

UNIVERSIDAD DE COSTA RICA
SISTEMA DE ESTUDIOS DE POSGRADO

**Análisis del impacto de los elementos genéticos móviles del genoma
accesorio y las mutaciones del genoma central en la microdiversificación
de la cepa NAP_{CR1} de *Clostridium difficile***

**Analysis of the impact of mobile genetic elements from the accessory
genome and mutations from the core genome in the microdiversification
of the *Clostridium difficile* NAP_{CR1} strain**

Tesis sometida a la consideración de la Comisión del Programa de Posgrado en
Microbiología, Parasitología, Química Clínica e Inmunología para optar al grado
y título de Maestría Académica en Microbiología

Sonia Tatiana Murillo Corrales

Ciudad Universitaria Rodrigo Facio, Costa Rica
2016

DEDICATORIA / DEDICATION

Mi trabajo se le dedico a mis padres y mi hermano, Sonia Corrales, Jorge Giovanni Murillo y José Pablo Murillo, quienes con su apoyo incondicional me han permitido perseguir mis sueños. A Dios por su apoyo y por mi familia.

I dedicate this work to my parents and my brother, Sonia Corrales, Jorge Giovanni Murillo and José Pablo Murillo for their unconditional support in order to achieve my dreams. And to God for the support and providing me with such an amazing family.

AGRADECIMIENTOS

A mi tutor César Rodríguez por su guía durante el proceso, siempre me ha motivado a dar lo mejor de mí y ha confiado en mis capacidades. Gracias por impulsarme a alcanzar mis metas.

Al equipo de trabajo del LIBA, Evelyn Rodríguez, María del Mar Gamboa, Diana López, Carlos Quesada, Pablo Vargas y Robin Cárdenas, por darme una de oportunidad laboral inolvidable y por los lindos momentos compartidos.

A Mariann González y Dirk Berkelmann, por creer en mí en todo momento, por su apoyo y por impulsarme a alcanzar mis sueños.

A Esteban Chaves y Caterina Guzmán por sus aportes en el desarrollo de este proyecto de investigación y por motivarme a realizarlo.

A Gabriel Ramírez, Andony Cordero y María Fernanda Ulate por sus aportes al proyecto de investigación.

Al Dr. Thomas Riedel del Leibniz-Institut DSMZ en Braunschweig por la colaboración para la secuenciación con la tecnología PacBio.

A todos los amigos de la Facultad de Microbiología y a los colegas de la Maestría que han convertido este proyecto en una linda travesía.

A mis profesores de la maestría por compartirme su conocimiento e impulsarme a ser mejor.

ACKNOWLEDGEMENTS

My advisor César Rodríguez for his guidance during the process, he constantly motivated me to give my best and trusted my abilities. Thank you for sponsoring me to achieve my goals.

The LIBA work group: Evelyn Rodríguez, María del Mar Gamboa, Diana López, Carlos Quesada, Pablo Vargas and Robin Cárdenas, for giving me an unforgettable work opportunity and for the nice moments we shared.

Mariann González and Dirk Berkelmann, for always believing in me and supporting me to accomplish my dreams.

Esteban Chaves and Caterina Guzmán for their contribution to the development of this project and their motivation.

Gabriel Ramírez, Andony Cordero and María Fernanda Ulate for their contribution to the project.

Dr. Thomas Riedel from the Leibniz-Institut DSMZ in Braunschweig for the PacBio sequencing of the isolates.

All my friends from the Faculty of Microbiology and my colleagues from the Master program who converted this project into an amazing experience.

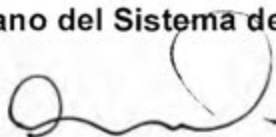
My professors from the master program for sharing their knowledge with me and motivating me to accomplish my best work.

"Esta tesis fue aceptada por la Comisión del Programa de Estudios de Posgrado en Microbiología, Parasitología, Química Clínica e Inmunología de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Académica en Microbiología"



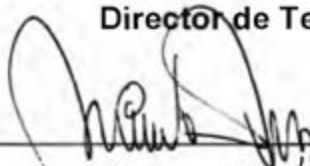
M.Sc. Carlos Chacón Díaz

Representante del Decano del Sistema de Estudios de Posgrado



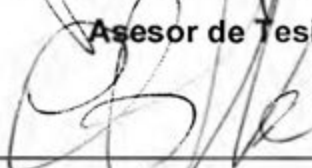
Dr. César Rodríguez Sánchez

Director de Tesis



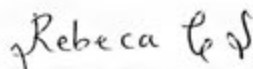
Dr. Esteban Chaves Olarte

Asesor de Tesis



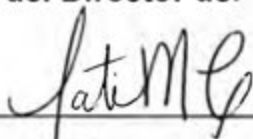
Dra. Caterina Guzmán Verri

Asesora de Tesis



Dra. Rebeca Campos Sánchez

Representante del Director del Programa de Posgrado



Sonia Tatiana Murillo Corrales

Candidata

INDEX

DEDICATORIA / DEDICATION	ii
AGRADECIMIENTOS.....	iii
ACKNOWLEDGEMENTS	iv
HOJA DE APROBACION	v
INDEX.....	vi
RESUMEN.....	viii
ABSTRACT.....	ix
LIST OF TABLES.....	x
LIST OF FIGURES	xi
LIST OF ABBREVIATIONS	xiii
1. INTRODUCTION.....	1
1.1 General characteristics of <i>Clostridium difficile</i> infections (CDI)	1
1.2 <i>Clostridium difficile</i> typing methods.....	2
1.3 Strains producing severe CDI cases.....	2
1.4 The epidemic lineage NAP _{CR1} and its relation to the reference strain CD630	3
1.5 Phylogenomics of <i>C. difficile</i>	4
1.6 Genome diversification in Bacteria	6
1.7 Mutations and recombination in <i>C. difficile</i>	7
1.8 The <i>C. difficile</i> mobilome and its biological effect	9
2. SCIENTIFIC QUESTION.....	12
3. JUSTIFICATION.....	12
4. HYPOTHESIS	12
5. MAIN AIM.....	13
5.1 Specific aims.....	13
6. MATERIALS AND METHODS.....	13
6.1 Bacterial isolates and WGS.....	13
6.2 SNPs calling and core SNP phylogeny	20
6.3 Analyses of feature frequency profiles (FFP).....	20
6.4 Proteome predictions and pangenome comparisons	21

6.5	Identification of unique gene clusters and MGE.....	21
7.	RESULTS	24
7.1	NAP _{CR1} and NAP1 isolates have slightly different SNP densities and dN/dS rates	24
7.2	The core genome of the NAP _{CR1} isolates is more heterogenous than that of NAP1 isolates.....	29
7.3	The accessory genome of the NAP _{CR1} isolates is more diverse than that of the NAP1 isolates.....	33
7.4	The NAP _{CR1} pulsotype has a larger accessory genome and more gene clusters than NAP1 strains	35
7.5	NAP _{CR1} isolates have more distinctive mobile genetic elements in their accessory genomes than NAP1 isolates	42
7.6	Annotation and functional characterization of the differential MGE.....	50
7.7	MGE have a greater effect in the microdiversification of isolates from the NAP _{CR1} pulsotype than in NAP1 isolates.	81
8	DISCUSSION.....	86
9	CONCLUSIONS.....	93
10	REFERENCES	93
11	APPENDIX.....	105

RESUMEN

Clostridium difficile, el principal agente productor de diarreas a nivel hospitalario, se caracteriza por tener un genoma central pequeño y un mobiloma muy diverso. Análisis previos de la cepa epidémica NAP1/ST01 concluyeron que la especie es clonal. Sin embargo, estudios recientes sugieren que otros linajes de *C. difficile* pueden diversificar por recombinación y/o adquisición de elementos genéticos móviles (EGM) en lugar de mutaciones discretas. Un grupo diverso de aislamientos con patrones de macrorestricción distintivos pero todos agrupados por PFGE y MLST como NAP_{CR1}/ST54 produjo, en conjunto con la cepa NAP1/ST1, un brote en Costa Rica por razones aún no conocidas. Para confirmar que la inusual diversidad de NAP_{CR1} es producto principalmente de la microdiversificación de su genoma accesorio y no de la acumulación de mutaciones en el genoma central, se compararon aislamientos clínicos de NAP_{CR1}/ST54 y NAP1/ST1 que coexistieron en espacio y tiempo. A esta colección de secuencias genómicas se les determinó el número y tipo de SNPs en el genoma central, la tasa dN/dS, el tamaño del pangenoma, la cantidad de grupos de genes únicos y la presencia de EGM diferenciales. Los resultados indicaron que ambos pulsotipos acumulan mutaciones pero que las cepas del pulsotipo NAP1/ST1 tienen más mutaciones no-sinónimas y que, por tanto, están bajo el efecto de la selección purificadora positiva. Por el contrario, las cepas NAP_{CR1}/ST54 tienen un genoma accesorio y un pangenoma más grande, diverso y con más EGM. Estos EGM se asociaron a microdiversificación porque su presencia/ausencia coincide con la topología de un árbol de máxima parsimonia generado con matrices de comparación pangenómica. Cuando las secuencias de algunos EGM fueron removidas artificialmente de los genomas NAP_{CR1}/ST54, las distancias de las ramas en un árbol de presencia/ausencia de grupos de genes en el pangenoma colapsaron. La secuencia de estos EGM incluye genes que, de verificarse su anotación, podrían incrementar el potencial patogénico y la capacidad epidémica de las cepas NAP_{CR1}/ST54, tales como factores de virulencia y genes de resistencia a los antibióticos.

ABSTRACT

Clostridium difficile, the most common causal agent of hospital-acquired diarrhea, has a small core genome and a highly diverse mobilome. Previous analyses on the epidemic NAP1/ST01 strain led to the conclusion that this bacterial species is clonal. However, recent studies have suggested that other *C. difficile* lineages may be diversifying through recombination and/or acquisition of mobile genetic elements (MGE) instead of discrete mutations. A group of diverse isolates with distinct macrorestriction patterns which all grouped in PFGE and MLST as NAP_{CR1}/ST54, produced an outbreak in Costa Rica together with NAP1/ST1, for unknown reasons. To confirm that the unusual diversity of the NAP_{CR1} pulsotype is driven by microdiversification of its accessory genome rather than the accumulation of mutations in its core genome, NAP_{CR1}/ST54 and NAP1/ST1 clinical isolates that coexisted in space and time were compared. The number and nature of their core genome SNPs, dN/dS rates, feature frequency profiles, pangenomes sizes, number of unique gene clusters, and carriage of functional MGE were determined in these collection of genomic sequences. Altogether, the results indicate both groups of strains accumulate mutations, but NAP1/ST1 has more non-synonymous mutations than NAP_{CR1}/ST54, therefore positive purifying pressure is driving its microdiversification. As anticipated, the accessory genome and pangenome of the NAP_{CR1}/ST54 isolates was larger, more diverse and contained more MGE than that of NAP1/ST1 isolates. These MGE were associated with microdiversification since its presence/absence coincided with the topology of the parsimony-based pangenomic tree generated from comparative pangenomic matrices. Additionally, when these MGE were artificially removed from the NAP_{CR1} genomes, the distances from the maximum-likelihood phylogenetic tree originated from the presence/absence plot of gene clusters in the pangenome, collapsed. The sequences of these MGE include genes, which once their annotation is verified, could increase the pathogenic potential and the epidemic capacity of NAP_{CR1}/ST54 by providing virulence factors and antibiotic resistance genes.

LIST OF TABLES

	PAGE
Table 1. PFGE typing, hospital, and year of isolation of the analyzed NAP _{CR1} and NAP1 isolates	15
Table 2. Statistics of the assemblies used in the study	17
Table 3. Primers, annealing temperatures, and predicted sizes of the amplicons used for detection of circular intermediates and excision of the MGE identified through pangenome comparisons	23
Table 4. Number of SNPs, SNP densities, and dN/dS rates calculated for NAP _{CR1} isolates	25
Table 5. Number of SNPs, SNP densities, and dN/dS rates calculated for NAP1 isolates	27
Table 6. Origin and amount of unique gene clusters of representative NAP _{CR1} isolates from each cluster	43
Table 7. Origin and amount of unique gene clusters of representative NAP _{CR1} isolates from each cluster	43
Table 8. Differential MGE present in the NAP _{CR1} pangenome.	45
Table 9. Differential MGE found in the NAP1 pangenome.	47
Table 10. Annotation of the <i>mobCksgA</i> element of NAP _{CR1} isolates	50
Table 11. Annotation of the Tn5397 from NAP _{CR1} isolates	52
Table 12. Annotation of the Tn4001-like element from NAP _{CR1} isolates	53
Table 13. Annotation of the <i>skin</i> ^{Cd} element from NAP _{CR1} isolates	55
Table 14. Annotation of a putative prophage from NAP _{CR1} isolates	56
Table 15. Annotation of a putative plasmid from the NAP _{CR1} isolate 6289	59
Table 16. Annotation of giant phage version 1 from NAP _{CR} isolates	63
Table 17. Annotation of giant phage version 2 from NAP _{CR1} isolates	70
Table 18. Annotation of a putative plasmid from NAP1 isolates	79

LIST OF FIGURES

	PAGE
Figure 1. Phylogenetic tree of the MLST <i>C. difficile</i> clades	5
Figure 2. Core genome SNPs analyses for isolates from the NAP _{CR1} and NAP1 pulsotypes. (A) Total amount of SNPs found in coding regions, (B) SNP density per kb, (C) dN/dS rates	28
Figure 3. Rooted (A) and unrooted (B) phylogenomic trees of NAP _{CR1} isolates generated with SNP distance matrices and the maximum likelihood method	31
Figure 4. Rooted (A) and unrooted (B) phylogenomic trees of NAP1 isolates generated with SNP distance matrices and the maximum likelihood method	32
Figure 5. Average root-to-tip distances of isolates from the NAP _{CR1} and NAP1 pulsotypes in SNP-based phylogenomic trees	33
Figure 6. Feature-frequency profile tree of NAP _{CR1} - (A) and NAP1-isolates (B)	34
Figure 7. Average root-to-tip distances of isolates from the NAP _{CR1} and NAP1 pulsotypes in feature frequency profiles-based trees	35
Figure 8. Pangenome comparison of NAP _{CR1} (A) and NAP1 isolates (B)	36
Figure 9. Presence-absence plot of gene clusters in the pangenome of NAP _{CR1} (A) and NAP1 isolates (B)	38
Figure 10. Parsimony-based pangenomic tree of NAP _{CR1} isolates generated with Get_Homologues	40
Figure 11. Parsimony-based pangenomic tree of NAP1 isolates generated with Get_Homologues	41
Figure 12. Average root-to-tip distances of isolates from the NAP _{CR1} and NAP1 pulsotypes in parsimony-based pangenomic trees	42
Figure 13. Location of the differential MGE in a parsimony-based pangenomic tree calculated for the NAP _{CR1} isolates	48

Figure 14. Location of the differential MGE in a parsimony-based pangenomic tree calculated for the NAP1 isolates	49
Figure 15. ACT comparison of Tn4001 and the Tn4001-like transposon of the NAP _{CR1} isolates 6276 and 6289	54
Figure 16. Insertion of a novel prophage in CTn5 of the NAP _{CR1} isolate 2945	58
Figure 17. ACT comparison of the two giant phage variants found among the NAP _{CR1} isolates	63
Figure 18. PCR products obtained for circular intermediates of (A) Tn5397 in isolates 6279 and 6289, (B) <i>skin</i> ^{Cd} in isolates 2945, 5761 and 6289, (C) putative plasmid in isolate 6289	78
Figure 19. Roary pangenome analysis of NAP _{CR1} and NAP _{CR1} WGS modified through manual removal of selected differential MGE	83
Figure 20. Roary pangenome analysis of NAP1 and NAP1 WGS modified through manual removal of selected differential MGE	85

LIST OF ABBREVIATIONS

CDI: *Clostridium difficile* infections

PaLoc: Pathogenicity Locus

SlpA: surface layer protein

PFGE: Pulsed Field Gel Electrophoresis

DNA: deoxyribonucleic acid

NAP: North American Pulsotype

RT: ribotype

MLST: Multilocus Sequence Typing

WGS: whole-genome sequencing

ST: sequence type

MGE: mobile genetic elements

SNPS: single nucleotide polymorphisms

Indel: insertion or deletion

dN: non-synonymous substitutions

dS: synonymous substitutions

dN/dS: rate of non-synonymous vs. synonymous substitutions

r: recombination

m: mutation

r/m: rate of recombination vs. mutation

Tn: transposon

CTn: conjugative transposon

phi: prophage

bp: basepairs

kpb: kilo basepairs

CDS: coding sequences

FFP: feature frequency profiles

PCR: polymerase chain reaction

1. INTRODUCTION

1.1 General characteristics of *Clostridium difficile* infections (CDI)

Clostridium difficile is a Gram-positive, anaerobic bacterium, capable of producing toxins and spores (1). CDI are the main cause of hospital-acquired diarrhea prompted by the use of antibiotics and the most common nosocomial infection in developed countries (2, 3). They have a high impact in healthcare costs and affect millions of patients worldwide (4, 5). Only in the United States, 250 000 people suffer from CDI every year and the associated medical costs add up to at least \$ 1 billion (6).

CDI varies from mild to moderate diarrhea with fever to severe clinical presentations, including pseudomembranous colitis, toxic megacolon, systemic complications and death (1, 2). These pathologies are mostly acquired through exposure to spores in the hospital environment, although the number of community cases of CDI is on the rise (1, 2, 7).

The large clostridial toxins TcdA and TcdB have been traditionally regarded as the main virulence factors of *C. difficile* (1, 8). They inactivate small monomeric GTPases through their glucosyltransferase activity and thereby damage the actin cytoskeleton of intestinal epithelial cells, among other deleterious host cell effects (9, 10). In most *C. difficile* strains, the genes encoding TcdA and TcdB are found in a so-called pathogenicity locus (PaLoc) composed of 5 genes: *tcdR*, *tcdB*, *tcdE*, *tcdA* and *tcdC*. TcdR is a sigma factor promoting toxin synthesis, TcdC is an anti-sigma factor that counteracts TcdR activity, and TcdE is a holin (1, 11, 12). Other virulence factors described in this species include the binary toxin CDT, which affects epithelial microtubules on account of its ADPribosyltransferase activity,

flagellin for host colonization, and the surface layer protein (SlpA), which has been linked to inflammation and host cell adherence (13–17).

1.2 *Clostridium difficile* typing methods

Given that the pathogenicity and epidemic potential of strains differ, several methods have been applied to type *C. difficile* isolates (18, 19). For instance, Pulsed Field Gel Electrophoresis (PFGE) provides an overview of the whole genome through digestion of genomic DNA by the endonuclease *Sma*I. This digestion produces a specific DNA fragmentation pattern that can later receive a designation through comparison with the North American Pulsotype (NAP) databases (20, 21). Ribotyping is a PCR-based method that classifies isolates based on variations in the size of their 16S-23S rRNA intergenic spacing regions. The results obtained are deposited in the database of the *C. difficile* Ribotyping Network from Public Health England to obtain the ribotype (RT) (22). A third method, termed Toxinotyping, involves a PCR amplification of certain PaLoc fragments and their subsequent enzymatic digestion. The resulting patterns are compared to those deposited in a database maintained by Dr. Maja Rupnik in Slovenia (23). Finally, Multilocus Sequence Typing (MLST) consists of the simultaneous analysis of the sequences of 7 essential housekeeping genes. Thus it requires PCR and Sanger-sequencing or whole-genome sequencing (WGS). The sequences are compared to allele lists deposited in public databases and allele combinations define a sequence type (ST) (24, 25). Ribotyping and PFGE are the main techniques used for epidemiology worldwide. However, MLST is the preferred methodology for most phylogenetic studies.

1.3 Strains producing severe CDI cases

After the year 2000, an increase in the number of nosocomial CDI outbreaks and morbidity and mortality rates linked to CDI was reported in the United States, Canada, and the United Kingdom (1, 26). Most of these infections were caused by a strain classified as NAP1/RT027 that is characterized by mutations leading to

increased toxin production and fluoroquinolone-resistance, a high sporulation rate, and carriage of CDT (1, 27, 28). Today, non-NAP1/RT027 strains with epidemical potential and producing severe CDI have emerged (29). For example, the strain NAP7/RT078 has been frequently isolated from patients with severe CDI despite the lack of evident risk factors or underlying diseases. Even though this strain is usually found in piglets and pigs, an increase of human cases was initially reported in The Netherlands and later in other European countries (30). The A-B+ strain NAP9/RT017, which does not produce TcdA, has also caused outbreaks and severe CDI cases, particularly in Asia but also in Costa Rica (31, 32). In this Central American country, a novel strain called NAP_{CR1}/RT012 caused hospital outbreaks in 2009 but its distribution started to fall from the year 2012 onwards. These observations clearly show that the epidemiology of CDI is changing (32, 33).

1.4 The epidemic lineage NAP_{CR1} and its relation to the reference strain CD630

During an outbreak of CDI in a Costa Rican hospital, a novel group of strains typed as ST54 and RT012 coexisted with the NAP1/RT027 strain. This group of strains was denominated NAP_{CR1}. It affected younger patients, produced higher leukocytosis, more severe cases, and an increase in recurrences, compared to NAP1/RT027 strains (33). NAP_{CR1}/RT012 strains have circulated in seven Costa Rican hospitals from 2003 until now, though they have not caused outbreaks since 2009 (32).

As indicated by its classification in at least in ten different *Smal* patterns, the NAP_{CR1}/ST54 strains are unusually diverse. Moreover, they are intriguingly closely related to the *C. difficile* reference strain CD630 (RT012/ST54), which was isolated from a patient with pseudomembranous colitis in 1982 in Zurich, Switzerland (34). The genome of CD630 has been sequenced twice and is one of the best annotated genomes of this pathogen (35–37). A comparison of NAP_{CR1} and CD630 genomes showed that the former has 10% more predicted coding sequences (CDS) and, interestingly, those genes unique to NAP_{CR1} are mainly associated with

mobile genetic elements (MGE). Indeed, an average NAP_{CR1} genome is at least 6% larger than that of strain CD630 and contains almost the double amount of phages, prophages, transposons and plasmids. Furthermore, the NAP_{CR1} strains are resistant to fluoroquinolones due to a Thr82Ile mutation in *gyrA*, just like the NAP1/RT027 strain (33).

1.5 Phylogenomics of *C. difficile*

Two different MLST schemes, with comparable results to those obtained with microarrays and single nucleotide polymorphisms (SNPs) studies, have been applied to *C. difficile*. The most popular one is the scheme of Griffiths *et al.* (2010) which focuses on the genes *adk* (adenosine kinase), *recA* (recombinase), *sodA* (superoxide dismutase), *dxr* (deoxy-xilulose-P-reductoisomerase), *glyA* (serine hidroximetiltransferase), *tpi* (triosafosfate isomerase) and *atpA* (alpha subunit of ATP synthase). When applied to 152 isolates, this method recognized 40 ST and distributed them in 5 clades (Figure 1). The majority of the isolates were assigned to Clade 1 (including NAP_{CR1}/ST54), Clade 2 included the hypervirulent strain NAP1/RT027, Clade 3 had a few less reported isolates, Clade 4 was typified by the NAP9/RT017 strain, and a potential 5th clade included the NAP7/RT078 strain (25). This clade definition was very similar to that reported by Stabler *et al.*, who allocated 75 isolates to four groups (HA1, HA2, HY and A-B+) using microarray hybridizations. Group HA1 is equivalent to MLST Clade 1 and included isolates of human and animal origin. Group HY relates to MLST Clade 2 and also contained the NAP1 strain. Group HA2 matches MSLT Clade 3, as it included isolates from human and animal origins unrelated to hypervirulence. Finally, A-B+ isolates, including NAP9 strains, correspond to Clade 4. A comparison of these clustering methods is presented in Figure 1 (25, 38). More recently, two new lineages of *C. difficile* have been reported. Similar to the *Escherichia coli* cryptic clades, Clade C-I is very divergent and could be a subspecies, a new species or ancestor. The second one cryptic clade is Clade 6, which appears in between Clade 1 and Clade 2, although some authors consider that it is a sublineage of Clade 1 (39, 40).

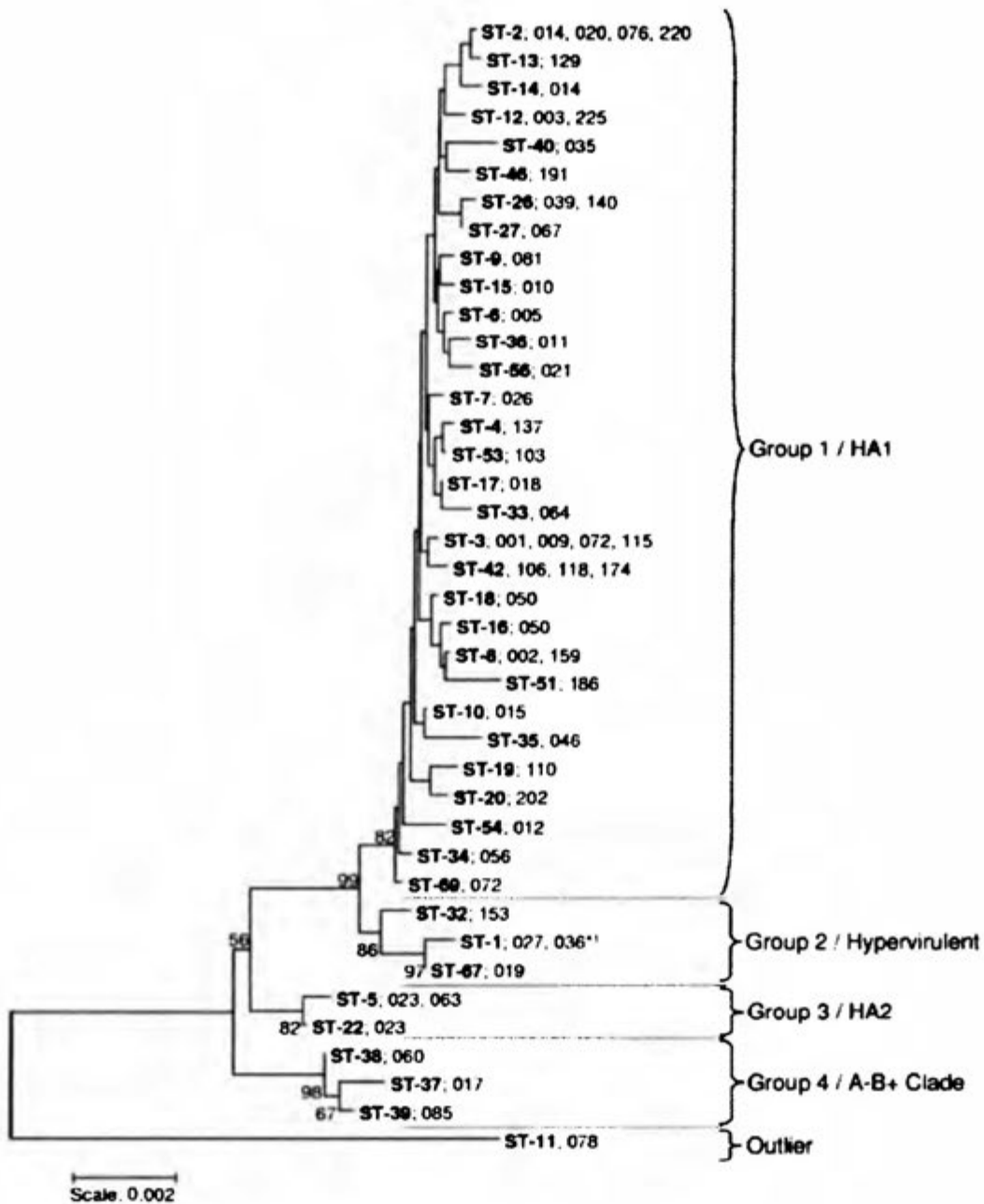


Figure 1. Phylogenetic tree of the MLST *C. difficile* clades. Taken from Griffiths *et al* (25).

Following the clades defined by Stabler *et al.*, one representative isolate from each clade was chosen by He *et al.* to study their divergence through core genome SNPs analyses (25, 41). From this comparison the authors concluded that there is

a great genetic diversity among the strains, with the NAP7 genotype being the more phylogenetically distant. Additionally, they found epidemic ribotypes associated with severe CDI in all clades, suggesting that virulence in *C. difficile* was acquired from a common ancestor (41).

A bacterial genome can be subdivided in a core genome and an accessory genome and the integration of all genomes of a species is known as pangenome. The core genome consists of all the genes shared by the isolates under analysis. In contrast, the accessory genome only includes those genes unique to each isolate or strain. The pangenome of a species contemplates both the core and the accessory genomes. Genes in the core genome are commonly involved in metabolic processes, biosynthesis, cell division, replication and gene regulation (18, 42). However, in *C. difficile* they also include genes for virulence factors, such as those related to adhesion and motility. Comparative genomic analyses of *C. difficile* have revealed that the core genome of this species is rather small (16% to 23%). Therefore, since the size of most *C. difficile* genomes ranges from 4.2 to 4.5 Mb, a great proportion of their genetic information is accessory and possibly linked to pathogenicity or adaptation. Indeed, the majority of the *C. difficile* pangenome is composed by MGE (38, 41).

1.6 Genome diversification in Bacteria

Bacterial genomes diversify by the effect of mutations or through recombination. In this regard, a mutation is defined as a change in a genetic sequence that produces a base substitution or an indel (insertion or deletion) (43) and recombination is regarded as the exchange of DNA fragments, as may occur when a recipient bacteria acquires DNA from the surrounding environment or from a donor organism. MGE are transferred by recombination (18).

Mutational and recombinational events can be distinguished through SNPs analyses because they impact the average SNP density of a genome differentially

(42). Moreover, the latter type of events can be identified through visual or bioinformatic sequence comparisons.

The outcome of a mutational event is determined by the intensity and direction of the natural selection, which can be neutral, positive purifying or negative purifying. Mutational events producing non-synonymous substitutions (dN) affect the coded protein and accumulate in a genome under the effect of positive purifying pressure. On the contrary, synonymous substitutions (dS) lead to silent mutations. Their accumulation indicate that negative purifying pressure is acting on coding regions in order to purge non-synonymous changes from the sequences. The dN/dS rate is used to estimate the effect of natural selection in coding sequences: while a dN/dS rate > 1 indicates changes in coding sequences because of positive purifying selection, rates < 1 signify that coding sequences are under the effect of negative purifying pressure. Finally, rates of 0 state that the selective pressure acting on coding sequences is neutral. Organisms suffering a clonal diversification are distinguished by dN/dS rates greater than 1 (18, 42, 44).

Another parameter used to weight the effect of mutations or recombination in genome evolution is the rate of nucleotide substitution from recombination (r) vs. mutation (m), r/m . Clonal species mainly evolve through acquisition of mutations and show r/m rates lower than 1. On the contrary, r/m rates above 1 typify organisms that diversify by recombination (18).

1.7 Mutations and recombination in *C. difficile*

Supporting the notion that mutation rather than recombination drives *C. difficile* evolution, He *et al.* analyzed the core genome of 9 representative isolates from each of the clades defined by Stabler *et al.* They calculated the dN/dS ratio for each strain and concluded that Clade 5 (NAP7) has diversified by negative purifying selection due to the high number of dS mutations detected in this group compared to the others. They also concluded that Clades 1 to 4 have diversified

mainly by positive purifying selection, as most point mutations were dN. These authors also determined that the $1/dN/dS$ rate was not linear, meaning that dN mutations were not efficiently purged in this species. Non-linear populations have already been reported for other species of the Firmicutes, such as *Streptococcus pyogenes* and the *Bacillus thurigiensis*, *B. anthracis*, *B. cereus* complex (41). This study also concluded that the role of recombination in *C. difficile* diversification is moderate, as indicated by the calculation of r/m rates ranging from 0.63 to 1.13 (41, 45). Similarly, Dingle *et al.* predicted a r/m ratio 0.08 for 77 STs from the five clades (11). Lemee *et al.* estimated that a *C. difficile* allele has 8-10 times greater possibility to diversify by mutation rather than by recombination (24). Further MLST analyses of virulence genes showed evidence of a clonal population with a possible coevolution of *tcdA*, *tcdB* and the binary toxin, with essential genes (46). Finally, Dingle *et al.* analyzed the MLST housekeeping genes of 1290 clinical isolates and calculated a dN/dS rate of < 1 , accounting for a negative purifying selection to preserve gene integrity (11).

