RESEARCH ARTICLE

Complete genome sequence of the *Clostridium difficile* laboratory strain 630Δerm reveals differences from strain 630, including translocation of the mobile element CTn5

Erika van Eijk1†, Seyed Yahya Anvar23†, Hilary P Browne4, Wai Yi Leung5, Jeroen Frank3, Arnoud M Schmitz3, Adam P Roberts6 and Wiep Klaas Smits1*

**Abstract**

**Background:** *Clostridium difficile* strain 630Δerm is a spontaneous erythromycin sensitive derivative of the reference strain 630 obtained by serial passaging in antibiotic-free media. It is widely used as a defined and tractable *C. difficile* strain. Though largely similar to the ancestral strain, it demonstrates phenotypic differences that might be the result of underlying genetic changes. Here, we performed a de novo assembly based on single-molecule real-time sequencing and an analysis of major methylation patterns.

**Results:** In addition to single nucleotide polymorphisms and various indels, we found that the mobile element CTn5 is present in the gene encoding the methyltransferase rumA rather than adhesin CD1844 where it is located in the reference strain.

**Conclusions:** Together, the genetic features identified in this study may help to explain at least part of the phenotypic differences. The annotated genome sequence of this lab strain, including the first analysis of major methylation patterns, will be a valuable resource for genetic research on *C. difficile*.

**Keywords:** Genome sequence, Conjugative transposon, Integrative and conjugative element, Single-molecule real-time sequencing

**Background**

*Clostridium difficile* is a Gram-positive, anaerobic bacterium that can asymptotically colonize the intestine of humans and other mammals. It was originally identified as part of the intestinal microbiota of healthy infants [1]. However, when the normal flora is disturbed – for instance as a result of antibiotic treatment – *C. difficile* can overgrow and cause potentially fatal disease [2,3]. The main virulence factors are toxins A and B, that are encoded on a chromosomal region called the pathogenicity locus (PaLoc) [4], but other factors are also likely to play a role [5]. Recent years have seen an increase in the incidence and severity of *C. difficile* infections, for reasons that are only partially understood [6,7].

In 2006, the first genome sequence of a *C. difficile* strain was published [8]. This multi-resistant strain, designated 630, was isolated from a patient with severe pseudomembranous colitis and caused an outbreak of diarrheal disease in a Swiss hospital [9]. Analysis of the 630 genome sequence revealed that approximately 11% consists of mobile genetic elements [8]. The majority of these elements are conjugative transposons of the Tn916 and Tn1549 families called CTns, which have the ability to excise from their genomic target sites and transpose intra- or intercellularly [8,10]. Exchange of mobile elements occurs frequently and contributes to the plasticity of the genome of *C. difficile* [8,11,12]. Functions encoded on conjugative transposons can contribute to environmental adaptation and antimicrobial resistance [10,13]. In *C.
difficile, transfer of the conjugative elements CTn1, CTn2, CTn4, CT5 and CTn7 from strain 630 into a non-toxogenic strain has been shown [10]. Transfer of CTn3 (Tn5397), harboring a tetracycline resistance gene, has been demonstrated between species [14,15]. CTn1, CTn3, CTn6 and CTn7 are related to Tn916, based on their conjugation module [8,13]. CTn2, CTn4 and CTn5 are all part of the Tn1549 family, based on DNA sequence homology, and their accessory modules code for uncharacterized ABC-transporters [8,10]. Recently it has been shown that these CTn’s may also be responsible for transfer of the PaLoc on large chromosomal fragments [16].

After the demonstration of conjugative transfer from DNA from Escherichia coli to C. difficile [17], genetic tools were developed for C. difficile. To facilitate the genetic manipulation, an erythromycin sensitive variant was derived from strain 630 by serial passaging [18]. This strain is particularly useful for generation of insertionally mutants using CloxTron that employs a retrotransposition activated erythromycin resistance marker (ermRAM) [19]). Recently, allelic exchange methods have been developed for C. difficile [20,21]. The efficiency of both methods depends on the accuracy of the genome sequence for selection of target sites and recombination events. However, no comprehensive mapping of differences between the lab- and reference strains has been published to date.

The most notable phenotypic difference between 630 and 630Δerm, erythromycin resistance, was found to be the result of a 2.4 kb deletion in the mobile genetic element Tn5398 that eliminates an ermB gene [18]. This explains at least in part the different behavior of the two strains in a Golden Syrian hamster model of acute disease [22], as animals are generally sensitized to C. difficile with a clindamycin treatment (ermB is an rRNA adenine N-6-methyltransferase that also confers resistance to clindamycin). At a genetic level, another difference between the two strains reported to date is a duplication in the master regulator of sporulation, spo0A, that is apparently without phenotypic consequences [23].

