Two Groups of Cocirculating, Epidemic *Clostridiodes difficile* Strains Microdiversify through Different Mechanisms

Tatiana Murillo[1], Gabriel Ramírez-Vargas[1], Thomas Riedel[2,3], Jörg Overmann[2,3], Joakim M. Andersen[4], Caterina Guzmán-Verri[5], Esteban Chaves-Olarte[1], and César Rodríguez[1,]*

[1]Facultad de Microbiología and Centro de Investigación en Enfermedades Tropicales (CIET), Universidad de Costa Rica, San José, Costa Rica

[2]Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany

[3]German Center for Infection Research (DZIF), Partner Site Hannover-Braunschweig, Braunschweig, Germany

[4]Department of Food, Processing and Nutritional Sciences, North Carolina State University

[5]Programa de Investigación en Enfermedades Tropicales (PIET), Escuela de Medicina Veterinaria, Universidad Nacional, Heredia, Costa Rica

*Corresponding author: E-mail: cesar.rodriguezsanchez@ucr.ac.cr.

Accepted: March 13, 2018

Data deposition: Sequencing data from the NAPCR1/ST54 and NAP1/ST01 isolates are available from the European Nucleotide Archive (Study PRJEB5034). Moreover, MGEs from selected NAPCR1/ST54 isolates were deposited under the accession numbers MF547662, MF547663, MF547664, MF547665, and MF547666.

Abstract

Clostridiodes difficile strains from the NAP$_{CR1}$/ST54 and NAP1/ST01 types have caused outbreaks despite of their notable differences in genome diversity. By comparing whole genome sequences of 32 NAP$_{CR1}$/ST54 isolates and 17 NAP1/ST01 recovered from patients infected with C. difficile we assessed whether mutation, homologous recombination (r) or nonhomologous recombination (NHR) through lateral gene transfer (LGT) have differentially shaped the microdiversification of these strains. The average number of single nucleotide polymorphisms (SNPs) in coding sequences (NAP$_{CR1}$/ST54 = 24; NAP1/ST01 = 19) and SNP densities (NAP$_{CR1}$/ST54 = 0.54/kb; NAP1/ST01 = 0.46/kb) in the NAP$_{CR1}$/ST54 and NAP1/ST01 isolates was comparable. However, the NAP1/ST01 isolates showed 3× higher average d*N*/d*S* rates (8.35) that the NAP$_{CR1}$/ST54 isolates (2.62). Regarding r, whereas 31 of the NAP$_{CR1}$/ST54 isolates showed 1 recombination block (3,301–8,226 bp), the NAP1/ST01 isolates showed no bases in recombination. As to NHR, the pangenome of the NAP$_{CR1}$/ST54 isolates was larger (4,802 gene clusters, 26% noncore genes) and more heterogeneous (644 ± 33 gene content changes) than that of the NAP1/ST01 isolates (3,829 gene clusters, ca. 6% noncore genes, 129 ± 37 gene content changes). Nearly 55% of the gene content changes seen among the NAP$_{CR1}$/ST54 isolates (355 ± 31) were traced back to MGEs with putative genes for antimicrobial resistance and virulence factors that were only detected in single isolates or isolate clusters. Congruently, the LGT/SNP rate calculated for the NAP$_{CR1}$/ST54 isolates (26.8 ± 2.8) was 4× higher than the one obtained for the NAP1/ST1 isolates (6.8 ± 2.0). We conclude that NHR-LGT has had a greater role in the microdiversification of the NAP$_{CR1}$/ST54 strains, opposite to the NAP1/ST01 strains, where mutation is known to play a more prominent role.

Key words: *Clostridioides difficile*, microdiversification, SNPs, core genome, accessory genome, LGT.

Introduction

Clostridioides difficile infections (CDI) are the main cause of hospital-acquired diarrhea after antibiotic treatment and the most common type of nosocomial infections in high-income countries (Slimings and Riley 2014; Knight et al. 2015). They vary from mild to moderate diarrhea to severe pseudomembranous colitis, toxic megacolon, and death (Hunt and Ballard 2013; Knight et al. 2015) and have a strong impact on health-care systems, affecting millions of patients worldwide (McGlone et al. 2012; Lessa et al. 2015). These infections are mostly acquired through the exposure of patients to spores in hospital environments, although the number of CDI community-acquired cases is also on the rise (Gupta and Khanna 2014; Knight et al. 2015).

The toxins TcdA and TcdB have been traditionally regarded as the main virulence factors of *C. difficile* (Hunt and Ballard 2013). They inactivate small GTPases through their glucosyltransferase activity and thereby damage the actin cytoskeleton of intestinal epithelial cells, among other deleterious host cell effects (Just et al. 1995; Chaves-Olarte et al. 1997). In most *C. difficile* strains, the genes encoding TcdA and TcdB are found in a so-called pathogenicity locus (PaLoc) (Braun et al. 1996). Other virulence factors described for this species include the binary toxin CDT, which affects the dynamics of epithelial microtubules as consequence of its ADP ribosyltransferase activity (Perelle et al. 1997; Schwan et al. 2009), as well as adhesins, fimbriae, and flagellin for host colonization (Goulding et al. 2009; Reynolds et al. 2011), and the surface layer protein (SlpA), which has been linked to inflammation and adherence to host cells (Calabi et al. 2001; Merrigan et al. 2013).

As the virulence and epidemic potential of strains differ significantly, several methods, including Pulsed Field Gel Electrophoresis (PFGE), ribotyping, toxinotyping, and Multilocus Sequence Typing (MLST) have been applied to type *C. difficile* isolates (Knight et al. 2015). Among the different types, NAP1/ST01 strains are particularly notorious and caused nosocomial outbreaks linked to high morbidity and mortality rates in the United States, Canada, the United Kingdom, Australia, and Latin America (Quesada-Gómez et al. 2010; Hunt and Ballard 2013). Other strains of *C. difficile*, such as the NAP$_{CR1}$/ST54, have also caused outbreaks (Quesada-Gómez et al. 2015).

The NAP$_{CR1}$/ST54 strains show high virulence in animal models despite their close phylogenetic relationship to nonepidemic ST54 isolates such as the *C. difficile* reference strain 630 (CD630) (Quesada-Gómez et al. 2015; López-Ureña et al. 2016). Moreover, they are multidrug-resistant and their genomes are unusually diverse, as indicated by their classification in at least 10 different *Sma*I macrorestriction patterns (López-Ureña et al. 2016; Ramírez-Vargas et al. 2017).

The rates at which different types of genomic change occur are of fundamental importance to understanding prokaryote genome evolution (Vos et al. 2015) and the emergence of new or more virulent strains (Knight et al. 2015). Bacterial genomes may evolve through accumulation of mutations (*m*), homologous recombination (*r*), or nonhomologous recombination (NHR) (Mugal et al. 2014; Vos et al. 2015). Mutations give rise to single nucleotide polymorphisms (SNPs), which are termed nonsynonymous (d*N*) if they affect the coded protein or synonymous (d*S*) if they do not (Kryazhimskiy and Plotkin 2008; Mugal et al. 2014). Homologous recombination, in turn, is the exchange of genetic information between identical or highly similar DNA molecules, even between the same bacterial chromosome (Vos and Didelot 2009; Hanage 2016). One recombination event, unlike a mutation, simultaneously substitutes several nucleotides (Guttman and Dykhuizen 1994; Hanage 2016). The *r*/*m* rate compares the effect of *r* and *m* in bacterial diversification by calculating the rate of nucleotides per generation substituted by each process (Guttman and Dykhuizen 1994; Croucher et al. 2015). NHR is the acquisition of dissimilar genetic content by mechanisms of lateral gene transfer (LGT), including transformation, conjugation, transduction, and gene transfer agents (Dagan et al. 2008; Darmon and Leach 2014). NHR is harder to measure, but can be detected and estimated through pangenome analyses and comparative genomics (Vos et al. 2015; McInerney et al. 2017).

Most studies on the diversification of *C. difficile* have so far focused on its core genome and only a few investigations have addressed the contribution of LGT to this process (He et al. 2010; Roberts et al. 2014) despite the recognized role of this parasexual process and the pangenome in bacterial niche adaptation and genome diversification (Hehemann et al. 2016; McInerney et al. 2017). Previous work on the NAP1/ST01 genome indicate that mutation, rather than homologous recombination, drives the microevolution of *C. difficile* (He et al. 2010; Dingle et al. 2011). In line with these studies, most *C. difficile* clades studied so far show d*N*/d*S* >1 and *r*/*m* rates below or close to 1 (He et al. 2010; Dingle et al. 2011). Other authors, by contrast, have reported that homologous recombination might play a strong role in the evolution of this species. For instance, Lemée et al. (2005) detected large SNPs blocks in *cwp66*, *slpA*, and flagellar genes among isolates from different MLST clades, Castillo-Ramírez et al. (2011) identified large recombinational blocks in NAP1/ST01 genomes, and Didelot et al. (2012) found that strains from certain STs have *r*/*m* ratios >2.

A major outbreak of CDI in a Costa Rican hospital was caused by NAP1/ST01 and NAP$_{CR1}$/ST54 strains (Wong-McClure et al. 2012; Quesada-Gómez et al. 2015). Thus, we compared the core and accessory genomes of 17 NAP1/ST01 and 32 NAP$_{CR1}$/ST54 isolates that cocirculated in Costa Rican hospitals to explore whether these two groups of strains display different signatures of mutation-, recombination-, and MGE-driven diversification in the context of genome evolution. Whereas the effect of mutation was appraised through the estimation of d*N*/d*S* rates, we calculated *r*/*m* rates and gene content changes to delimitate the contribution of homologous recombination and MGE-driven NHR in microdiversification, respectively. The results presented contribute to the ongoing debate about which and how evolutionary mechanisms shape microbial diversification processes and indicate that strains from the same species may microdiversify through different mechanisms.