Favoring the notion that recombination plays a stronger role than mutation in *C. difficile* diversification, Lemee *et al.* detected SNPs blocks and unusually large amount of polymorphisms in genes relevant for host colonization, such as those encoding flagella, the cell wall protein 66 (Cwp66) and SlpA (46). In addition, Didelot *et al.* reported that the r/m ratios of different STs may differ as does the diversity of *C. difficile* isolates from the same ST (47).

Microevolution studies have been performed only for the RT027 genotype. In a pivotal study by He *et al.*, the authors concluded that the RT027 strain suffered an expansion at the beginning of the century that coincides with the epidemic outbreaks. Additionally, they mention that it was hard to root their phylogeny analysis probably because of early recombination events (41). In another study on the same RT027 isolates, Castillo-Ramírez *et al.* detected 184 SNPs outside recombinational regions. Up to 64.7% of these variants were non-synonymous, 16.3% were synonymous and 19% were located in intergenic regions. When

recombination regions were included in the calculation, the total amount of SNPs detected increased almost 10 fold (n=1553), of which 47.5% were non-synonymous, 39.2% were synonymous and 13.3% were located in intergenic regions. This shows that the RT027 genomes include large recombinational blocks. Moreover, they conclude that the accumulation of synonymous SNPs in RT027 genomes relates to the acquisition of genes from foreign lineages (42).

1.8 The *C. difficile* mobilome and its biological effect

MGE are DNA fragments that codify for enzymes and other proteins capable of moving the fragment inside the genome (intracellular mobility) or between bacterial cells (intercellular mobility) and the mobilome is the group of MGE found in a species (18, 48).

Although MGE insertion in a genome may lead to gene acquisition, gene disruption or gene fusion, affect the nearby genomic regions through inversion, produce transcription breaks, and transactivate other MGE (18, 49, 50), studies on bacterial population genetics classically focuses on variants found in the core genome and disregard the accessory genome irrespective of the adaptive advantages that the latter can confer to its hosts.

Several types of MGE have been found in *C. difficile*. Of them, the best annotated come from strain CD630, which includes introns and *I*Strons, integrative and conjugative elements (ICE), *skin* (a prophage-like element inserted in the sporulation gene *sigK*) and several prophages and bacteriophages (35). These MGE have been found in different lineages from all known Clades and some of them are shared by phylogenetically distant strains (49, 51). The reasons behind the occurrence of a rather large and diverse repertoire of MGE in *C. difficile* is unknown, especially if one considers that it possesses the DNA repair system RecA and several CRISPR-Cas systems (clustered regularly interspaced short

palindromic repeats), which are defense mechanisms against bacteriophage infections and MGE insertions (52–54).

The transposable elements described in the pathogen usually contain antibiotic resistance genes. For instance, Tn5397 has *tetM* and Tn6164 contains *tet* (44), which encode for ribosomal protection proteins conferring resistance to tetracycline (55–57). Tn5398 and Tn6215 contain *ermB*, whose product is a methyltransferase of the 23S rRNA that modifies the target of clindamycin, erythromycin and streptogramin type B (58, 59). In addition to *ermB*, some Tn6218 elements also have *cfr*, which codes for a methyltransferase of the 23S rRNA and generates a multiresistance phenotype against chloramphenicol, lincosamides, oxazolidones, pleuromutilines and type A streptogramins (60). Specifically in CD630, the Tn4453a transposon includes the gene for an acetyltransferase conferring resistance against chloramphenicol termed *catD* (61). Moreover, other putative transposons are CTn1 to CTn7, which contain efflux pumps or ABC transporters conferring resistance to tetracycline, chloramphenicol, erythromycin and possibly other antibiotics (35). Some NAP7 isolates from human and porcine origin contain Tn6164 with resistance genes to tetracycline and aminoglycosides inserted in a genomic island with non-clostridial MGE (62). Finally, transposable elements similar to Tn916 and Tn1549 from *E. faecalis* have been found (63).

C. difficile has many prophages whose host range has not been determined. Nonetheless, they have been detected in different RT and in isolates from human and animal origin. Most of these prophages belong to the *Myoviridae* family with contractile non-flexible tails (phiCD119, phiC2, phiCD27, phiMMP02 and phiMM04), though some members of the *Siphoviridae* family with flexible non-contractile tails have been found as well (phiCD38-2 and phiCD6356). All of these phages have integrases, hence they are expected to undergo lysogeny. Some researchers have confirmed that *C. difficile* phages influence the virulence and/or adaptability of their hosts. For example, phiC2 mediates the transfer of Tn6215; phiCD119, phiCD38-2 and phiCD27 modulate toxin production, and phiCDHM1

possesses genes homologous to *agr* participating in quorum sensing. Finally, phiMMP02 and phiMMP04 have been isolated from patient samples, implying that they are induced during CDI (64–70).

He *et al.* produced a rooted phylogeny of strains 630, BI-9, CF5, M68, M120, CD196, BI-1, 2007855 and R20291, which represent the main MLST *C. difficile* Clades. They were able to trace back genomic insertions and deletions of MGE and reported a large proportion of putative conjugative transposons and bacteriophages in the *C. difficile* mobilome. Similar results were obtained in a study on 25 RT027 isolates (41). In a comparison between two RT027 strains, isolates CD196 and R20291, and strain CD630, 234 unique genes were found in at least 50 different genomic regions. These regions were a phage island, transposon genes, two-component response regulators, drug resistance genes, transporter genes and type I restriction enzyme/restriction modification genes. Moreover, in a comparison between strains R20291 and CD196, five genetic regions were unique to R20291, including a phage island (Stoke Mandeville phage island), a prophage with slight variations, the loss of 3 CDS for a putative protein and a region encoding genes for a multiantimicrobial extrusion family drug/sodium antiporters (71). In conclusion, differences between CD630 and RT027 isolates or between RT027 isolates mainly include MGE.

2. SCIENTIFIC QUESTION

Although both NAP_{CR1} and NAP1 strains coexisted during an outbreak in a Costa Rican hospital, a much larger diversity of the former group was observed. The mechanisms behind the genome diversification of the NAP_{CR1} lineage and its implications in virulence are unclear.

3. JUSTIFICATION

Although several authors agree that *C. difficile* expands clonally, this species is characterized by a high degree of genetic variability. The majority of studies have only included a few lineages, most notably NAP1 strains, and they have been restricted to the core genome, which is small in *C. difficile*. Moreover, most researchers have focused on the evolution of the PaLoc or other virulence factors, hence only few studies have compared the effect of mutation and recombination in the diversification of *C. difficile* at a genomic level. The results of these investigations are inconclusive and often contradictory.

This work focuses on determining the genetic mechanisms behind the diversification of NAP_{CR1} strains in the Costa Rican hospital environment. This knowledge has the potential to provide explanations for the increased diversity, virulence, and outbreak-causing capacity of this emerging lineage, which remains unknown.

4. HYPOTHESIS

The acquisition of MGE in the accessory genome, rather than the accumulation of mutations in its core genome, is a more important mechanism of diversification in NAP_{CR1} strains compared to NAP1 Costa Rican isolates.

5. MAIN AIM

To compare the effect of mutational events in the core genome and the acquisition of MGE in the accessory genome as mechanisms of microdiversification in NAP_{CR1} and NAP1 Costa Rican isolates.

5.1 Specific aims

- 5.1.1 To identify SNPs in the core genome of NAP_{CR1} and NAP1 isolates of *C. difficile* and estimate their contribution to the microdiversification of both lineages.
- 5.1.2 To estimate the diversity of the pangenome of NAP_{CR1} and NAP1 isolates of *C. difficile* to assess the contribution of the accessory genomes to their microdiversification.
- 5.1.3 To identify hypervariable genomic fragments and putative MGE in the accessory genomes of selected NAP_{CR1} and NAP1 isolates of *C. difficile* which could be related to microdiversification.
- 5.1.4 To define structurally and functionally the MGE that maximize the differentiation of NAP_{CR1} and NAP1 isolates and to estimate their contribution to the microdiversification of the lineages.

6. MATERIALS AND METHODS

6.1 Bacterial isolates and WGS

This study focused on 32 NAP_{CR1} and 17 NAP1/001 isolates from CDI patients that received attention in the following 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) between 2003 and 2012 (Table 1). Draft whole genome sequences (WGS) for all of the analyzed

isolates were obtained by sequence-by-synthesis at the Wellcome Trust Sanger Institute (UK). Some representative NAP_{CR1} isolates were resequenced using Single Molecule Real Time (SMRT) sequencing in a PacBio platform at the Leibniz-Institute DSMZ (Germany). The quality control of the Illumina reads included comparisons of their %GC, mapping to reference bacterial genomes, and determinations of the matching yields against *C. difficile* CD196. Illumina reads were assembled with Velvet (72) or Edena (73) then mapped back to assembly contigs to correct for misassemblies. Statistics for the Velvet assemblies are shown in Table 2. Edena assemblies were used only for the study of representative isolates. This sequencing data can be downloaded from the European Nucleotide Archive (Study PRJEB5034). SMRT reads, in turn, were assembled with HGAP 3 and error correction was done with Bridgemappper (74), no accession number is available yet for these assemblies. The analyzed isolates suffered no more than 5 culture passages before DNA extraction for WGS was performed. ORF prediction was done with Prodigal (75) and WGS were annotated with Prokka and custom *C. difficile* databases (76). The annotated genomes of *C. difficile* CD630 (AM180355) and *C. difficile* R20291 (FN545816) were used as reference genomes for NAP_{CR1} and NAP1 isolates, respectively.

Table 1. PFGE typing, hospital, and year of isolation of the analyzed NAP_{CR1} and NAP1 isolates.

PFGE	Smal pattern	Isolate	Hospital	Year
NAP _{CR1}	442	3147	HSJD	2003
	447	5701	HSJD	2009
		5711	HSJD	2009
		5767	HCG	2009
		5771	CENARE	2009
	448	2784	HBC	2003
		3125	HSJD	2003
		3137	HSJD	2003
		5434	HBC	2003
		5704	HSJD	2009
		5707	HSJD	2009
		5733	HSJD	2009
		5751	HMX	2009
		5774	HMX	2009
		6275	HMX	2011-2012
	449	3129	HSJD	2003
		5719	HSJD	2009
		5755	HMX	2009
		5772	HSVP	2009
		6276	HMX	2011-2012
		6289	HMX	2011-2012
	452	5734	HSJD	2009
	487	2945	CENARE	2009
		5763	HCG	2009
	488	2992	HCG	2009
	489	5761	HEB	2009
		5762	HEB	2009
	558	3145	HSJD	2003
		6285	HMX	2011-2012
	578	3144	HSJD	2003
3150		HSJD	2003	
5436		HBC	2003	

Table 1. PFGE typing, hospital, and year of isolation of the analyzed NAP_{CR1} and NAP1 isolates (continued).

PFGE	<i>Sma</i> I pattern	Isolate	Hospital	Year
NAP1	001	5700	HSJD	2009
		5703	HSJD	2009
		5705	HSJD	2009
		5706	HSJD	2009
		5708	HSJD	2009
		5709	HSJD	2009
		5710	HSJD	2009
		5713	HSJD	2009
		5714	HSJD	2009
		5718	HSJD	2009
		5720	HSJD	2009
		5749	HMX	2009
		5758	HBC	2009
		5759	HBC	2009
		5764	CENARE	2009
		5765	HSVP	2009
		5768	HCG	2009

Table 2. Statistics of the WGS assemblies used in the study.

Pulsotype	Isolate	Total length	Number of contigs	Contig mean length	Longest contig (bp)	Shortest contig (bp)	Number of n	Gaps	N50
	2784	4513122	62	72792.29	326372	317	3349	14	165946
	2945	4598748	87	52859.17	299452	311	3481	15	165906
	2992	4543631	91	49930.01	286334	362	2050	9	140269
	3125	4546460	72	63145.28	547813	395	3735	17	211769
	3129	4545745	81	56120.31	317892	325	5613	26	211721
	3137	4512451	66	68370.47	547832	329	3859	18	187283
	3144	4555159	67	67987.45	286240	135	7993	32	211702
	3145	4553141	67	67957.33	547733	307	7808	33	200755
	3147	4544596	59	77027.05	547735	316	5681	23	200355
	3150	4549284	63	72210.86	500457	308	8369	35	206405
NAP_{CR1}	5434	4518929	56	80695.16	575941	383	9605	38	213100
	5436	4550846	61	74604.03	286260	123	8800	38	206401
	5701	4512151	54	83558.35	547820	311	6405	25	209967
	5704	4549499	56	81241.05	547739	307	9311	40	212974
	5707	4507991	73	61753.3	286278	331	4323	16	164084
	5711	4537289	97	46776.18	547807	310	4879	20	159110
	5719	4539549	78	58199.35	446658	312	3509	16	159106
	5733	4548341	67	67885.69	370022	307	12354	49	168429
	5734	4513340	63	71640.32	370019	362	11083	46	171394
	5751	4548016	58	78414.07	370331	422	7034	31	206155
	5755	4550036	60	75833.93	351142	307	7642	29	171400

Table 2. Statistics of the WGS assemblies used in the study (continued).

Pulsotype	Isolate	Total length	Number of contigs	Contig mean length	Longest contig (bp)	Shortest contig (bp)	Number of n	Gaps	N50
NAP _{CR1}	5761	4500411	61	73777.23	547760	359	7734	31	211763
	5762	4499203	61	73757.43	547778	302	8732	34	213085
	5763	4609505	59	78127.2	583899	316	6529	30	213619
	5767	4548673	63	72201.16	370075	334	7340	28	213537
	5771	4554433	62	73458.6	548268	307	8822	37	211716
	5772	4552363	67	67945.72	369959	307	6729	26	213016
	5774	4551618	64	71119.03	547835	302	7940	30	206409
	6275	4521469	51	88656.25	547716	307	9376	38	213644
	6276	4535622	73	62131.81	370015	307	7498	36	185759
	6285	4553447	58	78507.71	556230	307	8444	35	216422
6289	4622692	63	73376.06	547797	302	8918	36	211712	
NAP1	5700	4181811	52	80419.44	407848	355	2675	11	217205
	5703	4180657	46	90883.85	424833	355	1385	7	218229
	5705	4125379	52	79334.21	408227	375	2567	9	148986
	5706	4123808	51	80858.98	408300	336	836	4	201810
	5708	41311956	46	89825.13	425101	452	4338	18	236318
	5709	4128976	51	80960.31	407399	355	2170	8	191539
	5710	4129568	51	80971.92	762773	360	1785	7	187329
	5713	4127874	53	77884.42	425115	311	3813	14	149047

Table 2. Statistics of the WGS assemblies used in the study (continued)

Pulsotype	Isolate	Total length	Number of contigs	Contig mean length	Longest contig (bp)	Shortest contig (bp)	Number of n	Gaps	N50
	5714	4090142	46	88916.13	407835	452	1629	7	186488
	5718	4128653	54	76456.54	407368	355	1722	7	149086
	5720	4179942	48	87082.12	425148	401	2312	10	217637
	5749	4132034	51	81020.27	424529	379	1527	6	149040
NAP1	5758	4126645	51	80914.61	353728	353	3015	14	149122
	5759	4128317	45	91740.38	425370	452	3292	15	217746
	5764	4097786	50	81955.72	426234	379	1461	7	236324
	5765	4136133	51	81100.65	469362	405	2356	9	217981
	5768	4135683	44	92992.8	425661	355	2018	10	289109

6.2 SNPs calling and core SNP phylogeny

The pipeline Breseq was used to determine core genome SNPs, the number of SNPs of each isolate, SNP densities per kb (amount of SNPs/genome size *1000), and to classify SNPs in coding sequences (CDS) as synonymous or non-synonymous mutations and thereby estimate dN/dS rates. The pipeline maps short sequences to the reference genome using Bowtie2. It performs mapping through two steps, the first one in stringent conditions to obtain perfect matches and a second one aligns previous unmapped reads in a more relax phase. The output includes SNP location, coverage and annotation to facilitate interpretation. The pipeline was run with the default parameters and a minimum threshold of 25 reads was used to call SNPs. Mutations in intergenic regions, large deletions, new junctions, and MGE-associated SNPs were discarded from downstream analyses. The results obtained for NAP_{CR1} and NAP1 isolates did not had a normal distribution, thus they were compared using Mann–Whitney *U* test.

The CFSAN SNP pipeline (79) and Seaview (80) were used to generate core SNPs alignments and maximum likelihood bootstrapped trees. CFSAN maps short reads to a reference genome using Bowtie2, generate pileups of the files with SAMtools, calls variant sites with VarScan and produces a SNP list from which the SNP matrix is generated. The pipeline was run with default parameters, thus with a coverage of 20 reads for SNP calling and a *p*- value of 0.98. When required, a NAP4 isolate (LIBA-2812), a NAP1 isolate (LIBA-5750), or a NAP_{CR1} isolate (LIBA-6289) were used as outgroups. Root-to-tip distances were calculated for NAP_{CR1} and NAP1 isolates to estimate SNP distances. These values were compared using Mann–Whitney *U* tests, since they did not have a normal distribution.

6.3 Analyses of feature frequency profiles (FFP)

To determine differences in the accessory genome and MGE found outside the core genome that could be related to the microdiversification of the strains, WGS were compared using FFP (81). This is an alignment-free method that detects differences

in relative *l*-mer frequencies to calculate distance scores and can be applied even when the WGS under study do not share genes with high similarity. Here, *l*-mers of 20 nt were used to find a compromise between discrimination potential and computational capacity. The rest of the analysis was done with the default parameters of the pipeline. Comparison matrices from the derived distance of a multistate unordered characters model of feature frequency were transformed with the neighbor-joining method into trees in which distances represent the number of character feature changes and were visualized using FigTree (<http://tree.bio.ed.ac.uk/software/figtree/>). Root-to-tip distances calculated for NAP_{CR1} and NAP1 isolates were compared using Mann–Whitney *U* tests, since they did not had a normal distribution.

6.4 Proteome predictions and pangenome comparisons

Roary (82) and Get_Homologues (83) were used to predict unique gene/sequence clusters/proteins as an input to detect MGE and to facilitate the comparison of the NAP_{CR1} and NAP1 pangenomes. In detail, Roary was employed to estimate the size of the core- and accessory genomes to generate a gene presence-absence spreadsheet, and an approximate-maximum-likelihood phylogenetic tree from the accessory genome. It classifies genes in four categories according to their frequency of occurrence: core genes (99% ≤ strains ≤ 100%), soft core genes (95% ≤ strains < 99%), shell genes (15% ≤ strains < 95%) and cloud genes (0% ≤ strains < 15%) Get_Homologues, in turn, produces pangenome matrices from which parsimony-based pangenomic trees can be derived. Trees were visualized with FigTree and root-to-tip distances obtained for NAP_{CR1} and NAP1 isolates were compared using Mann–Whitney *U* tests, since they did not have a normal distribution. Roary was run with the default parameters of the pipeline. Get_Homologues was run with the default parameters with the exception of considering all the possible clusters, including sequences from a single genome.

6.5 Identification of unique gene clusters and MGE

Based on the results of the Get_Homologues pipeline, four NAP_{CR1} and six NAP1 isolates were selected for further analyses according to their cluster location and their

branching distances. The unique gene clusters predicted for these representative isolates were highlighted in their genomes with Artemis (77) and their contigs were compared to cognates from reference genomes using WebACT/ACT (78) to spot unshared regions that resembled MGE. Criteria such as presence of genes from known MGE (phage proteins or recombinases), %GC deviations, and NCBI databases searches were used to define MGE. Putative MGE were annotated using Prokka (76), and BLAST (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>) or Interpro (<https://www.ebi.ac.uk/Interpro/search/sequence-search>) searches. A list of differential MGE was raised according to their presence-absence in the representative isolates. To confirm their role in the microdiversification of the NAP_{CR1} genotype, the Roary analyses were repeated with WGS in which these discriminatory MGE were deliberately removed.

6.6 Functional studies of selected MGE of NAP_{CR1} isolates

To prove that the discriminative MGE are functional, a PCR-based approach for detection of circularization and excision events was performed. The primers and temperatures used are detailed in Table 3. The reaction was done with the Platinum SuperFi DNA Polymerase (ThermoFisher), and genomic DNA obtained from overnight cultures in BHI medium in anaerobic conditions. A typical amplification program consisted of: 2 min at 95°C, 10 s at 95°C, 10 s at the respective annealing temperature, 45 s at 68°C and 5 min at 68°C.

Table 3. Primers, annealing temperatures, and predicted sizes of the amplicons used for detection of circular intermediates and excision of the differential MGE identified through pangenome comparisons.

Isolate(s)	MGE ^a	Forward (5' - 3')	Reverse (5' - 3')	Annealing temperature	Predicted amplicon size (bp)
2945	<i>mobCksgA_C</i>	CCGTCTGGTTCTCGGCTAAT	AGCTATGACGAGAACGGCAC	56°C	300
2945	<i>mobCksgA_E</i>	CGTTGATGTCAAGAAACATGGA	GTAGGTGCAGGACTTGGAGC	56°C	400
5761/6289	<i>mobCksgA_C</i>	TCAATGGCATTCCGCAACAC	TCATGGAAGTGTGCGCAGAC	56°C	520
5761/6290	<i>mobCksgA_E</i>	ACCGTATCAAAAAGCCCCGT	AGACCCTTGTTGTTGCCCTC	56°C	500
6276	<i>mobCksgA_C</i>	TCATGGAAGTGTGCGCAGAC	TCAATGGCATTCCGCAACAC	56°C	520
6276	<i>mobCksgA_E</i>	AGACCCTTGTTGTTGCCCTC	ACCGTATCAAAAAGCCCCGT	56°C	500
2945/5761/6289	<i>skin_C</i>	CCTATACAGGTGCTTTCCTA	ACCATGATTCAGATTCCCTTGG	52°C	1400
2945	<i>skin_E</i>	AGCCATAAGGAGTTAACCCA	ACATCAATAGCTTCCTCAACAC	52°C	1300
2945/5761/6291	<i>skin_E</i>	AGCCATAAGGAGTTAACCCA	AGAGATGGAGGAACTAAGAT	51°C	1300
2945	Tn5397_C	TGAACAAGCAGAGGTAGTGCA	ACGATTTTATCCTCGCCAGCA	57°C	650
2945	Tn5397_E	AGACACCTGCTAAGAACCGC	TCTTCTGTTGCTGATAGAGT	52°C	600
6276/6289	Tn5397_C	AGCAGAGGTAGTGCAAAGCT	ACCGATTTTGTAGCCCTCGG	56°C	1000
6276/6290	Tn5397_E	AGACACCTGCTAAGAACCGC	AGGCTCTTGATGTTCTTCCA	54°C	350
6276	Tn4001-like_C	GCACCCTCTGCAAATTTTGTCT	GAACCATAACCTTTGTCTTG	52°C	520
6276	Tn4001-like_E	GCAACATTCAAAGCTGCCCA	TGGCTAGATAGTATAGTTGGAG	56°C	400
6289	Tn4001-like_C	GAACCATAACCTTTGTCTTG	TCTTCGCCTTGTTCAAACCTCA	52°C	400
6289	Tn4001-like_E	TTTGTCAAGGGCTTGTTGCG	CCGTAAAGTCTTTGCACAGT	54°C	450
2945	Prophage	GGGAACTTGCCATATCGTGC	TCGTACACGGTATCGCATGG	58°C	330
2945	Prophage	GCAAAAAGCCGCCGAAAAGG	AGCTGCAAGAGAATCAACCCT	58°C	500
6289	Putative plasmid	ACTTCCTTTTTGTTGTGCCA	TGACATTGCAATGACTGATG	52°C	500

^a C= circularization, E=excision.

6.7 Amplicon sequencing and analysis

The amplicons obtained were purified with the QIAquick PCR Purification Kit (Qiagen) and capillary sequenced in both directions through traditional Sanger methods (Macrogen). Sequence editing, assembly, alignments and pairwise comparisons of the sequenced amplicons against the corresponding WGS were done with Geneious. Figures of the PCR products and the predicted MGE were generated with Geneious.

7. RESULTS

7.1 NAP_{CR1} and NAP1 isolates have slightly different SNP densities and dN/dS rates

Although the NAP_{CR1} genomes (4499203-4622692 bp) are in average 378 907 bp larger than their NAP1 cognates (4090142-4181811 bp), and that a much smaller proportion of the reads obtained for NAP_{CR1} isolates mapped to its reference sequence (87-91%) compared to NAP1 isolates (96-99%) (Tables 4 and 5). The former group of isolates only showed 5 more SNPs in average and a 10% higher average SNP density (0.55 vs. 0.50) than the NAP1 isolates (Figure 2). Despite these rather subtle differences, the dN/dS rate calculated for the NAP_{CR1} isolates (2.47) was three fold lower than that obtained for the NAP1 isolates (6.05) (Fig 2C). Both dN/dS rates were >1. These observations also hold true when NAP_{CR1} isolates from single *Sma*I patterns were pairwise compared to NAP1 isolates (Appendix 1). Interestingly, NAP_{CR1} isolates from the 487 *Sma*I pattern showed a greater dN/dS rate (4.95) than the rest of the isolates of the pulsotype (Appendix 1).

Table 4. Number of SNPs, SNP densities, and dN/dS rates calculated for NAP_{CR1} isolates.

<i>Smal</i> pattern	Isolate	Genome size (bp)	Mapped to CD630	Total number of SNPs	Average number of SNPs	Average SNP density (per kb)	Average SNP density	dN	dS	dN/dS	Average dN/dS rate
442	3147	4544596	90.2%	24	24	0.53	0.53	15	9	1.67	1.67
	5701	4512151	92.0%	27		0.60		18	9	2.00	
447	5711	4537289	90.5%	27	25	0.60	0.55	18	9	2.00	2.61
	5767	4548673	90.1%	23		0.51		17	6	2.83	
	5771	4554433	90.3%	23		0.51		18	5	3.60	
	2784	4513122	91.2%	23		0.51		17	6	2.83	
448	3125	4546460	90.4%	22	25	0.48	0.56	15	7	2.14	2.33
	3137	4512451	92.2%	23		0.51		15	8	1.88	
	5434	4513340	91.1%	25		0.55		18	7	2.57	
	5704	4549499	91.0%	27		0.59		17	10	1.70	
	5707	4507991	91.3%	29		0.64		18	11	1.64	
	5733	4548341	90.2%	28		0.62		20	8	2.50	
	5751	4548016	90.8%	24		0.53		16	8	2.00	
	5774	4551618	90.2%	22		0.48		17	5	3.40	
	6275	4521469	91.7%	29		0.64		21	8	2.63	
	449	3129	4545745	90.6%		22		25	0.48	0.54	
5719		4539549	89.5%	27	0.59	18	9		2.00		
5755		4550036	90.2%	23	0.51	18	5		3.60		
5772		4552363	90.5%	25	0.55	18	7		2.57		
6276		4535622	90.0%	26	0.57	18	8		2.25		
6289		4622692	89.9%	25	0.54	17	8		2.13		

Table 4. Number of SNPs, SNP densities, and dN/dS rates calculated for NAP_{CR1} isolates (continued).

<i>Smal</i> pattern	Isolate	Genome size (bp)	Mapped to CD630	Total number of SNPs	Average number of SNPs	Average SNP density (per kb)	Average SNP density	dN	dS	dN/dS	Average dN/dS rate
452	5734	4513340	91.1%	27	27	0.60	0.60	18	9	2.00	2.00
487	2945	4598748	87.0%	25	25	0.54	0.54	22	3	7.33	4.95
	5763	4609505	88.3%	25		0.54		18	7	2.57	
488	2992	4543631	90.0%	23	23	0.51	0.51	15	8	1.88	1.88
489	5761	4500411	90.4%	26	27	0.58	0.60	20	6	3.33	2.92
	5762	4499203	91.3%	28		0.62		20	8	2.50	
558	3145	4553141	90.3%	25	26	0.55	0.57	16	9	1.78	1.89
	6285	4553447	90.2%	27		0.59		18	9	2.00	
578	3144	4555159	90.0%	24	24	0.53	0.52	16	8	2.00	1.87
	3150	4549284	90.5%	26		0.57		16	10	1.60	
	5436	4550846	90.0%	21		0.46		14	7	2.00	
Average		4550197	90.4%	24	25	0.54	0.56	18	8	2.5	2.58

Table 5. Number of SNPs, SNP densities, and dN/dS rates calculated for NAP1 isolates.

NAP1 isolate	Genome size	Mapped to R20291	Number of SNPs	Average SNP density (per kbp)	dN	dS	dN/dS
5700	4181811	96.20%	26	0.62	21	3	7.00
5703	4180657	96.20%	21	0.50	24	3	8.00
5705	4125379	98.10%	17	0.41	18	3	6.00
5706	4123808	97.00%	21	0.51	18	3	6.00
5708	4131956	96.50%	23	0.56	18	5	3.60
5709	4128976	97.80%	21	0.51	16	5	3.20
5710	4129568	97.40%	22	0.53	17	5	3.40
5713	4127874	96.40%	24	0.58	19	5	3.80
5714	4090142	99.30%	19	0.46	17	2	8.50
5718	4128653	97.10%	19	0.46	17	2	8.50
5720	4179942	95.20%	22	0.53	20	2	10.00
5749	4132034	96.90%	21	0.51	16	5	3.20
5758	4126645	98.00%	17	0.41	15	2	7.50
5759	4128317	97.70%	23	0.56	17	6	2.83
5764	4097786	99.50%	20	0.49	18	2	9.00
5765	4136133	97.60%	19	0.46	17	2	8.50
5768	4135683	97.40%	19	0.46	15	4	3.75
Average	4134433	97.31%	21	0.50	18	3	6.05

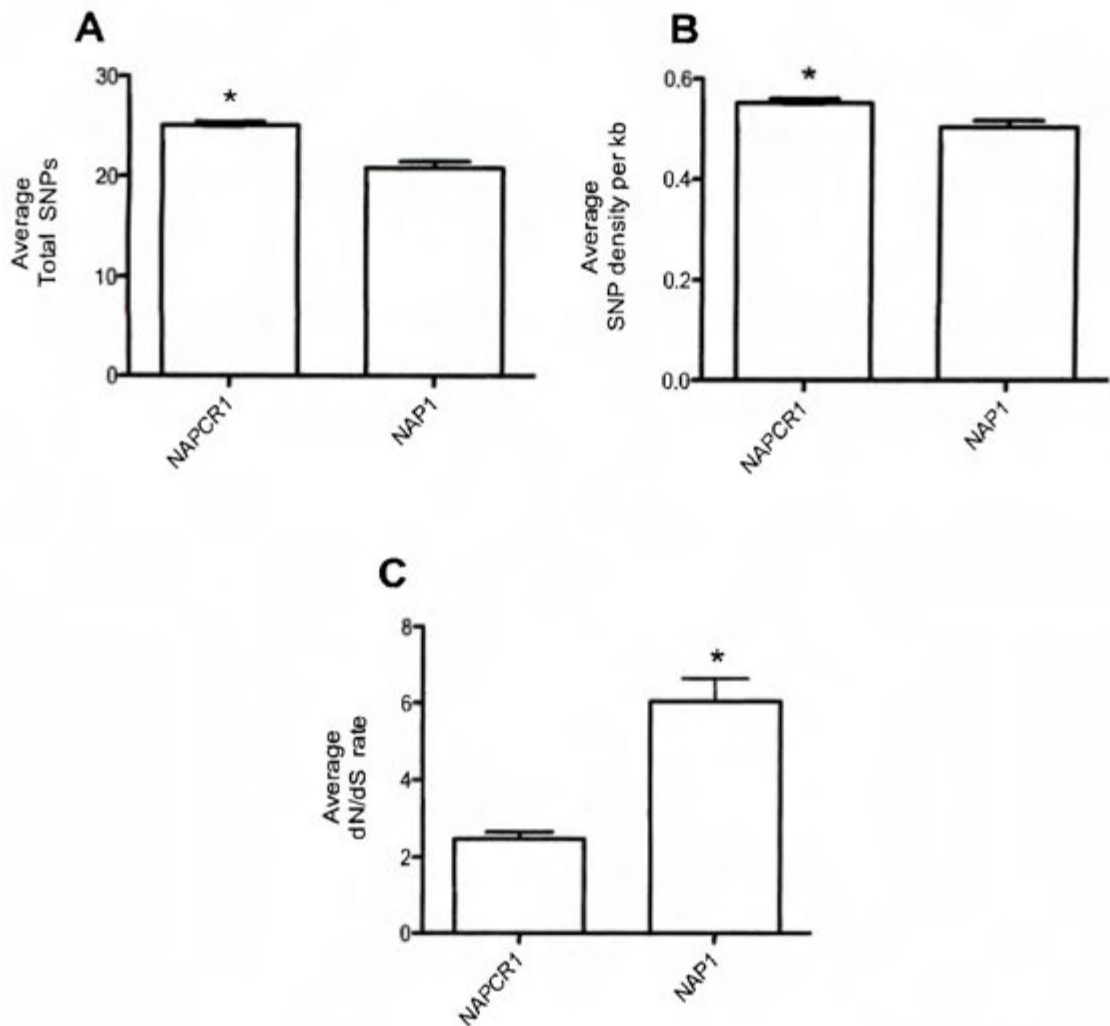


Figure 2. Core genome SNPs analyses for isolates from the NAP_{CR1} and NAP1 pulsotypes. (A) Total amount of SNPs found in coding regions, (B) SNP density per kb, (C) dN/dS rates. *C. difficile* 630 and R20291 were used as reference strains for NAP_{CR1} and NAP1 isolates, respectively. Asterisks above bars depict differences at a level of significance of $P < 0.05$ as indicated by Mann–Whitney U tests.

