CRISPR Diversity and Microevolution in *Clostridium difficile*

Joakim M. Andersen[1], Madelyn Shoup[2], Cathy Robinson[2], Robert Britton[3], Katharina E.P. Olsen[4], and Rodolphe Barrangou[1,*]

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

[2]Department of Microbiology and Molecular Genetics, Michigan State University, MI

[3]Department of Molecular Virology and Microbiology, Center for Metagenomics and Microbiome Research, Baylor College of Medicine, TX

[4]Microbial Competence Centre, Novo Nordisk, Bagsværd, Denmark (Former Employment: Department of Microbiology & Infection Control, Statens Serum Institut, Copenhagen, Denmark)

*Corresponding author: E-mail: rbarran@ncsu.edu.

Accepted: August 12, 2016

Abstract

Virulent strains of *Clostridium difficile* have become a global health problem associated with morbidity and mortality. Traditional typing methods do not provide ideal resolution to track outbreak strains, ascertain genetic diversity between isolates, or monitor the phylogeny of this species on a global basis. Here, we investigate the occurrence and diversity of clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated genes (*cas*) in *C. difficile* to assess the potential of CRISPR-based phylogeny and high-resolution genotyping. A single Type-IB CRISPR-Cas system was identified in 217 analyzed genomes with *cas* gene clusters present at conserved chromosomal locations, suggesting vertical evolution of the system, assessing a total of 1,865 CRISPR arrays. The CRISPR arrays, markedly enriched (8.5 arrays/genome) compared with other species, occur both at conserved and variable locations across strains, and thus provide a basis for typing based on locus occurrence and spacer polymorphism. Clustering of strains by array composition correlated with sequence type (ST) analysis. Spacer content and polymorphism within conserved CRISPR arrays revealed phylogenetic relationship across clades and within ST. Spacer polymorphisms of conserved arrays were instrumental for differentiating closely related strains, e.g., ST1/RT027/B1 strains and pathogenicity locus encoding ST3/RT001 strains. CRISPR spacers showed sequence similarity to phage sequences, which is consistent with the native role of CRISPR-Cas as adaptive immune systems in bacteria. Overall, CRISPR-Cas sequences constitute a valuable basis for genotyping of *C. difficile* isolates, provide insights into the micro-evolutionary events that occur between closely related strains, and reflect the evolutionary trajectory of these genomes.

Key words: CRISPR-Cas, CRISPR phylogeny, CRISPR typing, BI/NAP1/RT027/ST1Type-IB.

Introduction

Epidemic *Clostridium difficile* outbreaks have increased over the last two decades, both qualitatively in terms of increased virulence, and quantitatively in terms of the number of documented cases, posing an increasingly significant healthcare threat and medical resources burden (Dubberke and Olsen 2012; Mergenhagen et al. 2014). Indeed, *C. difficile* is the most common cause of antibiotic-associated diarrhea and can lead to more serious complications such as pseudomembranous colitis and ileus (Boyanova et al. 2015). Furthermore, 10-20% of patients that contract *C. difficile* infection will have recurring episodes of the disease that can, in some cases, last for years (Ofosu 2016). The main determining factor for *C. difficile* pathogenicity is the presence of the *tcdA* and *tcdB* genes (encoding Toxin A and B) encoded in a pathogenicity locus (PaLoc) together with the binary toxin, encoded by *cdtA* and *cdtB* in a discrete locus, associated with increased host mortality (Bacci et al. 2011). Insight into the evolution of the PaLoc through genomics analysis showed how the PaLoc occurs and varies throughout the species, facilitated by both horizontal transfer and genetic reorganization (Dingle et al. 2014). The complexity and diversity of *C. difficile* pose a significant challenge for efficient typing of clinical isolates, especially with regards to associating genotypes with virulence factors and clinical outcome (Sirard et al. 2011; Knetsch et al. 2013).

Assessing the genetic diversity of epidemic strains is critical for understanding the phylogenetic distribution

© The Author(s) 2016. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

of causative agents, monitoring outbreaks, and devising management strategies. Currently, clinical isolates are routinely being classified by pulse-field gel electrophoresis, PCR ribotyping or multilocus sequence typing (MLST) to group related strains and track outbreak isolates (Kuijper et al. 2009). Recently, we have observed an increase in the availability of whole and draft *C. difficile* genomes and sequence-based typing methods. PCR-based ribotyping and MLST have established phylogenetic lineages useful for large-scale analysis (Knetsch et al. 2012) and to display the distribution of *C. difficile* isolates from various locations and sources (Stabler et al. 2012). Yet, even within regional investigations *C. difficile* transmissions were found to be originating from diverse sources (Eyre et al. 2013). Whole genome sequencing of reference and outbreak strain collections (He et al. 2013; Knetsch et al. 2014; Cairns et al. 2015) have become the "gold standard" for strain-typing resolution, but this approach is still inaccessible to most laboratories and lacks the expediency and convenience required for routine screens (Huber et al. 2013). Furthermore, analyses are complex and compounded by the fact that *C. difficile* has a relatively low core genome SNP occurrence rate among clinical isolates (Knetsch et al. 2014; Steglich et al. 2015). Currently, PCR ribotyping and MLST are the most widely used technique but virulence cannot be strictly ascribed to specific PCR and sequence-types, emphasizing the need for novel, high-resolution typing approaches that correlate with virulence (e.g., toxins and antibiotic resistance), transmission and source attribution (Smits 2013).

For bacterial phylogenetic analysis, CRISPR-Cas [Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and CRISPR-associated sequences (*cas*)] systems constitute valuable genetic targets for high-resolution typing and micro-evolution studies in some bacteria, including human pathogens (Shariat and Dudley 2014). CRISPR-Cas immune systems provide adaptive resistance against invasive genetic elements such as phages (Barrangou et al. 2007) and plasmids (Marraffini and Sontheimer 2008). These loci consist of repeat-spacer arrays composed of partial palindromic repeats interspersed by unique DNA sequences that are derived from foreign genetic elements in a linear, time-resolved manner. This can be exploited to gaze into the time-series of exposures of a strain to invasive elements, establish phylogenetic relationships between strains that share common ancestral spacers, and display divergent evolutionary paths over time through unique vaccination events. Thus, the ordinal and divergent nature of CRISPR arrays makes them valuable for typing applications, especially in species with active CRISPR-Cas systems.

CRISPR-Cas typing has especially proved efficient for outbreak investigations of bacterial pathogens where other typing methods may be insufficient to differentiate highly clonal isolates. Detailed analyses using CRISPR-Cas genotyping have already led to outbreak tracking of, e.g., *Yersinia pestis* (Cui et al. 2008; Barros et al. 2014) and *Salmonella enterica* subsp. *enterica* (Timme et al. 2013; Pettengill et al. 2014). Of note, CRISPR genotypes can also provide insights into important phenotypes with which they correlate, such as the occurrence of antibiotic-resistance cassettes in enterococci (Palmer and Gilmore 2010), or the presence of prophages in *Streptococcus pyogenes* genomes (Nozawa et al. 2011). These correlations reflect the role of CRISPR-Cas systems in controlling horizontal gene transfer, and the uptake and dissemination of particular genes and operons involved in bacterial adaptation and pathogenesis (Louwen et al. 2014), and hence the specific species evolution.

The presence of CRISPR-Cas loci in *C. difficile* has been documented (Sebaihia et al. 2006), yet the widespread occurrence and diversity has not been systematically explored in the species despite previous genomic studies (He et al. 2013; Mullany et al. 2015). The *C. difficile* species is known to harbor prophages (Sekulovic et al. 2014), and phage infections are known to impact the behavior of the host cell (Sekulovic and Fortier 2014), whereas other mobile genetic elements are likewise known to be widespread and may confer antibiotic resistance (Wasels et al. 2014; Amy et al. 2015). Recently the *C. difficile* CRISPR-Cas system in the strain 630 (ST54/RT012) was shown to exhibit CRISPR spacer sequence similarity to *C. difficile* phages and plasmids, suggesting CRISPR interference against these mobile elements (Hargreaves et al. 2014). Later this observation was experimentally confirmed in the R20291 strain (ST1/RT027/B1) (Boudry et al. 2015) indicating how clinically important *C. difficile* strains encode active CRISPR-Cas systems. Because these loci generate mature interfering crRNAs (Soutourina et al. 2013; Boudry et al. 2015), they may be functionally active and involved in canonical CRISPR-encoded immunity, and thus may also be instrumental in CRISPR-based phylogenetic analysis. This sets the stage for analysis of the CRISPR-Cas systems within this important pathogenic species in terms of occurrence, abundance and phylogenetic diversity, to assess its genotyping potential. The aim of our study was to systematically investigate the occurrence and diversity of CRISPR-Cas systems in *C. difficile* genomes across clades and sequence types and use this information for species-wide phylogenetic analyses. Specifically, our objective was to investigate potential conservation of loci and diversity of spacers to explore the genotyping potential of CRISPR sequences in both distant and closely related *C. difficile* strains.

Materials and Methods

Sequence Collection

Deposited whole and draft *C. difficile* genomes sequences were downloaded from the NCBI and Wellcome Trust Sanger institute databases (November 2015). Genome

analysis was performed in Geneious (Biomatters) and whole genome alignment was performed using progressive MAUVE (Darling et al. 2004). Clostridia phage sequences were obtained from the EBI database. Quality control of *C. difficile* genome assemblies was performed to exclude genomes originating from single-end short reads as these genomes showed a high tendency to lack or have misassembled CRISPR arrays, likely due to the abundance of repeats and nature of CRISPR-Cas systems.

