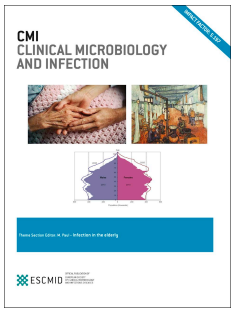
Journal Pre-proof



"The recent emergence of a highly related virulent *Clostridium difficile* clade with unique characteristics"

Helen Alexandra Shaw, Mark D. Preston, Karuna E.W. Vendrik, Michelle D. Cairns, Hilary P. Browne, Richard A. Stabler, Monique J.T. Crobach, Jeroen Corver, Hanna Pituch, Andre Ingebretsen, Munir Primohammed, Alexandra Faulds-Pain, Esmeralda Valiente, Trevor D. Lawley, Neil F. Fairweather, Ed J. Kuijper, Brendan W. Wren

PII: S1198-743X(19)30489-6

DOI: https://doi.org/10.1016/j.cmi.2019.09.004

Reference: CMI 1775

To appear in: *Clinical Microbiology and Infection*

Received Date: 2 June 2019

Revised Date: 6 September 2019

Accepted Date: 7 September 2019

Please cite this article as: Shaw HA, Preston MD, Vendrik KEW, Cairns MD, Browne HP, Stabler RA, Crobach MJT, Corver J, Pituch H, Ingebretsen A, Primohammed M, Faulds-Pain A, Valiente E, Lawley TD, Fairweather NF, Kuijper EJ, Wren BW, "The recent emergence of a highly related virulent *Clostridium difficile* clade with unique characteristics", *Clinical Microbiology and Infection*, https://doi.org/10.1016/j.cmi.2019.09.004.

ORIGINAL ARTICLE

Title: "The recent emergence of a highly related virulent *Clostridium difficile* clade with unique characteristics"

Helen Alexandra Shaw[1,2], Mark D. Preston[1,3], Karuna E. W. Vendrik[4], Michelle D. Cairns[1,5], Hilary P. Browne[6], Richard A. Stabler[1], Monique J. T. Crobach[4], Jeroen Corver[4], Hanna Pituch[7], Andre Ingebretsen[8,9], Munir Primohammed[10], Alexandra Faulds-Pain[1], Esmeralda Valiente[1], Trevor D. Lawley[6], Neil F. Fairweather[11], Ed J. Kuijper[4] & Brendan W. Wren[1*]

[1] Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK

[2] Division of Bacteriology, National Institute for Biological Standards and Controls (NIBSC), Blanche Lane, South Mimms, Potters Bar, Hertfordshire, EN6 3QG, UK

[3] Analytical Biological Service Division, National Institute for Biological Standards and Controls (NIBSC), Blanche Lane, South Mimms, Potters Bar, Hertfordshire, EN6 3QG, UK

[4] National Reference Laboratory for CDI surveillance, Department of Medical Microbiology and RIVM, Leiden University Medical Centre, Leiden, the Netherlands

[5] Public Health Laboratory London, Division of Infection, The Royal London Hospital, London, E1 2ES, UK

[6] Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SA, UK

[7] Department of Medical Microbiology, Medical University of Warsaw, Warsaw, Poland

[8] Department of Microbiology, Oslo University Hospital, Oslo, Norway

[9] Department of Infection Prevention, Oslo University Hospital, Oslo, Norway

[10] Department of Molecular and Clinical Pharmacology, The University of Liverpool, Liverpool, L69 3GL, UK

[11] CMBI, Imperial College London, South Kensington Campus, London, SW7 2AZ, UK

* Correspondence: Professor Brendan Wren, London School of Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK

Telephone: +44 (0) 20 7927 2288; brendan.wren@lshtm.ac.uk

Keywords: NGS, *Clostridium difficile*, CDI, RT023

Abstract word count: 238

Text word count: 2638

ABSTRACT:

Objectives

Clostridium difficile is a major global human pathogen divided into five clades, of which clade 3 is the least characterised and consists predominantly of PCR ribotype (RT) 023 strains. Our aim was to analyse and characterise this clade.

Methods

In this cohort study the clinical presentation of *C. difficile* RT023 infections was analysed in comparison with known "hypervirulent" and non-hypervirulent strains, using data from the Netherlands national *C. difficile* surveillance programme. European RT023 strains of diverse origin were collected and whole-genome sequenced to determine the genetic similarity between isolates. Distinctive features were investigated and characterised.

Results

Clinical presentation of *C. difficile* RT023 infections show severe infections akin to those seen with "hypervirulent" strains from clades 2 (RT027) and 5 (RT078) (35%, 29% and 27% severe CDI respectively), particularly with significantly more bloody diarrhoea than RT078 and non-hypervirulent strains (RT023 8%, other RTs 4%, p=0.036). The full genome sequence of strain CD305 is presented as a robust reference. Phylogenetic comparison of CD305 and a further 79 previously uncharacterised European RT023 strains of diverse origin revealed minor genetic divergence with >99.8% pairwise identity between strains. Analyses revealed distinctive features among clade 3 strains, including conserved PaLoc, CDT and phage insertion toxin genotypes, glycosylation of S-layer proteins, presence of the RT078 four gene trehalose cluster and an esculinase negative genotype.

Conclusions

Given their recent emergence, virulence and genomic characteristics, the surveillance of clade 3 strains should be more highly prioritised.

INTRODUCTION

C. difficile remains a major global pathogen; disease severity and relapse incidence have not abated, and community acquired infections have increased (1). *C. difficile* can be divided into five clades of virulent strains (2). The most understudied is clade 3, dominated by PCR ribotype (RT) 023 strains (2). RT023 has been reported primarily in Europe (3) and is amongst the top ten most common *C. difficile* PCR ribotypes in England (4) (CDRN report 2013-2015) and the Netherlands (unpublished data of the Dutch *C. difficile* Reference Laboratory). RT023 infections are not associated with increased mortality despite causing a high level of deleterious biomarkers (e.g. neutrophil counts) in patients and having toxin profiles similar to clade 2 (RT027) and clade 5 (RT078) strains (5, 6). However, disease severity with RT023 has been reported as similar to "hypervirulent strains", particularly in elderly patients (7), and is frequently associated with a relapse of CDI (3).