With a few exceptions, the non-synonymous mutations detected in the NAP_{CR1} and NAP1 isolates differed qualitatively. In the former group of isolates, mutations were observed in the genes encoding the DNA gyrase subunit A, putative exosporium glycoproteins, a putative transcriptional regulator activator Mor, an ABC-type

transport system, an oligopeptide-family ATP-binding protein, a putative penicillin-binding protein, Rnase Y, a putative drug/sodium antiporter from the MATE family, a two-component response regulator, a two-component sensor histidine kinase, glyceraldehyde-3-phosphate dehydrogenase, a transporter from the MFS superfamily, glucose-specific IIBC and lichenan-specific IIA/IIC components of the PTS system, and the precursor of the S-layer protein (Appendix 2). By contrast, the NAP1 isolates showed non-synonymous mutations in genes for a glucosamine-fructose-6-phosphate aminotransferase, a chromate transporter, the beta subunit of an electron transfer flavoprotein, an aconitate hydratase, a drug/sodium antiporter, a N-acetylmuramoyl-L-alanine amidase, a aminoacid ABC transporter ATP-binding protein, the subunit IIC2 of the glucitol/sorbitol-specific transporter of the PTS system, a two-component sensor histidine kinase, a UDP-N-acetylmuramoylalanine-D-glutamate ligase, the subunits IIA and IIABC of the glucose-specific transporter of the PTS system, and the S-layer protein (Appendix 3). SNPs located in non-coding regions are found in Appendix 4 for NAP_{CR1} and Appendix 5 for NAP1.

7.2 The core genome of the NAP_{CR1} isolates is more heterogenous than that of NAP1 isolates

With three exceptions (isolates 2945, 5763, and 5766), a SNP-based tree using outgroups was not useful to discriminate the NAP_{CR1} isolates despite its confident topology (Fig. 3A). A second tree produced without outgroups (Fig. 3B) confirmed the high divergence of isolates 2945 and 5763 from *SmaI* pattern 487 and separated with high confidence one of the *SmaI* 578 isolates (3144) from the rest of the isolates. The remaining isolates were distributed in various clusters that did not display temporospatial trends and included bacteria from different macrorestriction patterns (Fig. 3B). Once again, a rooted tree based on core SNPs failed to differentiate the NAP1 isolates (Fig. 4A) and various groups of low confidence were defined when the outgroup was removed (Fig 4B). As indicated by the average number of substitutions per site of the trees without outgroups (NAP_{CR1}=0.02 vs. NAP1=0.01), the NAP_{CR1} isolates are more diverse than the NAP1 isolates. This observation was confirmed by

the significantly larger root-to-tip distance calculated for the NAP_{CR1} isolates, which doubled that obtained for the NAP1 isolates (Fig 5).

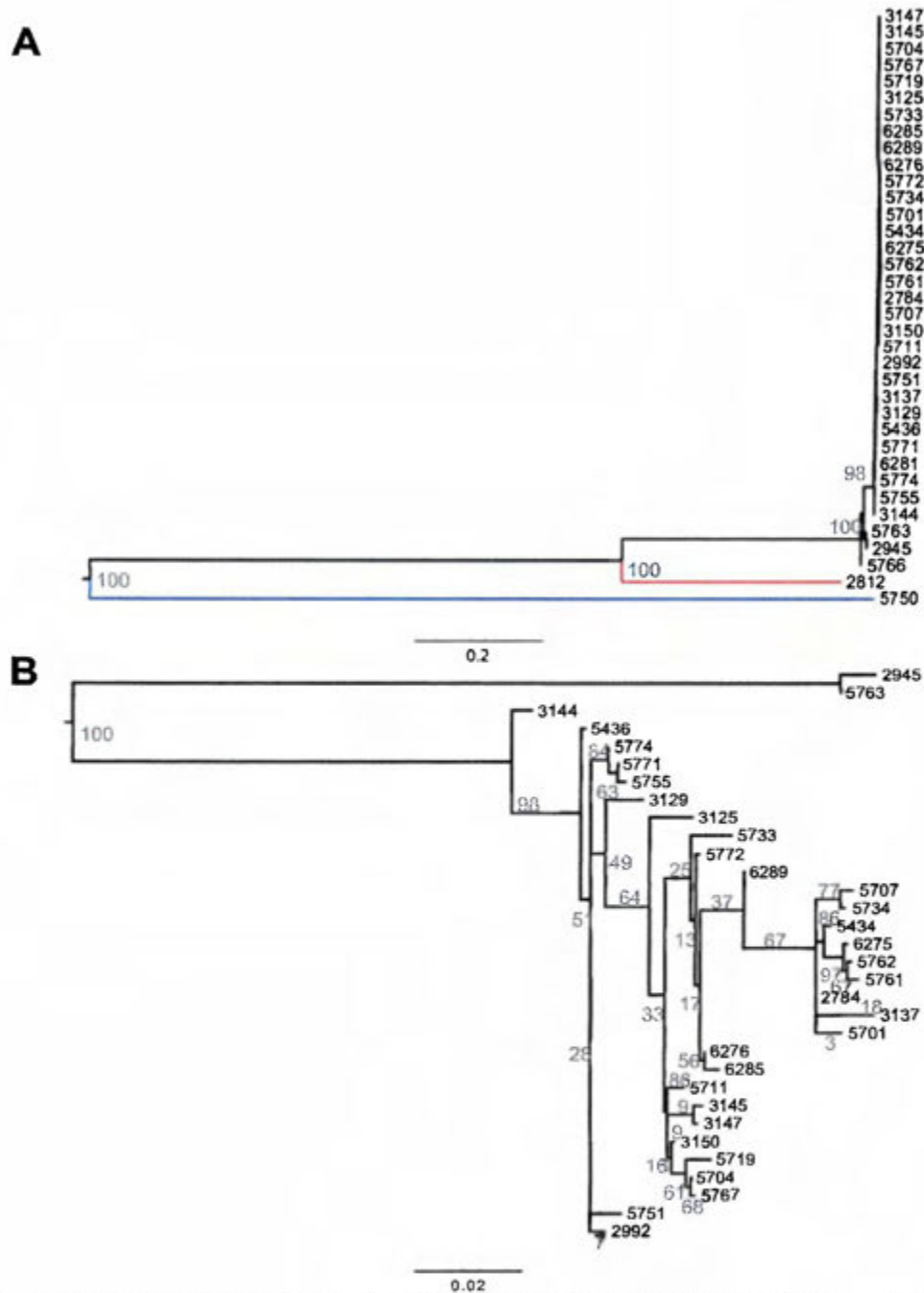


Figure 3. Rooted (A) and unrooted (B) phylogenomic trees of NAP_{CR1} isolates generated with SNP distance matrices and the maximum likelihood method. The NAP4 isolate 2812 from MLST Clade I (blue) and the NAP1 isolate 5750 from MSLT Clade 2 (red) were included in A as outgroups. Bootstrap values are indicated in gray numbers. Scales correspond to the average number of substitution per site.

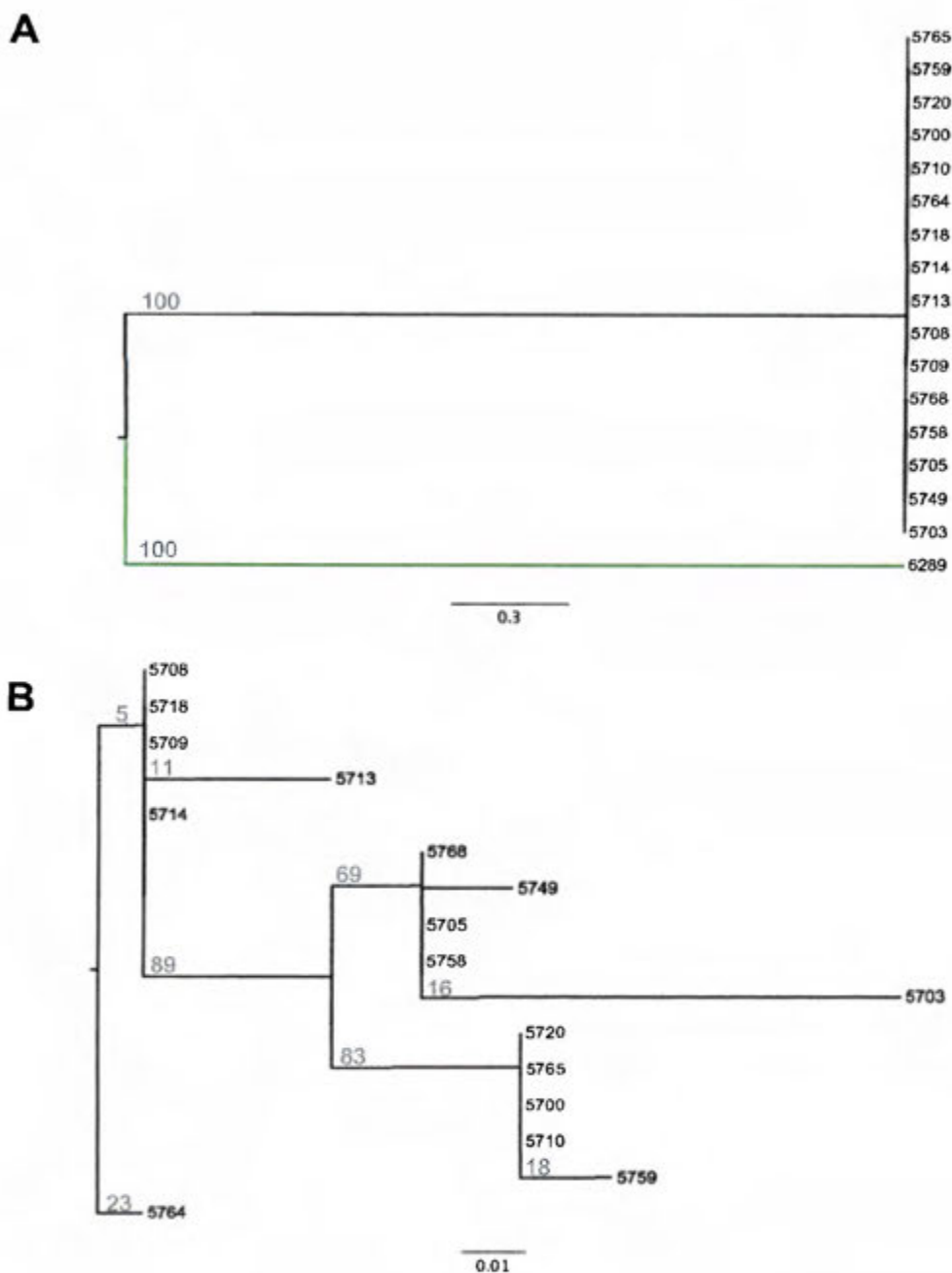


Figure 4. Rooted (A) and unrooted (B) phylogenomic trees of NAP1 isolates generated with SNP distance matrices and the maximum likelihood method. NAP_{CR1} isolate 6289 from MLST Clade I (green) was included as an outgroup. Bootstrap values are indicated in gray numbers. Scales correspond to average number of substitutions per site.

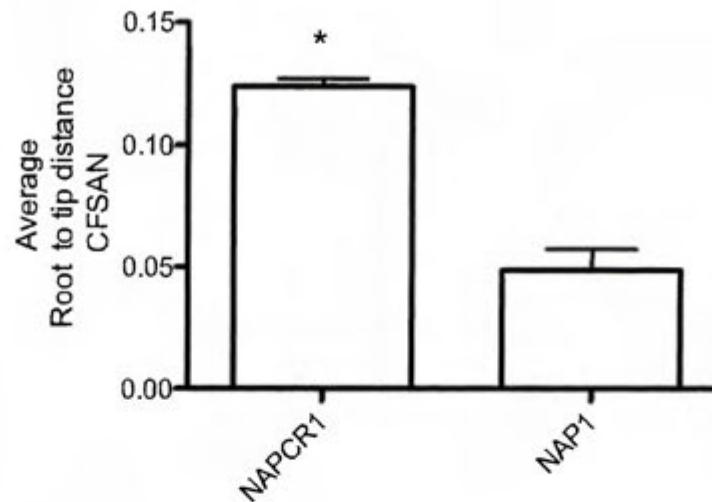


Figure 5. Average root-to-tip distances of isolates from the NAP_{CR1} and NAP1 pulsotypes in SNP-based phylogenomic trees. The asterisk depicts differences at a level of significance of $P < 0.05$ as indicated by a Mann–Whitney U test.

7.3 The accessory genome of the NAP_{CR1} isolates is more diverse than that of the NAP1 isolates

Though both scales were equal, a ffp-tree that takes into consideration the entire genome revealed more differences in the accessory genomes of the NAP_{CR1} isolates compared to NAP1 isolates according to the branch distance of each isolate (Fig 6A and 6B). In the NAP_{CR1} tree, isolates from the 487 macrorestriction pattern appeared separated. Additionally, isolates 6289, 5761, and 5762 outstood from the main clustering (Fig 6A). In the NAP1 ffp-tree, three clusters with very similar isolates were formed (Fig 6B). This conclusion was supported by the finding of significantly larger average root-to-tip distances for NAP_{CR1} isolates than for NAP1 isolates (Fig 7).

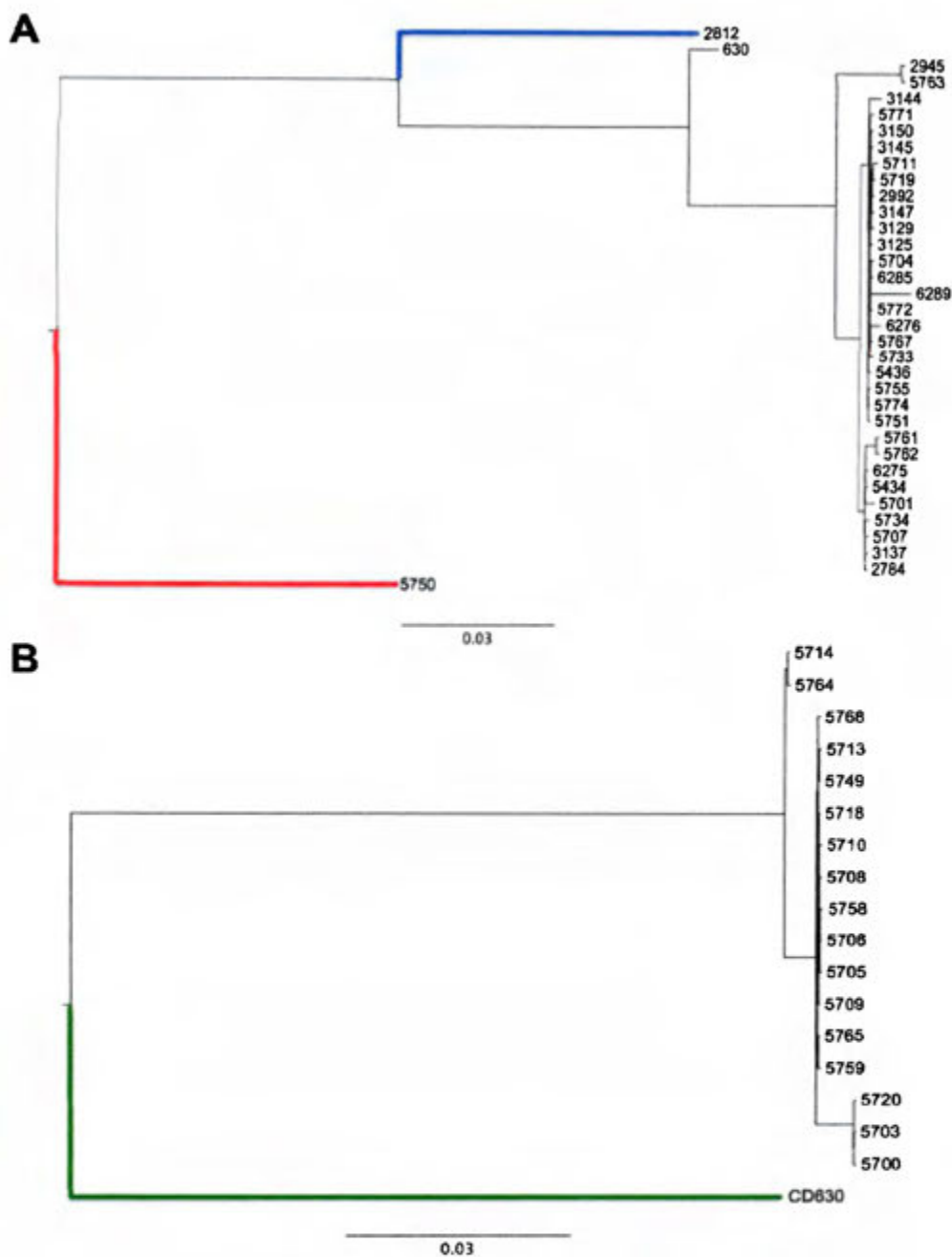


Figure 6. Feature-frequency profile tree of NAP_{CR1}-(A) and NAP1-isolates (B). The NAP4 isolate 2812 from MLST Clade I (blue) and the NAP1 isolate 5750 from MSLT Clade 2 (red) were included in A as outgroups. In B, strain CD630 (green) was included as an outgroup. The scales correspond to number of character feature changes.

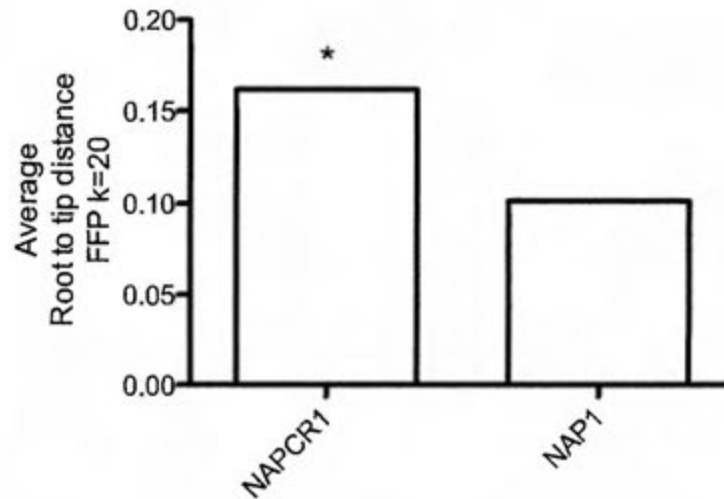


Figure 7. Average root-to-tip distances of isolates from the NAP_{CR1} and NAP1 pulsotypes in feature frequency profiles-based trees. The asterisk depicts differences at a level of significance of $P < 0.05$ as indicated by a Mann–Whitney U test.

7.4 The NAP_{CR1} pulsotype has a larger accessory genome and more gene clusters than NAP1 strains

For the NAP_{CR1} isolates Roary predicted 4802 gene clusters and a core genome of 3547 gene clusters, which accounts for 74% of the genome. Seven percent of the gene clusters were found in the soft-core genome, 8% in the shell genome and 11% in the cloud genome. Hence, the second genome category after the core genome having more gene clusters was the cloud (Fig 8A). In contrast, the same program predicted 3829 gene clusters and a core of 3588 gene clusters (94%) for the NAP1 isolates. In this group, the shell genome only contained 5% of the predicted gene clusters and a cloud genome of 1% (Fig 8B). Coinciding with the FFP results, this data shows that the NAP_{CR1} isolates have more genes in the accessory genome than the NAP1 pulsotype. Moreover, NAP1 has more genes shared by all of the isolates studied.

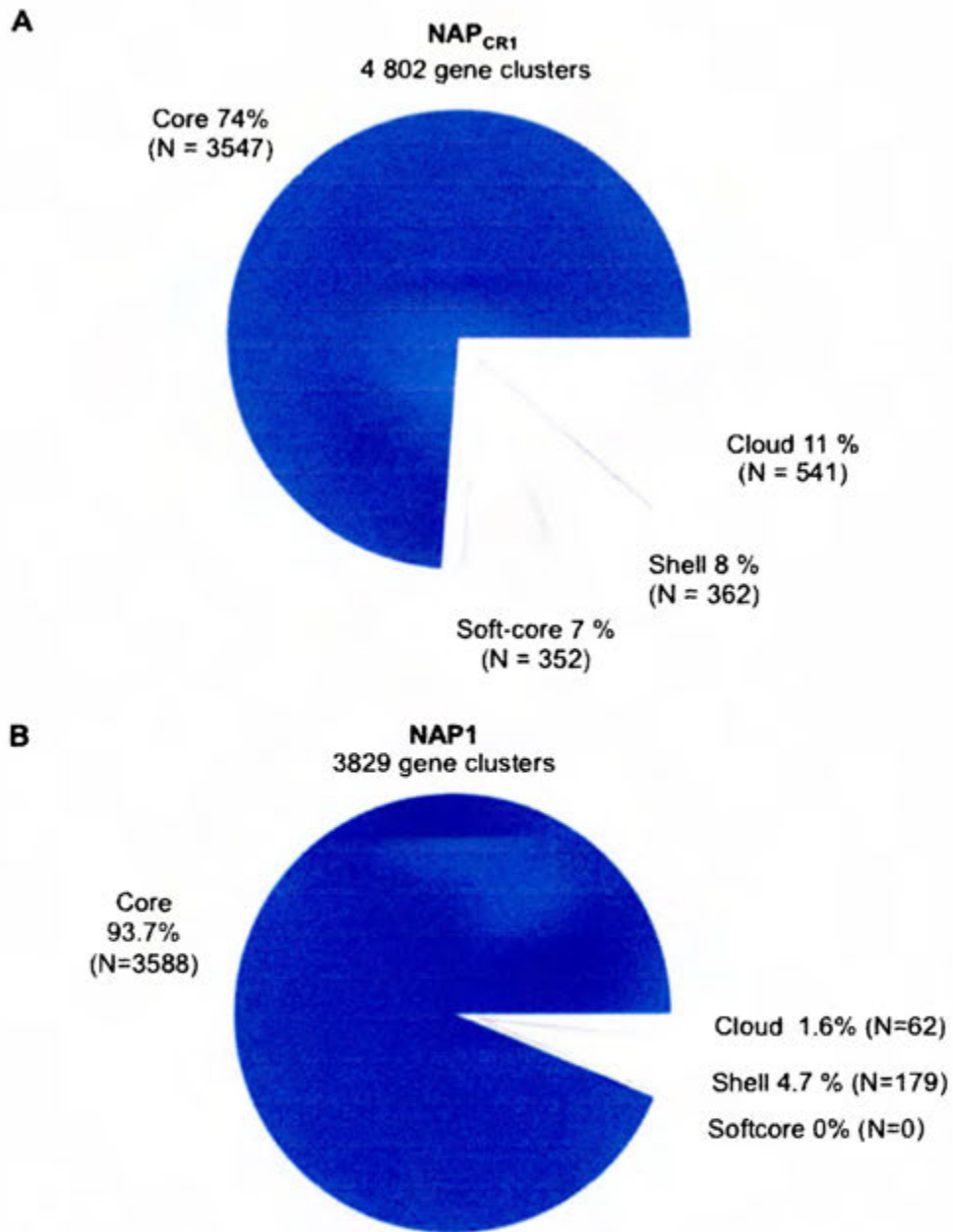


Figure 8. Pangenome comparison of NAP_{CR1} (A) and NAP1 isolates (B). Based on their occurrence rates, gene clusters are organized as core (99% ≤ strains ≤ 100%), soft core (95% ≤ strains < 99%), shell (15% ≤ strains < 95%) or cloud (0% ≤ strains < 15%).

Most NAP_{CR1} isolates showed unique gene clusters in the gene presence-absence spreadsheet generated by Roary (Fig 9A). However, they were not as evidently defined as the unique array that distinguished isolates 2945, 5763 and 6289. Two of these isolates, interestingly, were closely related to strain CD630 in a divergent cluster. As to the NAP1 isolates, although most gene clusters were ubiquitous and present in the reference strain R20291, two distinct groups of isolates could be identified (Fig 9B). The first group included isolates sharing a unique set of gene clusters (5720, 5700 and 5703) and the second group joined isolates (5714, 5759 and 5764) lacking a set of gene clusters shared by all other isolates.

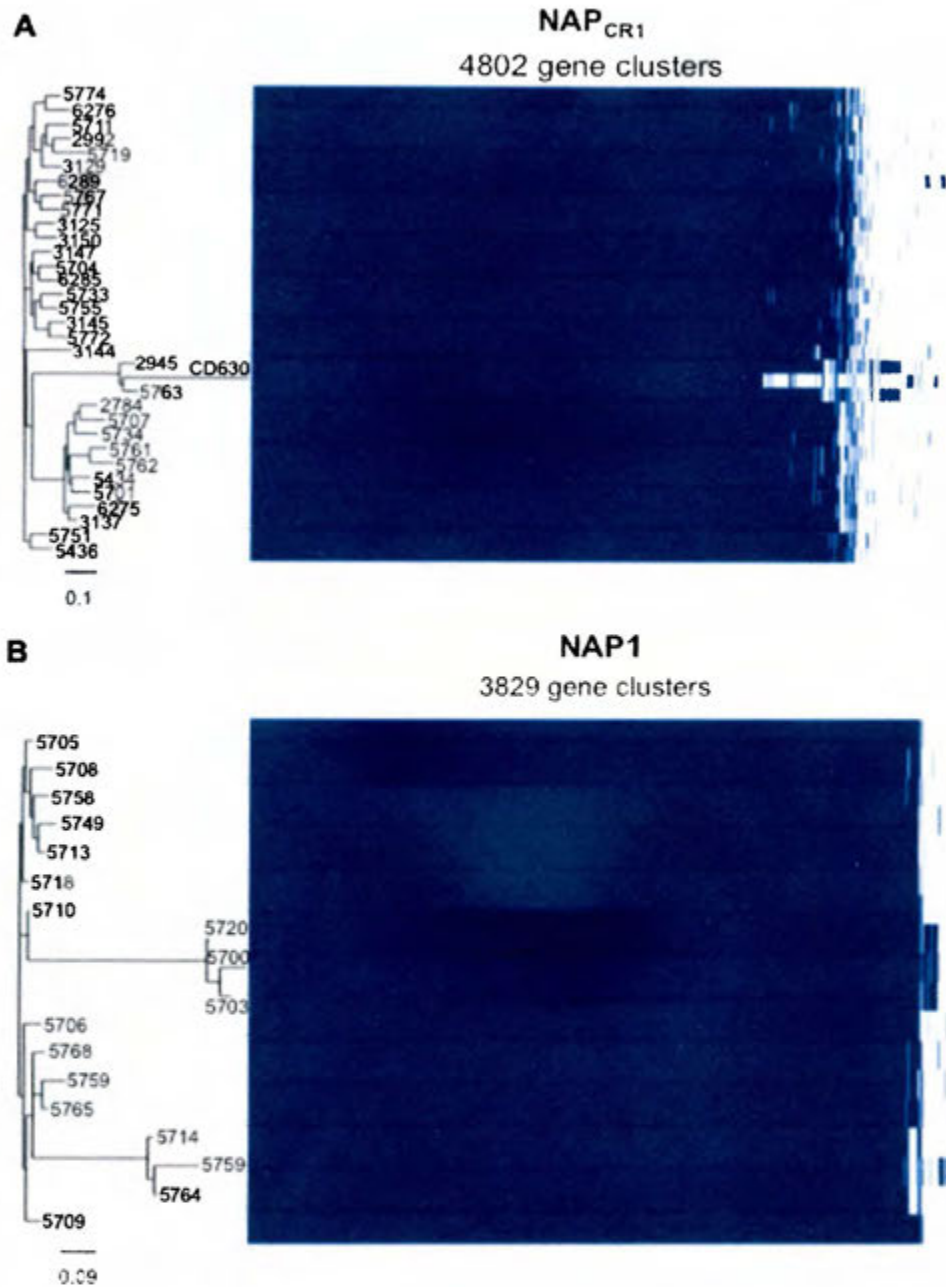


Figure 9. Presence-absence plot of gene clusters in the pangenome of NAP_{CR1} (A) and NAP1 isolates (B). Tree scales were generated from a binary matrix and indicate presence-absence of gene clusters. Blue bars indicate presence of gene cluster.

A parsimony-based pangenomic tree distributed the NAP_{CR1} isolates in three clearly *defined clusters* (Fig 10). The most distant is cluster I (purple) with isolates 2945 and 5763, cluster II (green) has nine isolates, and cluster III (blue) embraced the rest of the isolates. The most phylogenetically distant isolates from each cluster, thus having longer branches, were selected for further analyses, namely, isolates 2945 (cluster I), 5761 (cluster II) and isolates 6276 and 6289 from cluster III. Six different clusters, some of which including single isolates, were defined in the parsimony-based pangenomic tree of NAP1 (Fig 11). From this set, isolates 5764 (cluster I), 5714 (cluster II), 5759 (cluster III), 5708 (cluster IV), 5703 (cluster V) and 5710 (cluster VI) were further studied.