Materials and Methods

Isolates and Whole Genome Sequences

Clostridiodes difficile clinical isolates of the NAP$_{CR1}$/ST54 ($n = 32$) and NAP1/ST01 ($n = 17$) groups were recovered

Table 1
SNPs in the Core Genome of the NAP$_{CR1}$/ST54 Isolates

Isolate	PFGE *Sma*I Pattern	Genome Size (Mb)	% Reads Mapped to CD630	Total Number of SNPs	Average Number of SNPs	SNP Density (per 100 kb)	Average SNP Density (per *Sma*I)	Number of Nonsynonymous Mutations (d*N*)	Number of Synonymous Mutations (d*S*)	d*N*/d*S* Rate	Average d*N*/d*S* Rate
3147	442	4.54	90.2	24	24	0.53	0.53	16	8	2.00	2.00
5701	447	4.51	92.0	28	24	0.62	0.53	18	10	1.80	2.77
5711		4.54	90.5	23		0.51		17	6	2.83	
5767		4.55	90.1	23		0.51		17	6	2.83	
5771		4.55	90.3	23		0.51		18	5	3.60	
2784	448	4.51	91.2	23	24	0.51	0.53	16	7	2.29	2.59
3125		4.55	90.4	22		0.48		15	7	2.14	
3137		4.51	92.2	23		0.51		16	7	2.29	
5434		4.51	91.1	24		0.51		18	6	3.00	
5704		4.55	91.0	25		0.55		17	8	2.13	
5707		4.51	91.3	24		0.53		17	7	2.43	
5733		4.55	90.2	25		0.55		19	6	3.17	
5751		4.55	90.8	23		0.51		16	7	2.29	
5774		4.55	90.2	23		0.51		18	5	3.60	
6275		4.52	91.7	29		0.64		21	8	2.63	
3129	449	4.54	90.6	22	24	0.48	0.53	16	6	2.67	2.76
5719		4.54	89.5	26		0.57		19	7	2.71	
5755		4.55	90.2	23		0.51		18	5	3.60	
5772		4.55	90.5	25		0.55		18	7	2.57	
6276		4.53	90.0	25		0.55		18	7	2.57	
6289		4.62	89.9	24		0.52		17	7	2.43	
5734	452	4.51	91.1	24	24	0.53	0.60	18	6	3.00	3.00
2945	487	4.60	87.0	24	25	0.52	0.53	20	4	5.00	4.08
5763		4.61	88.3	25		0.54		19	6	3.17	
2992	488	4.54	90.0	23	23	0.51	0.51	15	8	1.88	1.88
5761	489	4.50	90.4	26	28	0.58	0.61	19	7	2.71	2.67
5762		4.50	91.3	29		0.64		21	8	2.63	
3145	558	4.55	90.3	25	26	0.55	0.56	16	9	1.78	2.01
6285		4.55	90.2	26		0.57		18	8	2.25	
3144	578	4.55	90.0	22	22	0.48	0.48	16	6	2.67	2.41
3150		4.55	90.5	25		0.55		16	9	1.78	
5436		4.55	90.0	19		0.42		14	5	2.80	
Average		4.54	90.4	24	24	0.53	0.54	17	7	2.62	2.62

between 2003 and 2012 from patients with CDI in the Costa Rican hospitals San Juan de Dios (HSJD), México (HMX), Blanco Cervantes (HBC), Calderón Guardia (HCG), San Vicente de Paul (HSVP), and the National Centre for Rehabilitation (CENARE) (supplementary tables 1 and 2, Supplementary Material online). Whole genome sequences (WGS) for these isolates were obtained at the Wellcome Trust Sanger Institute using HiSeq 2500 instruments (Illumina). Velvet v.1.1 (Zerbino 2010) or Edena V3.131028 (Hernandez et al. 2008) were used for sequence assembly and the corresponding assembly statistics are presented in the supplementary table 3, Supplementary Material online. To resolve the structure of some MGEs, the genomes of selected NAP$_{CR1}$ isolates were also sequenced using Single Molecule Real Time (SMRT) sequencing technology at the Leibniz Institute DSMZ (Germany). To this end, PacBio *RSII* long-read sequencing reads (P6 chemistry) were assembled with the "RS_HGAP_Assembly.3" protocol included in the SMRT Portal version 2.3.0. Sequencing data from the NAP$_{CR1}$/ST54 and NAP1/ST01 isolates is available from the European Nucleotide Archive (Study PRJEB5034). Moreover, MGEs from selected NAP$_{CR1}$/ST54 isolates were deposited under the accession numbers MF547662, MF547663, MF547664, MF547665, and MF547666.

Core Genome SNP Analyses

Breseq (Barrick et al. 2014) was used to call core genome SNPs using the annotated genomes of *C. difficile* R20291 (Acc. No.: FN545816) and *C. difficile* 630 (Acc. No.: AM180355) as

Table 2

SNPs in the Core Genome of the NAP1/ST01 Isolates

Isolate	Genome Size (Mb)	% Reads Mapped to R20291	Total Number of SNPs	Average Number of SNPs	SNP Density (per 100 kb)	Average SNP Density	Number of Nonsynonymous Mutations (d*N*)	Number of Synonymous Mutations (d*S*)	d*N*/d*S* Rate	Average d*N*/d*S* Rate
5700	4.18	96.2	20	19	0.48	0.46	18	2	9.00	8.35
5703	4.18	96.2	21		0.50		19	2	9.50	
5705	4.12	98.1	17		0.41		15	2	7.50	
5706	4.12	97.0	21		0.51		18	3	6.00	
5708	4.13	96.5	19		0.46		17	2	8.50	
5709	4.13	97.8	19		0.46		17	2	8.50	
5710	4.13	97.4	19		0.46		17	2	8.50	
5713	4.13	96.4	21		0.51		19	2	9.50	
5714	4.09	99.3	19		0.46		17	2	8.50	
5718	4.13	97.1	19		0.46		17	2	8.50	
5720	4.18	95.2	20		0.48		18	2	9.00	
5749	4.13	96.9	18		0.44		16	2	8.00	
5758	4.13	98.0	17		0.41		15	2	7.50	
5759	4.13	97.7	19		0.46		17	2	8.50	
5764	4.10	99.5	20		0.49		18	2	9.00	
5765	4.14	97.6	19		0.46		17	2	8.50	
5768	4.13	97.4	17		0.41		15	2	7.50	
Average	4.13	97.3	19	19	0.46	0.46	17	2	8.35	8.35

reference genomes for the NAP1/ST01 and NAP$_{CR1}$/ST54 isolates, respectively. A minimum coverage of 20 reads was used to define a SNP to avoid errors from misassemblies or bad alignments. Blocks of two or more SNPs, SNPs located within MGEs, and SNPs in intergenic regions were excluded from downstream analyses to disregard the influence of SNPs arising from recombination and to focus only on coding sequences (CDS). After this selection, we calculated the total number of SNPs and the SNP density for each isolate and classified the SNPs as d*N* or d*S* to estimate d*N*/d*S* rates. These d*N*/d*S* rates were compared using Mann–Whitney *U* tests. Additionally, we constructed maximum-likelihood bootstrapped trees from concatenated core SNP alignments generated by the CFSAN SNP pipeline (Gouy et al. 2010) with Seaview (Davis et al. 2015), and visualized with FigTree (http://tree.bio.ed.ac.uk/software/figtree/; last accessed March 21, 2018). The results of the Breseq and CFSAN pipelines differ because of thresholds used for SNP detection, since CFSAN considers SNPs in intergenic regions and MGE. All software was run with default parameters.

Analysis of Feature Frequency Profiles

Illumina WGS were compared using feature frequency profiles (FFP) to detect differences at the pangenome level. FFP is an alignment-free method that calculates distance scores based on differences in relative *l*-mer frequencies, being an *l*-mer a string of a defined amount of nucleotides. Since it is an alignment free method, it can be applied to WGS with dissimilar gene content and therefore used to determine differences in accessory genomes (Sims and Kim 2011). We used *l*-mers of 20 nt to reach a compromise between discrimination potential and computational capacity. Comparison matrices were transformed with the neighbor-joining method into trees that were visualized with FigTree.

Homologous Recombination Analyses in Core Genome

The alignments generated with the CFSAN SNP pipeline were analyzed with Gubbins (Croucher et al. 2015) to identify recombination blocks, detect SNPs within recombination blocks, and calculate *r*/*m* rates.