This study investigates the clinical presentation and phylogeny of *C. difficile* clade 3, uncovering and characterising unique features of these strains.

METHODS

Clinical data collection and analysis

A cohort study was performed. Clinical data from the Dutch national CDI sentinel surveillance from May 2009 until February 2018 were used to analyse the clinical characteristics of CDI episodes due to RT023. For this sentinel surveillance all hospitalized patients >2 years old, with clinical signs or symptoms of CDI in combination with a positive test for *C. difficile* toxins or toxigenic *C. difficile,* in Dutch participating hospitals, are registered. The indication for testing on CDI and the assay or algorithm that is used to diagnose CDI is chosen by the local laboratory.

Using classification criteria based on expert opinion that were previously used (8), CDI is classified as severe if one or more of the following conditions were present; fever (temperature of 38°C or

higher) and leucocytosis (>15 × 10^9/L), diarrhoea with hypoalbuminemia (<20 g/L) and/or dehydration, pseudomembranous colitis and/or bloody diarrhoea. A complicated course is defined as the need for surgical procedure, admission to intensive care unit and/or mortality (CDI- or non-CDI-related) within 30 days after CDI diagnosis (8).

Our primary aim was to test the null hypothesis that RT023 causes the same proportion of severe CDI as non-hypervirulent ribotypes. Therefore, clinical characteristics and 30-day outcome of CDI episodes due to RT023 were compared to CDI episodes due to other ribotypes (excluding hypervirulent strains RT027 and RT078/126). Thereafter, the results of the RT023-group were compared to the results of 4 pre-specified groups; RT027 and RT078/126, which are well-known hypervirulent strains, and RT001 and RT014/020/295, which are non-hypervirulent strains that are common in the Netherlands. Each time, results of the RT023-group were compared with the results of one other group. Some ribotypes were merged into one group since they are hard to distinguish with PCR ribotyping. Further details are in the web-only Supplementary Material.

Data are presented as number of cases (percentage). Age is presented as media [first quartile, third quartile], because of the skewed distribution. Categorical variables were compared by a Pearson's Chi square test and numerical variables were compared by a Wilcoxon rank-sum test. To identify the effect of RT023 on CDI severity, a multivariable logistic regression analysis was performed with age and sex as covariates. A p-value of <0.05 was considered statistically significant. STATA SE version 12.1 statistical software (StataCorp, Texas, USA) was used for statistical analysis.

Ethics

This was an observational study, using data that are already collected in the Dutch national CDI surveillance. This national surveillance program exists since 2009 and collects microbiological and clinical data from all hospitalized patients with CDI in the participating hospitals in the Netherlands. The surveillance has been developed by our National Institute of Public Health. There were no

additional data or isolates/materials specifically for this study collected and no actions were requested from patients.

Whole-genome sequencing

CD305 genomic DNA was sequenced using 454 pyrosequencing (GS-FLX pyrosequencing) to generate 3 kb paired-end libraries and Illumina GAII paired-end libraries of 400 bp insert size and 108 bp read length. The resulting sequence was assembled using Newbler and Velvet and the assemblies were combined using Newbler (9, 10). CDS identification and annotation was generated using PROKKA (11) with a bespoke *C. difficile* library. The assembled and annotated genome is available at ERS2502454. For 79 study isolates genomic DNA libraries were created using a Nextera XT kit (Illumina, CA, USA) and data obtained using the MiSeq sequencing system (Illumina, CA, USA).

Whole-genome bioinformatics analysis

The sequence data were processed according to a standard protocol as previously described (12) (Detail in web-only Supplementary Materials). SNP loci were identified with a samtools Q-score >= 30, coverage >= 10 and 80% of contributing reads. Pipeline, phylogenetic and post-analyses were carried out using Perl, R and RAxML (13).

Glycoprotein detection

Glycosylated proteins were detected using Pierce™ Glycoprotein Staining Kit according to the manufacturer's instructions (Detail in web-only Supplementary Material).

RESULTS

CDI in hospitalised patients due to RT023 strains is severe comparable with RT027 and RT078 strains

Between May 2009 and February 2018, 5359 samples from hospitalised patients in twenty-four hospitals in the Netherlands were PCR-ribotyped within the context of the national *C. difficile* surveillance program. Clinical data were complete in 4387 cases. RT023 accounted for 141 cases of CDI, a mean proportion of 2.4% (95% CI 2.0-2.8), which remained consistent within the study period. Demographic data, clinical characteristics and 30-day outcome of patients with CDI due to RT023 were compared to data of five other pre-specified ribotype groups, shown in Table 1. There were no significant differences in age and sex between the RT023 group and the other groups, except for higher age in the RT001 group.

The primary question was whether CDI due to RT023 was more severe when compared to all non-hypervirulent ribotypes, which was confirmed by our results (p=0.000: 35% (27-44) vs 22% (21-23)), also after correcting for sex and age. No significant differences of severity were found when RT023 was compared to "hypervirulent" strains RT027 and RT078/126 (p=0.310 and p=0.065 respectively, RT023: 35% (27-44), RT027: 29% (20-38), RT078/126: 27% (24-31)), also not after correction for sex and age. Of note, bloody diarrhoea was more frequently reported in RT023 infections compared with RT078/126 infections (p=0.031), RT014/020/295 (p=0.036) and RT001 (p=0.037) (RT023: 8% (3-13), RT078/126, 4% (2-5), RT014/020/295 4% (3-5), RT001 4% (2-5)). When compared to non-hypervirulent RT001 and RT014/020/295 isolates, with or without correcting for sex and age, RT023 presented with significantly more severe symptoms (p=0.000 for both, RT023: 35% (27-44), RT001: 16% (13-19), RT014/020/295: 21% (18-23)), such as more frequent diarrhoea with dehydration and/or hypoalbuminemia. However, the outcomes of CDI due to RT023 in terms of a complicated course, including mortality, were comparable with outcomes of CDI due to RT001, RT014/020/295 and all non-hypervirulent ribotypes. RT027 and RT078/126 infections showed higher overall mortality than RT023 (p=0.032, p=0.049 respectively, RT023: 9% (4-14), RT027: 19% (11-27),

RT078/126: 16% (13-19)) but CDI attributable mortality was similar between these groups (p=0.293, p=0.152 respectively, RT023: 2%(-1-4), RT027: 4% (0-8), RT078/126: 5%(3-7)). There were significantly more complicated courses in patients with CDI due to RT027 compared to RT023 (p=0.038, 23% (14-31) vs 12% (6-18) respectively), but no significant differences were observed between RT078/126 and RT023 (p=0.144, RT078/126 17% (14-20)).