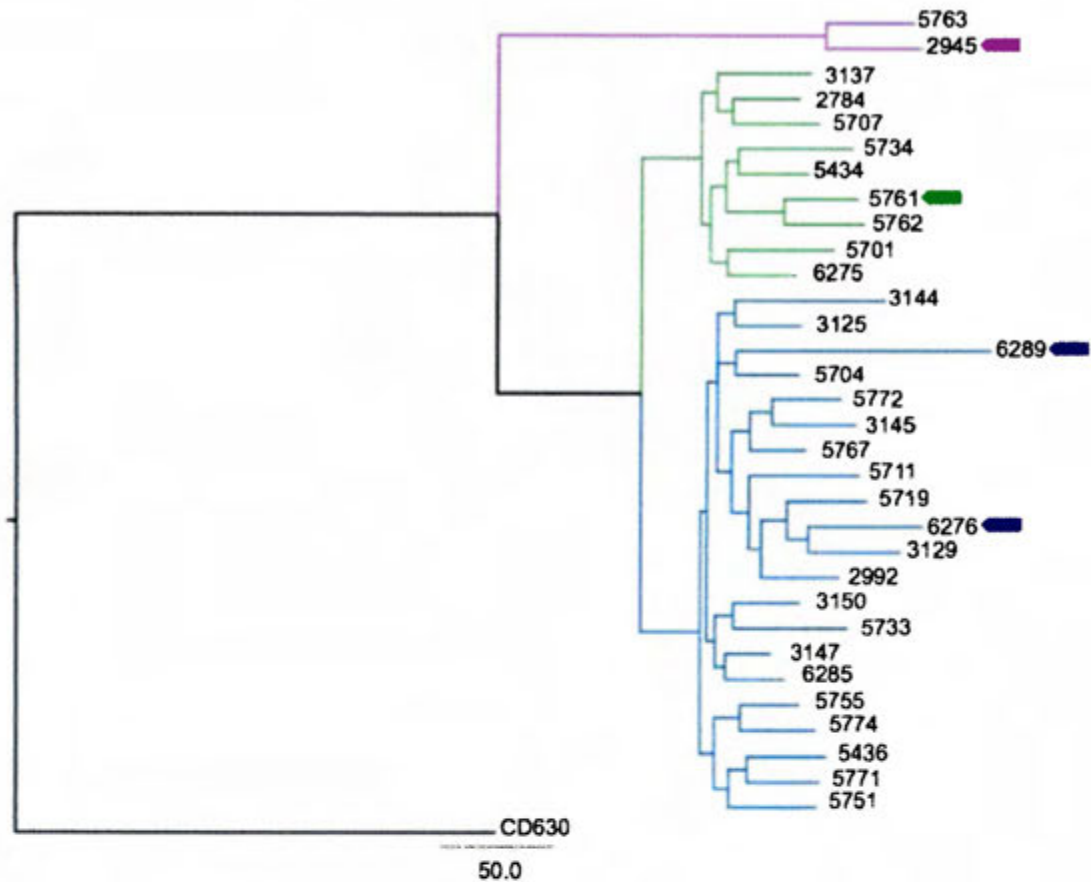


Figure 10. Parsimony-based pangenomic tree of NAP_{CR1} isolates generated with Get_Homologues. The tree was rooted with strain CD630. Three distinct groups were defined: cluster I (purple), cluster II (green) and cluster III (blue) and selected isolates from each group are depicted with arrows. Tree scales are generated from a binary matrix and indicate presence-absence of gene clusters.

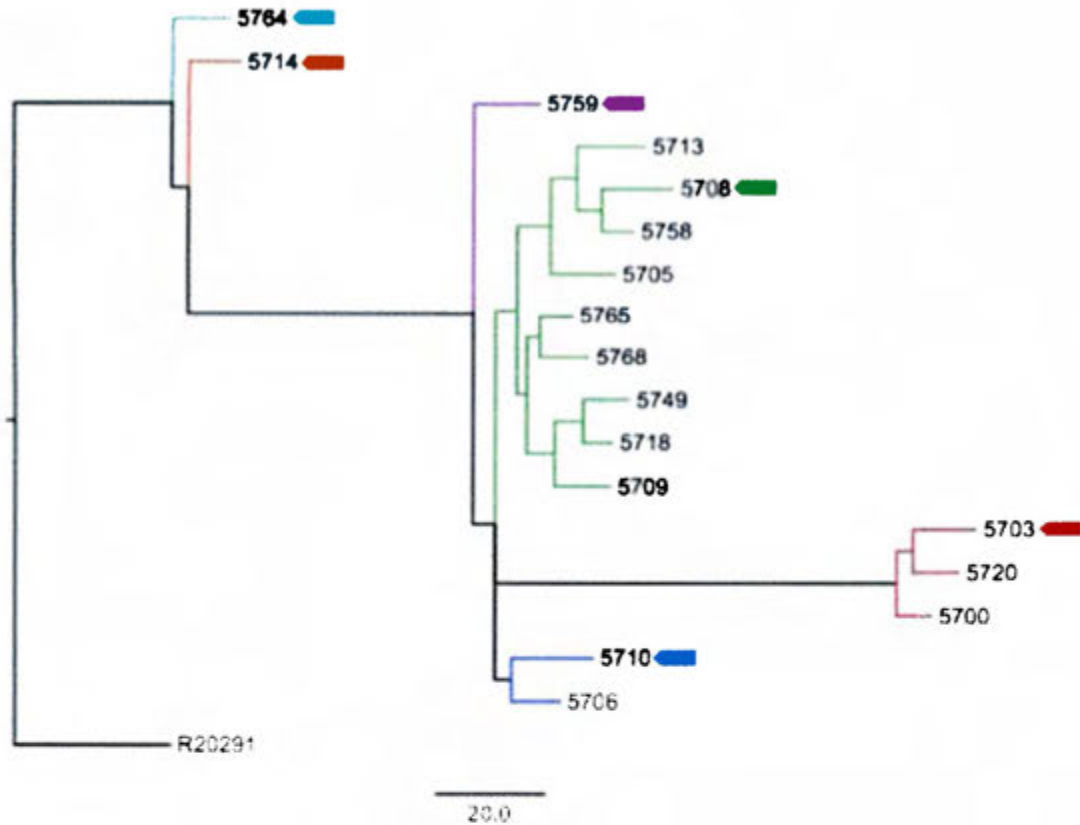


Figure 11. Parsimony-based pangenomic tree of NAP1 isolates generated with Get_Homologues. The tree was rooted with strain R20291. Six distinct groups were defined denominated cluster I (teal), cluster II (brown), cluster III (purple), cluster IV (green), cluster V (red) and cluster VI (blue) and selected isolates from each group are depicted with arrows. Tree scales are generated from a binary matrix and indicate presence-absence of gene clusters.

When the parsimony-based pangenomic trees of NAP_{CR1} and NAP1 were compared according to the scales, NAP_{CR1} (50.0) isolates have greater distances than NAP1 isolates (20.0). This result was confirmed by the average root-to-tip distance that characterized the NAP_{CR1} isolates, which was two-fold higher than the calculated for isolates of the NAP1 pulsotype (Fig 12). Therefore, the pangenome of the NAP_{CR1}

isolates is larger than that of the NAP1 pulsotype as a consequence of a greater *accessory genome*.

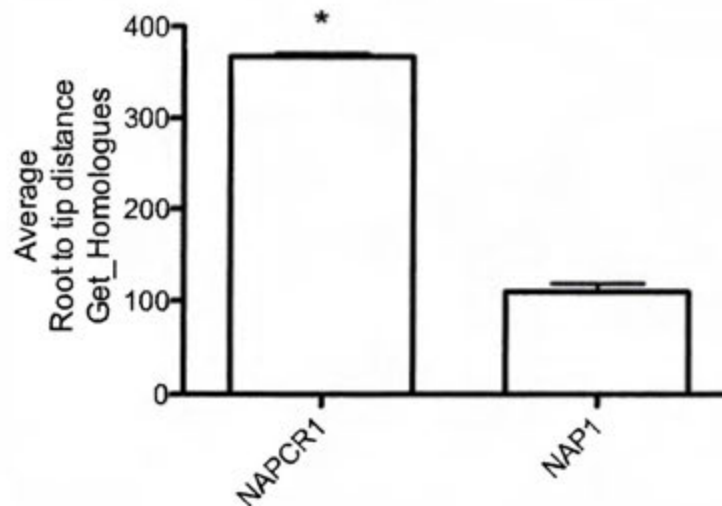


Figure 12. Average root-to-tip distances of isolates from the NAP_{CR1} and NAP1 pulsotypes in parsimony-based pangenomic trees. The asterisk depicts differences at a level of significance of $P < 0.05$ as indicated by a Mann–Whitney U test.

7.5 NAP_{CR1} isolates have more distinctive mobile genetic elements in their accessory genomes than NAP1 isolates

Two of the four NAP_{CR1} isolates selected in the parsimony-based pangenomic belong to the same pulsotype and were isolated in the same hospital during 2011-2012 (Table 6). The other two, which were derived from Clusters I and II, represent distinct pulsotypes and were isolated in different hospitals in 2009. In agreement with the tree shown in Figure 10, isolate 2945 from Cluster I showed the greater amount of unique gene clusters ($n=376$), followed by the isolates from Cluster III ($n=104$) and isolate X from cluster II ($n=62$).

Table 6. Origin and amount of unique gene clusters of representative NAP_{CR1} isolates from each cluster.

Cluster	Isolate	<i>Sma</i> I pattern	Year	Hospital	Unique gene clusters
I	2945	487	2009	CENARE	376
II	5761	489	2009	Hospital Raúl Blanco Cervantes	62
III	6276	449	2011-2012	Hospital México	104
III	6289	449	2011-2012	Hospital México	104

As to the NAP1 isolates that represent the six clusters defined in the parsimony-based pangenomic tree shown above, isolates 5714, 5708, 5703 and 5710 from Clusters II, IV, V and VI were derived from the same hospital during a year in which a CDI outbreak took place. Isolates 5764 and 5759 from Clusters I and III were recovered at different hospitals in 2009. As already indicated by the tree shown in Figure 10, isolate 5703 from Cluster V had the largest amount of unique gene clusters (n=85). All other representative NAP1 isolates had between 10 and 17 unique gene clusters. This figure is small when compared to the results obtained for the NAP_{CR1} isolates. Results are shown in Table 7.

Table 7. Origin and amount of unique gene clusters of NAP1 representative isolates from each cluster.

Cluster	Isolate	<i>Sma</i> I pattern	Year	Hospital	Unique gene clusters
I	5764	001	2009	CENARE	17
II	5714	001	2009	Hospital San Juan de Dios	13
III	5759	001	2009	Hospital Raúl Blanco Cervantes	17
IV	5708	001	2009	Hospital San Juan de Dios	10
V	5703	001	2009	Hospital San Juan de Dios	85
VI	5710	001	2009	Hospital San Juan de Dios	14

Most of the unique genes of the NAP_{CR1} isolates encode MGE-related and form part of novel MGE absent in CD630. Among these distinctive MGE there is a putative plasmid found only in isolate 6289 from Cluster III, an element similar to the Tn4001

of *S. aureus* present in most of the isolates but not in the *Sma*I patterns 487 and 488 from Clusters I and III, respectively, as well as a putative prophage exclusively found in the *Sma*I pattern 487. Interestingly, two versions of a putative pseudolysogenic phage of 130 kb were found. The first variant of this potential giant phage (Giant phi V.1) is present in all of the *Sma*I patterns except in 487. By contrast, isolates 2945 and 5763 from the *Sma*I pattern 487 had another variant (Giant phi V.2). Additionally, three isolates lacked two well described MGE from CD630 and other *C. difficile* genotypes: isolate 6276 from Cluster III lacks the *skin*^{Cd} element that is supposed to be important for the sporulation of this pathogen (79). Moreover, isolates 5761 and 5762 from Cluster II do not have Tn5397. Finally, another mobilizable element denominated *mobCksgA* was found in all NAP_{CR1} isolates but not in CD630. This element is characterized by *ksgA*, antibiotic resistance determinant. These results are summarized in Table 8.

Table 8. Differential MGE present in the NAP_{CR1} pangenome.

Smal pattern	Isolate	Putative plasmid	<i>mobCksgA</i>	Tn5397	<i>skin</i>	Tn4001-like	Prophage	Giant phi V.1	Giant phi V.2
	CD630	-	-	+	+	-	-	-	-
442	3147	-	+	+	+	+	-	+	-
447	5701	-	+	+	+	+	-	+	-
	5711	-	+	+	+	+	-	+	-
	5767	-	+	+	+	+	-	+	-
	5771	-	+	+	+	+	-	+	-
448	2784	-	+	+	+	+	-	+	-
	3125	-	+	+	+	+	-	+	-
	3137	-	+	+	+	+	-	+	-
	5434	-	+	+	+	+	-	+	-
	5704	-	+	+	+	+	-	+	-
	5707	-	+	+	+	+	-	+	-
	5733	-	+	+	+	+	-	+	-
	5751	-	+	+	+	+	-	+	-
	5774	-	+	+	+	+	-	+	-
	6275	-	+	+	+	+	-	+	-
449	3129	-	+	+	+	+	-	+	-
	5719	-	+	+	+	+	-	+	-
	5755	-	+	+	+	+	-	+	-
	5772	-	+	+	+	+	-	+	-
	6276	-	+	+	-	+	-	+	-
	6289	+	+	+	+	+	-	+	-

Table 8. Differential MGE present in the NAP_{CR1} pangenome (continued).

Smal pattern	Isolate	Putative plasmid	<i>mobCksgA</i>	Tn5397	<i>skin</i>	Tn4001-like	Prophage in CTn5	Giant phi V.1	Giant phi V.2
452	5734	-	+	+	+	+	-	+	-
487	2945	-	+	+	+	-	+	-	+
	5763	-	+	+	+	-	+	-	+
488	2992	-	+	+	+	-	-	+	-
489	5761	-	+	-	+	+	-	+	-
	5762	-	+	-	+	+	-	+	-
558	3145	-	+	+	+	+	-	+	-
	6285	-	+	+	+	+	-	+	-
578	3144	-	+	+	+	+	-	+	-
	3150	-	+	+	+	+	-	+	-
	5436	-	+	+	+	+	-	+	-

Presence (+, light blue), absence (-, light green).

A very different picture was derived from the comparison of NAP1 WGS (Table 9). Here, only isolates 5703, 5720 and 5700 from Cluster V had a differential MGE. This element was absent in strain R20291, but gave a perfect identity match to a previously reported plasmid of NAP1 (36).

Table 9. Differential MGE found in the NAP1 pangenome.

NAP1 isolate	Putative plasmid
5764	-
5714	-
5759	-
5713	-
5708	-
5758	-
5705	-
5765	-
5768	-
5749	-
5718	-
5709	-
5703	+
5720	+
5700	+
5710	-
5706	-
R20291	-

Presence (+, light blue), absence (-, light green).

The differential NAP_{CR1} MGE were distributed in the parsimony-based pangenomic tree according to potential gain/loss events (Fig 13). Starting from the root, all of the isolates have the *mobCksgA* mobilizable transposon. When the branches divide, the isolates of Cluster I were the only to acquire the variant 2 of the giant phage and the prophage inserted in CTn5 of CD630. In the other branch, all of the isolates from Clusters II and III have the Tn4001-like element and the first variant of the giant phage. Two isolates from Cluster II lack Tn5397 (5761 and 5762) and isolates 6289 and 5761 from Cluster II have a putative plasmid or lack the *skin* element, respectively.

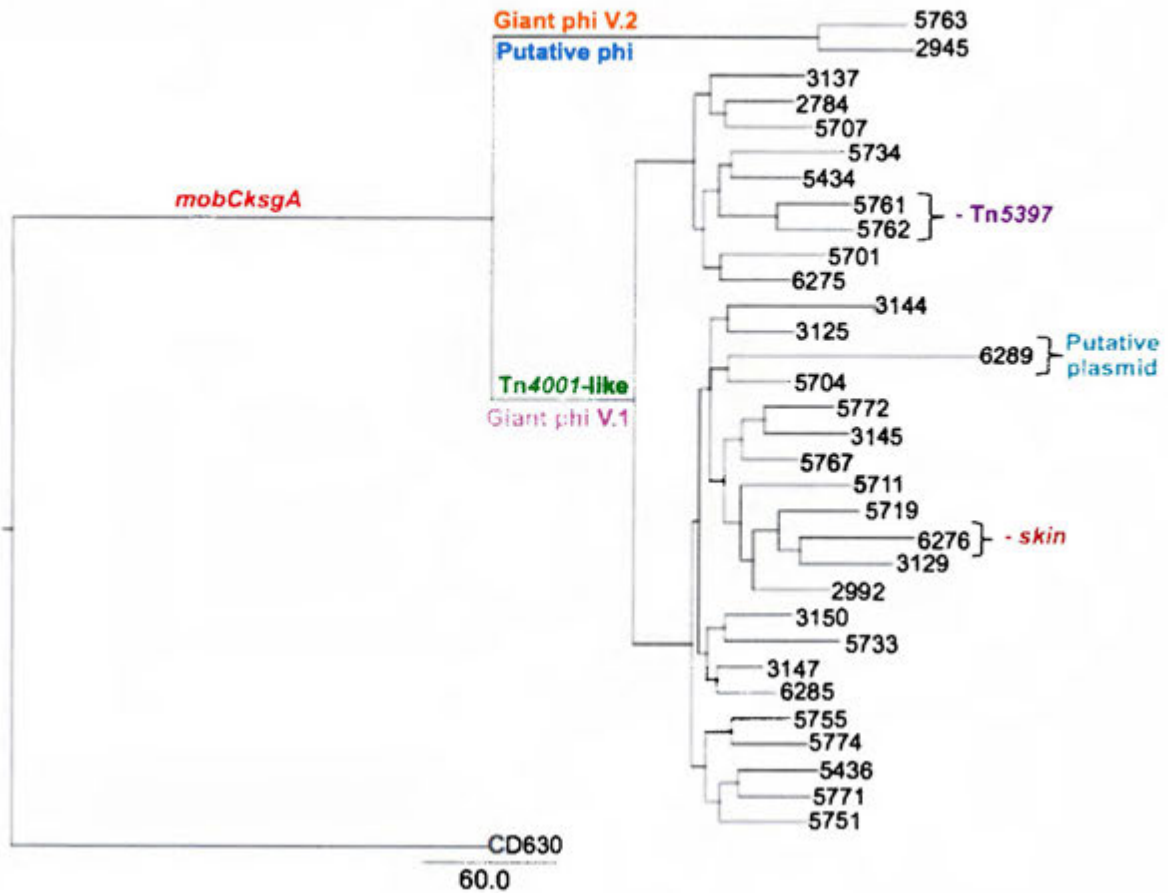


Figure 13. Location of the differential MGE in a parsimony-based pangenomic tree calculated for the NAP_{CR1} isolates. The tree was rooted with strain CD630 and the differential MGE were highlighted with colors. Tree scales are generated from a binary matrix and indicate presence-absence of gene clusters.

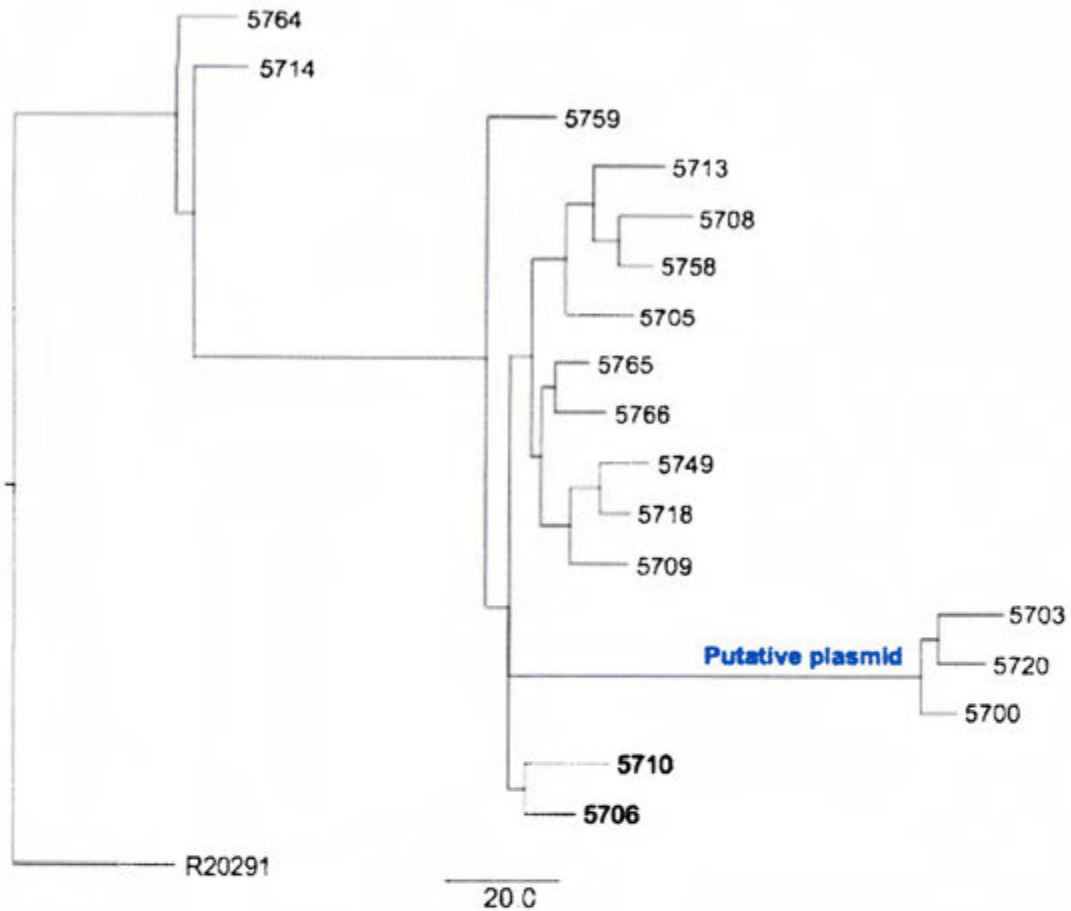


Figure 14. Location of the differential MGE in a parsimony-based pangenomic tree calculated for the NAP1 isolates. The tree was rooted with strain R20291. Tree scales are generated from a binary matrix and indicate presence-absence of gene clusters.

Summarizing, the pangenome comparisons detected more differential MGE in NAP_{CR1} ($n=6$) rather than in NAP1 isolates ($n=1$). This agrees with other results, that demonstrate that the NAP_{CR1} isolates have a greater pangenome and a smaller core genome.

7.6 Annotation and functional characterization of the differential MGE

With a size of 10 815 bp, the novel *mobCksgA* transposon (Table 10) was present in all NAP_{CR1} isolates but in two different genomic locations. It was found into CTn2 in isolates from Cluster I or inserted into a Tn916-like element in isolates from Clusters II and III. This element resembles a mobilizable transposon, as it has a recombinase, a bacterial mobilization protein MobC, and a replication initiation protein RepA (Table 9). Interestingly, it encodes for a kasugamycin dimethyltransferase (KsgA).

Table 10. Annotation of the *mobCksgA* element of NAP_{CR1} isolates.

ORF	ORR length (bp)	Prokka Annotation	Best Blast hit	Best Interpro hit
(+1)	444	HTH transcriptional regulator	Transcriptional regulator	Lambda repressor-like, DNA-binding domain, Cro/C1-type helix-turn-helix domain and putative zinc ribbon domain
(+1)	423	RNA polymerase sigma factor	DNA-directed RNA polymerase sigma-70 factor	RNA polymerase sigma factor, region 3/4
(+1)	930	Dimethyladenosine transferase (S-adenosylmethionine-6-N', N'-adenosyl (rRNA) dimethyltransferase) (16S rRNA dimethylase) (High level kasugamycin resistance protein ksgA) (Kasugamycin dimethyltransferase)	Ribosomal RNA adenine dimethylase family protein	Ribosomal RNA adenine methyltransferase KsgA/Erm and S-adenosyl-L-methionine-dependent methyltransferase
(+1)	486	Glycerol-3-phosphate cytidyltransferase	Glycerol-3-phosphate cytidyltransferase	Rossmann-like alpha/beta/alpha sandwich fold and cytidyltransferase-like domain
(+2)	612	(Alpha)-aspartyl dipeptidase	Hypothetical protein	Class I glutamine amidotransferase-like
(+3)	477	Hypothetical protein	Hypothetical protein	No prediction
(+1)	399	Bacterial mobilisation protein (MobC)	Bacterial mobilization protein (MobC)	Bacterial mobilisation
(+3)	1617	Endonuclease relaxase	Relaxase/mobilization nuclease domain protein	Endonuclease relaxase, MobA/VirD2
(+2)	852	Replication initiation factor	Replication Initiation factor A	P-loop containing nucleoside triphosphate hydrolase and IstB-like ATP-binding protein
(+3)	843	ATP-binding protein	Primosomal protein DnaI	Replication initiator A, N-terminal
(+1)	1620	TndX/TnpX recombinase	Recombinase	Resolvase, N-terminal catalytic domain, DNA-binding recombinase domain and recombinase zinc beta ribbon domain

Tn5397 originally denominated as CTn3 in CD630, was present in all of the NAP_{CR1} isolates, except for isolates 5761 and 5762. This transposon of 20 333 bp encodes conjugative transposon proteins and transcriptional regulators (Table 11). Moreover, it contains the gene *tetM*, whose product confers resistance to tetracycline, and a group II intron. The designed primers amplified a fragment of 1000-1500 bp in isolates 6275 and 6289 whose sequence aligned to the ends of the element (Fig 19A), confirming that it forms a circular intermediate.

Table 11. Annotation of the Tn5397 from NAP_{CR1} isolates.

ORF	ORF length (bp)	Pfam Annotation	Best Blast hit	Best Interpro hit
(+3)	315	Conjugative transposon protein	Conjugal transfer protein	No prediction
(+3)	387	Conjugative transposon protein	Transposase	No prediction
(+3)	1386	Cell division FtsK/SpoII-family protein	Cell division protein FtsK	FtsK domain
(+1)	153	Conjugative transposon protein	Transposase	No prediction
(+1)	1209	Replication initiation factor	Phage replication initiation and Cro/C1 family transcriptional regulator	Lambda repressor-like, DNA-binding domain and Cro/C1-type helix-turn-helix domain
(+1)	222	Conjugative transposon protein	Hypothetical protein	No prediction
(+2)	498	Antirestriction protein	Antirestriction protein ArdA	Antirestriction
(+2)	393	Conjugative transposon protein	Conjugal transfer protein (Tcpe family protein)	Tcpe family protein
(+1)	2448	ATPase	ATPase	ATPase
(+2)	2169	Membrane protein	Membrane protein	No prediction
(+3)	1113	Conjugative transposon protein	Peptidase P60	Lysozyme-like domain and endopeptidase, NLPC/P60 domain
(+1)	1830	Reverse transcriptase/maturase/endonuclease, Group II intron	Reverse transcriptase/maturase/endonuclease Group II intron	Group II intron reverse transcriptase/maturase
(+2)	1113	Conjugative transposon protein	Peptidase P60	Lysozyme-like domain and endopeptidase, NLPC/P60 domain
(+3)	933	Conjugative transposon protein	Conjugal transfer protein	Conjugative transposon protein Tcpe
(+1)	1920	Tetracycline resistance protein	TetM	GTPase activity
(+2)	195	Conjugative transposon protein	Conjugal transfer protein	Cysteine-rich KTR
(-2)	354	HTH-type transcriptional regulator	Transcriptional regulator	Lambda repressor-like, DNA-binding domain and Cro/C1-type helix-turn-helix domain
(+1)	66	Conjugative transposon protein	Transposase	No prediction
(+1)	426	RNA polymerase sigma factor	DNA-binding protein	RNA polymerase sigma factor
(+2)	231	Conjugative transposon protein	Hypothetical protein	Helix-turn-helix, conjugative transposon-like
(+3)	141	Conjugative transposon protein	Hypothetical protein	No prediction
(+1)	1602	Recombinase site-specific resolvase family protein	Recombinase family protein	Resolvase

A novel MGE resembling the Tn4001 of *S. aureus* was found among NAP_{CR1} isolates from Clusters II and III. This element, for now denominated as Tn4001-like, has size of 6 526 bp and contains only six genes (Table 12), including a recombinase, two transposases, a bifunctional acyl-CoA N-acyltransferase/aminoglycoside phosphotransferase, and a acetyltransferase from the GNAT family, which likely confer resistance to aminoglycosides. This Tn4001-like element has an extra recombinase when compared to Tn4001 (Figure 16).

Table 12. Annotation of the Tn4001-like element from NAP_{CR1} isolates.

ORF	ORF length (bp)	Prokka Annotation	Best Blast hit	Best Interpro hit
(+2)	231	Conjugative transposon site-specific recombinase	Resolvase, N-terminal domain protein	Resolvase, N-terminal catalytic domain
(+1)	1173	Transposase	IS256 transposase	Transposase, mutator type
(-1)	1440	Putative aminoglycoside phosphotransferase	Bifunctional AAC/APH (AAC(6'): 6'-aminoglycoside N-acetyltransferase and APH(2''): 2''-aminoglycoside phosphotransferase	Acyl-CoA N-acyltransferase and aminoglycoside phosphotransferase
(-1)	405	Ribosomal-protein-alanine acetyltransferase	GNAT family acetyltransferase	Acyl-CoA N-acyltransferase
(-3)	117	Transposase	IS256 transposase	Transposase, mutator type
(+3)	1446	Conjugative transposon site-specific recombinase	Resolvase	Resolvase, N-terminal catalytic domain and recombinase zinc beta ribbon domain

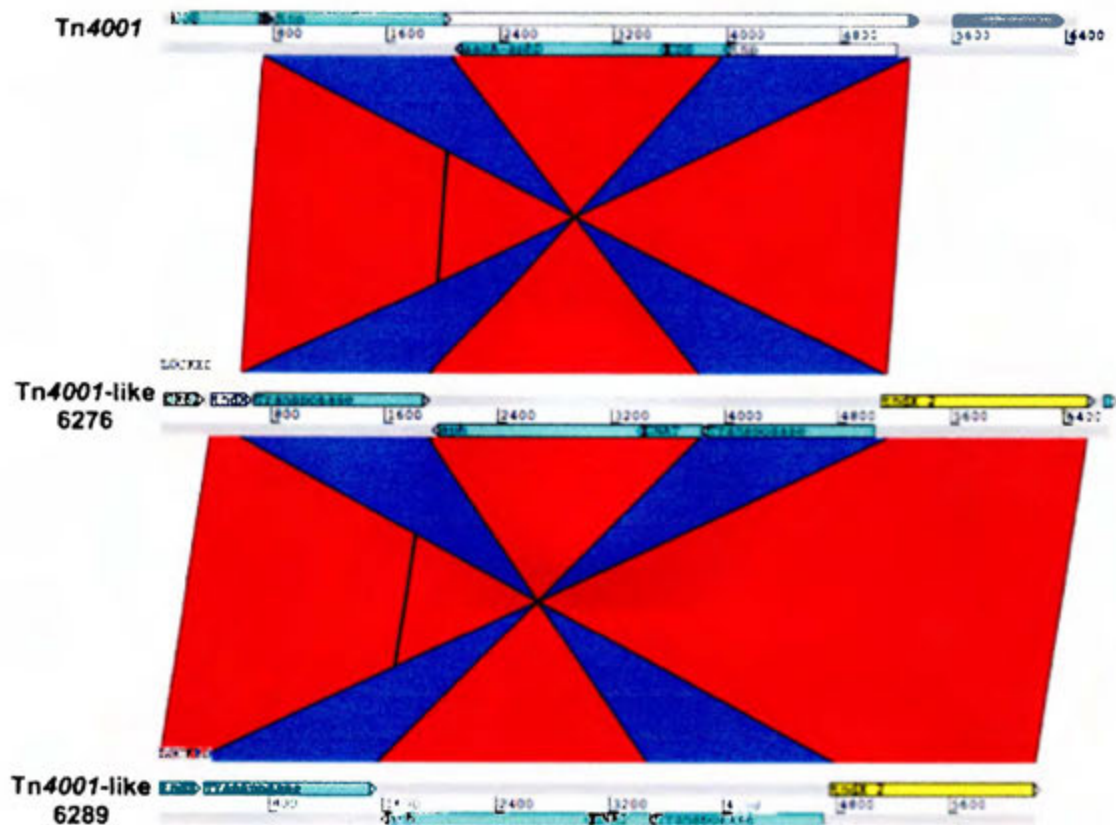


Figure 15. ACT comparison of Tn4001 and the Tn4001-like transposon of the NAP_{CR1} isolates 6276 and 6289. The latter element has an extra recombinase gene compared to the original Tn4001. Red depicts shared regions and blue inversions.

The *skin^{Cd}* element differentiated the NAP_{CR1} isolates because it is missing in isolate 6276 from Cluster III. It mainly includes phage proteins. Additionally, it has a recombinase, a beta-lactam repressor and the *vanZ* gene, encoding for a teicoplanin resistance protein (Table 13). The PCR analysis was positive for isolates 2945, 5761 and 6289, but not for isolate 6276, which lacks the element (Fig 19B). This sequence of this amplicon coincides with the ends of the *skin^{Cd}* element, confirming its circularization.