In another Gram-positive bacterium, Bacillus subtilis, phenotypic differences between the ancestral strain NCIB3610 and widely used laboratory strains have been linked to specific genetic differences [24-26]. A detailed map of the genetic differences between the C. difficile strains 630 and 630Δerm could therefore not only facilitate genetic manipulation, but also form the basis for the investigation of phenotypic differences between these strains.

**Results and discussion**

**Reference assembly of the 630Δerm genome reveals four breakpoints**

We set out to investigate differences between the laboratory strain 630Δerm and reference strain 630 by performing short-read next generation sequencing on the Illumina HiSeq platform. Based on the report that the erythromycin sensitivity of strain 630Δerm is due to a 2.4 kb deletion in Tn5398, we examined this region of the reference alignment. The analysis revealed the absence of reads mapping to the CD2007A and CD2008 genes which are located in the expected deletion [18]. Reads that mapped to CD2007 (erm2(B)/ermB1), the main erythromycin resistance determinant in strain 630 [18] are likely due to the fact that this gene shares 100% nucleotide identity with CD2010 (erm1(B)/ermB), which is still present. This is supported by the observation that the coverage of both these genes is approximately 2-fold lower than the immediate surrounding regions (Figure 1A). Notably, the reference assembly failed to identify the previously identified duplication in spo0A [23] (data not shown).

A further analysis of the reference assembly against a linearized 630 genome revealed four breakpoints (regions with discordantly mapped read-pairs). The first breakpoint is consistent with a deletion of ~70 bp. The remaining breakpoints are consistent with a transposition event, in which the transposed sequence is reinserted elsewhere in the genome and in the inverse orientation compared to the reference (Figure 1B).

**De novo assembly of the 630Δerm genome using third generation sequencing**

Based on the identification of a potential transposition event, and our previous finding that indels may have occurred that are difficult to detect using short reads, we decided to perform an unbiased, de novo, assembly of the 630Δerm genome using single-molecule real-time sequencing. The Pacific Biosciences RSII system is capable of generating large reads, and with sufficient coverage, can generate high quality single contigs for bacterial genome sequences. We sequenced a genomic library of strain 630Δerm on two SMRT cells, and validated the resulting single contig with a third SMRT cell. The resulting genome consists of 4,293,049 basepairs, with an average GC content of 29.08% and an estimated coverage of 158× (Figure 2A). We generated an annotated version of this genome by transferring the most recent version of the 630 annotation [EMBL:AM180355] [27], updating it with recent gene annotations from literature and incorporating qualifiers in the file to indicate specific features of 630Δerm. The annotated sequence has been deposited under accession number EMBL:LN614756.

Satisfyingly, our unbiased approach identified the 18-bp duplication in the spo0A gene, encoding the master regulator of sporulation, which we previously found [23] (Figure 2B). This demonstrates that the third generation sequencing approach is superior to Illumina in identifying this type of difference. In addition, we could confirm
the expected 2.4 kb deletion in Tn5398 (Figure 2C). The sequence of Tn5398ΔE, which we determined, shows 4 Single Nucleotide Polymorphisms (SNPs) compared to an in silico generated theoretical sequence of Tn5398ΔE (based on Hussain et al.) [18]. As a result of these differences, a progressive MAUVE [28] alignment of the Tn5398ΔE element from our strain with Tn5398 of strain 630 demonstrates the deletion of CD2010 (ermB1/erm(1)B), CD2009A (ORF3), CD2009 (fragment of a putative topoisomerase), CD2008 (ORF298) and most of CD2007A. This effectively removes the region between the two copies of ermB. The most likely scenario by which this occurred is through recombination between the two ermB genes or their immediate surrounding region; the sequence information is unable to determine the exact site of recombination, as these regions are identical, and the copies of ermB and ORF3 in 630Δerm may therefore represent hybrids of CD2007/CD2010 or CD2006A/CD2009A, respectively. To reflect the results of the alignment as well as the mechanism described above, we have chosen to rename the ermB gene of strain 630Δerm CD2007B/ermB (locustag: CD630Derm_20072) and ORF3 as CD2006B (locustag: CD630Derm_20062). The resulting arrangement suggests that CD2007B is potentially expressed, as it is fused to the promoter region of CD2010/ermB1 at the exact same location, though the strain remains erythromycin sensitive. This discrepancy has been noted since the isolation of 630Δerm [18], and cannot be resolved using the sequence information from our study.

We also identified short tandem repeats (≥90% nucleotide identity) up to 500 bp. Strikingly, the genome analysis revealed two regions of high repeat density (Figure 2A). The first region (approximately 0.6 Mb-0.9 Mb) includes the PaLoc that encodes toxins A and B. This region was found to be capable of transfer by a conjugation-like mechanism [16] and it is tempting to speculate that the high repeat density may contribute to this phenomenon. The second region (approximately 3.6 Mb-3.75 Mb) contains many genes involved in sugar metabolism, but does not seem to be associated with annotated or characterized mobile elements. Large repeats (≥95% identity and >500 bp in length) generally coincide with regions of high-GC content, and mainly reflect ribosomal gene clusters.