Comparison of RT023 with all groups in this study revealed that the onset of symptoms of CDI due to RT023 was more frequently at home and less often in healthcare facilities (p=0.000 compared to all other groups). Subgroup analysis of community and hospital onset CDI can be found in the web-only Supplementary Material. The number of episodes that were recurrences of a previous CDI episode 2-8 weeks earlier was the same in RT023 episodes compared to all other groups (Table 1).

Clade 3 strains are highly related

A high-quality (14) draft genome of strain CD305 (RT023) was generated and is presented here as a robust reference for this lineage. Further strains were sourced from across Europe (Supplementary Table S1), with this study comprising 86 strains: CD305 (reference); 79 (out of 170 WGS strains); and 6 published clade 3 strains (15, 16) (Supplementary Table S2), the largest RT023 genomic collection. MLST were identified *in silico* from *de novo* assemblies. The six published strains matched their published MLST with new strains composed of 68 ST005, 10 ST022, and one novel ST (strain OUS23024) (Figure 1).

The 79 core strains were aligned to the CD305 reference strain and a set of 19,262 (<0.5% of the 4.2 Mbp genome) high quality SNP loci identified. The individual strains were very closely related with only between 58 and 7,876 pair-wise SNP differences, with a mean of 1,767 SNPs (mean: 9.2% of 19262 SNPs; max: 40.9%) equating to >99.8% pairwise identity between strains. A phylogeny was created from all 86 strain's SNPs that reinforces the conclusion of little genetic diversity within clade

3 strains (Figure 1). From our 80 strains there are two outliers: strains 91 and 108698, which are not RT023 (Figure 1A, Supplementary Figure S3, Supplementary Text). The unassigned MLST strain (OUS23024) diverged slightly from the main population (Figure 1B). No significant relationship was found with any phenotypes including the infection date (2007-2014) or geographic origin (Supplementary Table S1, Supplementary Figure S1). Detail on MLST and ribotype divergence can be found in the web-only Supplementary Material.

There is high conservation in all 86 strains of larger clade-specific genetic features such as the pathogenicity locus (PaLoc), binary toxin CDT, PaLoc phage insertionand type B flagella glycosylation cluster (Supplementary Tables S2 and S3). The only common antibiotic resistance marker is *gyrB* (V426D) related to fluoroquinolone resistance. Analysis of twelve Polish RT023 strains for fluoroquinolone resistance revealed resistance to ciprofloxacin but sensitivity to moxifloxacin (Supplementary Table S4).

A unique trehalose metabolism genotype is present in clade 3 strains

Analysis of clade 3 strains for two trehalose clusters described to be important in global dissemination and virulence of *C. difficile* (17) showed a trehalose genotype unique to these strains. The primary cluster, in which SNP L172I defines increased metabolism in RT027 (clade 2) (Figure 2a), was absent from all clade 3 genomes analysed. This coincides with polymorphisms and a large deletion in sugar metabolism genes in clade 3, including beta-glucosidase genes (Supplementary Text). However, the RT078 (clade 5) second cluster (Figure 2b) was observed in all strains. Polymorphisms exist between the RT078 cluster in M120 cluster and RT023 CD305, with the most significant difference being a truncation of *treX* (Figure 2c). Between clade 3 strains there are only a small number of SNPs, predominantly in strain 91 (Supplementary Text).

Clade 3 have a glycosylated surface

SlpA is the major surface protein of *C. difficile* comprised of high and low molecular proteins (HMW and LMW SLP) (18). A putative glycosylation cluster within the *slp* gene island (Figure 3a) for S-layer cassette type 11, SLCT11 (18), has been previously reported (19). 83 of the 86 strains contain this feature (Supplementary Table S2, Supplementary Figure S2). Strains 91, Ox2183 and WCHCD103 from which this feature is absent are genetically distinct from other strains within this clade, with alternate *slpA* genes. In RT023 the *slpA* gene encodes a smaller LMW SLP than in other clades, predicted at approximately 18 kDa (Figure 3b). S-layer extracts of representative strains from each of the five clades of *C. difficile* show two distinct bands of equimolar ratio representing the HMW and LMW SLPs in clades 1, 2, 4 and 5 by Coomassie brilliant blue staining (Figure 3c). Strain Ox247 (RT005, clade 1) containing SLCT11 (20) along with S-layer preparations from three representative RT023 strains show an alternative pattern of SLPs. HMW SLP migrates at its expected molecular weight, but a band at 18 kDa for LMW SLP is absent. A periodic acid-Schiff assay to stain for glycans on S-layer preparations showed glycosylated proteins at ~45 kDa only in strains containing the glycosylation cluster, demonstrating the presumed functionality of the cluster and glycosylation of S-layer proteins.

DISCUSSION

This study provides a comprehensive analysis of clade 3 strains of *C. difficile* with an extensive report of RT023 CDI and detailed WGS analysis. The clinical characteristics of hospitalised patients with CDI due to RT023 showed CDI severity similar to the “hypervirulent” RT027 and RT078/126, with comparable CDI-related mortality, though overall mortality was lower in RT023 as previously reported (6). The phylogeny of clade 3 strains is compact, barring six distinct outliers. In contrast to clade 2 strains (RT027), clade 3 strains show great similarity consistent with a recently emerged clade under little selective pressure to evolve (21). WGS analysis revealed a unique trehalose

genotype and conserved incorporation of a glycosylation cassette into the clade 3 genomes which was demonstrated to glycosylate the S-layer.