Estimation of NHR through Pangenome Comparisons

To compare the pangenomes of the NAP$_{CR1}$/ST54 and NAP1/ST01 isolates and to facilitate MGE detection, Roary (Page et al. 2015) and Get_Homologues (Contreras-Moreira and Vinuesa 2013) were used to predict unique gene clusters. A unique gene cluster was defined a group of genes found only in a certain isolate. Additionaly, Roary was employed to estimate the size of the core and accessory genomes and to generate gene presence/absence spreadsheets and maximum likelihood phylogenetic trees from the accessory genomes. This pipeline classifies genes in four categories according to their frequency of occurrence in the data set: core genes (>99% of the isolates), soft-core genes ($95\% \leq$ isolates $< 99\%$), shell genes ($15\% \leq$ isolates $< 95\%$), and cloud genes ($0\% \leq$ isolates $< 15\%$). Get_Homologues, in turn, produces pangenomic matrices from which

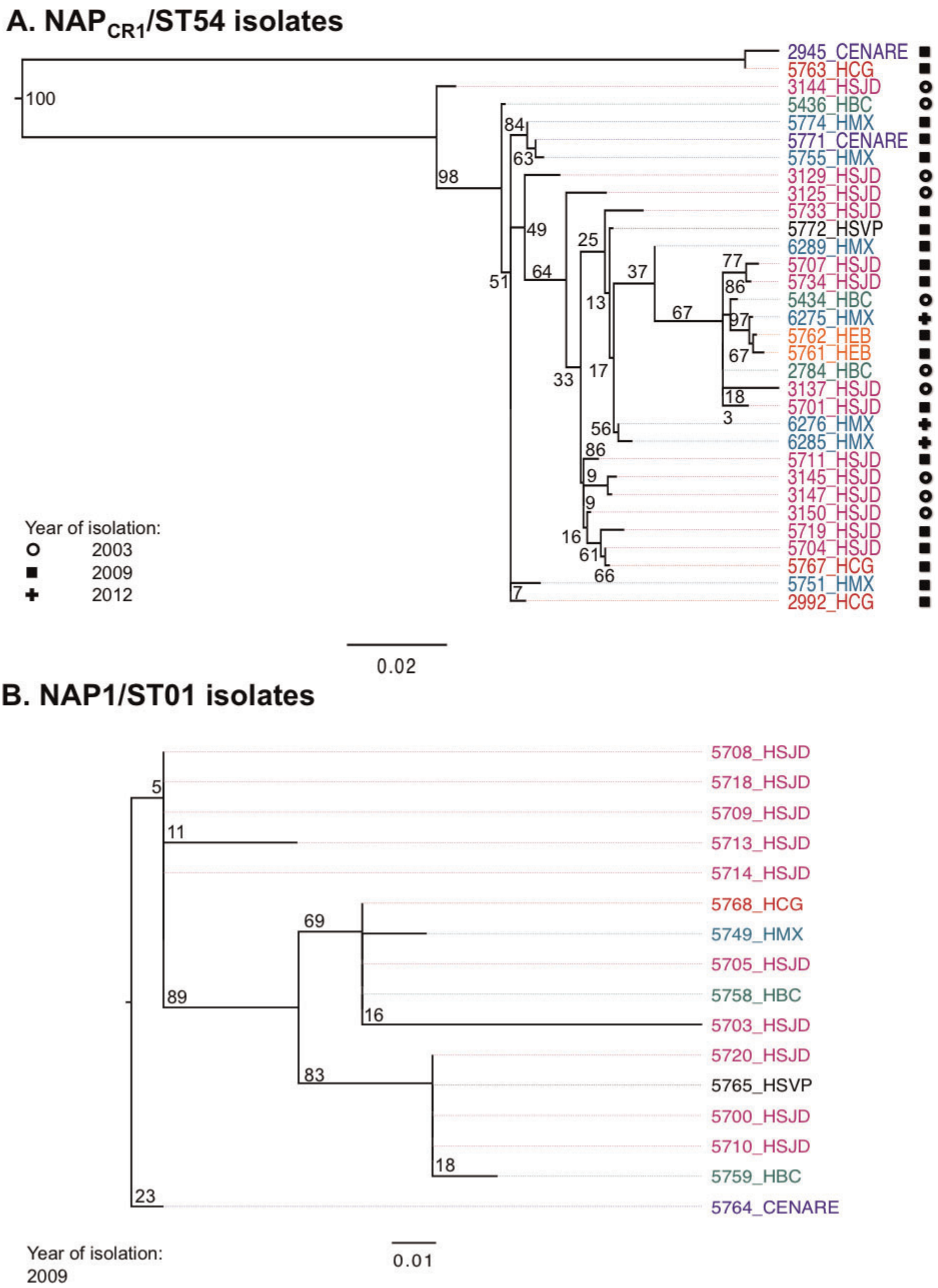


Fig. 1.—Unrooted phylogenomic maximum likelihood trees of NAP$_{CR1}$/ST54 (*A*) and NAP1/ST01 (*B*) isolates generated from core genome SNP alignments. Compared with the NAP1/ST01 isolates, the NAP$_{CR1}$/ST54 isolates showed larger distances and were supported by higher bootstrap values. Core genome SNPs were called and aligned using the CFSAN SNP pipeline. The resulting distance matrixes were used as input by Seaview to build trees using the PhyML algorithm and a bootstrap value of 100. Bootstrap values are indicated above the branches. Scales correspond to average number of substitutions per site. Different hospitals are shown by different colors. Different symbols denote different years of isolation.

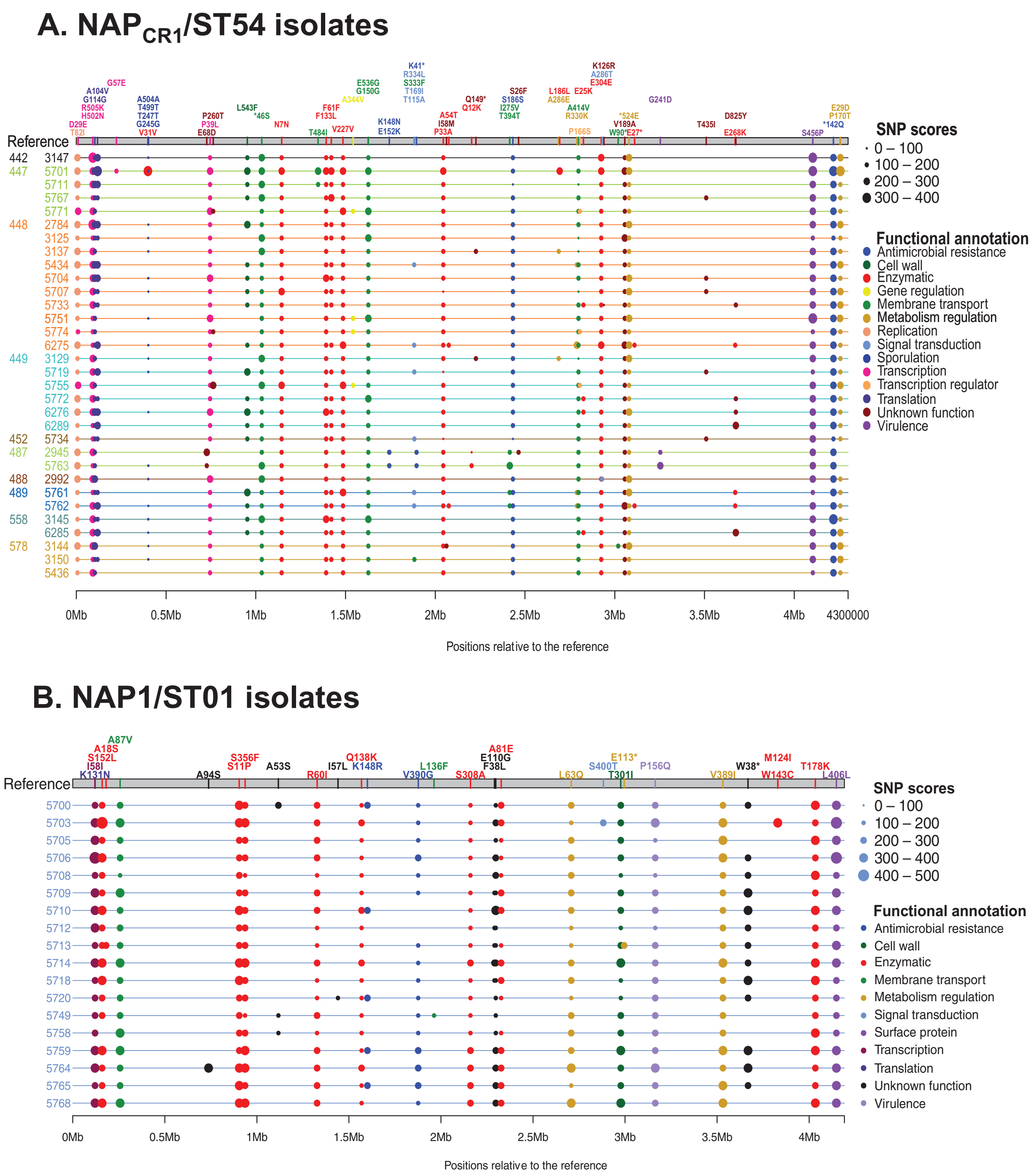


Fig. 2.—Genomic localization, score quality, and predicted function of nonsynonymous SNPs detected in core genome of the NAP$_{CR1}$/ST54 (*A*) and the NAP1/ST01 (*B*) isolates. The genomes of strains 630 or R20291 were used as references for the NAP$_{CR1}$/ST54 and the NAP1/ST01 isolates, respectively. The *Sma*I pattern of the isolates in panel *A* is shown in the *Y* axis. The diameter of the circles represents the score assigned by Breseq to each SNP and the different colors depict the predicted function of the genes with SNPs. The used color code refers to the functional annotation.

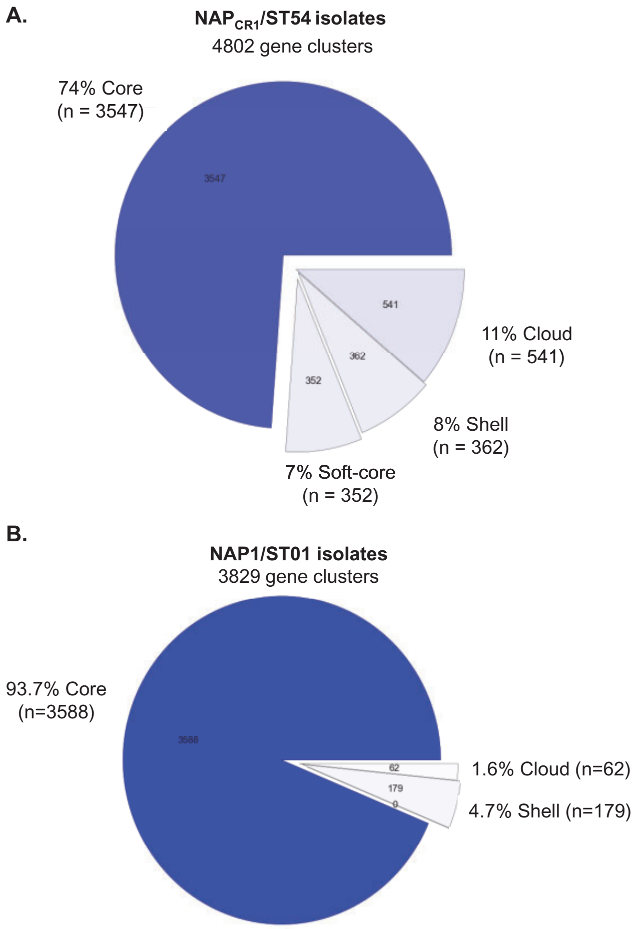


FIG. 3.—Comparison of the pangenomes of the analyzed NAP_{CR1}/ST54 (*A*) and NAP1/ST01 isolates (*B*). According to their frequency of finding, these Roary pie charts show the amount of genes clustered in the categories core ($99\% \leq$ strains $\leq 100\%$), soft-core ($95\% \leq$ strains $< 99\%$), shell ($15\% \leq$ strains $< 95\%$), and cloud ($0\% \leq$ strains $< 15\%$).

parsimony-based pangenomic trees can be derived. These trees were visualized as described earlier.