Multi-Locus Sequence Typing

In silico MLST of available *C. difficile* genomes was performed using the method by Griffiths et al. (2010). The typing scheme utilizes seven regions within conserved household genes (*adk, atpA, dxr, glyA, recA, soda, tpi*) and known sequences for each typing region was obtained from MLST database (www.pubmlst.org/cdifficile, last accessed June 2016). The sequence collection for each allele was aligned using MAFFT (PAM1K = 1 and default settings) and the resulting consensus sequence was generated. The consensus sequence for each gene was subsequently used to identify sequence homologs in all available genomes by a cutoff of minimum 90% DNA sequence identity. Identified loci were then extracted from the genomes and assigned alleles, sequence type (ST) and clade designations by the homology search tool available at the MLST database (www.pubmlst.org/cdifficile), if all seven loci could be identified.

The identified alleles were extracted from the analyzed genomes encoding all seven loci, ordered and concatenated. The resulting combined alleles were aligned using MAFFT with the aforementioned settings. For phylogenetic analysis a Neighbor-Joining tree was constructed using MEGA6 (Tamura et al. 2013) and tested with 1,000 bootstrap replications.

Prediction of CRISPR-Cas Loci

CRISPR spacer arrays were predicted using the CRISPR Recognition Tool (Bland et al. 2007) with the following setting: minimum repeats per array 3, repeat length 29–30 nt, spacer length 19–48 and search window 9 nt. A more inclusive CRISPR array prediction was attempted using a repeat length query of 28–40 nt but only yielded false-positive random repeat loci when manually inspected. In parallel, *cas* genes were predicted by similarity searching using the *C. difficile* strain 630 *cas* genes as query sequences and a minimum of 75% DNA sequence identity. Two *cas* gene clusters were annotated in strain 630 and we annotated them as either group A (*cas3*, *cas5*, *cas7*, *cas8* and *cas6* encoded as locus tag: CD630_24510, CD630_24520, CD630_24530, CD630_24540 and CD630_24550, respectively) or group B (*cas2*, *cas1*, *cas4*, *cas3*, *cas5*, *cas7*, *cas8* and *cas6* encoded as locus tag: CD630_29750, CD630_29760, CD630_29770, CD630_29780, CD630_29790, CD630_29800, CD630_29810 and CD630_29820, respectively).

Identification and Clustering of Ancestral CRISPR Spacers

CRISPR arrays conserved among the tested strains were identified by annotating each ancestral spacer in the collection of predicted CRISPR arrays. The curation was done in several iterations until ancestral spacers were identified in >90% of all CRISPR arrays and hence accounting for most CRISPR arrays. Lastly, any annotated ancestral spacers that were encoded elsewhere in other arrays were removed to create a nonredundant collection of unique ancestral spacers, per genome. The occurrence of each ancestral spacer per genome was then utilized to create a binary matrix, composed of the 217 genomes versus the 110 identified ancestral spacers, which was subjected to two-way hierarchical clustering analysis (Fast Ward) using JMP genomics (SAS).

Visualization of CRISPR Arrays

Visual representation of CRISPR arrays was done as previously described (Horvath et al. 2008) using conserved spacers as anchoring points to compare CRISPR arrays across genomes. In more detail, for each array the repeat sequences were removed and the list of spacers was oriented with the ancestral spacer on the right hand side. Each spacer within the array was visually represented by a box, as previously established (Horvath et al. 2008). This allowed comparison of conserved arrays by aligning spacers from the ancestral end of the arrays. Conservation among arrays was thus estimated from shared spacers, where distantly related arrays shared fewer spacers towards the ancestral end and none at the leader end of the arrays. This enabled grouping of CRISPR arrays by spacer order and extent of shared spacers among the listed arrays. Endogenous spacer deletions were estimated from array comparison where arrays shared both ancestral and recently acquired leader end spacers to anchor the spacer sequences within the arrays. Thus, by having shared origin, a spacer deletion would suggest a recent diversification of the two arrays.

Identification of CRISPR Protospacer Matches in Foreign and Chromosomal DNA

The source of *C. difficile* CRISPR spacers was analyzed by identification of protospacers in available clostridia phage sequences. All spacer sequences from one representative genome per ST were used for homology searching to find potential protospacers with >90% sequence identity. A table of the number of spacer matches (protospacers) for each genome in each phage was constructed and hierarchical clustering analysis (Complete clustering algorithm) was performed.

Results

Identification of CRISPR-Cas Systems in *C. difficile*

We investigated the presence and diversity of CRISPR-Cas systems in a collection of 217 *C. difficile* genomes (supplementary table S1, Supplementary Material online) to establish the utility for phylogenetic analysis and potential for resolution of CRISPR-based genotyping.

To assess the overall phylogenetic diversity of the species, we applied the multi-locus sequence typing (MLST) method by Griffiths et al. (2010) to identify clades and sequence types (ST) in *C. difficile* (fig. 1). From the analysis (supplementary table S1, Supplementary Material online), we identified 33 STs spanning clade 1 (148 isolates), clade 2 (15 isolates), clade 3 (1 isolate), clade 4 (8 isolates), clade 5 (3 isolates) and the recently defined clade C-1 (4 isolates). A total of 38 strains could not be assigned a ST nor a clade, as not all seven alleles were identified, or the MLST produced a novel allele combination yet to be assigned a ST number. The distribution of STs showed a clear grouping of the five main clades, coinciding with previous core genome phylogenetic analysis (Dingle et al. 2014), thus providing a basis for analysis between and within ST groups and clades of clinical importance otherwise poorly resolved by MLST (e.g., clade 1).

We showed the occurrence and diversity of CRISPR-Cas systems based on the polythetic nomenclature of CRISPR-Cas systems (Makarova et al. 2011, 2015), and system type and sub-type were determined using the sequence and arrangement of universal and signature *cas* genes. A Type-IB system (Makarova et al. 2013, 2015) was observed in all queried genomes of *C. difficile*, which is consistent with earlier reports (He et al. 2010; Hargreaves et al. 2014; Boudry et al. 2015). Through our analysis, we identified two majorly conserved clusters of *cas* genes in the queried genomes (supplementary fig. S1, Supplementary Material online). The first *cas* gene cluster, termed *casA* (minimum nucleotide identity 87% across genomes based on 630 sequences), appeared to encode a truncated *cas* gene cluster lacking *cas1*, *cas2* and *cas4*. Another cluster, *casB*, encoded the full set of Type-IB *cas* genes (*cas1*–*cas8*; minimum sequence identity 89%). Interestingly, *casA* was present in all genomes except the distant/diverging CD160, RA09-70 and CD10-165 strains. Remarkably, the full Type-IB gene cluster B was only identified in 82% of all tested genomes suggesting that the *casB* lacking strains, due to the lack of *cas1* and *cas2*, also lack the ability to acquire novel spacers. Other Type-IB *cas* gene clusters (complete or partial) were identified among the genomes (including CD160, RA09-70 and CD10-165) with gene organization and high sequence conservation only within a limited subset of strains, making *cas* gene clusters less useful for phylogenetic purposes. Analysis of the *C. difficile* CRISPR arrays identified a total of 1,865 arrays displaying a typical CRISPR repeat sequence (mostly 29-nt long, but occasionally 30 nt), across all analyzed genomes. This resulted in an astounding average of 8.5 CRISPR arrays per genome (ranging from 4 to 14 CRISPR arrays per genome after manual curation), which is substantially higher than other systematic investigations of bacterial CRISPR-Cas systems encoding one to three arrays typically (Cady et al. 2011; Shariat et al. 2013; Yin et al. 2013; Karah et al. 2015). In combination with the aforementioned *cas* gene cluster arrangements, this indicates that in this species, most CRISPR-Cas systems are split into several repeat-spacer arrays that putatively share *cas* genes in trans, which is more commonly typical of archaea than bacteria. Based on the remote link of clade C-1 to the main five clades (Knetsch et al. 2012) and paucity of genome data, we subsequently focused on the phylogenetic CRISPR-Cas analysis within the five established clades.

Sequence analysis of the Type-IB signature gene *cas3*, found in all *cas* gene clusters unlike *cas1*, and CRISPR repeat sequences among selected strains across the clades confirmed high sequence conservation, both for *cas* gene and CRISPR repeat sequences. This can be construed as a common core in a diverse set of sequences (supplementary fig. S2, Supplementary Material online). These observations show how the *C. difficile* CRISPR-Cas system encodes a markedly higher number of CRISPR arrays compared with other organisms through the CRISPR database, and also how the diversity of CRISPR-Cas systems in *C. difficile* is to be found within CRISPR arrays, in terms of both distribution and their spacer content.

No CRISPR-Cas system was found in known clostridial plasmids nor phage sequences from the EBI and NCBI databases, though *C. difficile* prophages have previously been shown to encode CRISPR spacer arrays (Hargreaves and Clokie 2014). Hence our subsequent analyses are focused on chromosomal encoded loci.