Considering previous investigations, the severity of disease is likely due to the production of binary toxin and the TcdC stop codon in RT023 (5). Recurrent infections due to RT023 were similar to other ribotypes. This contrasts with an earlier study, where RT023 was dominating among recurrent cases (22). We also observed more community acquisition of RT023 symptoms, but current reports cannot explain this observation. Circulating strains unlikely to be the source of RT023 with no representation of RT023 in a small group of *C. difficile* carriers (23) and a low representation in *C. difficile* infections in the community (24). The low proportion (2.4%) of CDI due to RT023 observed in this study in the Netherlands is consistent with a previous study on CDI in Europe (3).

Strengths of this study are the high sample size, multicenter design with high number of hospitals in different geographic regions, and 10 years of available data, making the data generalizable for hospitalized patients. Similarly, a sample size of over 80 strains across 8 years from a variety of pan-European sources for WGS, as well as published strains including Chinese strains, enabled us to understand the phylogeny of clade 3 in much greater detail. Limitations of the clinical data include the location of symptoms onset being documented but not the location of *C. difficile* acquisition. Furthermore, there was no data available regarding comorbidity, which might affect the outcome. Regarding severity of disease, occasionally not all laboratory parameters needing lab results were measured and included.

It has recently been shown that S-layer glycosylation is important for adherence to Caco-2 intestinal epithelial cells but not biofilm formation (20). Therefore, glycosylation of the S-layer in clade 3 may be important for colonisation but not persistence, explaining a low level of carriage and recurrence of these strains. Despite severe clinical presentation this clade is not as widely disseminated as other clades. The emergence of RT027 and RT078 strains has been linked to an increased ability to metabolise the food additive trehalose (17). RT023 strains contain the second four gene cluster,

corroborated by a recent study of trehalose genes in all clades of *C. difficile*. The presence of only the secondary cluster and the SNPs between RT023 and RT078 may result in a difference in uptake and metabolism of trehalose between these strains, which could explain the relatively reduced prevalence of RT023 strains compared with RT078 and RT027 strains globally. No link between trehalose and adverse disease outcomes has been suggested (25). Meanwhile, the emergence of epidemic clade 2 strains has also been linked to environmental spore contamination and the acquisition of fluoroquinolone resistance, which is less pronounced for clade 3 strains (21). More analysis on sporulation in clade 3 is required as reduced sporulation efficiency and survival outside the human host has been reported (26), however, a recent study highlighted a clade 3 strain in China which had a high sporulation and germination rate (27).

It remains to be determined why evolutionary distinct clades of *C. difficile* are emerging simultaneously to cause disease in human populations, or if *C. difficile* is evolving into subspecies (28). Our study suggests that a heightened awareness and continued surveillance of RT023 strains globally should be a current imperative.

DATA AVAILABILITY

Sequence data that supports the findings of this study have been deposited in EMBL Nucleotide Sequence Database (ENA) with accession code PRJEB26893 and CD305 reference genome ERS2502454.

TRANSPARENCY DECLARATION

The authors declare no conflicts of interest. The work was supported by The Wellcome Trust (Grant Reference 102979/Z/13/Z and 098051) and the Medical Research Council (Grant Reference MR/K000551/1).

ACKNOWLEDGEMENTS

We thank Ed Kuijper, Andre Ingebretsen, Hanna Pituch, Munir Primohammed, Paul Roberts (Royal Liverpool Hospital) and Neil Fairweather for the supply of strains to this study. We acknowledge the London *C. difficile* Ribotyping Laboratory for help with PCR ribotyping and supply of strains and Dr. Piotr Obuch-Woszczatynski for assistance with Polish RT023 strains. We thank all laboratories for helping to collect the data for the National Surveillance program in the Netherlands.

AUTHOR CONTRIBUTIONS

Concept and design of study: H.A.S., M.D.C. and B.W.W. Genomic assembly and annotation: M.D.P., H.P.B. and R.A.S. Genomic analysis: H.A.S. and M.D.P. Phenotypic experiments: H.A.S. Fluoroquinolone testing: H.P. Clinical analysis: K.E.W.V., M.J.T.C and E.J.K. The manuscript was drafted by H.A.S, M.D.P., K.E.W.V. and B.W.W., and revised by all authors.

REFERENCES

1. Khanna S, Pardi DS, Aronson SL, Kammer PP, Orenstein R, St Sauver JL, et al. The epidemiology of community-acquired Clostridium difficile infection: a population-based study. The American journal of gastroenterology. 2012;107(1):89-95.
2. Stabler RA, Dawson LF, Valiente E, Cairns MD, Martin MJ, Donahue EH, et al. Macro and micro diversity of Clostridium difficile isolates from diverse sources and geographical locations. PloS one. 2012;7(3):e31559.
3. Bauer MP, Notermans DW, van Benthem BH, Brazier JS, Wilcox MH, Rupnik M, et al. Clostridium difficile infection in Europe: a hospital-based survey. Lancet (London, England). 2011;377(9759):63-73.
4. Wilcox MH, Shetty N, Fawley WN, Shemko M, Coen P, Birtles A, et al. Changing epidemiology of Clostridium difficile infection following the introduction of a national ribotyping-based surveillance scheme in England. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2012;55(8):1056-63.
5. Dingle KE, Griffiths D, Didelot X, Evans J, Vaughan A, Kachrimanidou M, et al. Clinical Clostridium difficile: clonality and pathogenicity locus diversity. PloS one. 2011;6(5):e19993.