MGE Detection

According to their location and branching distances in the trees generated with Get_Homologues, draft genomes of four NAP_{CR1}/ST54 isolates and six NAP1/ST01 were selected for further analyses. To spot unshared regions resembling MGEs, contigs containing unique gene clusters were compared with cognate contigs from reference genomes using WebACT/ACT (Carver et al. 2005). For MGE delimitation, we considered criteria such as presence of genes from known MGEs (i.e., phage-related proteins or recombinases), % GC skews, and atypical codon usages. Putative MGEs were annotated using Prokka v.1.11 (Seemann 2014) and manually curated using BLAST (Gish and States 1993) or InterPro (Finn et al. 2017) searches. A list of differentially distributed MGEs was created, and to measure their role in microdiversification, the Roary analyses were repeated with modified WGS in which these discriminative MGEs were deliberately removed.

Calculation of Gene Content Changes and LGT/SNPs Rates

The pangenome comparisons done with Roary and Get_Homologues provide a list of all accessory genes and the isolates in which they are present. These lists were used to calculate the number of gene changes (gain or loss) between the isolates and their corresponding reference genome. We also determined the amount of gene content changes linked to the MGEs that show a differential distribution among each group of isolates (MGE-driven LGT). To calculate LGT/SNP and MGE-driven LGT/SNP rates, we divided the number of gene content changes by the number of SNPs calculated from the Breseq output.

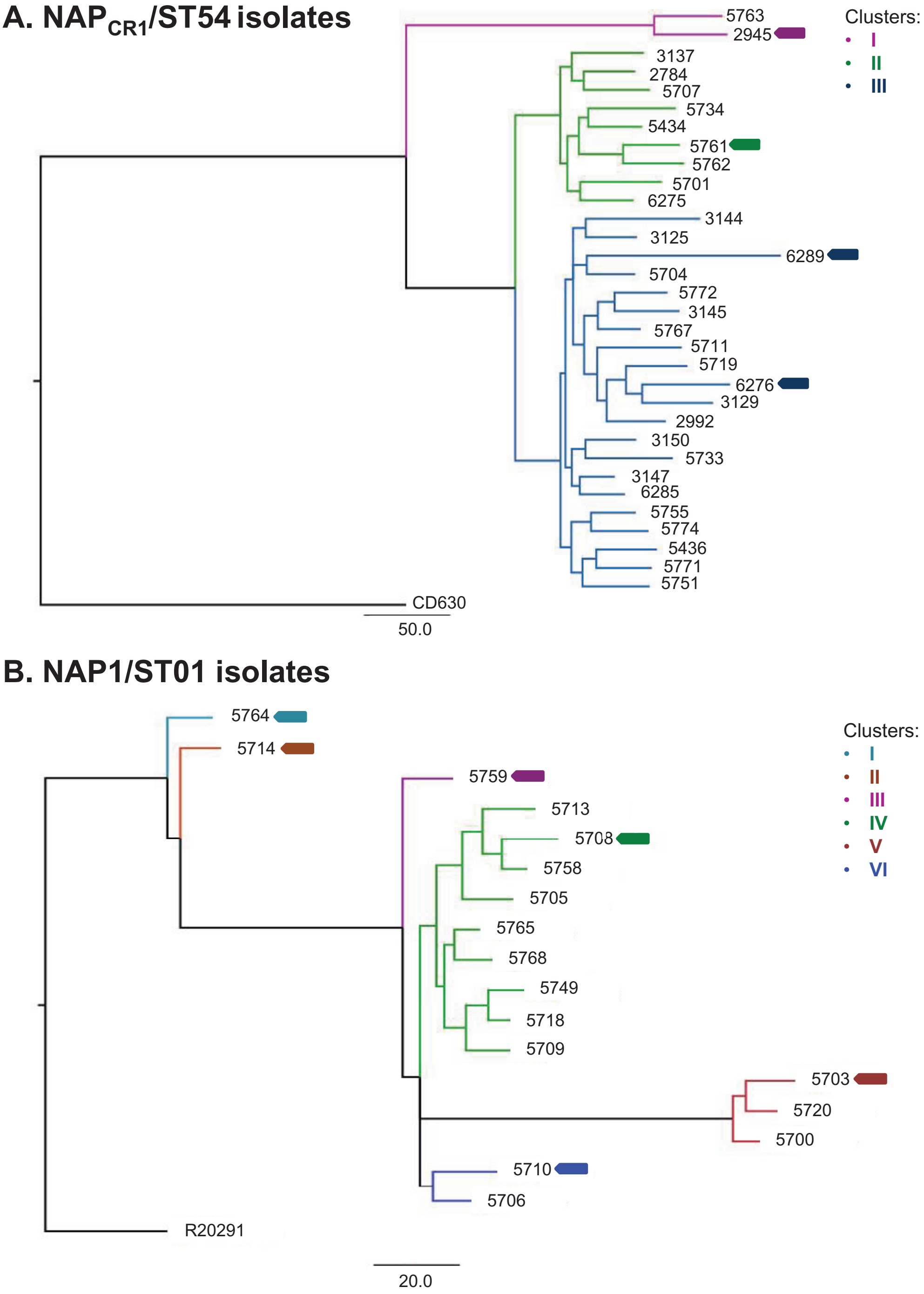
Comparison of CRISPR Arrays

CRISPR spacer arrays were predicted using the CRISPR Recognition Tool and thereafter manually curated for false positive repeats (Andersen 2016). In short, CRISPR loci were predicted and manually curated for false positive repeats. CRISPR loci were visualized by representing spacers with unique colored boxes that contain icons representing different spacer lengths. CRISPR loci were numbered from the ancestral end at the right hand side. Spacer deletions, showed by a crossed-out box, were deduced through spacer ordering across strains.

Results

SNPs Analyses

Compared with the NAP1/ST01 isolates, a smaller proportion of the reads obtained for the NAP_{CR1}/ST54 isolates mapped to the corresponding reference genome (tables 1 and 2). Though both sets of genomes consist of very closely related isolates separated by only dozens of SNPs, the NAP_{CR1}/ST54 isolates showed more SNPs in CDS and a higher average SNP density in their core genome than the NAP1/ST01 isolates (tables 1 and 2). The branching distance calculated for the NAP_{CR1}/ST54 isolates in an unrooted SNP-based tree was almost 2.5-fold higher than that obtained for the NAP1/ST01 isolates, confirming that the core genome of the NAP_{CR1}/ST54 group of isolates is more diverse (fig. 1). The topology of this tree did not match metadata such as the year or hospital of isolation and presented low bootstrap values. Most nonsynonymous SNPs identified in the NAP_{CR1}/ST54 and NAP1/ST01 isolates (fig. 2; supplementary tables 4 and 5, Supplementary Material online) were found in genes encoding metabolic enzymes (NAP_{CR1}: $n = 7$; NAP1: $n = 11$) or antibiotic resistance or virulence traits (NAP_{CR1}: $n = 6$; NAP1: $n = 3$).


A. NAP_{CR1}/ST54 isolates

B. NAP1/ST01 isolates

Fig. 4.—Unrooted parsimony-based pangenomic trees calculated for NAP_{CR1}/ST54 (*A*) and NAP1/ST01 isolates (*B*). Three distinct groups of NAP_{CR1}/ST54 and six distinct groups of NAP1/ST01 isolates were defined, respectively. These groups appear purple (I), green (II), and blue (III) in panel *A* or in teal (I), brown (II), purple (III), green (IV), dark red (V), and blue (VI) in panel *B*. Trees were derived from binary matrixes summarizing the presence–absence of gene clusters in proteome predictions generated with Get_Homologues. Isolates selected for downstream pangenome analyses were marked with block arrows.

The average d*N*/d*S* rates calculated for both groups of isolates were >1 (tables 1 and 2), but the rate calculated from the NAP_{CR1}/ST54 isolates was 3.2 times lower than that obtained for for the NAP1/ST01 isolates. An exception to this observation was the d*N*/d*S* rate of NAP_{CR1}/ST54 isolates from the 487 *Sma*I pattern, which was 4.08 and therefore unusually high (table 1).

Table 3
Presence–Absence Matrix of MGEs Differentially Distributed among the NAP_{CR1}/ST54 Isolates

Strain/Isolate	PFGE *Sma*I Pattern	MGE						
		56-kb Prophage	Big Phage (variant 1)	Big Phage (variant 2)	Putative Plasmid	mobTn withTn*4001*	*skin*^{Cd}	Tn*5397*
CD630		–	–	–	–	–	+	+
3147	442	–	+	–	–	+	+	+
5701, 5711, 5767, 5771	447	–	+	–	–	+	+	+
2784, 3125, 3137, 5434, 5704, 5707, 5733, 5751, 5774, 6275	448	–	+	–	–	+	+	+
3129, 5719, 5755, 5772, 6276, 6289	449	–	+	–	–[a]	+	+[b]	+
5734	452	–	+	–	–	+	+	+
2945, 5763	487	+	–	+	–	–	+	+
2992	488	–	+	–	–	+	+	+
5761, 5762	489	–	+	–	–	+	+	–
3145, 6285	558	–	+	–	–	+	+	+
3144, 3150, 5436	578	–	+	–	–	+	+	+

[a] Present in isolate 6289.
[b] Absent in isolate 6276.