Distribution and Conservation of CRISPR Spacer Arrays in *C. difficile*

The *C. difficile* CRISPR-Cas system represents a species-wide conserved genetic system, yet is constituted by a complex multi-locus architecture of CRISPR spacer arrays at various locations within each genome and in a variable manner between strains. We explored the CRISPR spacer array distribution in a qualitative comparative genomic analysis to test whether the distribution is random or if the distribution follows distinct patterns across ST and clades. Specifically, we analyzed local gene synteny to investigate co-occurrence of CRISPR loci genetically conserved across selected genomes representing the MLST distribution (fig. 1) to identify local genetic commonalities in the context of Type-IB systems. We used multiple genome alignment visualization of syntenic genome fragments (locally collinear blocks, LCB) along and across the analyzed genomes encoding CRISPR loci (fig. 2). The LCBs reflect fragmented genomes likely caused by

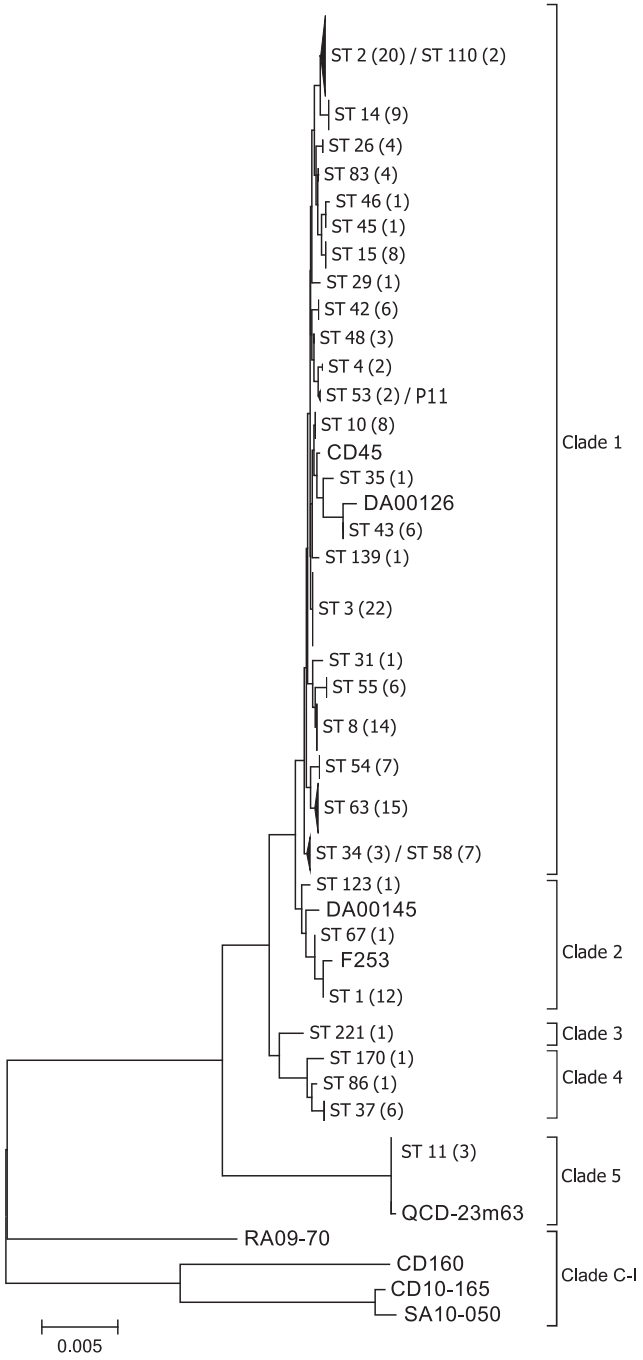

Fig. 1.—MLST-based phylogenetic overview of *Clostridium difficile* based on 217 genome sequences. Sequence types and clade were assigned as per standard nomenclature, and ST clusters were condensed with the number of strains per ST given in parentheses. For strains with unsigned ST number, the strain name was used.

horizontal gene transfer, as previously reported in *C. difficile* (He et al. 2010). Nonetheless, regions of LCB fragments with conserved gene synteny, coupled with shared common repeat sequences, were instrumental to deduce how CRISPR arrays are conserved in varying degrees across STs and revealed conserved parts of the pan-genome encoding CRISPR loci. Through our analysis, we annotated CRISPR spacer arrays and *cas* gene clusters (supplementary table S2, Supplementary Material online) onto conserved and mosaic regions within *C. difficile* (fig. 2). The occurrence and positions of CRISPR arrays across the nine representative genomes were found to be mainly confined to certain LCBs, of which multiple

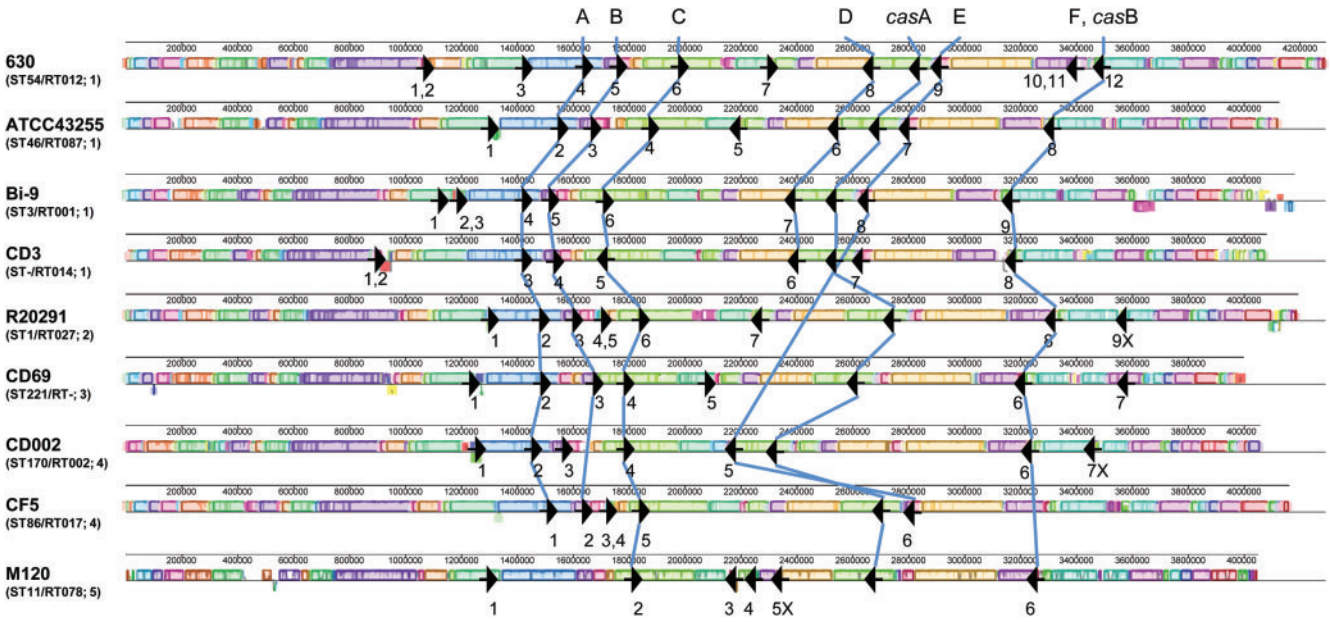

Fig. 2.—Whole genome alignment of nine *Clostridium difficile* genomes with vertical highlight of conserved CRISPR spacer arrays across the strains. The genomes (listed by strain name and ST/RT, and clade) are shown with color coding of the major co-linear blocks (locally region of the genomes that shared similarity) calculated by progressive MAUVE showing main conserved blocks of the genomes and uniquely found regions within each genome. The CRISPR locus position, coding direction (arrows), and numbering of identified CRISPR spacer arrays are shown horizontally across each strain. Conserved CRISPR arrays across genomes are depicted vertically by blue lines and labeled array A–F. The *casA* gene cluster is shown across genomes with blue lines whereas the *casB* gene cluster co-encoded with array F is marked. Additional *cas* gene clusters (column C in supplementary fig. S1, Supplementary Material online) found exclusive to each strain is highlighted at the corresponding array number with a X (eg. R20291 9X for the *cas* gene encoded with CRISPR array9).

apparently not conserved CRISPR arrays were found in putative mobile genetic regions (supplementary table S2, Supplementary Material online). We then investigated whether location-conserved CRISPR spacer arrays also shared common repeat sequences and ancestral spacers, reflecting a common evolutionary origin of the CRISPR array. Intriguingly, the analysis revealed shared CRISPR spacer array locations (fig. 2, arrays A–F).

Building on the observation of partial conservation of CRISPR arrays across the species and on conserved ancestral CRISPR spacer content, we applied a quantitative approach to outline the distribution of conserved CRISPR arrays, identified by their ancestral spacer content, in all 217 strains. Hence, a conserved ancestral end implies commonality among strains whereas the later acquired spacer(s) may differ between even closely related strains due to differential exposure to foreign invasive DNA over time. We found 110 unique ancestral spacers, each representing sets of conserved CRISPR arrays that were present in 1,660 out of 1,865 (89%) CRISPR arrays predicted. The remaining 11% of CRISPR loci are most likely to be unique arrays. The distribution of identified ancestral spacers per genome was subjected to two-way hierarchical clustering to show profiling of strains with shared and divergent CRISPR arrays, and co-occurrence of particular CRISPR arrays across STs (fig. 3).