6. Walker AS, Eyre DW, Wyllie DH, Dingle KE, Griffiths D, Shine B, et al. Relationship between bacterial strain type, host biomarkers, and mortality in Clostridium difficile infection. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2013;56(11):1589-600.
7. Vanek J, Hill K, Collins J, Berrington A, Perry J, Inns T, et al. Epidemiological survey of Clostridium difficile ribotypes in the North East of England during an 18-month period. The Journal of hospital infection. 2012;81(3):209-12.
8. Goorhuis A, Bakker D, Corver J, Debast SB, Harmanus C, Notermans DW, et al. Emergence of Clostridium difficile infection due to a new hypervirulent strain, polymerase chain reaction ribotype 078. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2008;47(9):1162-70.
9. Bonfield JK, Smith K, Staden R. A new DNA sequence assembly program. Nucleic acids research. 1995;23(24):4992-9.
10. Zerbino DR. Using the Velvet de novo assembler for short-read sequencing technologies. Current protocols in bioinformatics. 2010;Chapter 11:Unit 11.5.
11. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics (Oxford, England). 2014;30(14):2068-9.
12. Cairns MD, Preston MD, Hall CL, Gerding DN, Hawkey PM, Kato H, et al. Comparative Genome Analysis and Global Phylogeny of the Toxin Variant Clostridium difficile PCR Ribotype 017 Reveals the Evolution of Two Independent Sublineages. Journal of clinical microbiology. 2017;55(3):865-76.
13. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics (Oxford, England). 2014;30(9):1312-3.
14. Chain PS, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, et al. Genomics. Genome project standards in a new era of sequencing. Science (New York, NY). 2009;326(5950):236-7.
15. Dingle KE, Elliott B, Robinson E, Griffiths D, Eyre DW, Stoesser N, et al. Evolutionary history of the Clostridium difficile pathogenicity locus. Genome biology and evolution. 2014;6(1):36-52.
16. Chen R, Feng Y, Wang X, Yang J, Zhang X, Lu X, et al. Whole genome sequences of three Clade 3 Clostridium difficile strains carrying binary toxin genes in China. Scientific reports. 2017;7:43555.
17. Collins J, Robinson C, Danhof H, Knetsch CW, van Leeuwen HC, Lawley TD, et al. Dietary trehalose enhances virulence of epidemic Clostridium difficile. Nature. 2018.
18. Kirk JA, Banerji O, Fagan RP. Characteristics of the Clostridium difficile cell envelope and its importance in therapeutics. Microbial biotechnology. 2017;10(1):76-90.
19. Dingle KE, Didelot X, Ansari MA, Eyre DW, Vaughan A, Griffiths D, et al. Recombinational switching of the Clostridium difficile S-layer and a novel glycosylation gene cluster revealed by large-scale whole-genome sequencing. The Journal of infectious diseases. 2013;207(4):675-86.
20. Richards E, Bouche L, Panico M, Arbeloa A, Vinogradov E, Morris H, et al. The S-layer protein of a Clostridium difficile SLCT-11 strain displays a complex glycan required for normal cell growth and morphology. The Journal of biological chemistry. 2018;293(47):18123-37.
21. He M, Miyajima F, Roberts P, Ellison L, Pickard DJ, Martin MJ, et al. Emergence and global spread of epidemic healthcare-associated Clostridium difficile. Nature genetics. 2013;45(1):109-13.
22. Sandell S, Rashid MU, Jorup-Ronstrom C, Ellstrom K, Nord CE, Weintraub A. Clostridium difficile recurrences in Stockholm. Anaerobe. 2016;38:97-102.
23. Zomer TP, E VAND, Wielders CCH, Veenman C, Hengeveld P, W VDH, et al. Prevalence and risk factors for colonization of Clostridium difficile among adults living near livestock farms in the Netherlands. Epidemiology and infection. 2017;145(13):2745-9.
24. Hensgens MP, Dekkers OM, Demeulemeester A, Buiting AG, Bloembergen P, van Benthem BH, et al. Diarrhoea in general practice: when should a Clostridium difficile infection be considered?

Results of a nested case-control study. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases. 2014;20(12):O1067-74.
25. Eyre DW, Didelot X, Buckley AM, Freeman J, Moura IB, Crook DW, et al. Clostridium difficile trehalose metabolism variants are common and not associated with adverse patient outcomes when variably present in the same lineage. EBioMedicine. 2019;43:347-55.
26. Connor M, Flynn PB, Fairley DJ, Marks N, Manesiotis P, Graham WG, et al. Evolutionary clade affects resistance of Clostridium difficile spores to Cold Atmospheric Plasma. Scientific reports. 2017;7:41814.
27. Li C, Harmanus C, Zhu D, Meng X, Wang S, Duan J, et al. Characterization of the virulence of a non-RT027, non-RT078 and binary toxin-positive Clostridium difficile strain associated with severe diarrhea. Emerg Microbes Infect. 2018;7(1):211.
28. Kumar N, Browne HP, Viciani E, Forster SC, Clare S, Harcourt K, et al. Adaptation of host transmission cycle during Clostridium difficile speciation. Nature genetics. 2019.

FIGURE LEGENDS

Figure 1: Phylogenetic Tree by MLST

Phylogenetic tree of 86 strains generated from analysis of high-quality SNPs and coloured by MLST. A: full tree, with two cohort outliers (samples 91 and 108676), Ox2183 and three Chinese strains. B: the large, temporally indistinguishable main cluster, with reference CD305 and novel MLST strain OUS23024 indicated.

Figure 2: Clade 3 show a unique trehalose genotype

Schematic demonstrating the three trehalose metabolism genotypes observed in *C. difficile* with clade 3 strains lacking the primary trehalose metabolism cluster. A: RT012 630 and RT027 R20291 genotypes of a primary trehalose cluster, with the L172I SNP associated with increased metabolism of trehalose. B: RT078 M120 genotype with primary and secondary trehalose metabolism gene clusters observed. C: RT023 CD305 trehalose genotype with only the secondary cluster including a truncated *treX* gene.

Figure 3: Insertion of a glycosylation cluster results in S-layer glycosylation

RT023 contains a glycosylation cluster within the *slp* gene island. A: Genomic organisation of the *slp* gene island in 630 (Clade 1) and CD305 (Clade 3) showing loss of Cwp2 and acquisition of a gene cluster comprising putative glycosylation genes (adapted from Kirk *et al* (18)). B: Structure of SlpA in 630 and CD305 showing Cwp84 cleavage sites and truncated LMW (light grey) in CD305. C: Coomassie staining of S layer protein preparations from representative strains from each clade showing characteristic double banding for HMW and LMW SLP (grey arrows). D: Periodic acid-Schiff staining of glycans in S layer preparations.