Homologous Recombination Analyses

With a single exception (isolate 5763), the NAP_{CR1}/ST54 isolates showed between 3,301 and 8,226 bases in recombination (supplementary table 6, Supplementary Material online). Additionally, the WGS of the NAP_{CR1}/ST54 isolates 3150 and 5734 had one recombination block of 13 or 12 SNPs each and were therefore linked to *r*/*m* ratios >3 (supplementary table 6, Supplementary Material online). The NAP1/ST01 genomes, by contrast, did not have bases in recombination or recombination blocks.

Accessory Genome and Pangenome Comparisons for Assessment of NHR

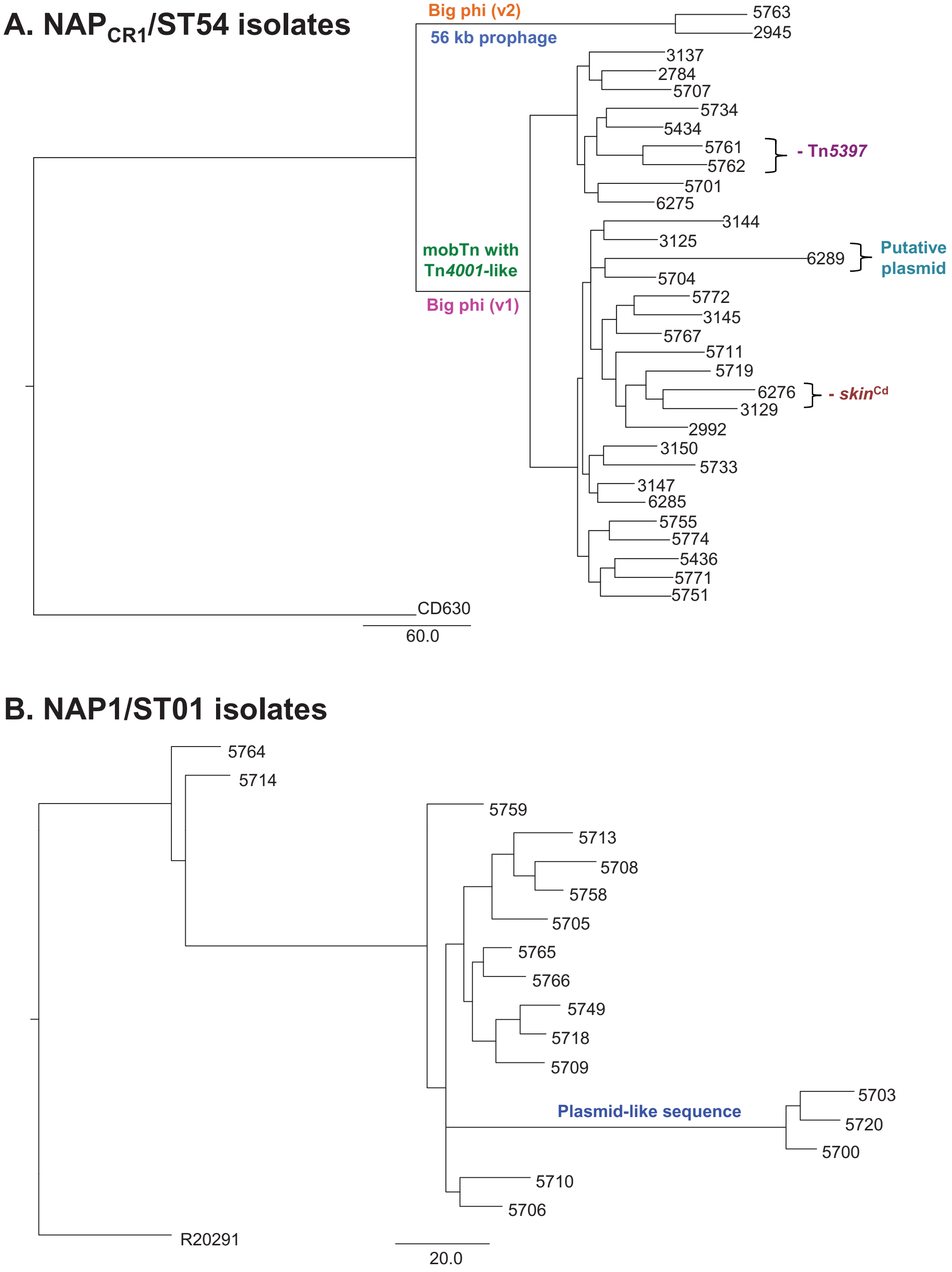
The NAP_{CR1}/ST54 genomes (4.50–4.62 Mb) were on an average 0.41 Mb larger than their NAP1/ST01 counterparts (4.09–4.18 Mb) (tables 1 and 2). Roary predicted 4,802 gene clusters for the NAP_{CR1}/ST54 isolates of which 74% were catalogued as core genome, 7% as soft-core genome, 8% as shell genome, and 11% as cloud genome (fig. 3*A*). In contrast, only 3,829 gene clusters and a core genome of 94% was predicted for the NAP1/ST01 isolates (fig. 3*B*). The shell and cloud genomes of this group of isolates only included 4.7% and 1.6% of the predicted gene clusters, respectively.

The root-to-tip distance of the NAP_{CR1}/ST01 isolates in a parsimony-based pangenomic tree was larger than that determined for isolates of the NAP1 pulsotype (fig. 4). Comparable results were observed in FFP-based trees (supplementary fig. 1, Supplementary Material online).

Based on the clustering of the isolates in the parsimony-based pangenomic tree, the accessory genomes of the isolates depicted with block arrows in figure 4 were studied in more detail with regard to the detection of unique gene clusters and therefore possible NHR. The selected NAP_{CR1}/ST54 isolates had many more unique gene clusters than the NAP1/ST01 isolates. In the NAP_{CR1} pulsotype, isolate 2945 from Cluster I showed the greatest number of unique gene clusters ($n = 376$), followed by isolates 6276 and 6289 from Cluster III ($n = 104$), and isolate 5761 from Cluster II ($n = 62$). Within the NAP1 genotype, isolate 5703 from Cluster V had the largest number of unique gene clusters ($n = 85$) (supplementary table 7, Supplementary Material online). All other representative NAP1 isolates only had between 10 and 17 unique gene clusters (supplementary table 7, Supplementary Material online).

Role of Differentially Distributed MGEs in Microdiversification

The majority of the unique genes of the NAP_{CR1}/ST54 isolates were associated with MGEs, which are absent in the closely related strain *C. difficile* 630 (table 3). The MGEs that were differentially represented among the NAP_{CR1}/ST54 isolates include: 1) a putative prophage of 56 kb in isolates with the *Sma*I pattern 487, 2) two putative big phages related to phiCDIF1296T (Wittmann et al. 2015), 3) a putative plasmid of 69 kb exclusively found in isolate 6289 from Cluster III, and 4) a mobilizable transposon similar to Tn*4001* not seen in isolates with the *Sma*I pattern 487 (Cluster I). Three NAP_{CR1}/ST54 isolates lacked two well-described MGEs from CD630 and other *C. difficile* genotypes, namely, isolate 6276 from Cluster III, which lacks the *skin*^{Cd} element, and isolates 5761 and 5762 from Cluster II, which do not encode Tn*5397* (fig. 5*A*) (Haraldsen and Sonenshein 2003; Dannheim et al. 2017). A very different picture was derived from the comparison of the NAP1/ST01 pangenomes, as only the isolates


A. NAP$_{CR1}$/ST54 isolates

B. NAP1/ST01 isolates

Fig. 5.—Localization of discriminative MGEs of NAP$_{CR1}$/ST54 (*A*) and NAP1/ST01 isolates (*B*) in unrooted, parsimony-based pangenomic trees. MGEs found in certain but not all isolates were highlighted with colors. The NAP$_{CR1}$/ST54 isolates from Cluster I were characterized by the carriage of a putative big phage (v2, orange) and a putative prophage of 56 kb (blue). Isolates from Clusters II and III have another type of big phage (v1, pink) and a predicted mobilizable transposon with a Tn*4001*-like element (green). Isolates 5761 and 5762 from Cluster II lack Tn*5397*. Moreover, isolate 6289 has a putative conjugative plasmid (teal) and isolate 6276 lacks the *skin*^{Cd} element (brown). Only the NAP1/ST01 isolates from Cluster V have a differentially distributed MGE. This element gave a perfect BLAST hit to an episomal sequence with bacteriophage functions previously found in the *Clostridiodes difficile* type strain DSM 1296^{T}. These trees were derived from binary matrixes summarizing the presence–absence of gene clusters in proteome predictions generated with Get_Homologues.