The clustering of strains by occurrence of common ancestral spacers largely correlates with ST profiling (fig. 1 and supplementary table S1, Supplementary Material online) and it is noteworthy from our analysis that some of these CRISPR loci were not only conserved within, but also partially between ST. The CRISPR array profiling extends into clade 1 showing clear divergence of strains only marginally differentiable by MLST. Curiously, ST3 strains display a peculiar CRISPR array profile (fig. 3) being divided into well-defined distinct groups (ST3A and ST3B) where the ST3A group was found to lack the *casB* gene cluster and associated CRISPR array. For the distribution of CRISPR arrays, it is noteworthy that a small set of five conserved CRISPR arrays are broadly conserved within the establish clades, suggesting that these arrays represent core CRISPR arrays that may be compared across distant strains within the species (fig. 3). The five broadly conserved CRISPR arrays account for 899/1,660 (54%) of the identified arrays and correspond to arrays A, B, C, D and F in figure 2, compounding the results obtained through both our qualitative and quantitative analyses. In contrast, CRISPR arrays that were either unique within or only occasionally shared among STs may hold detailed information about more recent or specific strain differentiation. The distribution of partially conserved CRISPR loci is consistent with the aforementioned universal nature of the Type-IB system and may be useful for phylogenetic analyses

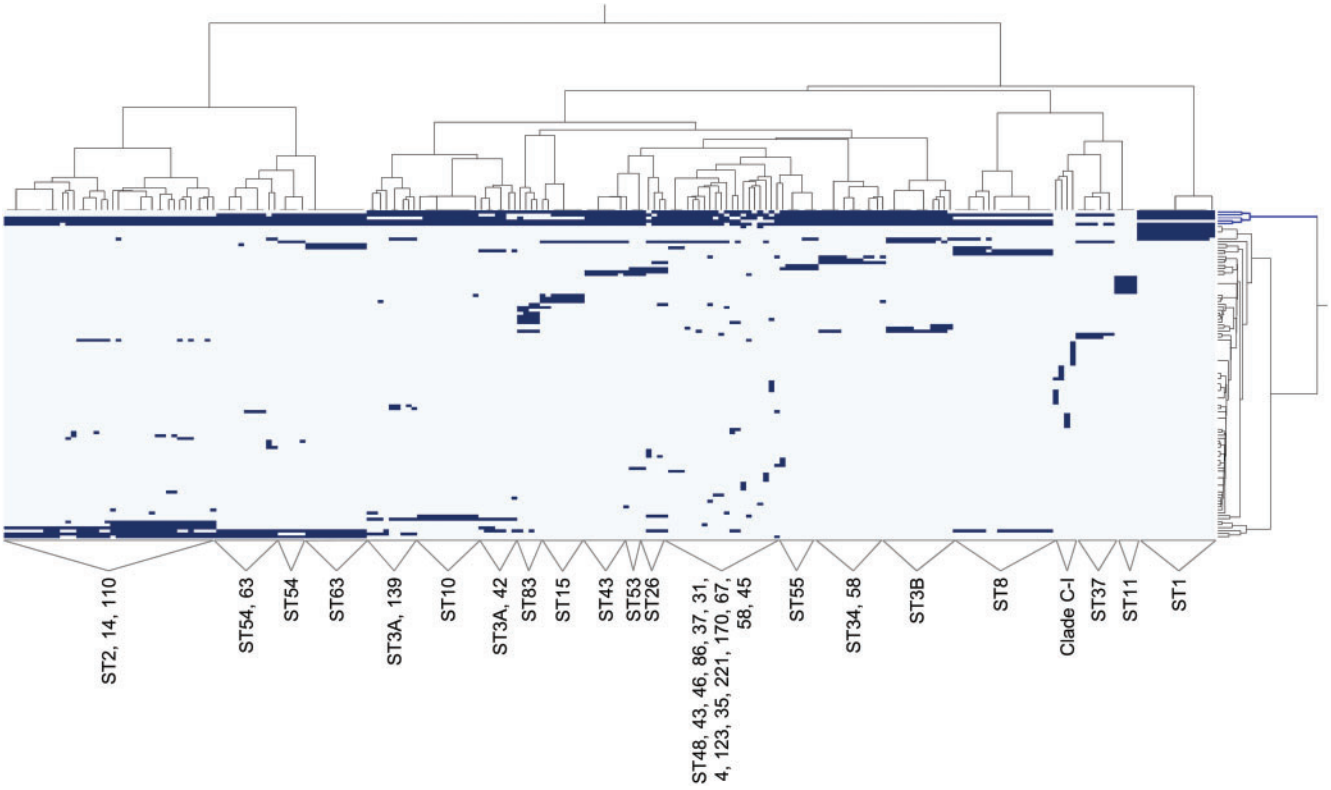

Fig. 3.—Clustering profiles of *Clostridium difficile* strains based on ancestral spacer content across 217 genomes and 110 nonredundant ancestral spacers. The distinct clustered profiles of strains are shown in dark blue for the presence of identified ancestral spacers within each genome across the 217 genomes whereas the light blue background denotes absence of the particular spacer. Two-way hierarchical clustering was performed to identify groups of genomes with comparable ancestral spacer composition (listed by ST vertically) and to group co-occurrence of spacers. Five ancestral spacers were found to be majorly widespread across clades 1–5 and are marked by the blue branch on the right-hand side clustering tree.

of *C. difficile* strains based on the occurrence of CRISPR loci and spacer content within conserved arrays.

Conserved CRISPR Spacers Reveal Microevolution among Strains

The observation of conserved CRISPR arrays across clades of *C. difficile* allows in-depth analysis of CRISPR spacer content and polymorphism to analyze strain diversity and infer an role of CRISPR loci in *C. difficile* evolution. We explored the spacer content of the five most widespread CRISPR arrays (fig. 4 and supplementary fig. S3, Supplementary Material online) to assess the similarities across and within ST. Notably, one of the CRISPR arrays (array F) was found to be encoded adjacent to the conserved full *cas* gene cluster (*casB*), harboring *cas1* and *cas2* essential for spacer acquisition in pre-existing arrays. Hence, we hypothesized that the co-occurrence of the combined locus to be an ancestral component of the *C. difficile* CRISPR-Cas system. Analysis of the CRISPR spacer sequences revealed how the array holds information to distribute strains both among clades and within ST (fig. 4). For example, all stains, but the ST11 strains (clade 5), share the two ancestral spacers (spacer 1 and 2, fig. 4) and within ST55 all strains share 39 spacers whereas strain P32 has acquired an additional spacer (spacer 40). For the ST2, 14 and 110 strains all share spacer 1–27 whereas the strains Y401, CD12, DA00306 and LIBA-5784 have acquired additional spacers however spacer deletions within the 27 conserved spacers have occurred and thus allows differentiation of strains within the group.

In clade 1, ST2, ST14 and ST110 showed a strong grouping and relatedness to ST53 and ST43 whereas part of ST3 (the remaining strains of ST3 was lacking the CRISPR-Cas locus) grouped with ST10, 42, 29, 31 and 35 although with a lesser degree of ancestral end spacer conservation. Likewise, in the globally spreading and likely hyper-virulent clade 2, the CRISPR spacer content enabled grouping of strains. Interestingly, spacer acquisitions and deletions were abundant and allowed strain differentiation into three CRISPR genotypes as seen for ST1 and the cluster of ST2, 14 and 110 where 22 CRISPR genotypes were assigned to 33 strains as detailed below. Thus, we have systematically cataloged the CRISPR loci abundance and complexity. We identified diversity of the Type-IB

Fig. 4.—CRISPR Spacer content and polymorphisms across the conserved CRISPR array associated with the *casB* gene cluster. Spacers are shown as squares uniquely colored by spacer sequence and with different icons representing spacer length and in case of a putative spacer deletion a ☒ is inserted to represent the gap to continuous align comparable spacers. Spacer numbering is initiated at the ancestral end (right) towards the most recently acquired spacers per strain towards the left. Strains are listed by name and ST whereas grouping of strains was based on shared spacers including strain with no measureable ST profiles but with CRISPR arrays readily analyzed. For clade 5 strains, the ancestral spacer differed in sequence at position 1 but the conserved location of array F was confirmed by genome comparison (fig. 2). Spacers assumed to be lacking due to the CRISPR array found on separate contigs are represented by ? as it cannot be stated if they are not assembled or if a spacer deletion has occurred.

CRISPR-Cas elements on multiple levels: (1) the number of CRISPR loci was generally conserved only among related ST based on the distribution of ancestral spacers; (2) shared common ancestral spacers revealed distribution of conserved CRISPR loci among strains; (3) spacer content and polymorphism varied between strains within ST in an expected ordinal manner sufficient to differentiate related strains in the ST. In order to further explore strain relatedness across the four additional highly conserved arrays (fig. 3, array A, B, D and D), we analyzed the spacer content and found that each array has evolved independently through spacer acquisitions and deletions, notably at the arrays' ancestral end (supplementary fig. S3, Supplementary Material online). Hence, each array represents the species, or a subset of related strains, evolution over time. The finding of five widely conserved CRISPR arrays positioned in syntenic genetic regions, unique to each of the five arrays, constitute convenient and valuable candidate genetic loci (generally 0.5–2 kb in length) for PCR amplification and sequencing as a novel strategy to type strains within the five main *C. difficile* clades based on presence of the CRISPR arrays and comparison of spacer content. Additionally, we showed how CRISPR arrays are highly conserved within ST groups, which renders spacer diversity useful for high resolution typing.