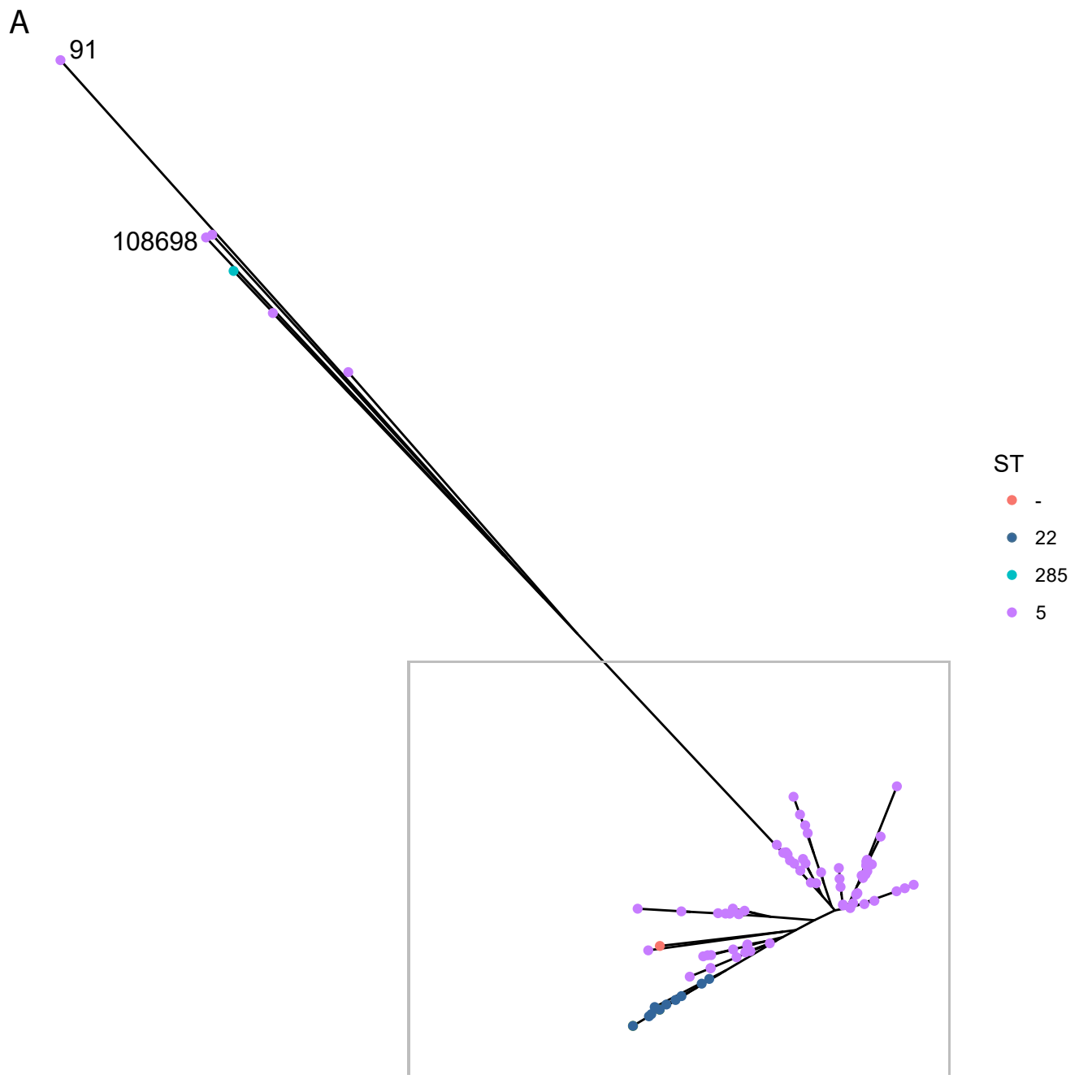
407

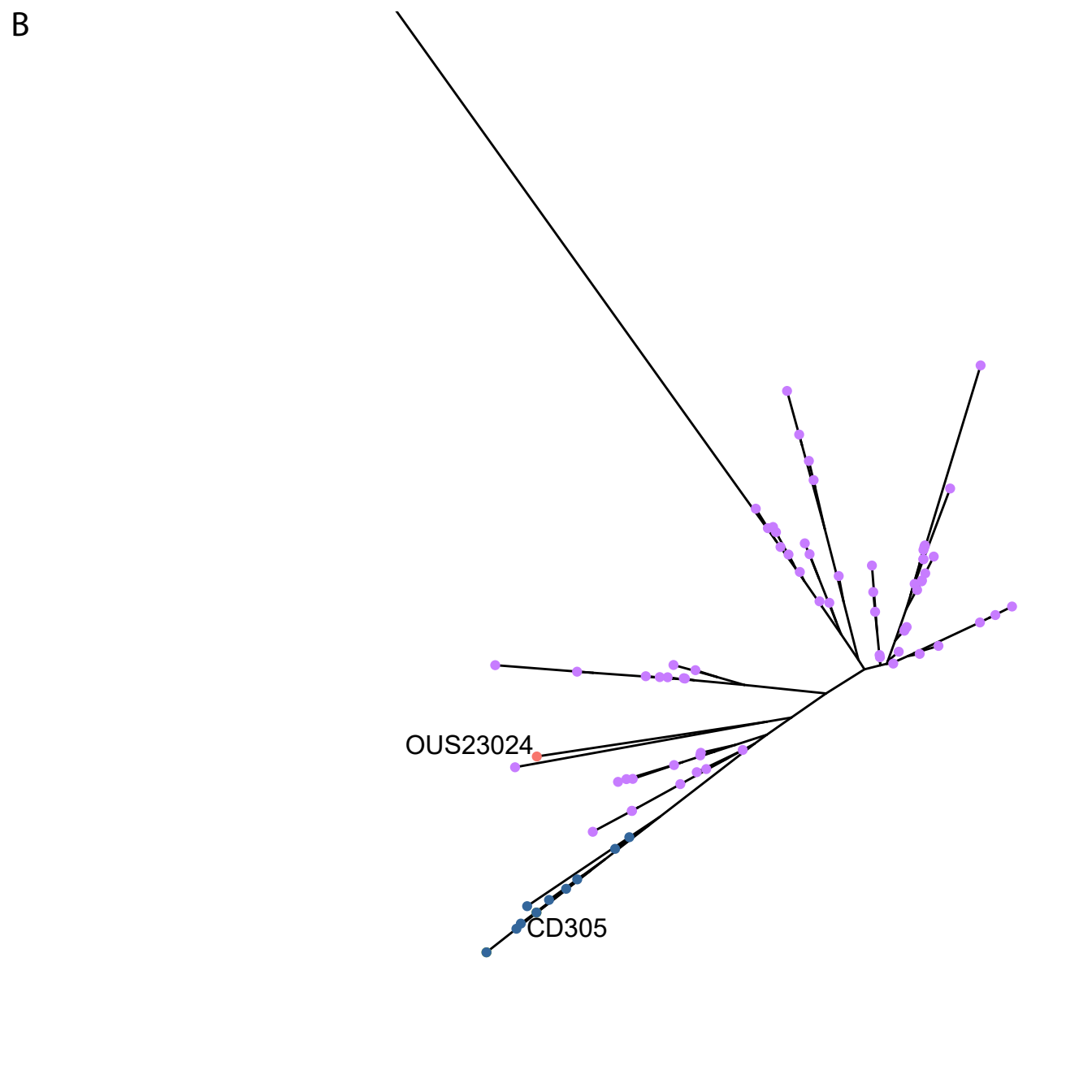
Table 1. Comparison of clinical characteristics of patients with RT023 versus other ribotypes (excluding RT027 and RT078/126), RT027, RT078/126, RT014/020/295 and RT001.