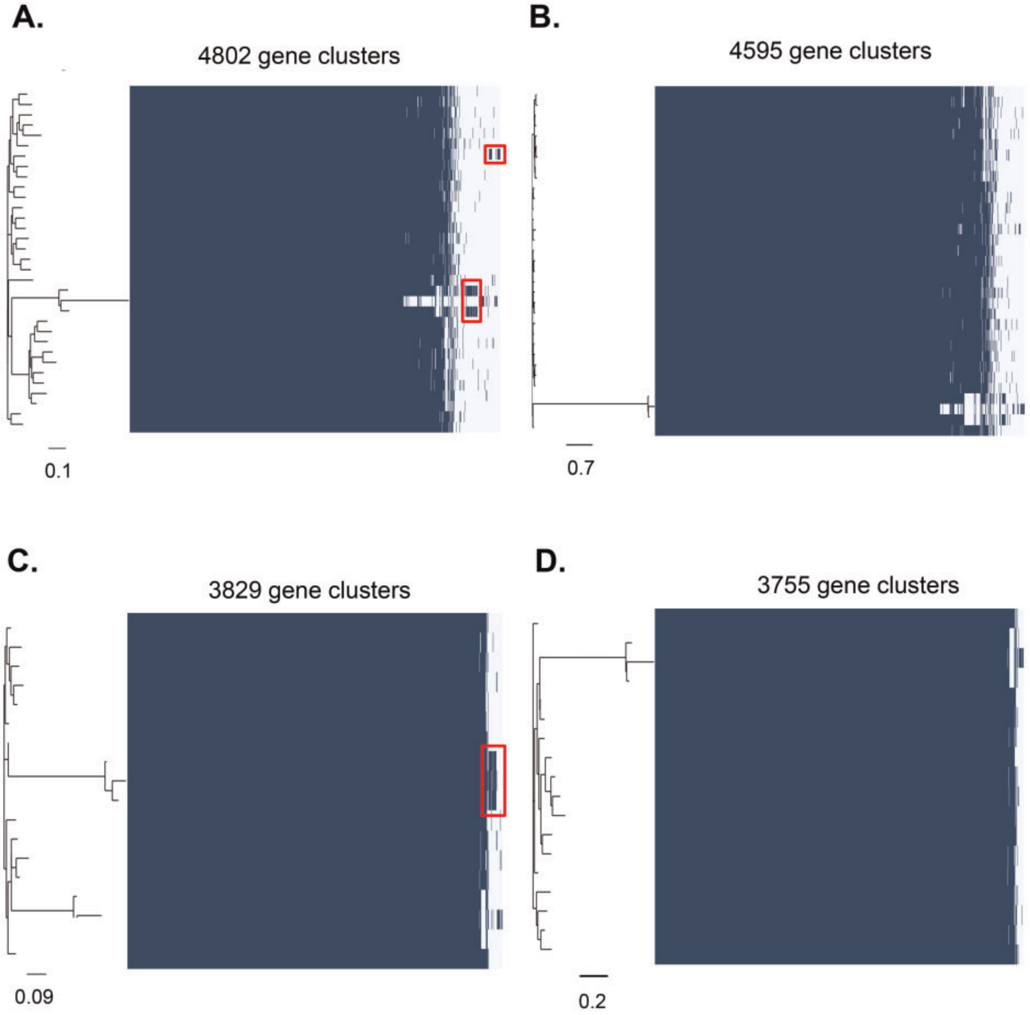


Fig. 6.—Roary analysis of WGS of NAP$_{CR1}$/ST54 and NAP1/ST01 strains with and without selected MGEs. (*A*) Original NAP$_{CR1}$/ST54 pangenome. (*B*) Pangenome analysis of NAP$_{CR1}$/ST54 genomes from which the putative plasmid, the two big phages, and the 56-kb prophage were removed. (*C*) Original NAP1/ST01 pangenome. (*D*) Pangenome analysis of NAP1/ST01 WGS lacking the putative plasmid-like sequence carrying bacteriophage genes. The trees show the clustering of isolates according to gene presence–absence matrixes. The blue and white bars represent shared and unshared gene clusters, respectively. Red squares delineate the gene clusters associated with the MGEs removed in the reanalysis. Tree distances were more notably reduced among the NAP$_{CR1}$ WGS when the differential MGE were eliminated.

5703, 5720, and 5700 from Cluster V had a distinctive MGE. This element is identical to a previously reported plasmid-like sequence of the *C. difficile* type strain DSM 1296^T (Riedel et al. 2015). The topology of the parsimony-based pangenomic trees mirrored the distribution of these MGEs in the data set (fig. 5*B*).

The MGEs that differentiate the clusters of NAP$_{CR1}$/ST54 isolates include genes linked to antibiotic resistance or virulence (supplementary table 8, Supplementary Material online). For instance, the putative conjugative plasmid of the NAP$_{CR1}$ isolate 6289 of Cluster III harbors a von Willebrand factor type A protein, a putative ADP-ribosyltransferase exoenzyme, and what seems to be a Fic/DOC toxin. Likewise, the mobTn with a Tn*4001*-like element and the 56 kb prophage inserted in some NAP$_{CR1}$ isolates, carry genes that likely confer resistance to aminoglycosides (Ramírez-Vargas et al. 2017).

When the sequences of the putative plasmid, the big bacteriophages, and the 56 kb prophage of the NAP$_{CR1}$/ST54 isolates were deliberatively removed from their draft WGS and the Roary pangenome calculations were repeated, the number of gene clusters in the NAP$_{CR1}$/ST54 WGS decreased 4% from 4,802 to 4,595 and, except for isolates from the 487 *Sma*I macrorestriction pattern, the branching of the resulting pangenomic tree collapsed (fig. 6, panels *A* and *B*). When this reanalysis was performed removing the putative plasmid-like sequence from the draft genomes of the NAP1/ST01 isolates 5700, 5703, and 5720, the number of predicted gene clusters was reduced by only 2%, from 3,829 to 3,755 (fig. 6, panels *C* and *D*).

Table 4
Gene Content Changes between the NAP_{CR1}/ST54 Isolates and the Reference Strain CD630

Isolate	PFGE *Sma*I Pattern	Gene Content Changes (*n*)	Average	Gene Content Changes Linked to Differentially Distributed MGEs (*n*)	Average	% of Gene Content Changes Linked to Differentially Distributed MGEs	Average
3147	442	649	649	346	346	53	53
5701	447	586	641	346	346	59	54
5711		663		346		52	
5767		657		346		53	
5771		658		346		53	
2784	448	599	624	346	346	58	56
3125		655		346		53	
3137		599		346		58	
5434		594		346		58	
5704		644		346		54	
5707		611		346		57	
5733		646		346		54	
5751		636		346		54	
5774		651		346		53	
6275		601		346		58	
3129	449	669	667	346	371	52	55
5719		657		346		53	
5755		640		346		54	
5772		642		346		54	
6276		664		346		52	
6289		730		494		68	
5734	452	604	604	346	346	57	57
2945	487	709	707	420	420	59	59
5763		704		420		60	
2992	488	664	664	346	346	52	52
5761	489	616	623	346	346	56	56
5762		629		346		55	
3145	558	646	644	346	346	54	54
6285		641		346		54	
3144	578	639	646	346	346	54	54
3150		654		346		53	
5436		645		346		54	
Average		644±33	644±33	355±31	355±31	55±3	55±3

Gene Content Changes and LGT to SNP Rates

When the WGS of the NAP_{CR1}/ST54 isolates were compared with the genome of the reference strain CD630, the number of gene content changes ranged between 586 and 730 (average: 644 ± 33) (table 4). Up to 55 ± 3% of this acquired genetic material ($n = 346$–494 CDS, 355 ± 31 on an average) was associated with the aforementioned discriminative MGEs (table 4). In agreement with this observation, the isolates that gained more genes (2945, 5763, 6289) had larger MGEs. A similar comparison of NAP1/ST01 isolates and the genome of the reference strain R20291 only revealed 68–194 gene content changes (average: 129 ± 37 CDS) (table 5).

As seen in tables 6 and 7, the NAP_{CR1}/ST54 isolates had a 4-fold higher average LGT/SNP rate (range: 20.7–33.9, average: 26.8 ± 2.8) than the NAP1/ST01 isolates (range: 3.4–11.0, average: 6.8 ± 2.0). Similar results were obtained when the calculation of LGT/SNP rates was restricted to gene content changes linked to the MGEs differentially distributed among both groups of strains (tables 6 and 7).

CRISPR Arrays

Based on the assumption that frequent exposure to MGEs will translate into a large diversity and number of CRISPR spacers, we compared the CRISPR-arrays of the NAP_{CR1}/ST54 isolates

Table 5

Gene Content Changes between the NAP1/ST01 Isolates and the Reference Strain R20291

Isolate	Gene Content Changes (*n*)	Gene Content Changes Linked to Differentially Distributed MGEs (*n*)	% of Gene Content Changes Linked to Differentially Distributed MGEs
5700	194	116	60
5703	184	116	63
5705	187	0	0
5706	118	0	0
5708	119	0	0
5709	120	0	0
5710	112	0	0
5713	119	0	0
5714	71	0	0
5718	116	0	0
5720	187	116	62
5749	121	0	0
5758	121	0	0
5759	125	0	0
5764	68	0	0
5765	117	0	0
5768	115	0	0
Average	129±37	20±46	10±24

and NAP1/ST01 isolates with those of the reference strains CD630 and R20291, respectively. The NAP$_{CR1}$/ST54 isolates had eight of the 12 CRISPR arrays of strain CD630 and showed spacer variations in the loci 3, 4, 5, 6, 7, 8, 9, and 12 (supplementary fig. 2*A*, Supplementary Material online). From the missing arrays, arrays 1 and 2 are reported to be present in MGEs (Sebaihia et al. 2006). On the contrary, the analyzed NAP1/ST01 isolates have the nine CRISPR arrays that characterize the reference strain *C. difficile* R20291 (supplementary fig. 2*B*, Supplementary Material online). In this data set, only isolates 5708 and 5709 deviated from the R20291 CRISPR profile, namely through to the lack of one spacer in locus 8 (supplementary fig. 2*B*, Supplementary Material online).

Discussion

Our results show that the acquisition/loss of MGEs and homologous recombination, rather than mutation, has had a stronger influence in the microdiversification of the NAP$_{CR1}$/ST54 isolates compared with the NAP1/ST01 isolates, which—as previously reported—is a pathogenic clone whose microdiversification is primarily driven by mutations in its core genome (He et al. 2010; Didelot et al. 2012) rather than by recombination (Dingle et al. 2011; Stabler et al. 2012).