CRISPR Genotyping for Resolving Highly Related Strains and Differentiating PaLoc Encoding Strains in Clade 1

Most CRISPR spacer-based divergence among ST was observed towards the ancestral end (supplementary fig. S3, Supplementary Material online), but in order to analyze the CRISPR typing potential of more recent acquired spacer within highly related strains, we explored the group of ST2, 14 and 110. This group of STs are poorly resolved by MLST and resolved as a single group (RT014) by PCR ribotyping (Janezic and Rupnik 2015). These STs showed high internal conservation of CRISPR arrays and we found seven strictly conserved CRISPR arrays (fig. 5) and additionally CRISPR arrays sporadically found among strains within the three ST groups. Based on CRISPR spacer acquisitions and deletions, we found 22 unique CRISPR genotypes among the 33 analyzed strains (CRISPR genotypes A–V, fig. 5). Notably, the CRISPR diversity in one of the conserved arrays (fig. 6, array B) was sufficient to differentiate ST2 and ST14. Both ST110 strains had unique CRISPR genotypes and could be differentiated from ST2 based on single spacer polymorphisms.

Interestingly, the ST3 genomes were found to be split, based on divergent CRISPR arrayprofiles (fig. 3). ST3 and ST7 were previously shown to contain both pathogenicity locus (PaLoc) encoding and noncoding strains (Dingle et al. 2014), however no ST7 genomes were available for our analysis. Upon correlation of the PaLoc gene organization across all available ST3 genomes and their CRISPR profiles (ST3A or ST3B, fig. 3) we observed strict segregation of the toxin (ST3B) versus nontoxin (ST3A) encoding strains based on the differential content of conserved CRISPR arrays (fig. 6). Notably ST3 was the only ST group with dichotomy based on the occurrence of toxin genes when identifying *tcdA* and *tcdB*, and the binary toxin encoded by *cdtA* and *cdtB* across the sampled genomes (supplementary table S1, Supplementary Material online). Thus, for this group, we observed a correlation between CRISPR groups and toxin groups.

CRISPR Diversity in the Hyper-Virulent Clade 2

We further explored the potential of using CRISPR diversity among the available hyper virulent clade 2 genomes, largely represented by ST1 (RT027/NAP1/B1) strains within our data set, to differentiate important clinical isolates related to the increase of epidemic strains associated with increased morbidity and mortality (Vedantam et al. 2012). Whole genome alignment showed how the strains within clade 2 have an overall more conserved genome structure (fig. 7A) as compared with the previous analysis of genomes across clades 1–5 (fig. 2) and earlier genomic analysis (He et al. 2010). Comparison of CRISPR array locations across the analyzed STs identified 4 CRISPR arrays positioned in the core LCBs shared among all analyzed genomes whereas those arrays that appeared to be unique among the strains were often found to be encoded in mobile genetic elements (insertion sequence or prophage regions) as previously found (supplementary table S2, Supplementary Material online) and reported for single strains (Sebaihia et al. 2006), or other loci exclusive to the ST. The shared CRISPR arrays show, again, how CRISPR arrays evolved independently within and among strains, and illustrate how the clade 2 genomes share spacer compositions and hence evolution history (fig. 7B). The shared arrays that display more divergent spacer content may be instrumental using CRISPR sequence information to correlate epidemiologic data to ST1 outbreaks and track strains displaying novel pathogenic phenotypes. Notably, within the available ST1 genomes CRISPR spacer polymorphism was observed and sufficient for differentiating the strains into three sub-genotypes (fig. 7C) with genotype 1 as the predecessor. Genotype 2 occurred by a single spacer deletion (spacer 39 in array F) whereas genotype 3 was found as a dual event by a deletion (spacers 4–7 in array D) and an acquisition (spacers 42 and 43 in array F). Notably the R20291 strain (RT027/ST1/B1) was found in the most recent CRISPR genotype 3 (fig. 7C) when compared and correlated to core genome SNP derived grouping of *C. difficile* (He et al. 2013). This shows how CRISPR-Cas typing has applicability to first differentiate strains from a clinical relevant clade, and second, reflect the evolution of highly related strains within ST1 (RT027/Bi1).

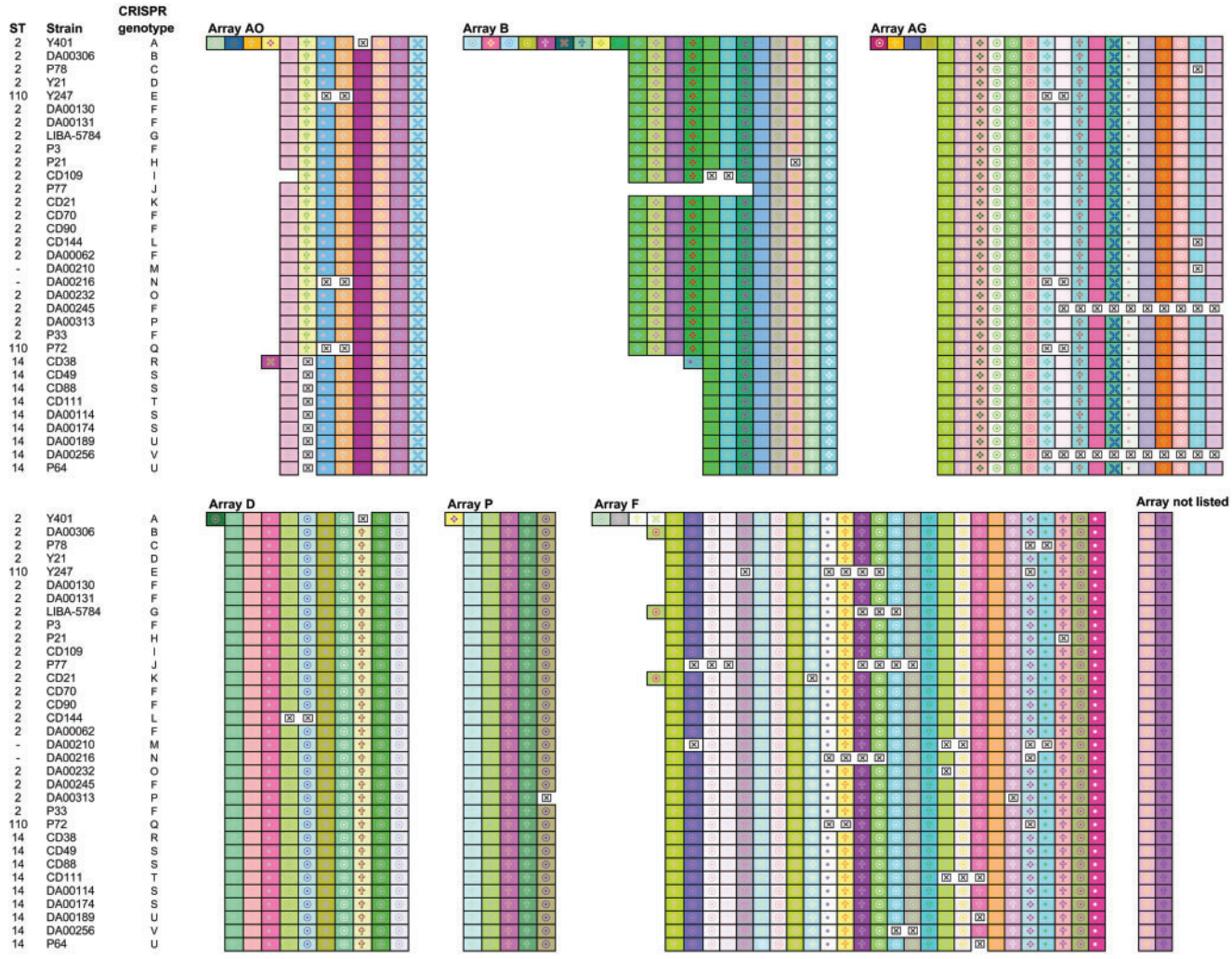

Fig. 5.—CRISPR spacer polymorphism from seven conserved CRISPR arrays show ST and strain differentiation within highly related clade 1 strains from ST2, 14 and 110. Spacer deletions, marked by ☒ or spacer acquisitions were used to establish CRISPR genotypes across the seven CRISPR arrays shown by letter designation (A–V).