Table 13. Annotation of the *skin*^{Cd} element from NAP_{CR1} isolates.

ORF	ORF length (bp)	Pfam Annotation	Best Blast hit	Best Interpro hit
(-1)	1518	Recombinase and resolvase	Serine recombinase	N-catalytic domain resolvase and recombinase DNA-binding domain
(-2)	822	Putative lipoprotein	Lipoprotein	Tetratricopeptide repeat-containing domain and helical domain
(+1)	1428	Putative cell surface protein cwp 26	Peptidase	Putative cell wall binding with PepSY domain
(+2)	153	Putative phage protein	Hypothetical protein	No prediction
(+2)	171	Conserved hypothetical protein	Conserved hypothetical protein	No prediction
(+3)	156	Putative phage regulator	Hypothetical protein	No prediction
(+3)	102	Putative phage protein	Putative phage protein	No prediction
(+1)	252	Putative phage protein	Hypothetical protein	No prediction
(+2)	114	Putative phage protein	Hypothetical protein	No prediction
(-2)	384	Putative phage protein	Phage protein	No prediction
(+2)	729	Fragment of putative phage protein	Hypothetical protein	No prediction
(+3)	729	Fragment of putative phage protein	Hypothetical protein	No prediction
(-2)	120	Conserved hypothetical protein	Hypothetical protein	No prediction
(-2)	318	Putative phage protein	Phage protein	No prediction
(+2)	174	Conserved hypothetical protein	Hypothetical protein	No prediction
(-3)	183	Fragment of conserved hypothetical protein	Hypothetical protein	No prediction
(+1)	390	Transcriptional regulator, beta-lactams repressor phage-type	Transcriptional regulator	Blal transcriptional regulator family
(+1)	510	Teicoplanin resistance protein (vanZ)	Teicoplanin resistance protein (vanZ)	vanZ-like
(-2)	195	Putative phage protein	DNA-directed RNA polymerase	No prediction

A prophage appeared inserted in in a DNA helicase of CTn5 in isolates 2945 and 5763 from Cluster I (Fig 17). It has a size of 56 600 bp and genes for phage related structural proteins and holins (Table 14). Interestingly, it contains recombinases previously described in *E. faecium* and putative antibiotic resistance gene such as a phosphotransferase and a GNAT acetyltransferase. It also has genes to escape the host recognition, including a DNA methylase N-4/N-6 domain-containing protein and DNA methyltransferases. This element is missing in all the other isolates as depicted in Figure 17.

Table 14. Annotation of a putative prophage from NAP_{CR1} isolates.

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(+3)	919	Hypothetical protein	Recombinase [Enterococcus faecium]	Recombinase
(-2)	1572	Resolvase domain protein	Recombinase [Enterococcus faecium]	Recombinase
(-3)	1767	Resolvase domain protein	Resolvase [Enterococcus faecium]	Recombinase
(-2)	423	Sigma-70 region 4 type 2	LIM domain protein [Enterococcus faecium]	DNA-binding
(-2)	972	N-acetylmuramoyl-L-alanine amidase family 2 protein	N-acetylmuramoyl-L-alanine amidase	N-acetylmuramoyl-L-alanine amidase domain
(-1)	534	Toxin secretion/phage lysis holin	Holin [Enterococcus faecium]	Bacteriophage holin family
(-3)	2475	Putative glycosyl hydrolase	Glycosyl hydrolase	Glycoside hydrolase
(-2)	846	Hypothetical protein	Hypothetical protein	No prediction
(-3)	2523	Phage minor structural protein	Phage minor structural protein	Phage minor structural protein
(-3)	768	Phage tail component	Tail protein / Hypothetical protein	Siphovirus type tail protein
(-3)	2670	Phage-like protein	Phage tail protein [Enterococcus faecium]	Armadillo-type fold and armadillo-type helical
(-2)	381	Hypothetical protein	Hypothetical protein	No prediction
(-3)	612	Phage major tail protein, phi13 family	Phage major tail protein [Enterococcus faecium]	Phage major tail protein
(-2)	432	Phage protein, HK97 gp10 family	Phage protein	Bacteriophage HK97 putative tail component
(-3)	1203	Phage major capsid protein, HK97 family	Phage capsid protein	Phage major capsid protein HK97
(+1)	813	Hypothetical protein	Hypothetical protein	No prediction
(-2)	1359	Phage portal protein, HK97 family	Phage portal protein	Phage portal protein
(-1)	1602	Phage terminase	Terminase	Phage terminase

Table 14. Annotation of a putative prophage from NAP_{CR1} isolates (continued).

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(-2)	159	Hypothetical protein	Hypothetical protein	No prediction
(-3)	477	AIG2 family protein	Gamma-glutamylcyclotransferase	Gamma-glutamylcyclotransferase
(-1)	951	Putative amidoligase enzyme	Amidoligase	Putative amidoligase
(-1)	318	Hypothetical protein	Hypothetical protein	No prediction
(-3)	720	Virulence-like protein	Virulence protein	No prediction
(-2)	1938	DNA-cytosine methyltransferase	DNA cytosine methyltransferase	C-5 cytosine methyltransferase
(-3)	1239	DNA methylase N-4/N-6 domain-containing protein	Lactate dehydrogenase and DNA modification methylase	DNA methylase, ParB domain-containing
(-2)	786	S-adenosylmethionine synthetase	S-adenosylmethionine synthetase	S-adenosylmethionine synthetase superfamily
(-3)	558	Phage terminase, small subunit	Terminase	No prediction
(-2)	384	HNH endonuclease	HNH endonuclease	HNH endonuclease
(-2)	1365	SNF2-related protein	DEAD/DEAH box helicase	SNF2-related, N-terminal domain and P-loop binding domain
(-3)	282	VRR-NUC domain-containing protein	Nuclease	VRR-NUC domain
(-2)	540	Hypothetical protein	Hypothetical protein	No prediction
(-3)	2340	Virulence-associated E family protein	Phage-like protein [Enterococcus faecium]	Virulence-associated E
(-2)	1941	Phage-related DNA polymerase	DNA-directed DNA polymerase or XRE transcriptional regulator	DNA-directed DNA polymerase
(-3)	576	Phage-like protein	Hypothetical protein	Nucleic acid-binding, OB-fold. Protein of unknown function.
(-1)	1128	Phage-like protein	Hypothetical protein	No prediction
(-2)	318	rRNA biogenesis protein rrp5	DNA ligase	No prediction
(-3)	423	Hypothetical protein	Hypothetical protein	No prediction
(+2)	978	Hypothetical protein	Hypothetical protein	No prediction
(+3)	1827	AAA domain protein	Hypothetical protein	P-loop containing nucleoside triphosphate hydrolase
(+1)	252	XRE family transcriptional regulator	Transcriptional regulator	Lambda repressor-like, DNA-binding domain
(+2)	1044	DNA-cytosine methyltransferase	DNA (cytosine-5-)-methyltransferase	S-adenosyl-L-methionine-dependent methyltransferase
(+3)	1050	DNA-methyltransferase Dcm	DNA (cytosine-5-)-methyltransferase	S-adenosyl-L-methionine-dependent methyltransferase

Table 14. Annotation of a putative prophage from NAP_{CR1} isolates (continued).

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(+2)	1092	Acetyltransferase (GNAT) family protein	GNAT family acetyltransferase	Acyl-CoA N-acyltransferase with GNAT domain
(+3)	1452	Aminoglycoside phosphotransferase	APH(2'')-Ih/Ih family aminoglycoside O-phosphotransferase	Acyl-CoA N-acyltransferase and aminoglycoside phosphotransferase
(+1)	798	Phosphotransferase enzyme family protein	Phosphotransferase enzyme family protein	Aminoglycoside phosphotransferase
(-3)	528	Hypothetical protein	Hypothetical protein	No prediction
(+3)	2250	GTPase subunit of restriction endonuclease	Hypothetical protein	ATP binding protein
(+1)	1476	Restriction endonuclease, type II, LlaJI	LlaJI family restriction endonuclease	Restriction endonuclease, type II, LlaJI
(+2)	501	Hypothetical protein	Hypothetical protein	No prediction

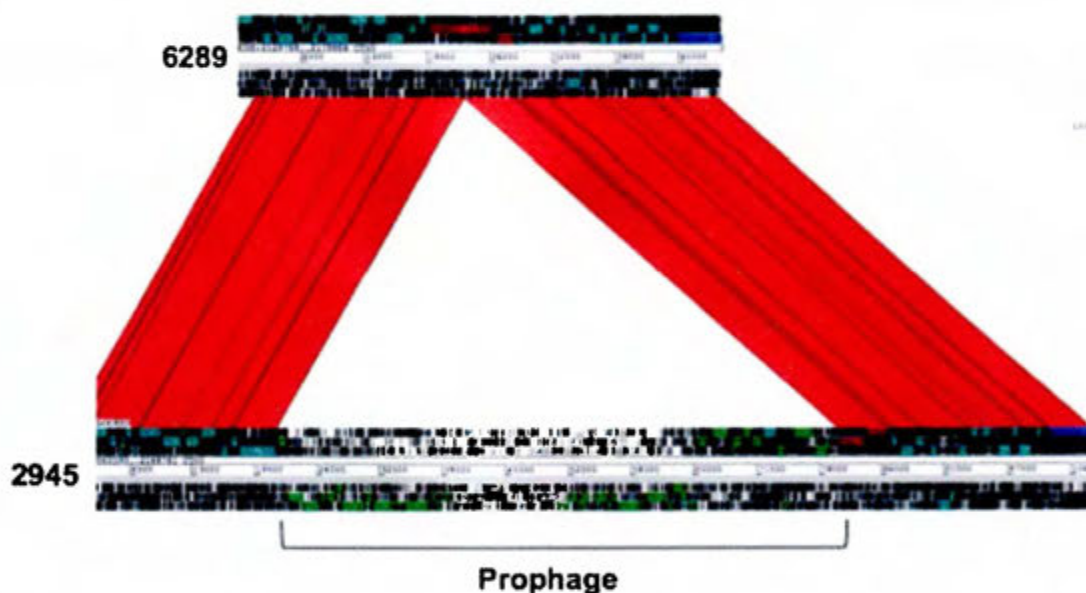


Figure 16. Insertion of a novel prophage in CTn5 of the NAP_{CR1} isolate 2945 from Cluster I (bottom). Isolate 6289 has intact the CTn5 (top).

A putative plasmid of approximately 69 kbp was only found in the NAP_{CR1} isolate 6289 of Cluster III. This element encodes a putative type IV secretion system, a prepilin type IV, DNA-binding proteins, and a partitioning protein ParA, thus it is

possibly a conjugative plasmid. Besides hypothetical proteins, this circular element harbors potential virulence factors, such as a putative adhesin (von Willebrand factor type A), a ADP-ribosyltransferase exoenzyme, a Fic/DOC protein (Table 15). The PCR products confirms it to be a circular plasmid (Fig 19C).

Table 15. Annotation of a putative plasmid from the NAP_{CR1} isolate 6289.

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(-1)	244	Hypothetical protein	Hypothetical protein	No prediction
(-3)	405	Hypothetical protein	Hypothetical protein	No prediction
(-1)	312	Hypothetical protein	Hypothetical protein	No prediction
(-1)	894	Hypothetical protein	Hypothetical protein	No prediction
(-3)	2103	Type IA DNA topoisomerase	DNA topoisomerase	DNA topoisomerase, type IA, core domain
(-1)	375	Putative single-strand binding protein	single-strand-binding family protein	Primosome PriB/single-strand DNA-binding
(-1)	774	Von Willebrand factor type A domain protein	VWA domain-containing protein	von Willebrand factor, type A
(-3)	438	Hypothetical protein	Hypothetical protein	No prediction
(-3)	258	Hypothetical protein	Hypothetical protein	No prediction
(-1)	1119	Cell-wall hydrolase	Bacteriophage peptidoglycan hydrolase	Lysozyme-like domain and endopeptidase, NLPC/P60 domain
(-3)	582	Hypothetical protein	Hypothetical protein	No prediction
(-2)	1917	Hydrolase	Type IV secretory pathway VirB4 component-like protein,conjugal transfer ATP-binding protein TraC,Type IV secretory pathway, VirB4 components,type-IV secretion system protein TraC,AAA-like domain	P-loop containing nucleoside triphosphate hydrolase
(-2)	606	Hypothetical protein	Hypothetical protein	No prediction
(-1)	378	Hypothetical protein	Hypothetical protein	No prediction
(-2)	2025	Hypothetical protein	Hypothetical protein	No prediction
(-2)	2175	Conjugative transfer protein	Type IV secretory system Conjugative DNA transfer family protein	Type IV secretion system protein TraG/VirD4
(-1)	231	Hypothetical protein	Hypothetical protein	No prediction
(-3)	159	Hypothetical protein	Hypothetical protein	No prediction

Table 15. Annotation of a putative plasmid from the NAP_{CR1} isolate 6289 (continued).

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(-1)	384	Hypothetical protein	Membrane protein	No prediction
(-2)	642	Hypothetical protein	Hypothetical protein	No prediction
(-3)	282	Hypothetical protein	Hypothetical protein	No prediction
(-1)	165	Hypothetical protein	Hypothetical protein	No prediction
(-1)	189	Hypothetical protein	Hypothetical protein	NifT/FixU
(-3)	1764	Hypothetical protein	Hypothetical protein	No prediction
(-3)	1035	Cell wall binding protein	Cell wall binding repeat 2 family protein	Putative cell wall binding repeat 2
(-2)	531	Hypothetical protein	Hypothetical protein	No prediction
(-3)	255	Putative IS transposase (OrfA)	merR HTH regulatory family protein	MerR-type HTH domain
(-2)	534	Hypothetical protein	Hypothetical protein	No prediction
(-1)	417	Hypothetical protein	Hypothetical protein	No prediction
(-1)	864	Hypothetical protein	Hypothetical protein	No prediction
(-1)	936	Hypothetical protein	Membrane protein	No prediction
(-2)	1383	Type IV pilus transporter system	Type II/IV secretion system family protein	Type II secretion system protein E and P-loop containing nucleoside triphosphate hydrolase
(-2)	771	CobQ/CobB/MinD/ParA nucleotide binding domain protein	chromosome partitioning protein ParA	CobQ/CobB/MinD/ParA nucleotide binding domain and P-loop containing nucleoside triphosphate hydrolase
(-2)	786	Hypothetical protein	Fli pilus assembly protein CpaB	Fli pilus assembly protein RcpC/CpaB domain
(-1)	555	Type 4 prepilin-like proteins leader peptide-processing enzyme	Type IV leader peptidase family protein	Prepilin type IV endopeptidase, peptidase domain
(-3)	255	Hypothetical protein	Hypothetical protein	No prediction
(-3)	90	Hypothetical protein	Hypothetical protein	No prediction
(-3)	114	Hypothetical protein	Putative membrane protein	No prediction
(-1)	711	Hypothetical protein	Hypothetical protein	No prediction
(-3)	543	Hypothetical protein	Hypothetical protein	No prediction
(-1)	318	Hypothetical protein	Hypothetical protein	No prediction
(-1)	960	RNA polymerase sigma factor SigX	Bacterial regulatory σ , luxR family protein	Winged helix-turn-helix DNA-binding domain
(-3)	210	Helix-turn-helix	Transcriptional regulator	Lambda repressor-like, DNA-binding domain and Cro/C1-type helix-turn-helix domain
(-3)	183	Hypothetical protein	Hypothetical protein	No prediction
(+2)	393	Helix-turn-helix domain protein	Transcriptional regulator or putative prophage repressor	Lambda repressor-like, DNA-binding domain and Cro/C1-type helix-turn-helix domain

Table 15. Annotation of a putative plasmid from the NAP_{CR1} isolate 6289 (continued).

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(-2)	132	Hypothetical protein	Hypothetical protein	No prediction
(+1)	261	Helix-turn-helix domain protein	Transcriptional regulator or putative prophage repressor	Lambda repressor-like, DNA-binding domain and Cro/C1-type helix-turn-helix domain
(+1)	150	Hypothetical protein	Hypothetical protein	No prediction
(-3)	1200	Hypothetical protein	Hypothetical protein	No prediction
(-1)	405	Helix-turn-helix domain protein	Transcriptional regulator	Lambda repressor-like, DNA-binding domain
(-1)	156	Putative membrane protein	Hypothetical protein	No prediction
(+3)	246	Hypothetical protein	Hypothetical protein	No prediction
(+2)	408	Hypothetical protein	Hypothetical protein	No prediction
(+3)	330	Hypothetical protein	Hypothetical protein	No prediction
(-1)	240	Helix-turn-helix	Transcriptional regulator	Lambda repressor-like, DNA-binding domain and Cro/C1-type helix-turn-helix domain
(-1)	258	Helix-turn-helix domain protein	Transcriptional regulator	Lambda repressor-like, DNA-binding domain and Cro/C1-type helix-turn-helix domain
(-1)	648	Putative resolvase	Resolvase	Resolvase, N-terminal catalytic domain
(-3)	705	ADP-ribosyltransferase exoenzyme	ADP-ribosyltransferase exoenzyme [Bacillus azotoformans MEV2011]	ADP ribosyltransferase
(-2)	891	Adenosine monophosphate-protein transferase VbnT	fic/DOC family protein	Fido domain
(-2)	216	Hypothetical protein	Hypothetical protein	No prediction
(-3)	543	Helix-turn-helix	Transcriptional regulator	Lambda repressor-like, DNA-binding domain and Cro/C1-type helix-turn-helix domain
(-3)	1365	Nucleoid occlusion protein	Chromosome partitioning protein ParB	ParB/Sulfiredoxin
(-3)	822	Sporulation initiation inhibitor	Chromosome partitioning protein ParA	P-loop containing nucleoside triphosphate hydrolase and AAA domain
(+2)	615	Region found in RelA/SpoT proteins	RelA/SpoT family protein	RelA/SpoT
(+1)	225	Hypothetical protein	Hypothetical protein	No prediction
(+2)	309	Hypothetical protein	Hypothetical protein	No prediction
(+2)	441	DNA repair protein RadC	radC-like JAB domain protein (DNA repair protein)	RadC-like JAB domain

Table 15. Annotation of a putative plasmid from the NAP_{CR1} isolate 6289 (continued).

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(+3)	867	DNA methylase	Restriction endonuclease subunit M	S-adenosyl-L-methionine-dependent methyltransferase
(+2)	351	HTH-type transcriptional regulator	Transcriptional regulator	Lambda repressor-like, DNA-binding domain and Cro/C1-type helix-turn-helix domain
(+1)	534	Hypothetical protein	Hypothetical protein	No prediction
(+2)	324	Putative transcriptional regulator	radC-like JAB domain protein	BlaI transcriptional regulatory family
(-1)	453	Hypothetical protein	Hypothetical protein	No prediction
(+1)	522	Hypothetical protein	Hypothetical protein	No prediction
(+1)	270	Hypothetical protein	Hypothetical protein	No prediction
(+2)	237	Hypothetical protein	Hypothetical protein	No prediction
(+2)	168	Hypothetical protein	Hypothetical protein	No prediction
(-3)	11103	Lectin C-type domain protein	Lectin C-type domain protein	C-type lectin-like
(-3)	201	Hypothetical protein	Hypothetical protein	No prediction
(-2)	1380	Anitrestriction protein	DNA repair protein, Antirestriction protein, Domain of unknown function	HTH domain, IrrE-type
(-3)	369	Hypothetical protein	Transcriptional regulator	No prediction
(-2)	630	Hypothetical protein	Hypothetical protein	Metallopeptidase, catalytic domain
(-1)	204	Hypothetical protein	Hypothetical protein	No prediction
(+2)	909	Hypothetical protein	Hypothetical protein	No prediction

Finally, two variants of a putative pseudolysogenic giant phage with a size of approximately 130 kbp was detected among the NAP_{CR1} isolates (Figure 18). Most of the predicted genes encode hypothetical proteins. Nonetheless some phage proteins, transposases and transcriptional regulators were also predicted. Some of the predicted proteins in which the phages differed are transposases, but phage version 1 is characterized by having a Fic/DOC family protein and a Cas3 protein (Tables 16 and 17). For these elements circularization assays were not performed.



Figure 17. ACT comparison of the two giant phage variants found among the NAP_{CR1} isolates. Red depicts shared regions.

Table 16. Annotation of giant phage version 1 from NAP_{CR} isolates.

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(+3)	861	Hypothetical protein	Hypothetical protein	No prediction
(+2)	1776	Terminase	Phage Terminase family protein	Terminase
(+1)	1485	Hypothetical protein	Hypothetical protein	No prediction
(+3)	237	Hypothetical protein	Hypothetical protein	No prediction
(+3)	1383	Hypothetical protein	Hypothetical protein	No prediction
(+1)	489	Hypothetical protein	Hypothetical protein	No prediction
(+1)	972	Hypothetical protein	Hypothetical protein	No prediction
(+2)	594	Hypothetical protein	Hypothetical protein	No prediction
(+1)	636	Hypothetical protein	Hypothetical protein	No prediction
(+3)	1047	Hypothetical protein	Hypothetical protein	No prediction
(+1)	450	Hypothetical protein	Hypothetical protein	No prediction
(+2)	804	Hypothetical protein	Hypothetical protein	No prediction
(+3)	819	Hypothetical protein	Hypothetical protein	No prediction
(+2)	786	Hypothetical protein	Hypothetical protein	Siphovirus-type tail component
(+1)	450	Hypothetical protein	Hypothetical protein	No prediction

Table 16. Annotation of giant phage version 1 from NAP_{CR} isolates (continued).

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(+1)	438	Hypothetical protein	Hypothetical protein	No prediction
(+1)	651	Phage repressor protein KilAC domain protein	Phage antirepressor KilAC domain protein	BRO N-terminal domain and antirepressor domain
(+2)	786	Hypothetical protein	Prophage antirepressor	BRO N-terminal domain
(+3)	714	Hypothetical protein	BRO family, N-terminal domain protein	BRO family, N-terminal domain protein
(+2)	705	Sensory transduction protein LytR	Sensory transduction protein LytR	CheY-like superfamily and LytTR DNA-binding domain
(+1)	360	Hypothetical protein	Hypothetical protein	No prediction
(+2)	288	Hypothetical protein	Hypothetical protein	No prediction
(+1)	159	Hypothetical protein	Hypothetical protein	No prediction
(+2)	3228	Type IIS restriction enzyme Eco57I	Putative type II restriction enzyme, methylase	S-adenosyl-L-methionine-dependent methyltransferase and TaqI-like C-terminal specificity domain
(-1)	450	HTH-type transcriptional regulator ImmR	Transcriptional regulator	Lambda repressor-like, DNA-binding domain
(+1)	1044	Phage antirepressor protein KilAC domain protein	Phage antirepressor KilAC domain protein	BRO N-terminal domain and antirepressor protein, C-terminal
(+3)	1161	Hypothetical protein	Hypothetical protein	No prediction
(+1)	711	Hypothetical protein	Hypothetical protein	SHOCT domain
(+3)	111	Hypothetical protein	Hypothetical protein	No prediction
(+3)	984	Hypothetical protein	Hypothetical protein	No prediction
(+2)	1479	Phage-related minor tail protein	Phage tail tape measure protein	No prediction
(+1)	6069	Hypothetical protein	Phage tail tape measure protein	Phage tail tape measure protein
(+2)	156	Hypothetical protein	Hypothetical protein	No prediction
(-1)	117	Hypothetical protein	Hypothetical protein	No prediction
(+2)	1716	Hypothetical protein	Prophage endopeptidase tail family protein (74% Id)	No prediction
(+1)	1929	Hypothetical protein	Chaperone of endosialidase	Intramolecular chaperone auto-processing domain

Table 16. Annotation of giant phage version 1 from NAP_{CR} isolates (continued).

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(+3)	156	Hypothetical protein	Hypothetical protein	No prediction
(+1)	2076	Hypothetical protein	Hypothetical protein	No prediction
(+3)	4884	Regulator of chromosom condensation (RCC1) repeat protein	Regulator of chromosom condensation (RCC1) repeat protein	Regulator of chromosome condensation 1/beta-lactamase-inhibitor protein II
(+2)	327	Hypothetical protein	Hypothetical protein	No prediction
(+2)	330	Hypothetical protein	Phage tail-collar fiber family protein	No prediction
(+1)	1716	Glycine rich protein	Glycine rich family protein	No prediction
(+3)	294	Hypothetical protein	Hypothetical protein	No prediction
(+1)	183	Hypothetical protein	Hypothetical protein	No prediction
(+2)	816	Sporulation-specific N-acetylmuramoyl-L-alanine amidase	N-acetylmuramoyl-L-alanine amidase	N-acetylmuramoyl-L-alanine amidase
(+3)	618	Mannosyl-glycoprotein endo-beta-N-acetylglucosaminidase	Mannosyl-glycoprotein endo-beta-N-acetylglucosaminidase	Mannosyl-glycoprotein endo-beta-N-acetylglucosaminidase
(+1)	252	Hypothetical protein	Membrane protein	Holin
(+1)	312	Hypothetical protein	Hypothetical protein	No prediction
(+3)	936	Tyrosine recombinase XerC	Phage integrase	Integrase
(+3)	354	HTH-type transcriptional regulator SinR	Transcriptional regulator	Lambda repressor-like, DNA-binding domain
(+1)	321	Methicillin resistance regulatory protein Mecl	Transcriptional regulator	BlaI transcriptional regulatory family
(+3)	1263	N-acetylmuramoyl-L-alanine amidase LytC precursor	Cell wall-binding repeat 2 family protein	Putative cell wall binding repeat 2
(+3)	222	Hypothetical protein	Hypothetical protein	No prediction
(-3)	132	Hypothetical protein	Hypothetical protein	No prediction
(-2)	1845	Hypothetical protein	Hypothetical protein	No prediction
(-1)	1113	Transposase from transposon Tn916	Transposase from transposon Tn916	DNA breaking-rejoining enzyme, catalytic core
(+3)	843	Exodeoxyrinonuclease X	Hypothetical protein	No prediction

Table 16. Annotation of giant phage version 1 from NAP_{CR} isolates (continued).

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(+1)	1311	Transposase	Transposase	Transposase IS204/IS1001/IS1096/IS1165
(-1)	474	Deoxyuridine 5'-triphosphate nucleotidohydrolase	Deoxyuridine 5'-triphosphate nucleotidohydrolase	Deoxyuridine triphosphate nucleotidohydrolase
(-3)	495	Holliday junction resolvase	Crossover junction endodeoxyribonuclease RuvC	Crossover junction endodeoxyribonuclease RuvC
(-2)	573	Ribose 1,5-bisphosphate phosphokinase PhnN	Guanylate kinase	Guanylate kinase/L-type calcium channel beta subunit
(-1)	363	Hypothetical protein	Putative phage protein	No prediction
(-2)	810	DNA adenine methyltransferase YhdJ	Site-specific DNA-methyltransferase	S-adenosyl-L-methionine-dependent methyltransferase
(-1)	285	Hypothetical protein	Hypothetical protein	No prediction
(-2)	255	Hypothetical protein	Hypothetical protein	No prediction
(-2)	666	Bis(5'-nucleosyl)-tetrakisphosphate, symmetrical	Serine/threonine protein phosphatase	Metallo-dependent phosphatase-like and calcineurin-like phosphoesterase domain, apaH type
(-3)	195	Hypothetical protein	Hypothetical protein	No prediction
(-1)	153	Hypothetical protein	Hypothetical protein	No prediction
(-3)	807	Hypothetical protein	Hypothetical protein	No prediction
(-2)	534	Hypothetical protein	Hypothetical protein	No prediction
(-1)	246	Hypothetical protein	Hypothetical protein	No prediction
(-1)	132	Hypothetical protein	Hypothetical protein	No prediction
(-1)	375	Hypothetical protein	Hypothetical protein	No prediction
(-2)	2181	Ribonucleoside-diphosphate reductase NrdZ	Ribonucleotide reductase, adenosylcobalamin-dependent	Ribonucleoside-diphosphate reductase, adenosylcobalamin-dependent
(-2)	1140	Hypothetical protein	DNA-sulfur modification-associated family protein	DNA sulphur modification protein DndB
(-1)	303	Hypothetical protein	Hypothetical protein	No prediction
(-2)	3333	DNA polymerase III subunit alpha	DNA polymerase III subunit alpha	Bacterial DNA polymerase III, alpha subunit

Table 16. Annotation of giant phage version 1 from NAP_{CR} isolates (continued).

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(-3)	1701	Putative SPBc2 prophage-derived single-strand DNA-specific exonuclease York	Putative SPBc2 prophage-derived single-strand DNA-specific exonuclease York	No prediction
(-3)	882	Hypothetical protein	Hypothetical protein	No prediction
(-1)	1593	Hypothetical protein	Chromosome partitioning protein ParA	DNA helicase DnaB, N-terminal/DNA primase DnaG, C-terminal and P-loop containing nucleoside triphosphate hydrolase
(-2)	558	Hypothetical protein	Hypothetical protein	No prediction
(-2)	576	ERF superfamily protein	Recombination protein, phage associated	Essential recombination function protein
(-3)	846	Hypothetical protein	Hypothetical protein	No prediction
(-2)	171	Hypothetical protein	Putative membrane protein	No prediction
(-3)	666	Hypothetical protein	Hypothetical protein	No prediction
(-2)	561	Hypothetical protein	Hypothetical protein	No prediction
(-3)	156	Hypothetical protein	Hypothetical protein	No prediction
(-3)	432	Hypothetical protein	Hypothetical protein	No prediction
(-2)	1122	Hypothetical protein	pcfJ-like family protein	PcfJ-like protein
(-2)	309	Hypothetical protein	Hypothetical protein	No prediction
(-1)	252	Hypothetical protein	Hypothetical protein	No prediction
(-3)	210	Hypothetical protein	Hypothetical protein	No prediction
(-2)	603	Hypothetical protein	Hypothetical protein	No prediction
(-1)	327	Hypothetical protein	Hypothetical protein	S-adenosyl-L-methionine-dependent methyltransferase
(-3)	135	Hypothetical protein	Hypothetical protein	No prediction
(-1)	285	Hypothetical protein	DUF5052 domain-containing protein	DUF5052 domain-containing protein
(-1)	396	Hypothetical protein	Hypothetical protein	No prediction
(-1)	351	Hypothetical protein	Hypothetical protein	No prediction
(-2)	183	Hypothetical protein	Hypothetical protein	No prediction
(-3)	210	Hypothetical protein	Hypothetical protein	No prediction
(-1)	177	Hypothetical protein	Hypothetical protein	PUB domain
(-1)	231	Hypothetical protein	Hypothetical phage protein	No prediction
(-3)	363	Hypothetical protein	Hypothetical protein	No prediction
(-1)	153	Hypothetical protein	Hypothetical protein	No prediction

Table 16. Annotation of giant phage version 1 from NAP_{CR} isolates (continued).