The d*N*/d*S* rates calculated for the core genomes of both groups of bacteria were >1 with the NAP1/ST01 having the higher values. Rates >1 can be attributed to purifying selection not having enough time to eliminate deleterious changes, and is a phenomenon usually seen in closely related lineages (Rocha et al. 2006; Castillo-Ramírez et al. 2011). Thus, the higher rates calculated in the NAP1/ST01 group could represent the greater proximity between the isolates as compared with the NAP$_{CR1}$/ST54 group, and not neccesarily positive selection. However, it is possible that the large number of d*N* mutations detected among the NAP1/ST01 isolates reflects a greater effect of mutation in its diversfication (Rocha et al. 2006; Kryazhimskiy and Plotkin 2008). Mainly, when previous publications have already stated that the NAP1/ST01 lineage is clonal and microdiversifies through accumulation of mutations in the core genomes rather than recombination (Dingle et al. 2011; Stabler et al. 2012). In addition, positive selection, which is likely to be favored by fine tuned pathogenic strains, has been proposed for other outbreak-causing *C. difficile* strains from the ST37 from Clade IV (Dingle et al. 2011; Didelot et al. 2012). By contrast, the high number of d*S* mutations seen among the NAP$_{CR1}$/ST54 isolates might have derived from unnoticed recombination events (Castillo-Ramírez et al. 2011).

In both groups of isolates, we identified SNPs that are noteworthy due to their potential influence on virulence or the regulation of virulence-related phenotypes. In particular, there were SNPs in the precursor of the S-layer protein SlpA, which is related to bacterial adhesion and immune response, and in putative exosporium proteins, which protect the spores in aerobic environments outside of the host as well as from the host immune system (Merrigan et al. 2013; Paredes-Sabja et al. 2014). In addition, we observed SNPs in genes related to the carbohydrate phosphotransferase system PTS, which is relevant for toxin production through catabolite repression (Martin-Verstraete et al. 2016).

Table 6
LGT/SNP Rates Calculated for the NAP_{CR1}/ST54 Isolates

Isolate	PFGE *Sma*I Pattern	LGT/SNP Rate[a]	Average	MGE-Driven LGT/SNP Rate[b]	Average
3147	442	27.0	27.0	14.4	14.4
5701	447	20.9	26.7	12.4	14.4
5711		28.8		15.0	
5767		28.6		15.0	
5771		28.6		15.0	
2784	448	26.0	26.0	15.0	14.4
3125		29.8		15.7	
3137		26.0		15.0	
5434		24.8		14.4	
5704		25.8		13.8	
5707		25.5		14.4	
5733		25.8		13.8	
5751		27.7		15.0	
5774		28.3		15.0	
6275		20.7		11.9	
3129	449	30.4	27.7	15.7	15.4
5719		25.3		13.3	
5755		27.8		15.0	
5772		25.7		13.8	
6276		26.6		13.8	
6289		30.4		20.6	
5734	452	25.2	25.2	14.4	14.4
2945	487	29.5	28.9	17.5	17.2
5763		28.2		16.8	
2992	488	28.9	28.9	15.0	15
5761	489	23.7	22.7	13.3	12.6
5762		21.7		11.9	
3145	558	25.8	25.2	13.8	13.6
6285		24.7		13.3	
3144	578	29.0	29.7	15.7	15.9
3150		26.2		13.8	
5436		33.9		18.2	
Average		26.8±2.8	26.8±2.8	14.8±1.7	14.8±1.7

[a] Defined as number of gene content changes with respect to strain CD630/number SNPs identified by Breseq.

[b] Defined as number of gene content changes linked to differentially distributed MGEs/number SNPs identified by Breseq.

Previous research has claimed that some sequence types from the Clade I microdiversify through homologous recombination (Stabler et al. 2012). For instance, Didelot et al. (2012) determined a higher *r/m* ratio for ST54 isolates from other geographic regions (2.54) than for ST01 isolates (0.04). Given that no bases in recombinations were detected for the NAP1/ST01 isolates, that the NAP_{CR1}/ST54 isolates had different amounts of bases in recombination, and that two of the NAP_{CR1}/ST54 isolates had an unshared recombination block, our results coincide with these previous reports. We therefore conclude that the effect of homologous recombination in microdiversification was greater for the NAP_{CR1}/ST54 isolates than for the NAP1/ST01 isolates.

The NAP_{CR1}/ST54 isolates gained more CDS, obtained a larger number of CDS through acquisition of differentially distributed MGEs, and were characterized by higher LGT/SNP rates than the NAP1/ST01 isolates. This indicates that the NAP_{CR1}/ST54 isolates are more prone than the NAP1/ST01 isolates to acquire genetic information by LGT. This trait is expected for organisms that thrive in heterogeneous and changing conditions, hence it seems likely that the NAP_{CR1}/ST54 and NAP1/ST01 strains take advantage of distinctive strategies to adapt and colonize the human gut and cause disease and/or outbreaks (Rouli et al. 2015; McInerney et al. 2017). Further supporting the concept that the NAP_{CR1} pangenome is open, the NAP_{CR1}/ST54 isolates were not distributed in the branches of a pangenomic tree according to their macrorestriction patterns or hospital/year of isolation. Instead, the topology of this tree was dictated by the gain or loss of MGEs that included most of the unique gene clusters.

We acknowledge that the disparity in the number of isolates from each genotype can affect our pangenome

Table 7

LGT/SNP Rates Calculated for the NAP1/ST01 Isolates

Isolate	LGT/SNP Rate[a]	MGE-Driven LGT/SNP Rate[b]
5700	9.7	5.8
5703	8.8	5.5
5705	11.0	NA
5706	5.6	NA
5708	6.3	NA
5709	6.3	NA
5710	5.9	NA
5713	5.7	NA
5714	3.7	NA
5718	6.1	NA
5720	9.4	5.8
5749	6.7	NA
5758	7.1	NA
5759	6.6	NA
5764	3.4	NA
5765	6.2	NA
5768	6.8	NA
Average	6.8±2.0	5.7±0.2

Note.—NA,not applicable.

[a] Defined as number of gene content changes with respect to strain R20291/ number SNPs identified by Breseq.

[b] Defined as number of gene content changes linked to differentially distributed MGEs/number SNPs identified by Breseq.

estimations. However, it is unlikely that the size of the NAP1/ST01 pangenome calculated for our isolates will depart from that of the global NAP1 population, as indicated by the lower SNP counts, the very high percentage of reads that mapped to the reference genome selected, and the already recognized clonality of this strain (Stabler et al. 2006, 2009).

MGEs are generally unstable and tend to be eliminated to reduce their burden (Karcagi et al. 2016; McInerney et al. 2017), yet under some circumstances greater pangenomes and the acquisition of MGEs provide advantageous traits for certain bacterial species (Vos et al. 2015; McInerney et al. 2017). Five of the differential MGEs found among the $\mathrm{NAP_{CR1}}$/ST54 isolates are absent in the closely related *C. difficile* strain 630, suggesting that the biological differences between this reference strain and the more virulent $\mathrm{NAP_{CR1}}$ genotype could be due to laterally transferred DNA (Quesada-Gómez et al. 2015). Although these MGEs await functional characterization, we hypothesize that they are mobilizable or conjugative based on the predicted functions of some of their genes.

Our data confirm the enhanced capability of the $\mathrm{NAP_{CR1}}$/ST54 isolates to acquire MGEs and explains the large size of the pangenomes of this clade. This feature is not fully understood, although it could be related to the accuracy and efficiency of restriction-modification systems, CRISPR-Cas systems, and DNA repair mechanisms to cite possible mechanisms (Darmon and Leach 2014). Whether the NAP1/ST01 isolates have active barriers for LGT that are absent in the $\mathrm{NAP_{CR1}}$ isolates remains to be determined.

Our results demonstrate that highly virulent, outbreak-causing *C. difficile* strains from two different ST groups and MLST clades microdiversify through different mechanisms and emphasize the importance of MGE as drivers of bacterial diversification also for ST54 isolates. Future studies addressing the evolution of *C. difficile* should consider the role of MGEs and the pangenome along with investigations of the core genome because accessory genes may mediate clinically relevant phenotypes such as antimicrobial resistance and virulence. We also acknowledge that the genomic plasticity of the $\mathrm{NAP_{CR1}}$/ST54 isolates poses a threat, as it suggests that MGE gain/loss events may lead to the emergence of non-NAP1 lineages with increased virulence and outbreak potential that cannot be distinguished from ordinary strains through MLST, ribotyping, or core genome-based typing.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Robin Cárdenas-Mora and Pablo Vargas-Dengo from the University de Costa Rica for their technical support, and Gabriel González for providing us with the R code used to generate figure 2. We also thank Boyke Bunk, Isabel Schober, Cathrin Spröer, Simone Severitt, and Nicole Heyer at the Leibniz Institute DSMZ for helpful advices and technical assistance. This work was supported by the Costa Rican Ministry of Science Technology and Telecommunications of Costa Rica (FORINVES grant 803-B4-510; T.M., G.R., C.G.V., E.C.O., and C.R.), the Vice Rectorate for Research of the University of Costa Rica (Recdif network 803-B5-77; T.M., G.R., C.G.V., E.C.O., and C.R.), the Federal State of Lower Saxony, Niedersächsisches Vorab (VWZN2889/3215; T.R. and J.O.) and the German Center for Infection Research (DZIF) (BMBF grant 8000-105-3; T.R. and J.O.).

Literature Cited

Andersen JM. 2016. CRISPR diversity and microevolution in *Clostridium difficile*. Genome Biol Evol. 8(9):2841–2855.

Barrick JE, et al. 2014. Identifying structural variation in haploid microbial genomes from short-read resequencing data using breseq. BMC Genomics 15:1039.