Analysis of CRISPR Spacer Sequences and Homology to Foreign and Chromosomal DNA

The CRISPR array composition and spacer polymorphisms were utilized for phylogenetic analysis within the *C. difficile* species, yet CRISPR spacer analysis may hold additional information enabling CRISPR-Cas based functional differentiation of strains. The *C. difficile* CRISPR-Cas system was recently shown to be active in strain R20291 (ST1/RT027/B1) for DNA interference (Boudry et al. 2015). Analysis of CRISPR spacers from strains 630 (ST54/RT012) and R20291 showed, through sequence homology, how the CRISPR-Cas system likely targets phages (Hargreaves et al. 2014). To extend the phylogenetic and functional diversification of *C. difficile*, we analyzed the putative phage targeting profiles representative for strains in each available ST (fig. 8, top), interpreted by the number of identified protospacers per phage sequence per ST through two-way hierarchical clustering. This indicated how strains across ST may hold highly divergent phage targeting profiles, even considering the low number of available phages sequences. Notably, there was little correlation of clustering of putative phage targeting profiles to the phylogenetic grouping of ST. For ST2, 14 and 110 distinct profiles were observed despite the overall shared set of core CRISPR arrays (fig. 5), indicating how the dispersed CRISPR arrays (fig. 3) likely impact the phage resistance profiles of *C. difficile*. This may be used as another tool for differentiation of *C. difficile* strains, based on encoded CRISPR spacers. A cluster of phages (38-2, 146, 111 and 6,356) displayed clear enrichment of protospacers. The distribution of protospacers appeared random throughout the phage sequences as exemplified for phage

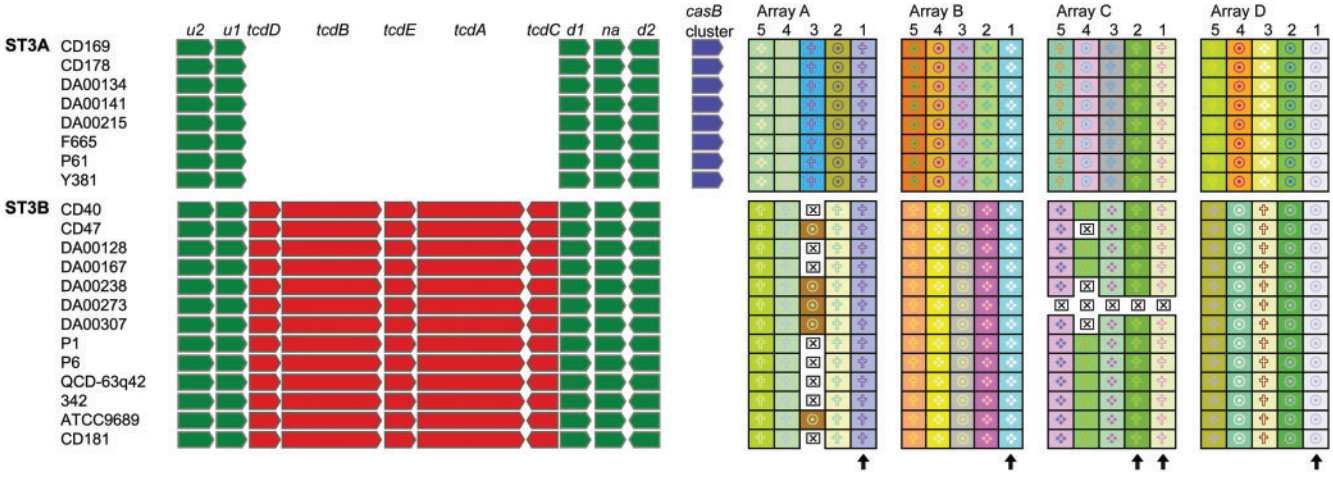

Fig. 6.—Differentiation of ST3 strains by CRISPR spacer composition in conserved arrays correlate to toxin gene occurrence. The PaLoc and flanking regions, colored in red and green, respectively, gene organization is depicted for all strains together with the presence of absence of the *casB* gene cluster (blue; all eight genes are represented by the marker), and the four conserved CRISPR arrays shown by the five ancestral spacers sufficient to differentiate ST3 into ST3A and ST3B.

phiCD146 (fig. 8, bottom) and a predicted protospacer adjacent motif was identified 5′-CC(A/T), as previously proposed (Boudry et al. 2015). Interestingly ST11, represented by strain M120 (RT078), encoded a markedly increased number of spacer with homology to the identified phages.

Discussion

Analysis of CRISPR occurrence and complexity in *C. difficile* reveal potential of CRISPR-based phylogeny across and within ST. Based on the occurrence and diversity of *cas* gene and CRISPR repeat sequences, and the number of CRISPR spacer arrays throughout the species, we propose that *C. difficile* harbors a diverse and distinguishable strictly conserved Type-IB system with internal variations sufficient for CRISPR-based phylogenetic and typing analyses. Furthermore, the CRISPR array abundance observed suggests an unprecedented complexity within one bacterial CRISPR-Cas system. Our results allow us to exploit conserved and variable CRISPR sequences as highly informative genotypes due to their inherent linear acquisition of novel spacers at the leader end of the CRISPR array. This makes comparison of conserved arrays able to both differentiate distantly related strains only sharing few ancestral spacers and highly related strain sharing most spacers and only diverging by recent spacer deletions or strain specific spacer acquisitions. Our analyses illustrate how CRISPR sequences can be exploited to visualize genetic diversity in *C. difficile* across clades, ST and highly clonal isolates.

Arguably, the most prominent feature of the *C. difficile* Type-IB system, is the unusually high number of CRISPR arrays per genome, which averages 8.5 arrays per genome, much higher than other bacterial systems in the CRISPR database (Grissa et al. 2007), and notably higher than the typical one to three CRISPR arrays found in other major pathogens (Cady et al. 2011; Shariat et al. 2013; Yin et al. 2013; Karah et al. 2015). Additionally, other comprehensive CRISPR-Cas based phylogenetic analysis in pathogenic bacteria have found multiple types of CRISPR-Cas systems present within a single species (van Belkum et al. 2015). The enumeration of CRISPR loci poses a novel challenge for sequence interpretation from the otherwise widely accepted paradigm of one *cas* gene cluster and one CRISPR array (Shariat and Dudley 2014), where conserved arrays are readily identified and comparable. The internal CRISPR-Cas diversity in *C. difficile* is best visualized through genome alignment for qualitative comparative analysis. The comprehensive profiling of ancestral CRISPR spacers allowed us to unravel the complexity and occurrence of multiple seemingly unrelated CRISPR arrays within and between strains. Overall, five distinct CRISPR arrays were shown to be conserved across clades 1 through 4, and partially in clade 5. Extensive spacer composition and ordering within the arrays enabled deduction of relatedness among ST and coupled with spacer polymorphism could be applied for differentiating ST. Spacer polymorphism within ST may be useful for differentiation of highly clonal strains as seen for ST2, 14 and 110 and for clade 2 with a focus on ST1 (fig. 5). Hence, CRISPR locus occurrence and (ancestral) spacer composition are sufficient to differentiate ST and even strains. Future studies of clinical isolates will establish their potential for blinded genotyping potential. Nevertheless, CRISPR-Cas systems are species and type dependent (Kupczok et al. 2015), and our analysis of the *C. difficile* Type-IB system shows promising phylogenetic and typing applications through the vast number of partially conserved CRISPR arrays that each evolve independently.

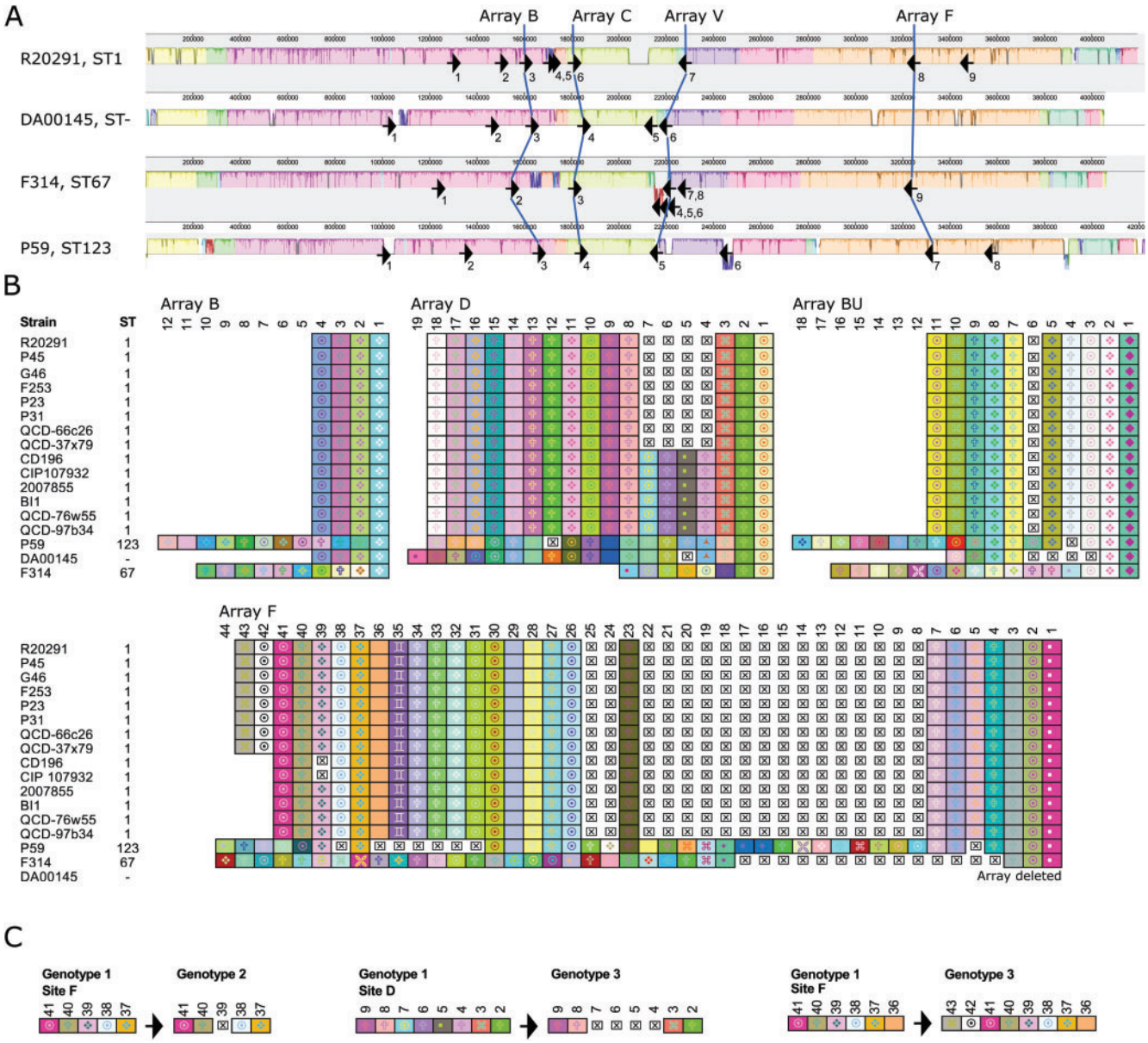

Fig. 7.—CRISPR array occurrence and spacer polymorphism based phylogenetic relationships in clade 2. Whole genome alignment of clade 2 strains, with CRISPR arrays marked as per fig. 2, was used to identify four conserved arrays highlighted with blue lines (A). The spacer composition of these four arrays (B) showed phylogenetic relationship clade-wide and how spacer polymorphism could be applied to group the ST1 strains into three CRISPR genotypes 1 through 3 (C).