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(-1)	180	Hypothetical protein	Hypothetical protein	No prediction
(-1)	1254	RNA-splicing ligase RtcB	RNA-splicing ligase RtcB	tRNA-splicing ligase, RtcB
(-2)	360	Hypothetical protein	Hypothetical protein	No prediction
(-3)	468	Hypothetical protein	Hypothetical protein	No prediction
(-2)	351	Hypothetical protein	Hypothetical protein	No prediction
(-3)	282	Hypothetical protein	Hypothetical protein	No prediction
(-1)	447	Hypothetical protein	Hypothetical protein	No prediction
(-3)	156	Hypothetical protein	Hypothetical protein	No prediction
(-3)	381	Peptidyl-tRNA hydrolase	Peptidyl-tRNA hydrolase PTH2 family protein	peptidyl-tRNA hydrolase PTH2 family protein
(-2)	468	Hypothetical protein	Hypothetical protein	No prediction
(-3)	111	Hypothetical protein	Hypothetical protein	No prediction
(-1)	345	Hypothetical protein	Hypothetical protein	No prediction
(-2)	285	Hypothetical protein	Hypothetical protein	No prediction
(-3)	369	Hypothetical protein	Hypothetical protein	No prediction
(-2)	537	Hypothetical protein	Hypothetical protein	No prediction
(-3)	396	Hypothetical protein	Hypothetical protein	No prediction
(-3)	1011	Hypothetical protein	Hypothetical protein	No prediction
(-2)	189	Hypothetical protein	Hypothetical protein	No prediction
(-2)	588	Hypothetical protein	Hypothetical protein	No prediction
(-1)	201	Hypothetical protein	Hypothetical protein	No prediction
(-1)	1284	DNA ligase	ATP dependent DNA ligase domain protein	DNA ligase, ATP-dependent, central
(-1)	186	Hypothetical protein	Hypothetical protein	No prediction
(-2)	390	Hypothetical protein	Hypothetical protein	No prediction
(-3)	327	Hypothetical protein	Hypothetical protein	No prediction
(-1)	600	Hypothetical protein	RNA 2'-phosphotransferase	Phosphotransferase KptA/Tpt1
(-2)	117	Hypothetical protein	Hypothetical protein	No prediction
(-3)	195	Hypothetical protein	Hypothetical protein	No prediction
(-3)	414	YopX protein	YopX family protein	YopX protein
(-3)	234	Hypothetical protein	Hypothetical protein	No prediction
(-1)	141	Hypothetical protein	Hypothetical protein	No prediction
(-2)	339	Hypothetical protein	Hypothetical protein	No prediction

Table 16. Annotation of giant phage version 1 from NAP_{CR} isolates (continued).

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(-3)	894	Hypothetical protein	Hypothetical protein	No prediction
(-1)	972	Tyrosine recombinase XerD	Site-specific tyrosine recombinase XerC	DNA breaking-rejoining enzyme, catalytic core
(-3)	369	Fic/DOC family protein	Death-on-curing family protein	Death on curing protein and FIDO domain
(-2)	174	Hypothetical protein	Hypothetical protein	No prediction
(-3)	573	Hypothetical protein	CopG family transcriptional regulator	No prediction
(-3)	2397	CRISPR-associated nuclease/helicase Cas3	CRISPR-associated helicase/endonuclease Cas3	CRISPR-associated Cas3
(-1)	525	Hypothetical protein	Hypothetical protein	No prediction
(-3)	534	Hypothetical protein	Hypothetical protein	Type IIA DNA topoisomerase subunit A, alpha-helical domain
(-3)	333	Hypothetical protein	Hypothetical protein	No prediction
(-2)	303	Hypothetical protein	Hypothetical protein	No prediction
(-2)	507	Hypothetical protein	Hypothetical protein	No prediction
(-3)	1083	Hypothetical protein	Hypothetical protein	No prediction
(-1)	117	Hypothetical protein	Hypothetical protein	No prediction
(-2)	486	Hypothetical protein	Hypothetical protein	No prediction
(-1)	585	Hypothetical protein	Hypothetical protein	No prediction
(-3)	339	Hypothetical protein	Hypothetical protein	No prediction
(-3)	759	Hypothetical protein	BRO family, N-terminal domain protein	BRO N-terminal domain
(-2)	267	Hypothetical protein	Hypothetical protein	No prediction
(-3)	345	Hypothetical protein	Hypothetical protein	No prediction
(-2)	141	Hypothetical protein	Hypothetical protein	No prediction
(-1)	258	Hypothetical protein	Hypothetical protein	No prediction
(-1)	189	Hypothetical protein	Antidote-toxin recognition MazE family protein	No prediction
(-2)	324	Hypothetical protein	Hypothetical protein	No prediction
(+3)	318	HTH-type transcriptional regulator ImmR	Transcriptional regulator	Lambda repressor-like, DNA-binding domain

Table 16. Annotation of giant phage version 1 from NAP_{CR} isolates (continued).

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(-3)	369	Hypothetical protein	Hypothetical protein	No prediction
(+2)	138	Hypothetical protein	Hypothetical protein	No prediction
(+3)	945	Hypothetical protein	StbA family protein	Plasmid segregation protein ParM/StbA
(-1)	627	Hypothetical protein	Helix-turn-helix domain protein	No prediction
(-1)	240	Hypothetical protein	Hypothetical protein	No prediction
(+3)	195	Hypothetical protein	Hypothetical protein	No prediction
(+2)	1203	Bacterial regulatory proteins, luxR family	LuxR family transcriptional regulator	Transcription regulator LuxR, C-terminal
(+3)	306	Bacterial DNA-binding protein	DNA-binding protein	Integration host factor (IHF)-like DNA-binding domain
(+1)	672	Hypothetical protein	Hypothetical protein	No prediction
(+3)	555	Hypothetical protein	Hypothetical protein	No prediction
(+1)	618	Hypothetical protein	Hypothetical protein	No prediction
(+1)	189	Hypothetical protein	Hypothetical protein	No prediction
(+2)	1194	Calcineurin-like phosphoesterase superfamily domain protein	Calcineurin-like phosphoesterase superfamily domain protein	Metallo-dependent phosphatase-like or calcineurin-like phosphoesterase domain, apaH type

Table 17. Annotation of giant phage version 2 from NAP_{CR1} isolates.

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(+3)	861	Hypothetical protein	Hypothetical protein	No prediction
(+2)	1776	Terminase	Phage Terminase family protein	Terminase
(+1)	1485	Hypothetical protein	Hypothetical protein	No prediction
(+3)	237	Hypothetical protein	Hypothetical protein	No prediction
(+3)	1383	Hypothetical protein	Hypothetical protein	No prediction
(+1)	489	Hypothetical protein	Hypothetical protein	No prediction
(+3)	972	Hypothetical protein	Hypothetical protein	No prediction
(+1)	594	Hypothetical protein	Hypothetical protein	No prediction
(+1)	636	Hypothetical protein	Hypothetical protein	No prediction
(+3)	1047	Hypothetical protein	Hypothetical protein	No prediction
(+1)	450	Hypothetical protein	Hypothetical protein	No prediction
(+2)	804	Hypothetical protein	Hypothetical protein	No prediction

Table 17. Annotation of giant phage version 2 from NAP_{CR1} isolates (continued).

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(+3)	819	Hypothetical protein	Hypothetical protein	No prediction
(+2)	786	Hypothetical protein	Hypothetical protein	Siphovirus-type tail component
(+1)	450	Hypothetical protein	Hypothetical protein	No prediction
(+1)	438	Hypothetical protein	Hypothetical protein	No prediction
(+1)	651	Phage repressor protein KilAC omain protein	Phage antirepressor KilAC domain protein	BRO N-terminal domain and antirepressor domain
(+2)	786	Hypothetical protein	Prophage antirepressor BRO family, N-terminal domain protein	BRO N-terminal domain
(+3)	714	Hypothetical protein	Prophage antirepressor BRO family, N-terminal domain protein	BRO family, N-terminal domain protein
(+2)	705	Sensory transduction protein LytR	Sensory transduction protein LytR	CheY-like superfamily and LytTR DNA-binding domain
(+1)	360	Hypothetical protein	Hypothetical protein	No prediction
(+2)	288	Hypothetical protein	Hypothetical protein	No prediction
(+1)	159	Hypothetical protein	Hypothetical protein	No prediction
(+2)	3228	Type IIS restriction enzyme Eco57I	Putative type II restriction enzyme, methylase	S-adenosyl-L-methionine-dependent methyltransferase and TaqI-like C-terminal specificity domain
(-1)	450	HTH-type transcriptional regulator ImmR	Transcriptional regulator	Lambda repressor-like, DNA-binding domain
(+1)	1044	Phage antirepressor protein KilAC domain protein	Phage antirepressor KilAC domain protein	BRO N-terminal domain and antirepressor protein, C-terminal
(+3)	1161	Hypothetical protein	Hypothetical protein	No prediction
(+1)	711	Hypothetical protein	Hypothetical protein	SHOCT domain
(+3)	111	Hypothetical protein	Hypothetical protein	No prediction
(+3)	984	Hypothetical protein	Hypothetical protein	No prediction
(+2)	1479	Phage-related minor tail protein	Phage tail tape measure protein	No prediction
(+1)	6069	Hypothetical protein	Phage tail tape measure protein	Phage tail tape measure protein
(+2)	156	Hypothetical protein	Hypothetical protein	No prediction
(+2)	1716	Hypothetical protein	Prophage endopeptidase tail family protein (74% Id)	No prediction
(+1)	1929	Hypothetical protein	Chaperone of endosialidase	Intramolecular chaperone auto-processing domain
(+3)	156	Hypothetical protein	Hypothetical protein	No prediction
(+1)	2076	Hypothetical protein	Hypothetical protein	No prediction

Table 17. Annotation of giant phage version 2 from NAP_{CR1} isolates (continued).

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(+3)	4884	Regulator of chromosome condensation (RCC1) repeat protein	Regulator of chromosome condensation (RCC1) repeat protein	Regulator of chromosome condensation 1/beta-lactamase-inhibitor protein II
(+2)	327	Hypothetical protein	Hypothetical protein	No prediction
(+2)	330	Hypothetical protein	Phage tail-collar fiber family protein	No prediction
(+1)	1716	Glycine rich protein	Glycine rich family protein	No prediction
(+3)	294	Hypothetical protein	Hypothetical protein	No prediction
(+1)	183	Hypothetical protein	Hypothetical protein	No prediction
(+2)	816	Sporulation-specific N-acetylmuramoyl-L-alanine amidase	N-acetylmuramoyl-L-alanine amidase	N-acetylmuramoyl-L-alanine amidase
(+3)	618	Mannosyl-glycoprotein endo-beta-N-acetylglucosaminidase	Mannosyl-glycoprotein endo-beta-N-acetylglucosaminidase	Mannosyl-glycoprotein endo-beta-N-acetylglucosaminidase
(+1)	252	Hypothetical protein	Membrane protein	Holin
(+1)	312	Hypothetical protein	Hypothetical protein	No prediction
(+3)	936	Tyrosine recombinase XerC	Phage integrase	Integrase
(+3)	354	HTH-type transcriptional regulator SinR	Transcriptional regulator	Lambda repressor-like, DNA-binding domain
(+1)	321	Methicillin resistance regulatory protein MecI	Transcriptional regulator	BlaI transcriptional regulatory family
(+3)	1263	N-acetylmuramoyl-L-alanine amidase LytC precursor	Cell wall-binding repeat 2 family protein	Putative cell wall binding repeat 2
(-1)	222	Hypothetical protein	Hypothetical protein	No prediction
(-3)	132	Hypothetical protein	Hypothetical protein	No prediction
(-2)	1845	Hypothetical protein	Hypothetical protein	No prediction
(-1)	1113	Transposase from transposon Tn916	Transposase from transposon Tn916	DNA breaking-rejoining enzyme, catalytic core
(-2)	750	DNA topoisomerase I	Topoisomerase DNA binding C4 zinc finger family protein	Nuclease-related domain, NERD and DNA topoisomerase, type IA, zn finger
(-1)	174	Hypothetical protein	Hypothetical protein	No prediction
(-3)	141	Hypothetical protein	Hypothetical protein	No prediction

Table 17. Annotation of giant phage version 2 from NAP_{CR1} isolates (continued).

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(-1)	474	Deoxyuridine 5'-triphosphate nucleotidohydrolase	Deoxyuridine 5'-triphosphate nucleotidohydrolase	Deoxyuridine triphosphate nucleotidohydrolase
(-2)	495	Holliday junction resolvase	Crossover junction endodeoxyribonuclease RuvC	Crossover junction endodeoxyribonuclease RuvC
(-1)	582	Guanylate kinase	Guanylate kinase	Guanylate kinase/L-type calcium channel beta subunit
(-2)	321	Hypothetical protein	Hypothetical protein	Putative phage protein
(-3)	840	PD-(D/E)XK nuclease superfamily protein	PD-(D/E)XK nuclease superfamily protein	Restriction endonuclease type II-like and exonuclease, phage-type/RecB, C-terminal
(-2)	210	Hypothetical protein	Hypothetical protein	No prediction
(-3)	258	Hypothetical protein	Hypothetical protein	No prediction
(-2)	255	Hypothetical protein	Hypothetical protein	No prediction
(-3)	528	Hypothetical protein	Spore protease YyaC	Peptidase HybD-like domain
(-1)	666	Bis(5'-nucleosyl)-tetraphosphatase, symmetrical	Serine/threonine protein phosphatase	Metallo-dependent phosphatase-like and calcineurin-like phosphoesterase domain, apaH type
(-3)	444	Hypothetical protein	Hypothetical protein	No prediction
(-2)	852	Hypothetical protein	Hypothetical protein	No prediction
(-1)	273	Hypothetical protein	Hypothetical protein	No prediction
(-2)	501	Pyruvate formate-lyase I-activating enzyme	Anaerobic ribonucleoside-triphosphate reductase activating protein	Ribonucleoside-triphosphate reductase activating, anaerobic
(-3)	2226	Anaerobic ribonucleoside-triphosphate reductase	Ribonucleoside-triphosphate reductase	Ribonucleoside-triphosphate reductase, anaerobic
(-1)	1140	Hypothetical protein	DNA-sulfur modification-associated family protein	DNA sulphur modification protein DndB
(-3)	303	Hypothetical protein	Hypothetical protein	No prediction
(-1)	3333	DNA polymerase III subunit alpha	DNA polymerase III subunit alpha	Bacterial DNA polymerase III, alpha subunit
(-2)	1701	Putative SPBc2 prophage-derived single-strand DNA-specific exonuclease YorK	Putative SPBc2 prophage-derived single-strand DNA-specific exonuclease YorK	No prediction

Table 17. Annotation of giant phage version 2 from NAP_{CR1} isolates (continued).

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(-3)	1593	Hypothetical protein	Chromosome partitioning protein ParA	DNA helicase DnaB, N-terminal/DNA primase DnaG, C-terminal and P-loop containing nucleoside triphosphate hydrolase
(-1)	558	Hypothetical protein	Hypothetical protein	No prediction
(-1)	576	ERF superfamily protein	Recombination protein, phage associated	Essential recombination function protein
(-2)	846	Hypothetical protein	Hypothetical protein	No prediction
(-1)	171	Hypothetical protein	Putative membrane protein	No prediction
(-2)	666	Hypothetical protein	Hypothetical protein	No prediction
(-1)	561	Hypothetical protein	Hypothetical protein	No prediction
(-2)	426	Hypothetical protein	Hypothetical protein	No prediction
(-1)	1122	Hypothetical protein	pcfJ-like family protein	PcfJ-like protein
(-1)	309	Hypothetical protein	Hypothetical protein	No prediction
(-3)	252	Hypothetical protein	Hypothetical protein	No prediction
(-2)	210	Hypothetical protein	Hypothetical protein	No prediction
(-1)	603	Hypothetical protein	Hypothetical protein	No prediction
(-3)	327	Hypothetical protein	Hypothetical protein	S-adenosyl-L-methionine-dependent methyltransferase
(-2)	135	Hypothetical protein	Hypothetical protein	No prediction
(-3)	285	Hypothetical protein	DUF5052 domain-containing protein	DUF5052 domain-containing protein
(-3)	396	Hypothetical protein	Hypothetical protein	No prediction
(-3)	351	Hypothetical protein	Hypothetical protein	No prediction
(-1)	183	Hypothetical protein	Hypothetical protein	No prediction
(-2)	210	Hypothetical protein	Hypothetical protein	No prediction
(-3)	177	Hypothetical protein	Hypothetical protein	PUB domain
(-3)	231	Hypothetical protein	Hypothetical phage protein	No prediction
(-2)	363	Hypothetical protein	Hypothetical protein	No prediction
(-3)	153	Hypothetical protein	Hypothetical protein	No prediction
(-3)	180	Hypothetical protein	Hypothetical protein	No prediction
(-3)	1254	RNA-splicing ligase RtcB	RNA-splicing ligase RtcB	tRNA-splicing ligase, RtcB

Table 17. Annotation of giant phage version 2 from NAP_{CR1} isolates (continued).

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(-1)	360	Hypothetical protein	Hypothetical protein	No prediction
(-2)	468	Hypothetical protein	Hypothetical protein	No prediction
(-1)	351	Hypothetical protein	Hypothetical protein	No prediction
(-2)	282	Hypothetical protein	Hypothetical protein	No prediction
(-3)	447	Hypothetical protein	Hypothetical protein	No prediction
(-2)	156	Hypothetical protein	Hypothetical protein	No prediction
(-2)	381	Peptidyl-tRNA hydrolase	Peptidyl-tRNA hydrolase PTH2 family protein	peptidyl-tRNA hydrolase PTH2 family protein
(-1)	468	Hypothetical protein	Hypothetical protein	No prediction
(-2)	111	Hypothetical protein	Hypothetical protein	No prediction
(-3)	345	Hypothetical protein	Hypothetical protein	No prediction
(-1)	285	Hypothetical protein	Hypothetical protein	No prediction
(-2)	369	Hypothetical protein	Hypothetical protein	No prediction
(-1)	537	Hypothetical protein	Hypothetical protein	No prediction
(-2)	396	Hypothetical protein	Hypothetical protein	No prediction
(-2)	1011	Hypothetical protein	Hypothetical protein	No prediction
(-1)	189	Hypothetical protein	Hypothetical protein	No prediction
(-1)	588	Hypothetical protein	Hypothetical protein	No prediction
(-3)	201	Hypothetical protein	Hypothetical protein	No prediction
(-3)	1284	DNA ligase	ATP dependent DNA ligase domain protein	DNA ligase, ATP- dependent, central
(-3)	186	Hypothetical protein	Hypothetical protein	No prediction
(-1)	390	Hypothetical protein	Hypothetical protein	No prediction
(-2)	327	Hypothetical protein	Hypothetical protein	No prediction
(-3)	600	RNA 2'- phosphotransferase	RNA 2'- phosphotransferase	Phosphotransferase KptA/Tpt1
(-1)	117	Hypothetical protein	Hypothetical protein	No prediction
(-2)	195	Hypothetical protein	Hypothetical protein	No prediction
(-2)	414	YopX protein	YopX family protein	YopX protein
(-2)	234	Hypothetical protein	Hypothetical protein	No prediction
(-3)	141	Hypothetical protein	Hypothetical protein	No prediction
(-1)	339	Hypothetical protein	Hypothetical protein	No prediction
(-2)	894	Hypothetical protein	Hypothetical protein	No prediction

Table 17. Annotation of giant phage version 2 from NAP_{CR1} isolates (continued).

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(-2)	972	Tyrosine recombinase XerD	Site-specific tyrosine recombinase XerC	DNA breaking-rejoining enzyme, catalytic core
(-3)	573	Hypothetical protein	CopG family transcriptional regulator	No prediction
(-1)	1164	Putative transposase DNA-binding domain protein	Transposase	Transposase IS605, OrfB, C-terminal
(-3)	402	Transposase IS200 like protein	Transposase	Transposase IS200-like
(+1)	300	Hypothetical protein	Hypothetical protein	No prediction
(-3)	486	Hypothetical protein	Hypothetical protein	No prediction
(-3)	333	Hypothetical protein	Hypothetical protein	No prediction
(-2)	348	Helix-turn-helix domain protein	Transcriptional regulator	Lambda repressor-like, DNA-binding domain
(-1)	381	Transcriptional repressor DicA	Transcriptional regulator	Lambda repressor-like, DNA-binding domain
(-2)	369	Helix-turn-helix domain protein	Transcriptional regulator	Lambda repressor-like, DNA-binding domain
(-3)	1446	Hypothetical protein	Hypothetical protein	No prediction
(-1)	165	Hypothetical protein	Hypothetical protein	No prediction
(-2)	453	Hypothetical protein	Hypothetical protein	No prediction
(-3)	159	Hypothetical protein	Hypothetical protein	No prediction
(-2)	498	Hypothetical protein	Hypothetical protein	No prediction
(-1)	585	Hypothetical protein	Hypothetical protein	No prediction
(-2)	264	Hypothetical protein	Hypothetical protein	No prediction
(-2)	612	Hypothetical protein	Hypothetical protein	No prediction
(-3)	237	Hypothetical protein	Hypothetical protein	No prediction
(-3)	138	Hypothetical protein	Hypothetical protein	No prediction
(-2)	258	Hypothetical protein	Hypothetical protein	No prediction
(-2)	189	Hypothetical protein	Antidote-toxin recognition MazE family protein	No prediction
(-3)	324	Hypothetical protein	Hypothetical protein	No prediction
(+3)	318	HTH-type transcriptional regulator ImmR	Transcriptional regulator	Lambda repressor-like, DNA-binding domain
(-1)	369	Hypothetical protein	Hypothetical protein	No prediction
(+1)	138	Hypothetical protein	Hypothetical protein	No prediction
(+2)	945	Hypothetical protein	StbA family protein	Plasmid segregation protein ParM/StbA
(-3)	627	Hypothetical protein	Helix-turn-helix domain protein	No prediction

Table 17. Annotation of giant phage version 2 from NAP_{CR1} isolates (continued).

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(-3)	240	Hypothetical protein	Hypothetical protein	No prediction
(+3)	195	Hypothetical protein	Hypothetical protein	No prediction
(+2)	1203	Bacterial regulatory proteins, luxR family	LuxR family transcriptional regulator	Transcription regulator LuxR, C-terminal
(+2)	306	Bacterial DNA-binding protein	DNA-binding protein	Integration host factor (IHF)-like DNA-binding domain
(+3)	672	Hypothetical protein	Hypothetical protein	No prediction
(+1)	555	Hypothetical protein	Hypothetical protein	No prediction
(+1)	618	Hypothetical protein	Hypothetical protein	No prediction
(+1)	189	Hypothetical protein	Hypothetical protein	No prediction
(+3)	1194	Calcineurin-like phosphoesterase superfamily domain protien	Calcineurin-like phosphoesterase superfamily domain protein	Metallo-dependent phosphatase-like or calcineurin-like phosphoesterase domain, apaH type

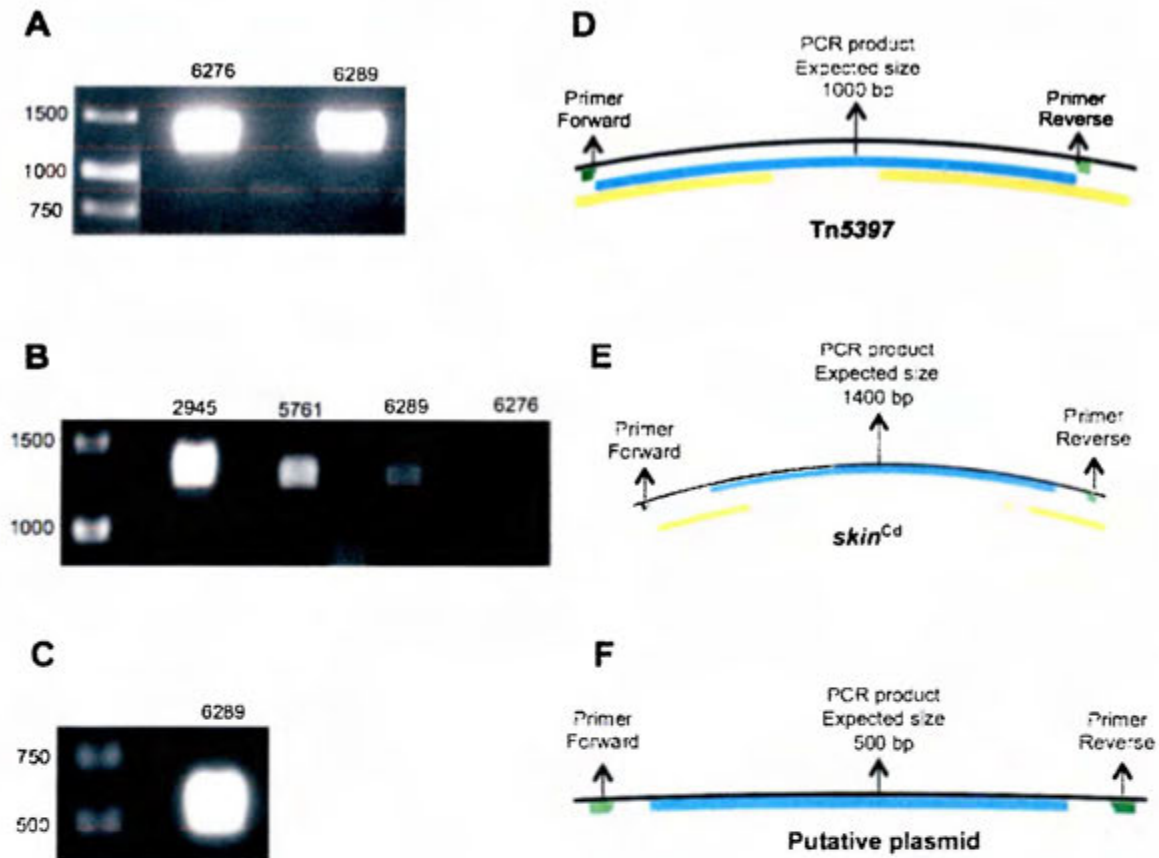


Figure 18. PCR products obtained for circular intermediates of (A) Tn5397 in isolates 6279 and 6289, (B) *skin*^{Cd} in isolates 2945, 5761 and 6289, (C) putative plasmid in isolate 6289. Sequences confirmed through bidirectional Sanger sequencing are shown as blue bars (D, E, F) with primer binding sites highlighted in green.

The only differential MGE found among the NAP1 isolates was previously annotated as a putative plasmid with phage proteins (36). This putative plasmid was found in isolates 5700, 5703 and 5720 from Cluster V. Once again the element consists mostly of hypothetical proteins. However, some of the predicted proteins are DNA-binding proteins, phage for GIY-YIG protein, the plasmid segregation protein StbA, structural phage proteins, resolvase, endopeptidases and a peptidoglycan binding protein LysM (Table 18). For this element circularization assays were not performed.

Table 18. Annotation of a putative plasmid from NAP1 isolates.

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(-1)	189	Hypothetical protein	Hypothetical protein	No prediction
(-1)	1740	Hypothetical protein	6-phospho-beta-glucosidase	No prediction
(-3)	459	Hypothetical protein	Transposase family protein	Homeodomain-like (DNA binding)
(-1)	645	Tyrosine recombinase XerD	Integrase	Integrase
(-1)	546	Hypothetical protein	Hypothetical protein	No prediction
(-2)	444	Hypothetical protein	RNA polymerase subunit sigma	Winged helix-turn-helix DNA-binding domain (DNA transcription factor)
(-2)	258	Hypothetical protein	Hypothetical protein	No prediction
(-3)	486	Hypothetical protein	Hypothetical protein	No prediction
(-3)	642	Hypothetical protein	HNH endonuclease	No prediction
(-3)	261	Hypothetical protein	Hypothetical protein	No prediction
(-3)	807	Hypothetical protein	Hypothetical protein	No prediction
(-2)	537	LemA family protein	LemA family protein	LemA domain
(-1)	327	Hypothetical protein	Hypothetical protein	No prediction
(-3)	159	Hypothetical protein	Hypothetical protein	No prediction
(-1)	546	Sporulation sigma factor SigF	RNA polymerase sigma factor, sigma-70 family protein	RNA polymerase sigma 70
(-2)	378	Single-stranded DNA-binding protein	single-stranded DNA-binding protein	Primosome PriB/single-strand DNA-binding
(-1)	183	Hypothetical protein	Hypothetical protein	No prediction
(-1)	705	Hypothetical protein	Hypothetical protein	No prediction
(-3)	387	Hypothetical protein	Hypothetical protein	No prediction
(-1)	1332	Replicative DNA helicase	Replicative DNA helicase	DNA helicase
(-1)	234	Hypothetical protein	Hypothetical protein	No prediction
(-1)	261	Hypothetical protein	Hypothetical protein	No prediction
(+3)	207	Hypothetical protein	Hypothetical protein	No prediction
(-1)	219	Hypothetical protein	Hypothetical protein	No prediction
(-3)	186	Hypothetical protein	Hypothetical protein	No prediction
(-1)	732	Hypothetical protein	Helix-turn-helix domain protein	No prediction
(-2)	216	Hypothetical protein	Hypothetical protein	No prediction
(-3)	162	Hypothetical protein	Hypothetical protein	Insect odorant-binding protein A10/Ejaculatory bulb-specific protein 3
(+1)	369	Helix-turn-helix domain protein	Helix-turn-helix domain protein, transcriptional regulator	Lambda repressor-like, DNA-binding domain and Cro/C1-type helix-turn-helix domain
(-1)	156	Hypothetical protein	Hypothetical protein	No prediction

Table 18. Annotation of a putative plasmid from NAP1 isolates (continued).

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(+1)	384	Penicillinase repressor	Penicillinase repressor family protein	Blal transcriptional regulatory family
(-1)	690	HTH domain protein	HTH domain protein	Winged helix-turn-helix DNA-binding domain
(+1)	1188	Initiator replication protein	RepB family plasmid replication initiator protein	Winged helix-turn-helix DNA-binding domain and initiator Rep protein
(+3)	801	Hypothetical protein	GIY-YIG catalytic domain	GIY-YIG nuclease superfamily
(+2)	735	GIY-YIG catalytic domain protein	Endonuclease	GIY-YIG nuclease superfamily
(+3)	153	Hypothetical protein	Hypothetical protein	No prediction
(+3)	303	Hypothetical protein	Hypothetical protein	No prediction
(+2)	459	Hypothetical protein	Hypothetical protein	No prediction
(+1)	204	Hypothetical protein	Hypothetical protein	No prediction
(+1)	906	StbA protein	StbA protein, ATPase and ppx/GppA phosphatase	Plasmid segregation protein ParM/StbA
(+1)	516	Hypothetical protein	Hypothetical protein	Homeodomain-like
(-1)	336	Penicillinase repressor	Penicillinase repressor	Blal transcriptional regulatory family
(-1)	543	Modification methylase DpnIIB	DNA methylase family protein	S-adenosyl-L-methionine-dependent methyltransferase
(-1)	156	Hypothetical protein	Putative phage DNA methylase	S-adenosyl-L-methionine-dependent methyltransferase
(-2)	837	N-acetylmuramoyl-L-alanine amidase	N-acetylmuramoyl-L-alanine amidase	Cell wall hydrolase/autolysin, catalytic
(-1)	432	Holin family protein	Toxin secretion/phage lysis holin family protein	Bacteriophage holin family
(-2)	219	Hypothetical protein	Hypothetical protein	No prediction
(-3)	183	Hypothetical protein	Phage protein	No prediction
(-2)	294	Hypothetical protein	Phage protein	No prediction
(-2)	1488	Hypothetical protein	Hypothetical protein	No prediction
(-3)	783	Hypothetical protein	Tail protein	Phage tail fibre protein
(-3)	615	Hypothetical protein	Phage protein	Bacteriophage Mu, Gp48
(-1)	1053	Baseplate J-like protein	Baseplate J-like family protein	Baseplate protein J-like
(-2)	435	Hypothetical protein	Phage protein	No prediction
(-3)	333	Hypothetical protein	Hypothetical protein	No prediction
(-1)	1677	Putative endopeptidase p60 precursor	Phage cell wall hydrolase (plasmid)	Endopeptidase, NLPC/P60 domain
(-3)	636	LysM domain/BON superfamily protein	Peptidoglycan-binding protein LysM	LysM domain

Table 18. Annotation of a putative plasmid from NAP1 isolates (continued).

