet* **38**, 779-786, (2006).
20 He, M. *et al.* Evolutionary dynamics of Clostridium difficile over short and long time scales. *Proc Natl Acad Sci U S A* **107**, 7527-7532, (2010).
21 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 Clostridium difficile. *PLoS ONE* **6**, e23014, (2011).