Genomic analysis suggested CRISPR loci can be encoded on mobile genetic elements and hence associated with lateral DNA transfer as previously documented (Sebaihia et al. 2006). This observation reflects potential benefits of DNA uptake for the receiving strain, which would gain immunity even prior to phage exposure through horizontal gene transfer (He et al. 2010). In contrast to the vivid diversification of CRISPR loci by spacer acquisition, we observed CRISPR spacer deletions as widely occurring in loci shared across and within ST (figs. 4–8), possibly happening as a response to the rapid expansion of CRISPR spacer content not through acquisition but perhaps lateral DNA uptake. This would suggest that the CRISPR array maintains its own form of genetic homeostasis that manifests in the balance of spacer acquisitions and deletions governing array size and total number of spacer per genome. The CRISPR loci and *cas* functions are likely evolving more rapidly through genetic transfer within the species than through spacer acquisitions and mutations occurring naturally.

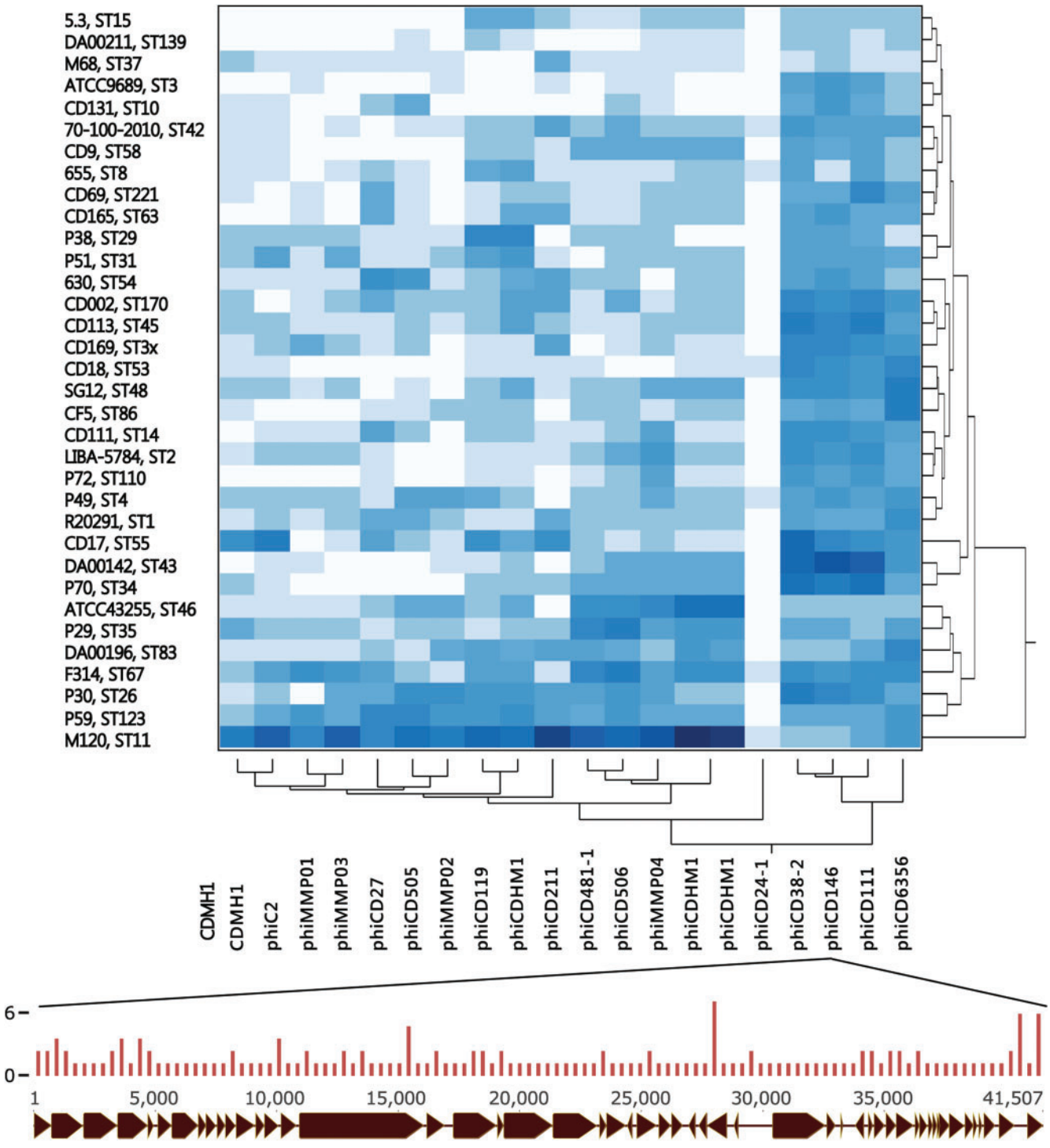

Fig. 8.—Hierarchical clustering analysis of spacer matches in *Clostridium difficile* phages per ST. The number of spacer sequence hits per phage genome is color dependent scaled from zero hits (white) to dark blue (16 hits). The location and distribution of spacer matches (protospacers) are shown as a bar chart as example for phage phiCD146 with gene organization shown.

This observation may explain how a species with relatively low mutation rates (approximately one nucleotide per genome per year in *C. difficile*; He et al. 2013; Knetsch et al. 2014) could mount CRISPR-Cas immunity so rapidly in terms of CRISPR array abundance. This is also an argument for the value of CRISPR-focused analysis of genome drafts to complement core genome SNP analysis (Huber et al. 2013).

The microevolution of *C. difficile* observed through horizontal genetic transfer of CRISPR loci is supported by comparing the differential spacer matches within each ST as it relates to phage targeting. This observation was proposed earlier within a smaller subset of PCR ribotypes and in connection with putatively mobile CRISPR loci (Hargreaves et al. 2014; Boudry et al. 2015). It is highly plausible that the *C. difficile* CRISPR-Cas systems both maintain some whole genome homeostasis within STs by limiting phages and plasmid exposure, but also drive selective, differential uptake of new DNA in a process both relying on transfer of entire CRISPR loci and spacer acquisition for sustainable phage immunity. These observations link the presence of complex CRISPR systems in *C. difficile* to historic differential genetic transfer amongst STs, and thus constitute an exploitable basis for evolutionary studies. Our analysis of CRISPR loci and the association to genetic transfer, both in terms of loci encoding CRISPR arrays and the prevention of phage and plasmid spread, add a novel perspective to mobile genetic elements in *C. difficile* and emphasize a need to increase our understanding of noncore parts of the genome, some of which related to and are important for pathogenicity (Brouwer et al. 2013). This should be valuable towards a definite genotyping method, and for comprehensive phylogenetic analyses.

The above observation expands our understanding the role of CRISPR-Cas in *C. difficile* linked both to phylogenetic interpretations and the role of foreign DNA in shaping this pathogenic species. We further showed the CRISPR typing potential within the significant epidemiologic group of ST1 (RT027/B1) strains where more clinical strains are needed to expand its epidemiological use, yet we showed strain evolution (He et al. 2010, 2013), by analyzing CRISPR spacer polymorphism.

In conclusion, our analysis of CRISPR-Cas loci occurrence and diversity in 217 *C. difficile* genomes has revealed a widespread, variable Type-IB CRISPR-Cas system conserved across the species. We resolved an unprecedented complexity of CRISPR loci, likely resulting from lateral DNA transfer, through profiling of conserved CRISPR arrays among and within MLST groups and clades. CRISPR-based phylogenetic analysis of the species showed common ancestry largely through five conserved CRISPR arrays in clades 1–5. ST group-specific arrays within clade 1 could further resolve highly similar ST2, 14 and 110 whereas spacer composition was instrumental to differentiate *tcdA* and *tcdB* encoded ST3 strains from nontoxigenic ST3 strains. The prominent ST1/RT027/B1 group was differentiable from other clade 2 genomes and was assigned three CRISPR genotypes within ST1/RT027/B1, based on CRISPR spacer polymorphism. CRISPR loci constitute a potential genetic target for epidemiological studies in *C. difficile*, with the flexibility to determine both common origin and recent divergence. Functionally, it appears CRISPR-Cas systems are often associated with mobile genetic elements, highlighting a plausible route of CRISPR immunization prior to actual infection by foreign DNA. Accordingly, CRISPR immunity is an important factor in *C. difficile* evolution, and may provide important insights into genome microevolution and genetic homeostasis.