ORF	ORF length (bp)	Prokka Annotation	Blast Annotation	Interpro Annotation
(-3)	477	Telomeric repeat-binding factor 2	Hypothetical protein	Immunoprotective extracellular, immunoglobulin-like domain
(-1)	3393	Phage-related minor tail protein	Phage tail tape measure protein	Phage tail tape measure protein
(-1)	150	Hypothetical protein	Hypothetical protein	No prediction
(-3)	417	Phage XkdN-like protein	Phage XkdN-like family protein	Clostridium phage phiCD119, XkdN
(-3)	441	Phage-like element PBSX protein XkdM	Phage-like element PBSX protein XkdM	Phage tail tube protein
(-3)	1056	Phage tail sheath protein	Phage portal protein	Tail sheath protein
(-3)	450	Hypothetical protein	Hypothetical protein	No prediction
(-2)	357	Hypothetical protein	Hypothetical protein	Bacteriophage HK97-gp10, putative tail-component
(-1)	348	Hypothetical protein	Hypothetical protein	No prediction
(-3)	381	Phage gp6-like head-tail connector protein	Phage gp6-like head-tail connector family protein	Phage gp6-like head-tail connector protein
(-1)	273	Hypothetical protein	Rho termination factor	Rho termination factor, N-terminal
(-2)	924	Hypothetical protein	Hypothetical protein	No prediction
(-3)	603	Hypothetical protein	Hypothetical protein	No prediction
(-1)	327	Hypothetical protein	Hypothetical protein	No prediction
(-2)	762	Phage Mu protein F like protein	Phage head morphogenesis, SPP1 gp7 family domain protein	Phage head morphogenesis domain
(-3)	264	Phage portal protein, SPP1 Gp6-like	Phage portal, SPP1 Gp6-like family protein	Portal protein, SPP1 Gp6-like

7.7 MGE have a greater effect in the microdiversification of isolates from the NAP_{CR1} pulsotype than in NAP1 isolates.

When the putative plasmid (60 kb), giant phage version 1 (130 kb), giant phage version 2 (130 kb) and prophage (56 kb) inserted in CTn5 were removed from the sequences, the amount of gene clusters detected in the resulting pseudomolecules were reduced from 4 802 (Fig 20A) to 4 595 (Fig 20B) and the distances between the isolates were diminished. This observation confirms that these MGE indeed play a role in the differentiation of the NAP_{CR1} isolates. Interestingly, the 487 isolates

remained distant from the other isolates and closer to CD630 despite this sequence edition. Thus, the selected MGE have an effect in the microdiversification of the NAP_{CR1} pulsotype.

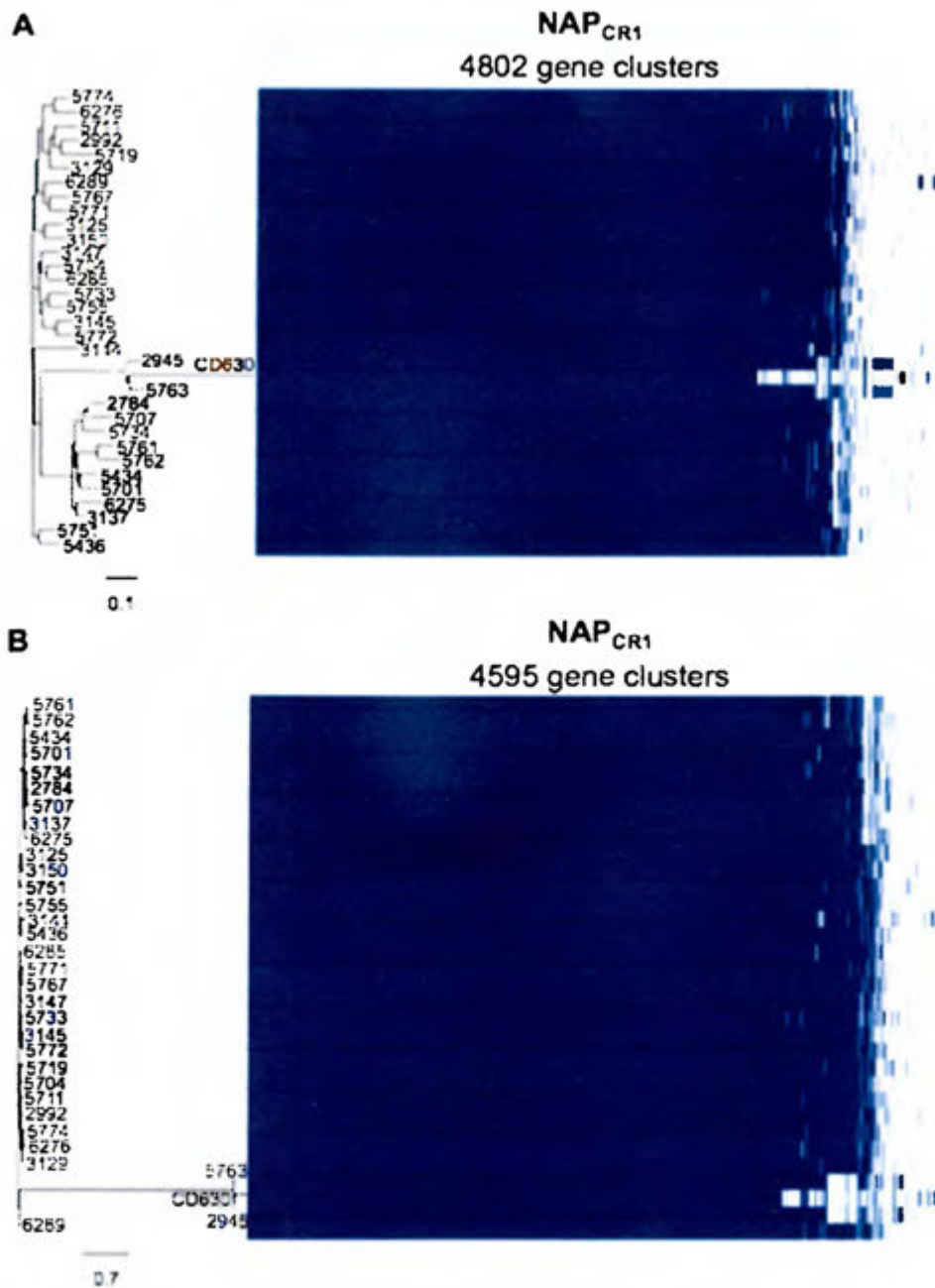


Figure 19. Roary pangenome analysis of NAP_{CR1} and NAP_{CR1} WGS modified through manual removal of selected differential MGE. Blue bars indicates presence of gene cluster. (A) Original NAP_{CR1} pangenome analysis. (B) Repetition of the analysis with WGS from which the putative plasmid, the giant phage v1, the giant phage v2 and the putative prophage was manually removed.

Since it was the only differential MGE among the NAP1 isolates, the putative plasmid was removed from the sequences of isolates 5700, 5703 and 5720 and the pangenome analysis was removed. This modification resulted in a reduction of the predicted gene clusters from 3 829 (Fig 21A) to 3 755 (Fig 21B). Additionally, the distance between Cluster V and the rest of the isolates was reduced (Figure 21B).

These results show that MGE do have an effect in the variations of the accessory genome for each pulsotype, but in the case of NAP_{CR1} the effect is greater because more differential MGE were found.

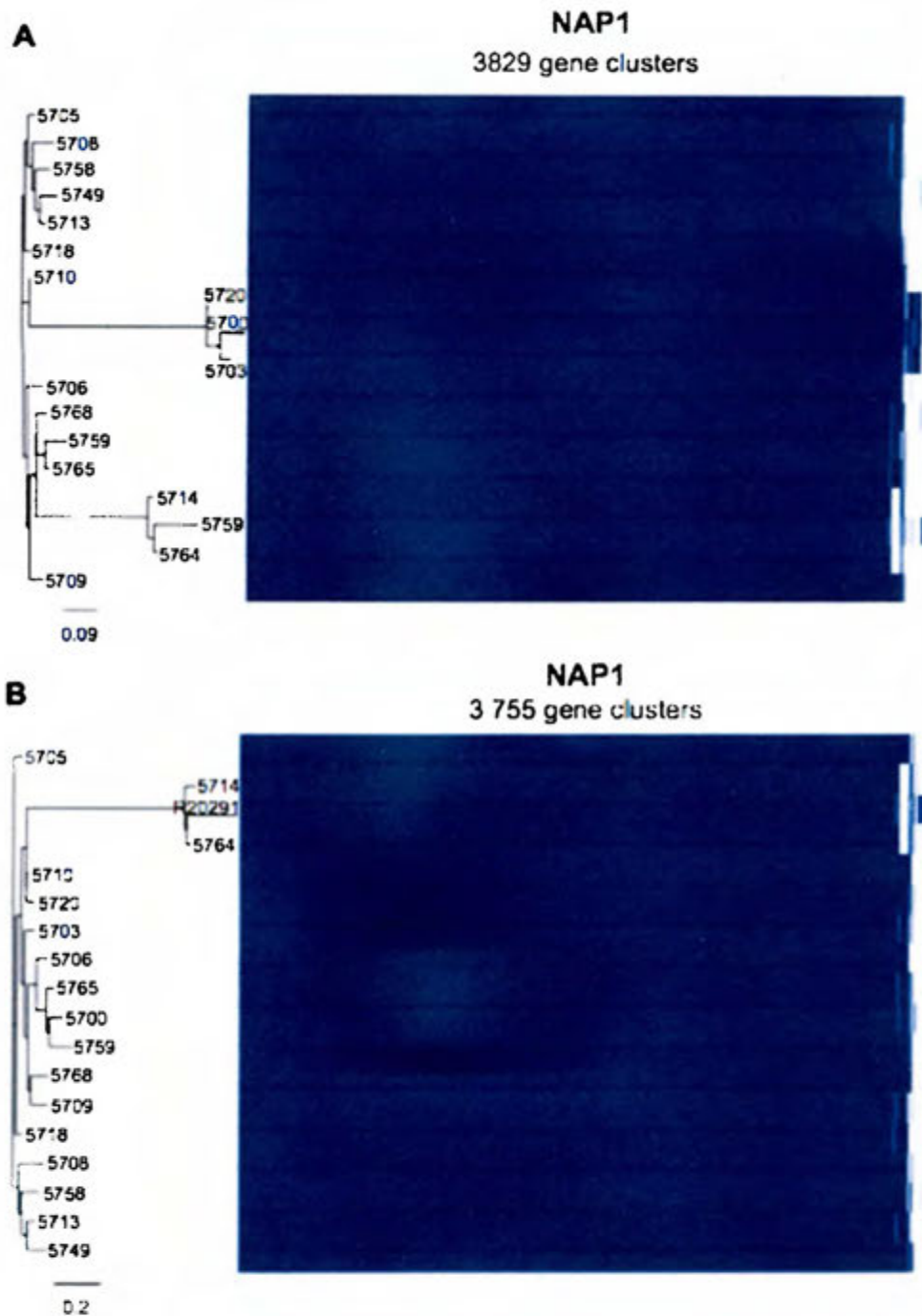


Figure 20. Roary pangenome analysis of NAP1 and NAP1 WGS modified through manual removal of selected differential MGE. Blue bar indicates presence of gene cluster. (A) Original NAP1 pangenome analysis. (B) Repetition of the analysis with WGS from which the putative plasmid was manually removed.

8 DISCUSSION

This study proved that acquisition of several MGE rather than core SNP accumulation was more relevant for the microdiversification of the NAP_{CR1} isolates. In the NAP1 isolates, by contrast, only one differential MGE was identified. In agreement with these notions, a smaller percentage of the NAP_{CR1} reads mapped to the selected reference, the size of the accessory genome and pangenome of the NAP_{CR1} isolates was larger, and they contained more unique gene clusters.

In regard to the core genome, although the absolute number of SNPs and the SNP density of both groups of strains did not differ markedly, it was clear that their mutations are under the effect of different selective pressures. The NAP_{CR1} isolates accumulated more synonymous SNPs, indicating that their genomes are purging non-synonymous mutations that may affect coding regions (44). This tendency is seen when core genomes undergo unnoticed recombination events (42). Congruently, a *r/m* ratio compatible with this scenario (2.54) was already calculated by Didelot *et al.* for other strains of the ST54 (47) and Stabler *et al.* reported that some STs from Clade I are in the process of microdiversification (80). Striking differences in the *dN/dS* rates were observed when the different subtypes of NAP_{CR1} isolates were analyzed separately. In this regard, isolates from the 487 macrorestriction pattern showed a much higher *dN/dS* rate and are likely under a greater positive purifying pressure. Congruently, these isolates formed a discrete cluster with a bootstrap of 100 in maximum likelihood trees generated from core SNPs alignments.

Also in relation to the core genome, the NAP1 isolates accumulated more non-synonymous SNPs than their NAP_{CR1} cognates, hence are under the effect of positive purifying pressure. However, it should be considered that when natural selection has not had enough time to act, non-synonymous SNPs accumulate (41, 42, 44). The *dN/dS* estimates >1 calculated for both pulsotypes indicate mutational diversification (44). Nevertheless, the *dN/dS* rate of the NAP1 isolates was three times higher than that of the NAP_{CR1} isolates meaning that this process has a greater relevance in the former group. This conclusion is in agreement with the very low *r/m* ratios that others

have calculated for NAP1 strains, which oscillate between 0.04 (47) and 0.25 (11). The fact that the NAP1 strains diversify through mutational events is further exemplified and confirmed by the emergence of two global lineages that in different time points acquired an identical mutation in the DNA gyrase A subunit conferring resistance to fluoroquinolones (81). Our results match conclusions raised by Didelot *et al.*, who detected variations in the *r/m* rates of several *C. difficile* STs. These authors stated that the diversification forces acting on the *C. difficile* genomes may vary between the lineages. Second, they reported that ST1 strains such as NAP1 are characterized by a low recombination rate and diversify by mutational processes. This strategy, which is likely to be favored by a fine-tuned pathogenic strain, has also been proposed for ST37 strains (A-B+), which as NAP1 strains produces outbreaks (47).

Some of the SNPs identified in both pulsotypes justify further studies, as they were present in genes related to metabolism, virulence factors, membrane transport, antibiotic resistance and transcriptional regulators. For instance, SNPs were detected in the precursor of the *slpA* gene in at least two NAP_{CR1} isolates. SlpA is a recognized virulence factor related to bacterial adhesion and immune response (14, 82). Also, SNPs were found in genes coding for putative exosporium proteins, which are part of the surface proteins of spores of *C. difficile*. Given that the composition of the spore surface plays an essential role in protecting this pathogen from insults in aerobic environments outside the host (83) as well as from the immune system inside the host, spore surface proteins are under a high selective pressure (42). Finally, some interesting genes containing SNPs were related to metabolic processes, including the carbohydrate phosphotransferase system (PTS) that is relevant for *C. difficile* virulence due to its relation to the bacterial catabolite repression system (CCR). In the presence of glucose, the HPr protein from the PTS system binds to the transcriptional regulator CcpA, which in turn interacts with *cre* sites located in several promoters of PaLoc genes repressing toxin production (84). Another gene expression regulation mechanism is related to Rnase Y, a gene coding for an endonuclease for mRNA processing and degradation (85, 86). The NAP1 isolates shared SNPs in the SlpA and PTS systems, but also had SNPs in two-component sensor histidine kinases,

which are essential for signal transductions in prokaryotes in response to environmental signals or quorum sensing (87).

In relation to the accessory and pangenomes, only 74% of the predicted NAP_{CR1} genes, compared to 93.4% NAP1 genes, were shared by all analyzed isolates. Even more outstanding is that 11% of the predicted NAP_{CR1} genes clusters are found in only 0% to 15% of the analyzed isolates. These results clearly show that the NAP_{CR1} isolates have an open pangenome, as expected for organisms living in highly heterogeneous and changing conditions (88). By contrast, bacteria thriving in narrow niches contain a small genome and a closed pangenome. Although further tests are required to reach a solid conclusion (90), this seems to be the case of the NAP1 isolates. This difference is of substantial biological relevance as it implies that the NAP_{CR1} and NAP1 are possibly confronted and specialized to different conditions in the human gut and outside of it. Our results also agree with the notion that larger genomes have an increased capacity to acquire MGE (88).

To further sustain that the NAP_{CR1} pangenome is open, these isolates were not distributed in the branches of a parsimony-based pangenomic tree according to their *Sma*I pattern or hospital of isolation. Instead, the topology of this tree was dictated by the gain or loss of certain MGE that included most unique gene clusters: at least eight differential MGE were identified in NAP_{CR1} isolates, whereas only one MGE was found in the NAP1 strains. MGE can be an important source of diversification in bacterial genomes and even in the generation of new pathogens (48). For instance, they commonly carry antibiotic resistance genes, promoting the spread of resistant variants among a bacterial population, for example the multiresistance genomic island SSC*mec* conferring methicillin resistance to *S. aureus* or transposons Tn4453a and Tn5397 from *C. difficile* conferring resistance to chloramphenicol and tetracycline, respectively (48, 49, 89). Also, clear examples exist of MGE carrying virulence factors increasing the pathogenicity of bacterial strains. For example, the toxins of *Bacillus anthracis* are harbored in a plasmid, the diphtheria and cholera toxins were acquired from bacteriophages by *Corynebacterium diphtheriae* and *Vibrio cholerae* respectively,

or the human pathogenic *Escherichia coli* strains that acquired many of their virulence factors from genomic islands (48, 89). Additionally, MGE have been related to the acquisition of genes for processing new substrates, degrading toxic components or membrane transport (89). These MGE also suffer from selection processes since bacteria need a balance between genome integrity and instability, thus avoiding the intake of genetic content that does not increase their fitness (90, 91). These tasks can be performed by systems like the Restriction-Modification, CRISPR-Cas and the DNA repair guided by RecA (91).

The defined differential MGE consists of known elements of *C. difficile* and new elements. In the case of NAP_{CR1}, the known elements were Tn5397 and *skin*^{Cd} present in other strains of this pathogen, including CD630 (79, 92). The PCR assays detected circular intermediates for both of these MGE, hence they can participate in horizontal gene transfer. The *skin*^{Cd} element has been reported to be excised from the chromosome during late sporulation in a process vital for the regulation of efficient sporulation (79). Therefore, it is expected that isolates that lack the element like NAP_{CR1}-6276 sporulate poorly. This hypothesis awaits verification in our isolate.

The other six differential MGE found in NAP_{CR1} are novel or gave partial hits with previously described elements. All of the them were not found in CD630, confirming that many of the differences between CD630 and NAP_{CR1} come from foreign DNA (33). The first one is the so-called *mobCksgA* found in all the isolates but on different locations. This element is a putative mobilizable transposon because it has a recombinase, a *mobC* and a *repA* (48, 93). Even though we were not able to detect a circular intermediate, the fact that it is found in different genomic locations indirectly shows its capacity to undergo mobilization. Additionally, it includes a gene similar to *ksgA* which encodes a methyltransferase essential for the assembly of the 30S subunits. Inactivation or lack of this protein in bacteria diminishes susceptibility to the aminoglycoside kasugamycin (94). Two further differential MGE carry confirmed or potential antibiotic resistance determinants. First, a variation of the Tn4001 transposon originally described in *S. aureus* (95) was found in our isolates. Second, a

prophage inserted in the CTn5 of CD630 was found only in the 487 macrorestriction pattern having a variety of resistant determinants like an aminoglycoside phosphotransferase and a GNAT acetyltransferase for resistance to aminoglycosides (96, 97). Antibiotics play an essential role in bacterial selection (98) and in some cases they have been shown to promote the uptake of foreign DNA through horizontal gene transfer. For instance, exposition of fecal phage to ciprofloxacin and ampicillin increased the expression of several antibiotic resistance genes, phage integration and phages host range in a mouse model (99). The mentioned MGE have aminoglycosides resistance determinants for which *C. difficile* is intrinsically resistant, but this represents a risk because *C. difficile* may serve as a gene reservoir for other bacteria of the gut microbiota. Interestingly, several of the phage proteins and recombinases showed BLAST hits to phages from *Enterococcus faecium*, suggesting that the NAP_{CR1} isolates share DNA with other intestinal Firmicutes. For example, lateral transfer of Tn5397 between *C. difficile* and *E. faecalis* has been accomplished in the laboratory and, in the other direction, anaerobic enterococci could transfer their vancomycin resistance to *C. difficile* (100), jeopardizing one of the last therapies available for multidrug resistant isolates. Additionally, all of the NAP_{CR1} isolates shared multiple MGE with antibiotic resistance determinants which were not part of this study. For example, the previously mentioned Tn4453a with *catD* conferring resistance to chloramphenicol, Tn5398 with an *ermB* gene for resistance to clindamycin, and a Tn916-like element with a putative SAM-radical protein similar to *cfr*, a gene known to confer simultaneous resistance to phenicols, lincosamides, pleuromutilins, streptogramin A and certain macrolides.

Some of the novel elements, including the giant phages and the plasmid could provide NAP_{CR1} isolates with virulence factors. For instance, one variant of the giant phages has a protein with a Fic/DOC domain, a protein family known to postranslationally modify the cytoskeleton and thereby interfere with intracellular traffic, signaling and translation pathways in eukaryotic cells (101). Likewise, the putative plasmid found only in isolate 6289 contains genes for a type IV secretion system, plasmid segregation and two putative novel virulence factors: an ADP

ribosyltransferase exoenzyme and a von Willebrand type A domain protein. We predict this plasmid to be conjugative and circular according to the annotation of a type IV secretion system and the PCR results, respectively. (48). The activity of the ADP ribosyltransferase exoenzyme remains to be determined but the general mechanism of these enzymes is based on a covalent linkage of ADP to host proteins disrupting intracellular signaling pathways (102). The von Willebrand type A domains, which are present in proteins from the extracellular matrix and integrins receptors, mediate adhesion, thus we predict this gene to code for an adhesin (103). The single differential MGE of the NAP1 isolates has been reported before as a putative plasmid (36). Nonetheless, we recommend revising this affirmation because it includes several phage proteins. Some of the predicted proteins in this element resemble SigF and LemA. The former is a transcriptional regulator that regulates sporulation (83) and proteins from the LemA family are predicted to be transcriptional regulators from two component systems, as described in *Pseudomonas syringae* pv. *syringae* (104). Both proteins have the potential to modulate the virulence of the NAP1 isolates that carry this putative plasmid/phage element.

Most of the genes found in the differential MGE of both groups of strains encode hypothetical proteins, thus there is plenty of unknown information. A more precise annotation of the MGE, based on functional studies, will allow us to gain insight into other potential adaptations of *C. difficile* to its environment and its virulent capacity. Nonetheless, this does not necessarily mean these proteins are expressed in the pathogen so further studies are needed.

From the constant exchange of MGE, the NAP_{CR1} pulsotype and, in particular the 487 pattern, is microdiversifying through the acquisition of antibiotic resistance determinants, virulence factors and metabolic advantages (105). This could end up with a better adapted pathogen capable of coexisting with the NAP1 strains, which, in contrast, protect their well-developed pathogenic strategy and avoid drastic modifications of their accessory genomes. Whether the NAP1 strains have active

barriers for lateral gene transfer that are not present in the NAP_{CR1} strains remains to be determined

To corroborate some of the findings, it is desirable to estimate the r/m rate of the isolates studied and thereby determine whether recombination or mutations are driving the microdiversification of their core genome. The expected result would be to obtain $r/m < 0$ for NAP1 isolates and $r/m > 0$ for the other group. Moreover, the novel MGE found should be further studied and better annotated as this improved information can clarify the advantages that they confer to NAP_{CR1} strains.

A limitation of the study is the difference in the number of isolates from each pulsotype studied. However, this was tolerated to have the chance to compare isolates that cocirculated in time and space and in this manner limit the effect of confounding factors. This decision might have affected the size of the NAP1 pangenome, but it is unlikely that it will depart significantly from that of the global NAP1 population as indicated by the very high percentage of reads that mapped to the reference and the clonality of this strain (38, 71). Another limitation comes from working with MGE, as they can be lost from bacterial genomes through constant passages or from exposition to antibiotics. To reduce this possibility, DNA extractions for WGS were performed from freshly thawed isolates.

CD630 was chosen as the reference genome for all NAP_{CR1} analyses because, from all sequenced *C. difficile* reference strains, it is the most closely related (33). For now, it is not known whether the NAP_{CR1} pulsotype diversified from CD630 or alternatively if they share a common ancestor. To answer this, their time of divergence should be calculated and further phylogeny analyses with other isolates from Clade I should be performed. Isolates from the 487 pattern were closer to CD630 than the rest of the NAP_{CR1} isolates, hence they could serve as useful tools to confirm their relationship.

9 CONCLUSIONS

C. difficile lineages can microdiversify through different genetic mechanisms.

The acquisition of MGE in the accessory genome is the main mechanism of microdiversification of NAP_{CR1} isolates when compared to NAP1. These MGE confer the NAP_{CR1} novel functions and antibiotic resistance genes, which could relate to its capacity to generate outbreaks.

The accumulation of non-synonymous mutations in the core genome was the main mechanism of microdiversification of NAP1 pulsotype when compared to NAP_{CR1}. In this way, NAP1 is able to accumulate potential adaptations without risking its fine-tuned pathogenic capacity.

Coinciding with the well known genetic diversity of this species, studies of the diversification of *C. difficile* should include the pangenome and not only the core genome.

10 REFERENCES

1. **Hunt JJ, Ballard JD.** 2013. Variations in virulence and molecular biology among emerging strains of *Clostridium difficile*. *Microbiol Mol Biol Rev* **77**:567–81.
2. **Leffler D, Lamont T.** 2015. *Clostridium difficile* Infection. *N Engl J Med* **372**:1539–1548.
3. **Slimings C, Riley T V.** 2014. Antibiotics and hospital-acquired *Clostridium difficile* infection: Update of systematic review and meta-analysis. *J Antimicrob Chemother.*
4. **Lessa FC, Mu Y, Bamberg WM, Beldavs ZG, Dumyati GK, Dunn JR, Farley MM, Holzbauer SM, Meek JI, Phipps EC, Wilson LE, Winston LG, Cohen J a, Limbago BM, Fridkin SK, Gerding DN, McDonald LC.** 2015. Burden of

- Clostridium difficile* Infection in the United States. N Engl J Med **372**:825–834.
5. **McGlone SM, Bailey RR, Zimmer SM, Popovich MJ, Tian Y, Ufberg P, Muder RR, Lee BY.** 2012. The economic burden of *Clostridium difficile*. Clin Microbiol Infect **18**:282–9.
 6. **Center for Disease Control and Prevention of the United States (CDC).** 2013. Antibiotic resistance threats in the United States, 2013 Report Antibiotic Resistance Threats.
 7. **Gupta A, Khanna S.** 2014. Community-acquired *Clostridium difficile* infection: An increasing public health threat. Infect Drug Resist **7**:63–72.
 8. **Lyerly DM, Lockwood DE, Richardson SH, Wilkins TD.** 1982. Biological activities of toxins A and B of *Clostridium difficile*. Infect Immun **35**:1147–50.
 9. **Just I, Selzer J, Wilm M, von Eichel-Streiber C, Mann M, Aktories K.** 1995. Glucosylation of Rho proteins by *Clostridium difficile* toxin B. Nature **375**:500–3.
 10. **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**:1734–41.
 11. **Dingle KE, Griffiths D, Didelot X, Evans J, Vaughan A, Kachrimanidou M, Stoesser N, Jolley K a, Golubchik T, Harding RM, Peto TE, Fawley W, Walker a S, Wilcox M, Crook DW.** 2011. Clinical *Clostridium difficile*: clonality and pathogenicity locus diversity. PLoS One **6**:e19993.
 12. **Dupuy B, Govind R, Antunes a, Matamouros S.** 2008. *Clostridium difficile* toxin synthesis is negatively regulated by TcdC. J Med Microbiol **57**:685–9.
 13. **Ausiello CM, Cerquetti M, Fedele G, Spensieri F, Palazzo R, Nasso M, Frezza S, Mastrantonio P.** 2006. Surface layer proteins from *Clostridium difficile* induce inflammatory and regulatory cytokines in human monocytes and dendritic cells. Microbes Infect **8**:2640–6.
 14. **Merrigan M, Venugopal A, Roxas J.** 2013. Surface-Layer Protein A (SlpA) Is a Major Contributor to Host-Cell Adherence of *Clostridium difficile*. PLoS One **8**:1–12.
 15. **Schwan C, Stecher B, Tzivelekidis T, van Ham M, Rohde M, Hardt W-D,**

- Wehland J, Aktories K.** 2009. *Clostridium difficile* toxin CDT induces formation of microtubule-based protrusions and increases adherence of bacteria. *PLoS Pathog* **5**:e1000626.
16. **Faulds-Pain A, Twine SM, Vinogradov E, Strong PCR, Dell A, Buckley AM, Douce GR, Valiente E, Logan SM, Wren BW.** 2014. The post-translational modification of the *Clostridium difficile* flagellin affects motility, cell surface properties and virulence. *Mol Microbiol* **94**:272–289.
 17. **Twine SM, Reid CW, Aubry A, McMullin DR, Fulton KM, Austin J, Logan SM.** 2009. Motility and flagellar glycosylation in *Clostridium difficile*. *J Bacteriol* **191**:7050–62.
 18. **Knight DR, Elliott B, Chang BJ, Perkins TT, Riley T V.** 2015. Diversity and Evolution in the Genome of *Clostridium difficile*. *Clin Microbiol Rev* **28**:721–741.
 19. **Knetsch CW, Lawley TD, Hensgens MP, Corver J, Wilcox MW, Kuijper EJ.** 2013. Current application and future perspectives of molecular typing methods to study *Clostridium difficile* infections. *Euro Surveill* **18**:20381.
 20. **Klaassen CHW, van Haren HA, Horrevorts AM.** 2002. Molecular fingerprinting of *Clostridium difficile* isolates: pulsed-field gel electrophoresis versus amplified fragment length polymorphism. *J Clin Microbiol* **40**:101–4.
 21. **Wren BW, Tabaqchali S.** 1987. Restriction Endonuclease DNA Analysis of *Clostridium difficile*. *Am Fam Physician* **25**:2402–2404.
 22. **Cartwright CP, Stock F, Beekmann SE, Williams EC, Gill VJ.** 1995. PCR Amplification of Ribosomal-Rna Intergenic Spacer Regions as a Method for Epidemiologic Typing of *Clostridium difficile*. *J Clin Microbiol* **33**:184–187.
 23. **Rupnik M, Janezic S.** 2016. An Update on *Clostridium difficile* Toxinotyping. *J Clin Microbiol* **54**:13–18.
 24. **Lemee L, Dhalluin A, Pestel-caron M.** 2004. Multilocus Sequence Typing Analysis of Human and Animal *Clostridium difficile* Isolates of Various Toxigenic Types. *J Clin Microbiol* **42**:2609–2617.
 25. **Griffiths D, Fawley W, Kachrimanidou M, Bowden R, Crook DW, Fung R, Golubchik T, Harding RM, Jeffery KJM, Jolley KA, Kirton R, Peto TE, Rees G, Stoesser N, Vaughan A, Walker AS, Young BC, Wilcox M, Dingle KE.**

2010. Multilocus Sequence Typing of *Clostridium difficile*. J Clin Microbiol **48**:770–778.
26. **Muto C, Pokrywka M, Shutt K.** 2005. A large outbreak of *Clostridium difficile*-associated disease with an unexpected proportion of deaths and colectomies at a teaching hospital following increased. Infect Control Hosp Epidemiol **26**:273–280.
27. **Warny M, Pepin J, Fang A, Killgore G, Thompson A, Brazier J, Frost E, McDonald LC.** 2005. Toxin production by an emerging strain of *Clostridium difficile* associated with outbreaks of severe disease in North America and Europe. Lancet **366**:1079–84.
28. **Carter GP, Douce GR, Govind R, Howarth PM, Mackin KE, Spencer J, Buckley AM, Antunes A, Kotsanas D, Jenkin G a, Dupuy B, Rood JI, Lyras D.** 2011. The anti-sigma factor TcdC modulates hypervirulence in an epidemic BI/NAP1/027 clinical isolate of *Clostridium difficile*. PLoS Pathog **7**:e1002317.
29. **Freeman J, Bauer MP, Baines SD, Corver J, Fawley WN, Goorhuis B, Kuijper EJ, Wilcox MH.** 2010. The changing epidemiology of *Clostridium difficile* infections. Clin Microbiol Rev **23**:529–549.
30. **Goorhuis A, Bakker D, Corver J, Debast SB, Harmanus C, Notermans DW, Bergwerff AA, Dekker FW, Kuijper EJ.** 2008. Emergence of *Clostridium difficile* infection due to a new hypervirulent strain, polymerase chain reaction ribotype 078. Clin Infect Dis **47**:1162–70.
31. **Du P, Cao B, Wang J, Li W, Jia H, Zhang W, Lu J, Li Z, Yu H, Chen C, Cheng Y.** 2014. Sequence variation in tcdA and tcdB of *Clostridium difficile*: ST37 with truncated tcdA is a potential epidemic strain in China. J Clin Microbiol **52**:3264–3270.
32. **López-Ureña D, Quesada-Gómez C, Montoya-Ramírez M, del Mar Gamboa-Coronado M, Somogyi T, Rodríguez C, Rodríguez-Cavallini E.** 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**:e42.
33. **Quesada-Gómez C, López-Ureña D, Acuña-Amador L, Villalobos-Zúñiga**

- M, Du T, Freire R, Guzmán-Verri C, Gamboa-Coronado MDM, Lawley TD, Moreno E, Mulvey MR, de Castro Brito GA, Rodríguez-Cavallini E, Rodríguez C, Chaves-Olarte E.** 2015. Emergence of an outbreak-associated *Clostridium difficile* variant with increased virulence. *J Clin Microbiol* **53**:JCM.03058–14.
34. **Wust J, Sullivan NM, Hardegger U, Wilkins TD.** 1982. Investigation of an outbreak of antibiotic-associated colitis by various typing methods. *J Clin Microbiol* **16**:1096–1101.
35. **Sebahia M, Wren BW, Mullany P, Fairweather NF, Minton N, Stabler R, Thomson NR, Roberts AP, Cerdeño-Tárraga AM, Wang H, Holden MTG, Wright A, Churcher C, Quail M a, Baker S, Bason N, Brooks K, Chillingworth T, Cronin A, Davis P, Dowd L, Fraser A, Feltwell T, Hance Z, Holroyd S, Jagels K, Moule S, Mungall K, Price C, Rabinowitsch E, Sharp S, Simmonds M, Stevens K, Unwin L, Whithead S, Dupuy B, Dougan G, Barrell B, Parkhill J.** 2006. The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. *Nat Genet* **38**:779–86.
36. **Riedel T (DSMZ), Bunk B (DSMZ), Thuermer A (Universität G, Sproer C (DSMZ), Brzuszkiewicz EB, Abt B, Gronow S, Liesegang H, Daniel R, Overmann J.** 2015. Genome Resequencing of the Virulent and Multidrug-Resistant Reference Strain *Clostridium difficile* 630 **3**:15–16.
37. **Monot M, Boursaux-Eude C, Thibonnier M, Vallenet D, Moszer I, Medigue C, Martin-Verstraete I, Dupuy B.** 2011. Reannotation of the genome sequence of *Clostridium difficile* strain 630. *J Med Microbiol* **60**:1193–9.
38. **Stabler R a, Gerding DN, Songer JG, Drudy D, Brazier JS, Trinh HT, Witney a a, Hinds J, Wren BW.** 2006. Comparative phylogenomics of *Clostridium difficile* reveals clade specificity and microevolution of hypervirulent strains. *J Bacteriol* **188**:7297–305.
39. **Dingle KE, Elliott B, Robinson E, Griffiths D, Eyre DW, Stoesser N, Vaughan A, Golubchik T, Fawley WN, Wilcox MH, Peto TE, Walker a S, Riley T V, Crook DW, Didelot X.** 2014. Evolutionary history of the *Clostridium difficile* pathogenicity locus. *Genome Biol Evol* **6**:36–52.