Braun V, Hundsberger T, Leukel P, Sauerborn M, von Eichel-Streiber C. 1996. Definition of the single integration site of the pathogenicity locus in *Clostridium difficile*. Gene 181(1–2):29–38.

Calabi E, et al. 2001. Molecular characterization of the surface layer proteins from *Clostridium difficile*. Mol Microbiol. 40(5):1187–1199.

Carver TJ, et al. 2005. ACT: the Artemis comparison tool. Bioinformatics 21(16):3422–3423.

Castillo-Ramírez S, et al. 2011. The impact of recombination on dN/dS within recently emerged bacterial clones. PLoS Pathog. 7(7):e1002129.

Chaves-Olarte E, Weidmann M, Eichel-Streiber C, Thelestam M. 1997. Toxins A and B from *Clostridium difficile* differ with respect to enzymatic potencies, cellular substrate specificities, and surface binding to cultured cells. J Clin Invest. 100(7):1734–1741.

Contreras-Moreira B, Vinuesa P. 2013. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol. 79(24):7696–7701.

Croucher NJ, et al. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 43(3):e15–e15.

Dagan T, Artzy-Randrup Y, Martin W. 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc Natl Acad Sci U S A. 105(29):10039–10044.

Dannheim H, et al. 2017. Manual curation and reannotation of the genomes of *Clostridium difficile* 630Δ*erm* and *C. difficile* 630. J Med Microbiol. 66(3):286–293.

Darmon E, Leach DRF. 2014. Bacterial genome instability. Microbiol Mol Biol Rev. 78(1):1–39.

Davis S, et al. 2015. CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data. PeerJ Comput Sci. 1:e20.

Didelot X, et al. 2012. Microevolutionary analysis of *Clostridium difficile* genomes to investigate transmission. Genome Biol. 13(12):R118.

Dingle KE, et al. 2011. Clinical *Clostridium difficile*: clonality and pathogenicity locus diversity. PLoS One 6(5):e19993.

Finn RD, et al. 2017. InterPro in 2017 – beyond protein family and domain annotations. Nucleic Acids Res. 45(D1):D190–D199.

Gish W, States DJ. 1993. Identification of protein coding regions by database similarity search. Nat Genet. 3(3):266–272.

Goulding D, et al. 2009. Distinctive profiles of infection and pathology in hamsters infected with *Clostridium difficile* strains 630 and B1. Infect Immun. 77(12):5478–5485.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. 27(2):221–224.

Gupta A, Khanna S. 2014. Community-acquired *Clostridium difficile* infection: an increasing public health threat. Infect Drug Resist. 7:63–72.

Guttman D, Dykhuizen D. 1994. Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. Science 266(5189): 1380–1383.

Hanage WP. 2016. Not So Simple After All: bacteria, Their Population Genetics, and Recombination. Cold Spring Harb Perspect Biol. 8(7):a018069.

Haraldsen JD, Sonenshein AL. 2003. Efficient sporulation in *Clostridium difficile* requires disruption of the sigK gene. Mol Microbiol. 48(3):811–821.

He M, et al. 2010. Evolutionary dynamics of *Clostridium difficile* over short and long time scales. Proc Natl Acad Sci U S A. 107(16):7527–7532.

Hehemann JH, et al. 2016. Adaptive radiation by waves of gene transfer leads to fine-scale resource partitioning in marine microbes. Nat Commun. 7:12860.

Hernandez D, François P, Farinelli L, Osterås M, Schrenzel J. 2008. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res. 18(5):802–809.

Hunt JJ, Ballard JD. 2013. Variations in virulence and molecular biology among emerging strains of Clostridium difficile. Microbiol Mol Biol Rev. 77(4):567–581.

Just I, et al. 1995. Glucosylation of Rho proteins by *Clostridium difficile* toxin B. Nature 375(6531):500–503.

Karcagi I, et al. 2016. Indispensability of Horizontally Transferred Genes and Its Impact on Bacterial Genome Streamlining. Mol Biol Evol. 33(5):1257–1269.

Knight DR, Elliott B, Chang BJ, Perkins TT, Riley TV. 2015. Diversity and Evolution in the Genome of *Clostridium difficile*. Clin Microbiol Rev. 28(3):721–741.

Kryazhimskiy S, Plotkin JB. 2008. The Population Genetics of dN/dS. PLoS Genet. 4(12):e1000304.

Lemée L, et al. 2005. Multilocus sequence analysis and comparative evolution of virulence-associated genes and housekeeping genes of *Clostridium difficile*. Microbiology 151(Pt 10):3171–3180.

Lessa FC, et al. 2015. Burden of *Clostridium difficile* Infection in the United States. N Engl J Med. 372(9):825–834.

López-Ureña D, et al. 2016. Predominance and high antibiotic resistance of the emerging *Clostridium difficile* genotypes NAPCR1 and NAP9 in a Costa Rican hospital over a 2-year period without outbreaks. Emerg Microbes Infect. 5(5):e42.

Martin-Verstraete I, Peltier J, Dupuy B. 2016. The Regulatory Networks That Control *Clostridium difficile* Toxin Synthesis. Toxins (Basel) 8(12):153.

McGlone SM, et al. 2012. The economic burden of *Clostridium difficile*. Clin Microbiol Infect. 18(3):282–289.

McInerney JO, McNally A, O'Connell MJ. 2017. Why prokaryotes have pangenomes. Nat Microbiol. 2:17040.

Merrigan MM, et al. 2013. Surface-Layer Protein A (SlpA) Is a Major Contributor to Host-Cell Adherence of *Clostridium difficile*. PLoS One 8(11):e78404–e78412,

Mugal CF, Wolf JBW, Kaj I. 2014. Why Time Matters: codon Evolution and the Temporal Dynamics of dN/dS. Mol Biol Evol. 31(1):212–231.

Page AJ, et al. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31(22):3691–3693.

Paredes-Sabja D, Shen A, Sorg J. a. 2014. *Clostridium difficile* spore biology: sporulation, germination, and spore structural proteins. Trends Microbiol. 22(7):406–416.

Perelle S, Gibert M, Bourlioux P, Corthier G, Popoff MR. 1997. Production of a complete binary toxin (actin-specific ADP-ribosyltransferase) by *Clostridium difficile* CD196. Infect Immun. 65(4):1402–1407.

Quesada-Gómez C, et al. 2010. Emergence of *Clostridium difficile* NAP1 in Latin America. J Clin Microbiol. 48(2):669–670.

Quesada-Gómez C, et al. 2015. Emergence of an outbreak-associated *Clostridium difficile* variant with increased virulence. J Clin Microbiol. 53(4):1216–1226.

Ramírez-Vargas G, et al. 2017. A *Clostridium difficile* lineage endemic to Costa Rican hospitals is multidrug-resistant by acquisition of chromosomal mutations and novel mobile genetic elements. Antimicrob Agents Chemother. 61:e02054-16.

Reynolds CB, Emerson JE, de la Riva L, Fagan RP, Fairweather NF. 2011. The *Clostridium difficile* cell wall protein CwpV is antigenically variable between strains, but exhibits conserved aggregation-promoting function. PLoS Pathog. 7(4):e1002024.

Riedel T, et al. 2015. Complete Genome Sequence of the *Clostridium difficile* Type Strain DSM 1296T. Genome Announc. 3:e01186–e01115.

Roberts AP, Allan E, Mullany P. 2014. The impact of horizontal gene transfer on the biology of *Clostridium difficile*. In: Poole, RK, editor. Advances in microbial physiology. Vol. 65. Amsterdam: Elsevier Ltd. p. 63–82.

Rocha EPC, et al. 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol. 239(2):226–235.

Rouli L, Merhej V, Fournier P-E, Raoult D. 2015. The bacterial pangenome as a new tool for analyzing pathogenic bacteria. New Microbes New Infect. 7:72–85.

Schwan C, et al. 2009. *Clostridium difficile* toxin CDT induces formation of microtubule-based protrusions and increases adherence of bacteria. PLoS Pathog. 5(10):e1000626.

Sebaihia M, et al. 2006. The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. Nat Genet. 38(7):779–786.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30(14):2068–2069.

Sims GE, Kim S-H. 2011. Whole-genome phylogeny of *Escherichia coli/Shigella* group by feature frequency profiles (FFPs). Proc Natl Acad Sci U S A. 108(20):8329–8334.

Slimings C, Riley TV. 2014. Antibiotics and hospital-acquired *Clostridium difficile* infection: update of systematic review and meta-analysis. J Antimicrob Chemother. 69(4):881–891.

Stabler RA, et al. 2006. Comparative phylogenomics of *Clostridium difficile* reveals clade specificity and microevolution of hypervirulent strains. J Bacteriol. 188:7297–7305.

Stabler RA, et al. 2009. Comparative genome and phenotypic analysis of *Clostridium difficile* 027 strains provides insight into the evolution of a hypervirulent bacterium. Genome Biol. 10:R102.

Stabler RA, et al. 2012. Macro and micro diversity of *Clostridium difficile* isolates from diverse sources and geographical locations. PLoS One 7:1–12.

Vos M, Didelot X. 2009. A comparison of homologous recombination rates in bacteria and archaea. ISME J. 3(2):199–208.

Vos M, Hesselman MC, te Beek TA, van Passel MWJ, Eyre-Walker A. 2015. Rates of Lateral Gene Transfer in Prokaryotes: high but Why? Trends Microbiol. 23(10):598–605.

Wittmann J, et al. 2015. Complete Genome Sequence of the Novel Temperate *Clostridium difficile* Phage phiCDIF1296T. Genome Announc. 3:e00839-15.

Wong-McClure R. a, et al. 2012. *Clostridium difficile* outbreak in Costa Rica: control actions and associated factors. Rev Panam Salud Publica 32(6):413–418.

Zerbino DR. 2010. Using the Velvet *de novo* assembler for short-read sequencing technologies. Curr Protoc Bioinformatics 18:1–13.

Associate editor: Tal Dagan