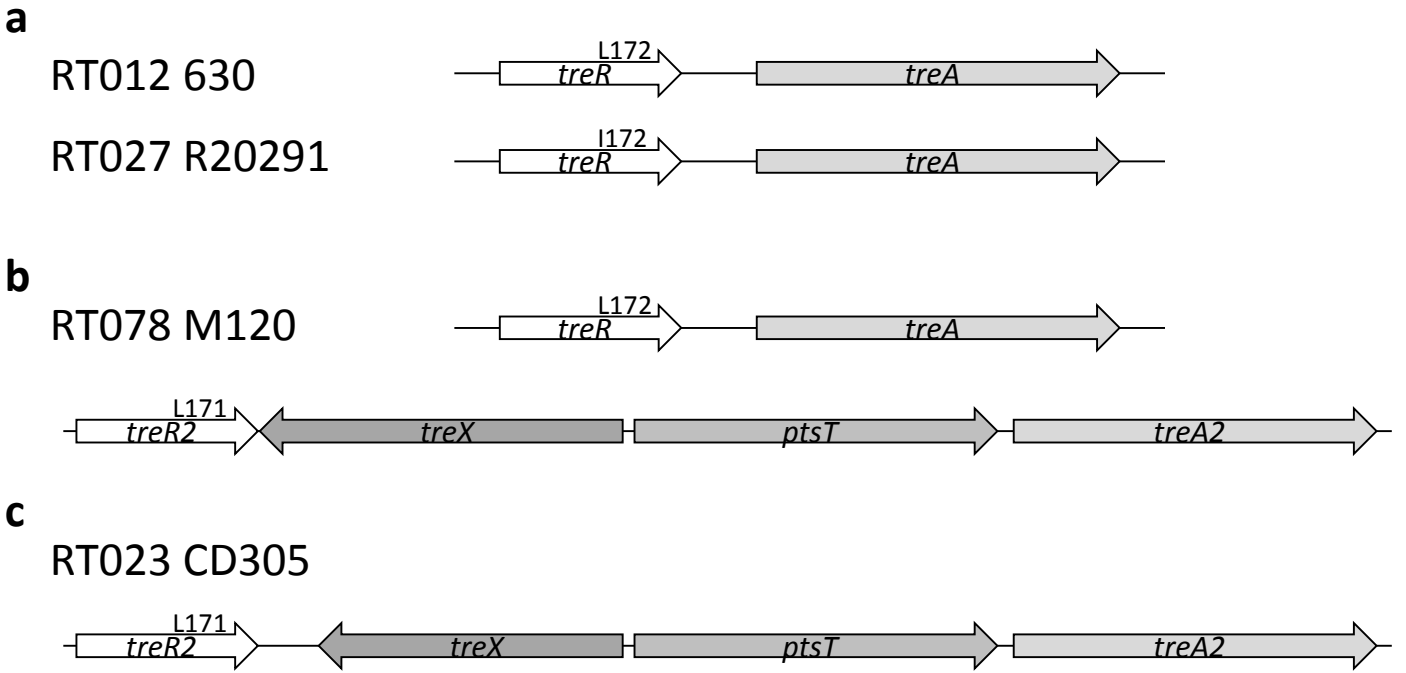
		Primary outcome	Hypervirulent strains		Non-hypervirulent strains		
	RT023, n=141	**Others, n=4368**	**RT027, n=116**	**RT078/126, n=734**	**RT014/020/295, n=962**	**RT001, n=699**	**All info available**
Age	71.4 [10.0, 97.7]	71.3 [1.9, 102.3]	73.2 [11.2, 91.5]	70.9 [5.2, 100.7]	70.4 [2.1, 99.2]	76.0 [3.3, 96.7]*	5359/5359
Men	71 (50)	2095 (48)	63 (54)	365 (50)	444 (46)	344 (49)	5356/5359
Severe CDI	45 (35)	880 (22)*	30 (29)	188 (27)	185 (21)*	104 (16)*	4948/5359
Dehydration and/or hypoalbuminemia	25 (20)	450 (11)*	14 (14)	100 (15)	97 (11)*	44 (7)*	4940/5359
Bloody diarrhoea	10 (8)	192 (5)	6 (6)	25 (4)*	34 (4)*	24 (4)*	4948/5359
Pseudomembranous colitis	8 (6)	159 (4)	6 (6)	41 (6)	28 (3)	21 (3)	4948/5359
Fever and leucocytosis	11 (9)	295 (7)	9 (9)	76 (11)	64 (7)	36 (6)	4940/5359
Complicated course	13 (12)	485 (14)	21 (23)*	104 (17)	78 (10)	95 (17)	4387/5359
Overall mortality	10 (9)	428 (12)	18 (19)*	98 (16)*	68 (9)	86 (15)	4387/5359
CDI mortality	2 (2)	104 (3)	4 (4)	29 (5)	16 (2)	27 (5)	4387/5359
Community onset	75 (54)	1545 (36)*	31 (27)*	272 (37)*	356 (38)*	155 (23)*	5283/5359
CDI last 8 weeks	22 (27)	684 (25)	12 (20)	133 (29)	161 (27)	115 (25)	3312/5359