Supplementary Material

Supplementary tables S1 and S2 and figures S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We thank Sarah O'Flaherty and Kurt Selle for skillful comments and discussions during article preparation. This work was supported by a personal *post doc* grant and a *Sapere Aude Research Talent* to J.M.A. from the Danish Council for Independent Research [DFF-4002-00176]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the article.

Literature Cited

Amy J, Johanesen P, Lyras D. 2015. Extrachromosomal and integrated genetic elements in *Clostridium difficile*. Plasmid 80:97–110.

Bacci S, et al. 2011. Binary toxin and death after *Clostridium difficile* infection. Emerg Infect Dis. 17(6):976–982.

Barrangou R, et al. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. Science 315(5819):1709–1712.

Barros MPS, et al. 2014. Dynamics of CRISPR loci in microevolutionary process of *Yersinia pestis* strains. PLoS One 9(9):e108353.

Bland C, et al. 2007. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics 8:209.

Boudry P, et al. 2015. Function of the CRISPR-Cas system of the human pathogen *Clostridium difficile*. mBio 6(5):e01112–e01115.

Boyanova L, Kolarov R, Mitov I. 2015. Recent evolution of antibiotic resistance in the anaerobes as compared to previous decades. Anaerobe 31:4–10.

Brouwer MS, et al. 2013. Horizontal gene transfer converts non-toxigenic *Clostridium difficile* strains into toxin producers. Nat Commun. 4:2601.

Cady KC, et al. 2011. Prevalence, conservation and functional analysis of *Yersinia* and *Escherichia* CRISPR regions in clinical *Pseudomonas aeruginosa* isolates. Microbiology 157(Pt 2):430–437.

Cairns MD, et al. 2015. Genomic epidemiology of a protracted hospital outbreak caused by a toxin A negative, *Clostridium difficile* sublineage PCR Ribotype 017 strain in London, England. J Clin Microbiol. 53(10):3141–3147.

Cui Y, et al. 2008. Insight into microevolution of *Yersinia pestis* by clustered regularly interspaced short palindromic repeats. PLoS One 3(7):e2652.

Darling ACE, et al. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14(7):1394–1403.

Dingle KE, et al. 2014. Evolutionary history of the *Clostridium difficile* pathogenicity locus. Genome Biol Evol. 6(1):36–52.

Dubberke ER, Olsen MA. 2012. Burden of *Clostridium difficile* on the healthcare system. Clin Infect Dis. 55(Suppl 2):S88–S92.

Eyre DW, et al. 2013. Short-term genome stability of serial *Clostridium difficile* ribotype 027 isolates in an experimental gut model and recurrent human disease. PLoS One 8(5):e63540.
Griffiths D, et al. 2010. Multilocus sequence typing of *Clostridium difficile*. J Clin Microbiol. 48(3):770–778.
Grissa I, Vergnaud G, Pourcel C. 2007. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics 8:172.
Hargreaves KR, Clokie MR. 2014. *Clostridium difficile* phages: still difficult?. Front Microbiol. 5:184.
Hargreaves KR, et al. 2014. Abundant and diverse clustered regularly interspaced short palindromic repeat spacers in *Clostridium difficile* strains and prophages target multiple phage types within this pathogen. MBio 5(5):e01045–e01013.
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.
He M, et al. 2013. Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. Nat Genet. 45(1):109–113.
Horvath P, et al. 2008. Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. J Bacteriol. 190(4):1401–1412.
Huber CA, et al. 2013. Challenges for standardization of *Clostridium difficile* typing methods. J Clin Microbiol. 51(9):2810–2814.
Janezic S, Rupnik M. 2015. Genomic diversity of *Clostridium difficile* strains. Res Microbiol. 166(4):353–360.
Karah N, et al. 2015. CRISPR-cas subtype I-Fb in *Acinetobacter baumannii*: evolution and utilization for strain subtyping. PLoS One 10(2):e0118205.
Knetsch CW, et al. 2012. Comparative analysis of an expanded *Clostridium difficile* reference strain collection reveals genetic diversity and evolution through six lineages. Infect Genet Evol. 12(7):1577–1585.
Knetsch CW, et al. 2013. Current application and future perspectives of molecular typing methods to study *Clostridium difficile* infections. Euro Surveill. 18(4):20381.
Knetsch CW, et al. 2014. Whole genome sequencing reveals potential spread of *Clostridium difficile* between humans and farm animals in the Netherlands, 2002 to. Euro Surveill. 19(45):20954–22011.
Kuijper EJ, van den Berg RJ, Brazier JS. 2009. Comparison of molecular typing methods applied to *Clostridium difficile*. Methods Mol Biol. 551:159–171.
Kupczok A, Landan G, Dagan T. 2015. The contribution of genetic recombination to CRISPR array evolution. Genome Biol Evol. 7(7):1925–1939.
Louwen R, et al. 2014. The role of CRISPR-Cas systems in virulence of pathogenic bacteria. Microbiol Mol Biol Rev. 78(1):74–88.
Makarova KS, et al. 2011. Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol. 9(6):467–477.
Makarova KS, et al. 2015. An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol. 13(11):722–736.
Makarova KS, Wolf YI, Koonin EV. 2013. The basic building blocks and evolution of CRISPR-CAS systems. Biochem Soc Trans. 41(6):1392–1400.
Marraffini LA, Sontheimer EJ. 2008. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science 322(5909):1843–1845.
Mergenhagen KA, Wojciechowski AL, Paladino JA. 2014. A review of the economics of treating *Clostridium difficile* infection. Pharmacoeconomics 32(7):639–650.
Mullany P, Allan E, Roberts AP. 2015. Mobile genetic elements in *Clostridium difficile* and their role in genome function. Res Microbiol. 166(4):361–367.
Nozawa T, et al. 2011. CRISPR inhibition of prophage acquisition in *Streptococcus pyogenes*. PLoS One 6(5):e19543.
Ofosu A. 2016. *Clostridium difficile* infection: a review of current and emerging therapies. Ann Gastroenterol. 29(2):147–154.
Palmer KL, Gilmore MS. 2010. Multidrug-resistant enterococci lack CRISPR-cas. MBio 1(4):e00227–10.
Pettengill JB, et al. 2014. The evolutionary history and diagnostic utility of the CRISPR-Cas system within *Salmonella enterica* ssp. *enterica*. PeerJ. 2:e340.
Sebaihia M, et al. 2006. The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. Nat Genet. 38(7):779–786.
Sekulovic O, et al. 2014. Characterization of temperate phages infecting *Clostridium difficile* isolates of human and animal origins. Appl Environ Microbiol. 80(8):2555–2563.
Sekulovic O, Fortier LC. 2014. Global transcriptional response of *Clostridium difficile* carrying the phiCD38-2 prophage. Appl Environ Microbiol. 81(4):1364–1374.
Shariat N, Dudley EG. 2014. CRISPRs: molecular signatures used for pathogen subtyping. Appl Environ Microbiol. 80(2):430–439.
Shariat N, et al. 2013. The combination of CRISPR-MVLST and PFGE provides increased discriminatory power for differentiating human clinical isolates of *Salmonella enterica* subsp. *enterica* serovar *Enteritidis*. Food Microbiol. 34(1):164–173.
Sirard S, Valiquette L, Fortier LC. 2011. Lack of association between clinical outcome of *Clostridium difficile* infections, strain type, and virulence-associated phenotypes. J Clin Microbiol. 49(12):4040–4046.
Smits WK. 2013. Hype or hypervirulence: a reflection on problematic *C. difficile* strains. Virulence 4(7):592–596.
Soutourina OA, et al. 2013. Genome-wide identification of regulatory RNAs in the human pathogen *Clostridium difficile*. PLoS Genet. 9(5):e1003493.
Stabler RA, et al. 2012. Macro and micro diversity of *Clostridium difficile* isolates from diverse sources and geographical locations. Plos One 7(3):e31559.
Steglich M, et al. 2015. Tracing the spread of *Clostridium difficile* ribotype 027 in Germany based on bacterial genome sequences. PLoS One 10(10):e0139811.
Tamura K, et al. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol. 30(12):2725–2729.
Timme RE, et al. 2013. Phylogenetic diversity of the enteric pathogen *Salmonella enterica* subsp. *enterica* inferred from genome-wide reference-free SNP characters. Genome Biol Evol. 5(11):2109–2123.
van Belkum A, et al. 2015. Phylogenetic distribution of CRISPR-Cas systems in antibiotic-resistant *Pseudomonas aeruginosa*. MBio 6(6):e01796–15.
Vedantam G, et al. 2012. *Clostridium difficile* infection: toxins and non-toxin virulence factors, and their contributions to disease establishment and host response. Gut Microbes 3(2):121–134.
Wasels F, et al. 2014. Inter- and intraspecies transfer of a *Clostridium difficile* conjugative transposon conferring resistance to MLSB. Microb Drug Resist. 20(6):555–560.
Yin S, et al. 2013. The evolutionary divergence of Shiga toxin-producing *Escherichia coli* is reflected in clustered regularly interspaced short palindromic repeat (CRISPR) spacer composition. Appl Environ Microbiol. 79(18):5710–5720.

Associate editor: Rotem Sorek