40. **Monot M, Eckert C, Lemire A, Hamiot A, Dubois T, Tessier C, Dumoulaud B, Hamel B, Petit A, Lalande V, Ma L, Bouchier C, Barbut F, Dupuy B.** 2015. *Clostridium difficile*: New Insights into the Evolution of the Pathogenicity Locus. *Sci Rep* **5**:15023.
41. **He M, Sebaihia M, Lawley TD, Stabler R a, Dawson LF, Martin MJ, Holt KE, Seth-Smith HMB, Quail M a, Rance R, Brooks K, Churcher C, Harris D, Bentley SD, Burrows C, Clark L, Corton C, Murray V, Rose G, Thurston S, van Tonder A, Walker D, Wren BW, Dougan G, Parkhill J.** 2010. Evolutionary dynamics of *Clostridium difficile* over short and long time scales. *Proc Natl Acad Sci U S A* **107**:7527–32.
42. **Castillo-Ramírez S, Harris SR, Holden MTG, He M, Parkhill J, Bentley SD, Feil EJ.** 2011. The impact of recombination on dN/dS within recently emerged bacterial clones. *PLoS Pathog* **7**.
43. **Loewe L.** 2008. Genetic Mutation. *Nat Educ* **1**:113.
44. **Kryazhimskiy S, Plotkin JB.** 2008. The Population Genetics of dN/dS. *PLoS Genet* **4**:e1000304.
45. **Vos M, Didelot X.** 2009. A comparison of homologous recombination rates in bacteria and archaea. *ISME J* **3**:199–208.
46. **Lemée L, Bourgeois I, Ruffin E, Collignon A, Lemeland J-F, Pons J-L.** 2005. Multilocus sequence analysis and comparative evolution of virulence-associated genes and housekeeping genes of *Clostridium difficile*. *Microbiology* **151**:3171–80.
47. **Didelot X, Eyre DW, Cule M, Ip CL, Ansari MA, Griffiths D, Vaughan A, O'Connor L, Golubchik T, Batty EM, Piazza P, Wilson DJ, Bowden R, Donnelly PJ, Dingle KE, Wilcox M, Walker a S, Crook DW, A Peto TE, Harding RM.** 2012. Microevolutionary analysis of *Clostridium difficile* genomes to investigate transmission. *Genome Biol* **13**:R118.
48. **Frost LS, Lepiae R, Summers AO, Toussaint A.** 2005. Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol* **3**:722–732.
49. **Roberts AP, Allan E, Mullany P.** 2014. The impact of horizontal gene transfer on the biology of *Clostridium difficile*. *Advances in microbial physiology*, 1st ed.

Elsevier Ltd.

50. **Brouwer MSM, Warburton PJ, Roberts AP, Mullany P, Allan E.** 2011. Genetic Organisation, Mobility and Predicted Functions of Genes on Integrated, Mobile Genetic Elements in Sequenced Strains of *Clostridium difficile*. *PLoS One* **6**:e23014.
51. **Amy J, Johanesen P, Lyras D.** 2015. Extrachromosomal and integrated genetic elements in *Clostridium difficile*. *Plasmid* **80**:97–110.
52. **Boudry P, Semenov E, Monot M, Datsenko KA, Lopatina A, Sekulovic O, Ospina-Bedoya M, Fortier LC, Severinov K, Dupuy B, Soutourina O.** 2015. Function of the CRISPR-cas system of the human pathogen: *Clostridium difficile*. *MBio* **6**:1–16.
53. **Hargreaves KR, Flores CO, Lawley TD, Clokie MRJ.** 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**:e01045–13.
54. **Walter BM, Rupnik M, Hodnik V, Anderluh G, Dupuy B, Paulič N, Žgur-Bertok D, Butala M.** 2014. The LexA regulated genes of the *Clostridium difficile*. *BMC Microbiol* **14**:88.
55. **Burdett V.** 1991. Purification and characterization of Tet(M), a protein that renders ribosomes resistant to tetracycline. *J Biol Chem* **266**:2872–7.
56. **Mullany P, Wilks M, Lamb I, Clayton C, Wren B, Tabaqchali S.** 1990. Genetic analysis of a tetracycline resistance element from *Clostridium difficile* and its conjugal transfer to and from *Bacillus subtilis*. *J Gen Microbiol* **136**:1343–1349.
57. **Roberts AP, Johanesen P a, Lyras D, Mullany P, Rood JI.** 2001. Comparison of Tn5397 from *Clostridium difficile*, Tn916 from *Enterococcus faecalis* and the CW459 tet(M) element from *Clostridium perfringens* shows that they have similar conjugation regions but different insertion and excision modules. *Microbiology* **333235**:1243–1251.
58. **Farrow K a., Lyras D, Rood JI.** 2001. Genomic analysis of the erythromycin resistance element Tn5398 from *Clostridium difficile*. *Microbiology* **147**:2717–

2728.

59. **Schmidt C, Löffler B, Ackermann G.** 2007. Antimicrobial phenotypes and molecular basis in clinical strains of *Clostridium difficile*. *Diagn Microbiol Infect Dis* **59**:1–5.
60. **Marín M, Martín A, Alcalá L, Cercenado E, Iglesias C, Reigadas E, Bouza E.** 2015. *Clostridium difficile* Isolates with High Linezolid MICs Harbor the Multiresistance Gene *cfr*. *Antimicrob Agents Chemother* **59**:586–589.
61. **Lyras D, Storie C, Huggins AS, Crellin PK, Bannam TL, Rood JI.** 1998. Chloramphenicol resistance in *Clostridium difficile* is encoded on Tn4453 transposons that are closely related to Tn4451 from *Clostridium perfringens*. *Antimicrob Agents Chemother* **42**:1563–1567.
62. **Corver J, Bakker D, Brouwer MSM, Harmanus C, Hensgens MP, Roberts AP, Lipman LJA, Kuijper EJ, van Leeuwen HC.** 2012. Analysis of a *Clostridium difficile* PCR ribotype 078 100 kilobase island reveals the presence of a novel transposon, Tn6164. *BMC Microbiol*.
63. **Roberts AP, Mullany P.** 2011. Tn916-like genetic elements: A diverse group of modular mobile elements conferring antibiotic resistance. *FEMS Microbiol Rev* **35**:856–871.
64. **Goh S, Hussain H, Chang BJ, Emmett W, Riley T V, Mullany P.** 2013. Phage C2 Mediates Transduction of Tn6215 , Encoding Erythromycin Resistance, between *Clostridium difficile* Strains. *MBio* **4**:e00840–13.
65. **Goh S, Ong PF, Song KP, Rily T V., Chang BJ.** 2007. The complete genome sequence of *Clostridium difficile* phage ϕ C2 and comparisons to ϕ CD119 and inducible prophages of CD630. *Microbiology* **153**:676–685.
66. **Govind R, Fralick J a, Rolfe RD.** 2006. Genomic Organization and Molecular Characterization of *Clostridium dificile* Bacteriophage PhiCD119. *J Bacteriol* **188**:2568–2577.
67. **Meessen-Pinard M, Sekulovic O, Fortier L-C.** 2012. Evidence of In Vivo Prophage Induction during *Clostridium difficile* Infection. *Appl Environ Microbiol*.
68. **Sekulovic O, Garneau JR, Néron A, Fortier L-C.** 2014. Characterization of Temperate Phages Infecting *Clostridium difficile* Isolates of Human and Animal

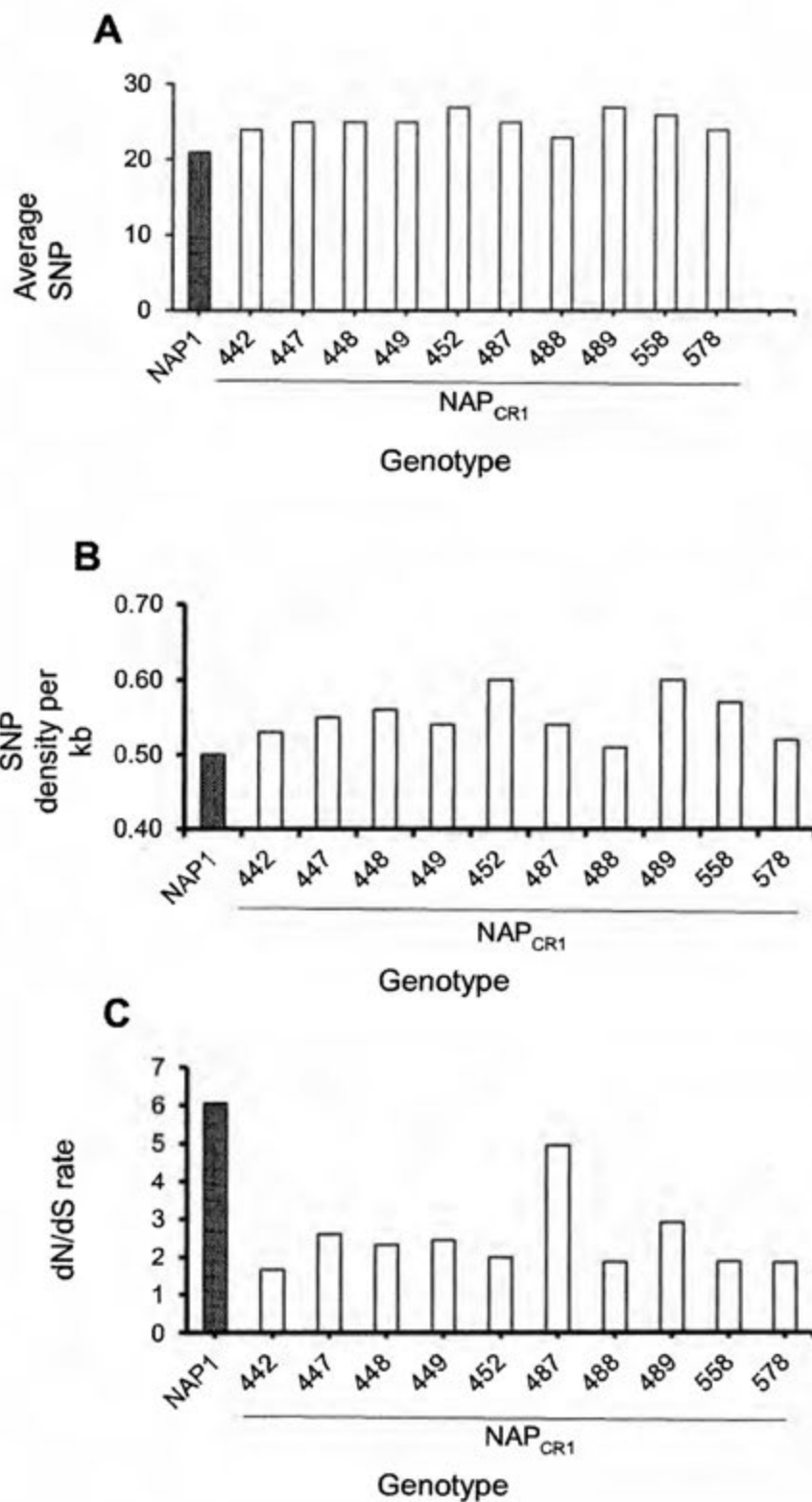
- Origins. *Appl Environ Microbiol* **80**:2555–63.
69. **Sekulovic O, Meessen-Pinard M, Fortier LC.** 2011. Prophage-stimulated toxin production in *Clostridium difficile* NAP1/027 lysogens. *J Bacteriol* **193**:2726–2734.
 70. **Hargreaves KR, Kropinski AM, Clokie MRJ.** 2014. What does the talking?: quorum sensing signalling genes discovered in a bacteriophage genome. *PLoS One* **9**:e85131.
 71. **Stabler R a, He M, Dawson L, Martin M, Valiente E, Corton C, Lawley TD, Sebahia M, Quail M a, Rose G, Gerding DN, Gibert M, Popoff MR, Parkhill J, Dougan G, Wren BW.** 2009. Comparative genome and phenotypic analysis of *Clostridium difficile* 027 strains provides insight into the evolution of a hypervirulent bacterium. *Genome Biol* **10**:R102.
 72. **Zerbino DR.** 2010. Using the Velvet de novo assembler for short-read sequencing technologies. *Curr Protoc Bioinforma* **Unit-11.5**:1–13.
 73. **Hernandez D, François P, Farinelli L, Østerås M, Schrenzel J.** 2008. De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. *Genome Res* **18**:802–809.
 74. **Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J.** 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**:563–569.
 75. **Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ.** 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**:119.
 76. **Seemann T.** 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**:2068–2069.
 77. **Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream M a, Barrell B.** 2000. Artemis: sequence visualization and annotation. *Bioinformatics* **16**:944–945.
 78. **Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J.** 2005. ACT: The Artemis comparison tool. *Bioinformatics* **21**:3422–

- 3423.
79. **Haraldsen JD, Sonenshein AL.** 2003. Efficient sporulation in *Clostridium difficile* requires disruption of the sigK gene. *Mol Microbiol* **48**:811–821.
 80. **Stabler RA, Dawson LF, Valiente E, Cairns MD, Martin MJ, Donahue EH, Riley T V., Songer JG, Kuijper EJ, Dingle KE, Wren BW.** 2012. Macro and micro diversity of *Clostridium difficile* isolates from diverse sources and geographical locations. *PLoS One* **7**:1–12.
 81. **He M, Miyajima F, Roberts P, Ellison L, Pickard DJ, Melissa J.** 2013. Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. *Nat Genet* **45**:109–113.
 82. **Calabi E, Calabi F, Phillips AD, Fairweather NF.** 2002. Binding of *Clostridium difficile* surface layer proteins to gastrointestinal tissues. *Infect Immun* **70**:5770–8.
 83. **Paredes-Sabja D, Shen A, Sorg J a.** 2014. *Clostridium difficile* spore biology: Sporulation, germination, and spore structural proteins. *Trends Microbiol* **22**:406–416.
 84. **Martin-Verstraete I, Peltier J, Dupuy B.** 2016. The Regulatory Networks That Control *Clostridium difficile* Toxin Synthesis. *Toxins (Basel)* **8**:153.
 85. **Deutscher J, Francke C, Postma PW, Deutscher J, Francke C, Postma PW.** 2006. How Phosphotransferase System-Related Protein Phosphorylation Regulates Carbohydrate Metabolism in Bacteria How Phosphotransferase System-Related Protein Phosphorylation Regulates Carbohydrate Metabolism in Bacteria. *Microbiol Mol Biol Rev* **70**:939–1031.
 86. **Chen Z, Itzek A, Malke H, Ferretti JJ, Kreth J.** 2013. Multiple roles of Rnase Y in *Streptococcus pyogenes* mRNA processing and degradation. *J Bacteriol* **195**:2585–2594.
 87. **Mitrophanov AYA, Groisman EE a.** 2008. Signal integration in bacterial two-component regulatory systems. *Genes Dev* **22**:2601–2611.
 88. **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.

89. **Juhas M.** 2013. Horizontal gene transfer in human pathogens. *Crit Rev Microbiol* **78**:1–8.
90. **Moran NA.** 2002. Microbial Minimalism: Genome Reduction in Bacterial Pathogens **108**:583–586.
91. **Darmon E, Leach DRF.** 2014. Bacterial genome instability. *Microbiol Mol Biol Rev* **78**:1–39.
92. **Wang H, Roberts AP, Lyras D, Rood JI, Wilks M, Mullany P.** 2000. Characterization of the ends and target sites of the novel conjugative transposon Tn5397 from *Clostridium difficile*: Excision and circularization is mediated by the large resolvase, TndX. *J Bacteriol* **182**:3775–3783.
93. **Zhang S, Meyer R.** 1997. The relaxosome protein MobC promotes conjugal plasmid mobilization by extending DNA strand separation to the nick site at the origin of transfer. *Mol Microbiol* **25**:509–516.
94. **Demirci H, Murphy F, Belardinelli R, Kelley AC, Ramakrishnan V, Gregory ST, Dahlberg AE, Jogi G.** 2010. Modification of 16S ribosomal RNA by the KsgA methyltransferase restructures the 30S subunit to optimize ribosome function. *RNA* **16**:2319–2324.
95. **Mahairas GG, Lyon BR, Skurray R a., Pattee P a.** 1989. Genetic analysis of *Staphylococcus aureus* with Tn4001. *J Bacteriol* **171**:3968–3972.
96. **Dyda F, Klein DC, Hickman AB.** 2000. GCN5-Related N-Acetyltransferases: A Structural Overview. *Annu Rev Biophys Biomol Struct* **29**:81–103.
97. **Wright GD, Thompson PR.** 1999. Aminoglycoside phosphotransferases: proteins, structure, and mechanism. *Front Biosci* **4**:D9–21.
98. **Hanage WP.** 2016. Not So Simple After All: Bacteria, Their Population Genetics, and Recombination. *Cold Spring Harb Perspect Biol* cshperspect.a018069–.
99. **Modi SR, Lee HH, Spina CS, Collins JJ.** 2013. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature* **499**:219–22.
100. **Jasni AS, Mullany P, Hussain H, Roberts AP.** 2010. Demonstration of conjugative transposon (Tn5397)-mediated horizontal gene transfer between

- Clostridium difficile* and *Enterococcus faecalis*. *Antimicrob Agents Chemother* **54**:4924–4926.
101. **Roy CR, Cherfils J.** 2015. Structure and function of Fic proteins. *Nat Rev Microbiol* **13**:631–40.
 102. **Simon NC, Aktories K, Barbieri JT.** 2014. Novel bacterial ADP-ribosylating toxins: structure and function. *Nat Rev Microbiol* **12**:599–611.
 103. **Whittaker C, Hynes R.** 2002. Distribution and Evolution of von Willebrand/Integrin A Domains: Widely Dispersed Domains with Roles in Cell Adhesion and Elsewhere. *Mol Biol Cell* **13**:3369–3387.
 104. **Hrabak EM, Willis DK.** 1992. The *lemA* Gene Required for Pathogenicity of *Pseudomonas-Syringae* Pv *Syringae* on Bean Is a Member of a Family of Two-Component Regulators. *J Bacteriol* **174**:3011–3020.
 105. **Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R.** 2005. The microbial pan-genome. *Curr Opin Genet Dev* **15**:589–594.

11 . APPENDIX



Appendix 1. Core genome SNPs analyses of various NAP_{CR1} macrorestriction patterns and NAP1 isolates. (A) Total amount of SNPs found in coding regions, (B) SNP density per kb, (C) dN/dS rates.

Appendix 2. SNPs in coding regions of the NAP_{CR1} isolates (continued).

SNP	Gene	Smal pattern	4	447					448						449						4	487	4	489	558	578										
		4	2	3	5	5	5	5	6	2	3	3	5	5	5	5	5	6	6	3	5	5	6	6	5	2	5	2	5	5	3	6	3	3	5	
400 547 (T247T)	Putative exosporium glycoprotein	+	-	-	-	-	-	-	-	-	+	-	+	+	+	+	-	-	-	+	-	-	-	-	+	-	-	+	-	-	+	+	-	+	-	
400 948 (V381A)	Putative exosporium glycoprotein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	
401 303 (T499T)	Putative exosporium glycoprotein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-
401 318 (A504A)	Putative exosporium glycoprotein	+	+	+	-	-	-	+	-	+	+	+	+	+	+	+	-	+	+	+	+	-	+	+	+	+	+	-	+	+	-	+	+	+	+	+
726 788 (E68D)	Fragment of conserved hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-
745 199 (P39L)	Putative transcriptional regulator, activator Mor	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	+	+	+	+
762 547 (P260T)	Conserved hypothetical protein	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
952 731 (L543F)	Putative penicillin-binding protein	+	+	+	+	-	-	+	-	-	+	+	+	+	+	-	-	+	-	+	+	+	+	+	+	-	-	-	+	+	+	+	-	-	-	-

Appendix 2. SNPs in coding regions of the NAP_{CR1} isolates (continued).

	Smal pattern	4 4 2	447	448	449	4 5 2	487	4 8 8	489	55 8	578	
SNP	Gene	3 1 4 7	5 5 5 5 6 7 7 7 7 2 0 1 6 7 8 1 1 7 1 1	2 3 3 5 5 5 5 5 6 7 1 1 4 7 7 7 7 2 8 2 3 3 0 0 3 5 7 7 4 5 7 4 4 7 3 1 4 5	3 5 5 5 6 6 1 7 7 7 2 2 2 1 5 7 7 8 9 9 5 2 6 9	5 7 3 4	2 5 9 7 4 6 5 3	2 9 9 2	5 5 7 7 6 6 1 2	3 6 1 2 4 8 5 5	3 3 5 1 1 4 4 5 3 4 0 6	
1 033 259 (*46S)	Fragment of ABC-type transport system, oligopeptide-family ATP-binding protein	+	+	+	+	+	+	+	+	+	+	+
1 143 858 (N7N)	Fragment of putative oxydoreductase	+	+	+	+	+	+	+	+	+	+	+
1 347 052 (T484I)	Putative penicillin-binding protein with a N-terminal TonB-box	-	+	+	-	-	-	-	-	-	-	-
1 391 850 (F133L)	Putative acetyltransferase	+	+	+	+	+	+	+	+	+	+	+
1 420 829 (F61F)	Phosphopentomutase	+	+	+	+	-	-	+	+	+	+	-
1 485 957 (V227V)	Cysteine desulfurase	+	+	+	+	+	+	+	+	+	+	+
1 542 162 (A344V)	Rnase Y	-	-	-	-	+	+	-	-	-	-	-

Appendix 2. SNPs in coding regions of the NAP_{CR1} isolates (continued).

SNP	Gene	4	447	448	449	4	487	4	489	558	578
		4 4 2	5 5 5 5 6	2 3 3 5 5 5 5 5 6	3 5 5 5 6 6	5	2 5	8	5 5	3 6	3 3 5
		3	5 5 5 5 6	2 3 3 5 5 5 5 5 6	3 5 5 5 6 6	5	2 5	8	5 5	3 6	3 3 5
		1	7 7 7 7 2	7 1 1 4 7 7 7 7 2	1 7 7 7 2 2	7	9 7	9	7 7	1 2	1 1 4
		4	0 1 6 7 8	8 2 3 3 0 0 3 5 7 7	2 1 5 7 7 8	3	4 6	9	6 6	4 8	4 5 3
		7	1 1 7 1 1	4 5 7 4 4 7 3 1 4 5	9 9 5 2 6 9	4	5 3	2	1 2	5 5	4 0 6
1 883 105 (T169I)	Two-component sensor histidine kinase	-	- - - - -	- - - - -	- - - - -	-	- - - - -	-	- - - - -	- - - - -	- - - - -
1 883 597 (S333F)	Putative oligopeptide transporter	-	- - - - -	- - - - -	- - - - -	-	- - - - -	-	- - - - -	- - - - -	- - - - -
1 883 600 (R334L)	Two-component sensor histidine kinase	-	- - - - -	- - - - -	- - - - -	-	- - - - -	-	- - - - -	- - - - -	- - - - -
1 895 820 (K41*)	Tellurium resistance protein terD2	-	- - - - -	- - - - -	- - - - -	-	- - - - -	-	- - - - -	- - - - -	- - - - -
2 044 514 (P33A)	Glyceraldehyde- 3-phosphate dehydrogenase	+	+ + + + +	+ + + + +	+ + + + +	+	+ + + + +	+	+ + + + +	+ + + + +	+ + + + +
2 062 399 (I58M)	Putative lipoprotein	-	- - - - -	- - - - -	- - - - -	-	- - - - -	-	- - - - -	- - - - -	- - - - -
2 075 244 (A54T)	Putative oxidoreductase	-	- - - - -	- - - - -	- - - - -	-	- - - - -	-	- - - - -	- - - - -	- - - - -
2 202 472 (Q12K)	Putative dCMP deaminase	-	- - - - -	- - - - -	- - - - -	-	- - - - -	-	- - - - -	- - - - -	- - - - -
2 226 440 (Q149*)	Putative membrane protein	-	- - - - -	- - - - -	- - - - -	-	- - - - -	-	- - - - -	- - - - -	- - - - -
2 414 821 (T394T)	Putative xanthine/uracil permease	-	- - - - -	- - - - -	- - - - -	-	- - - - -	-	- - - - -	- - - - -	- - - - -

Appendix 2. SNPs in coding regions of the NAP_{CR1} isolates (continued).

	Smal pattern	4 4 2	447	448	449	4 5 2	487	4 8 8	489	558	578
SNP	Gene	3 1 4 7	5 5 5 5 6 7 7 7 7 2 0 1 6 7 8 1 1 7 1 1	2 3 3 5 5 5 5 5 6 7 1 1 4 7 7 7 7 2 8 2 3 3 0 0 3 5 7 7 4 5 7 4 4 7 3 1 4 5	3 5 5 5 6 6 1 7 7 7 2 2 2 2 1 5 7 7 8 9 9 5 2 6 9	5 2 7 3 4	2 5 9 7 4 6 5 3	2 9 9 2	5 5 7 7 6 6 1 2	3 6 1 2 4 8 5 5	3 3 5 1 1 4 4 5 3 4 0 6
2 415 180 (I275V)	Putative xanthine/uracil permease	-	- - - - -	- - - - -	- - - - -	-	+ +	-	- - -	- - -	- - -
2 432 624 (S186S)	ABC-type transport system, multidrug-family ATP-binding protein	+	+ + + + +	+ + + + +	+ + + + +	+	- -	+	+ + +	+ + +	+ + +
2 464 157 (S26F)	Putative exported protein	-	- - - - -	- - - - -	- - - - -	-	+	-	- - -	- - -	- - -
2 789 083 (R330K)	Transcription antiterminator, PTS operon regulator	-	- - - - -	- - - + -	- - - - -	-	-	-	+ - -	- - -	- - -
2 797 931 (A414V)	Transporter, Major Facilitator Superfamily (MFS)	+	+ + + + +	+ + + + +	+ + + + +	+	+	+	+ + +	+ + +	+ + +
2 805 690 (P166S)	Transcriptional regulator, IclR family	-	- - - + +	- - - - -	- - - + -	-	-	-	- - -	- - -	- - -
2 825 335 (E25K)	Ribosomal protein L11 methyltransferase (L11 Mtase)	-	- - - - -	- - - + -	- - - + +	-	-	-	+ - -	- - +	- - -

Appendix 3. SNPs in coding regions of NAP1 isolates.

SNP	Gene	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
		7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
		0	0	0	0	0	0	1	1	1	1	1	2	4	5	5	6	6	6
		0	3	5	6	8	9	0	2	3	4	8	0	9	8	9	4	5	8
120 450 (K131N)	30S ribosomal protein S4	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
120 932 (I58I)	DNA-directed RNA polymerase subunit alpha	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
159 347 (S152L)	Glucosamine-fructose-6-phosphate aminotransferase	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
180 888 (A18S)	Glycosylasparaginase	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
257 041 (A87V)	Chromate transporter	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
737 617 (A94S)	Hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-
904 356 (S11P)	Electron transfer flavoprotein subunit beta	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
935 725 (S356F)	Aconitate hydratase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1 116 866 (A53S)	Membrane protein	+	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-
1 326 963 (R60I)	FMN-dependent dehydrogenase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1 440 182 (Y18F)	Hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-

Appendix 4. SNPs in non-coding regions of each NAP_{CR1} isolates.

	<i>Sma</i> I pattern	4 4 2	447					448					449					4 5 2	487	488	489	558	578							
SNP	Intergenic region	3 1 4 7	5 7 0 1	5 7 1 1	5 7 6 7	6 2 8 1	2 7 8 4	3 1 2 5	3 1 3 7	5 4 0 4	5 7 0 7	5 7 3 3	5 7 5 1	5 7 3 4	5 7 5 7	6 2 7 5	6 2 8 9	5 7 3 4	2 9 4 5	5 7 6 3	2 9 9 2	5 7 6 6	5 7 6 1	3 1 4 5	6 2 8 5	3 1 4 5	3 1 4 5	5 4 3 6		
12 583 (T>C)	tRNA-Ala/23S ribosomal RNA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
654 030 (A>G)	Putative penicillin-binding peptidase BlaR1-like M56 family/conserved hypothetical protein	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
682 619 (C>T)	Putative membrane protein/Threonyl-tRNA synthetase	+	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
690 658 (A>T)	Conserved hypothetical protein/Transporter , Major Facilitator Superfamily (MFS)	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1 143 810 (A>C)	Fragment of putative oxydoreductase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1 143 817 (G>A)	Fragment of putative oxydoreductase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1 143 837 (C>T)	Fragment of putative oxydoreductase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1 210 938 (G>A)	16S ribosomal RNA/23S ribosomal RNA	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	-	+	+	+	+	+	+	+