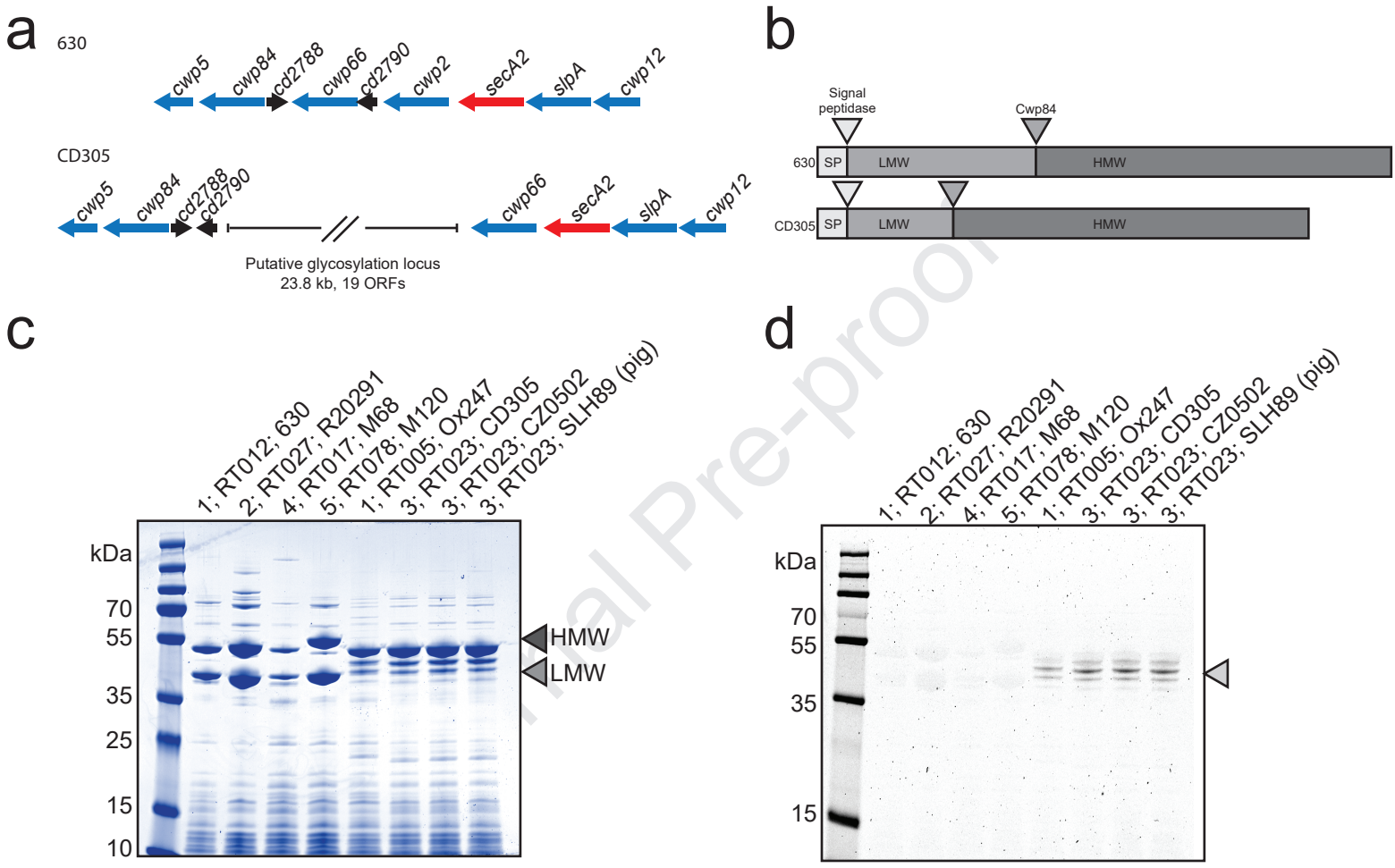
408 Data are presented as number of cases (percentage). Age is presented as median [first quartile, third quartile], because of the skewed distribution. Categorical variables were
409 compared by a Pearson's Chi square test and numerical variables were compared by a Wilcoxon rank-sum test. An asterisk (*) represents a *p*-value<0.05, when comparing
410 with RT023. *Abbreviations: LTCF: longtermcare facility, HCF: healthcare facility, RT: ribotype, CDI: Clostridium difficile infection*

411


A
91
108698
ST
-
22
285
5


B
OUS23024
CD305


a
RT012 630
L172
treR
treA
RT027 R20291
I172
treR
treA
b
RT078 M120
L172
treR
treA
L171
treR2
treX
ptsT
treA2
c
RT023 CD305
L171
treR2
treX
ptsT
treA2


a
630
cwp5
cwp84
cd2788
cwp66
cd2790
cwp2
secA2
slpA
cwp12
CD305
cwp5
cwp84
cd2788
cd2790
cwp66
secA2
slpA
cwp12
Putative glycosylation locus
23.8 kb, 19 ORFs
b
Signal peptidase
Cwp84
630
SP
LMW
HMW
CD305
SP
LMW
HMW
c
1; RT012; 630
2; RT027; R20291
4; RT017; M68
5; RT078; M120
1; RT005; Ox247
3; RT023; CD305
3; RT023; CZ0502
3; RT023; SLH89 (pig)
kDa
70
55
35
25
15
10
HMW
LMW
d
1; RT012; 630
2; RT027; R20291
4; RT017; M68
5; RT078; M120
1; RT005; Ox247
3; RT023; CD305
3; RT023; CZ0502
3; RT023; SLH89 (pig)
kDa
70
55
35
15