Analysis of m6A and m4C methylation patterns of C. difficile

In bacteria, post-replicative addition of a methyl group to a base by a DNA methyltransferase can result in the formation of N6-methyladenine (m6A), C5-methylcytosine (m5C) and N4-methylcytosine (m4C) [29,30]. These modified bases play a role in restriction/modification systems, or may regulate cellular processes (reviewed in [30-33]).

There is little information on methylation of chromosomal DNA in C. difficile. Five methylases have been identified in C. difficile 630 [34], but in vivo methylation patterns have not been characterized. We took advantage of the pulse profiles of the Pacific Biosciences RSII reads that hold information about base modifications [35,36] to generate the first comprehensive analysis of methylation patterns in C. difficile (Figure 3A).

m6A modifications can be identified with high confidence and the vast majority of the modifications (7288/7687 = 95%) were associated with the motif CAAAAA, in which the last adenine residue is
modified (Figure 3B). Previous studies identified a single methylase, M.Cdi25 (corresponding to CD2758) with homology to adenine specific methylases, but failed to identify its target site in restriction protection experiments [34]. We postulate that CD2758 recognizes and methylates last adenine residue the CAAAAA motif and that this is possibly the only adenine-methylase in C. difficile 630Δerm.

The pulse profiles of the Pacific Biosciences RSII reads also identify modified cytosines. Only a fraction of these are positively identified as $m^4$C, in part due effect of
modifications that are in close proximity to each other on the pulse profiles [36,37]. We did not further investigate m5C modifications, as they can only reliably be detected on the Pacific Biosciences platform after Tet1-treatment, by preparation of shorter library fragments that are not ideal for genome de novo assembly, and with much higher coverage than obtained in our experiment [36]. Unspecified modifications may therefore represent m4C, and possibly m5C or other modifications.

The SMRT Portal identified the motif GCGAGC, in which the first cytosine residue is modified, as overrepresented in the methylcytosine dataset (Figure 3B). This motif is remarkably similar to the GCWGC motif identified for the M. Cdi1226 methylase (CD3147) [34].

We could identify 146 instances of m4C methylation and 16 of those contained the motif (11%). When a DREME search was performed [38] using 41 bp sequences centered on m4C only, a highly similar motif (GCGAGC) was found in 33 instances. Moreover, none of the other motifs (see below) were specifically linked to m4C modifications, suggesting that many if not all of the m4C modifications are due to CD3147.

m4C and m6A methylations that were not associated with the overrepresented motifs seemed to correspond to regions of high GC-content, including the mobile elements CTnI, CTn2 and CTn4 (Figure 3).

We also evaluated motifs previously identified as putative target sites for the other three cytosine specific methylases of C. difficile, M. Cdi633 (CD0935), M. Cdi587 (CD0927) and M. Cdi824 (CD1109) [34]. CD0935 conferred partial protection against digestion with Bal31 (target site: TGGCCA). Our data did not show any modifications on cytosine or adenine residues of this motif anywhere in the genome (n = 396). Considering that we cannot reliably detect m5C modifications in our setup, it is possible that M. Cdi633 is an m5C specific methylase. CD0927 could confer protection against Sau96I (target site: GGNCC) in E. coli, but C. difficile chromosomal DNA is only partially resistant to Sau96I digestion [34]. We found only very low levels (~0.1%) of modified cytosines for this motif (n = 3824) in 630Δerm, which together with the earlier observations suggests that CD0927 is either minor m4C or a m3C methylase. CD1109 conferred protection against Smal (which recognizes CCCGGG). We found that 6/60 (10%) of the motifs contained a modified cytosine at the third position. These modifications are likely m4Cs that cannot be positively identified as m4C due to adjacent modified bases.
C. difficile chromosomal DNA is wholly resistant to Tsel (target site: GCWGCG) and Smal (target site (CCCGGG), though we only detected modifications for ~10% of the occurrences of these motifs. This may be due to only a fraction of the methyllysosine modifications being called by the Pacific Biosciences SMRT platform in our analyses.

The function of the methylases of C. difficile is unknown. None seem associated with an endonuclease, indicating they are not likely to be part of a restriction-modification system. Consistent with this, no effect on conjugation efficiency was observed [34]. CD0927 and CD0935 are part of prophage 1, and CD1109 is present on the CTnΔ element, suggesting they may play a role in the biology of mobile elements.

Comparison of the complete genome of 630Δerm with strain 630 reveals SNPs, indels and rearrangements

It is likely that more than the two previously identified differences (Δerm deletion and 18 bp duplication in spo0A) exist between strain 630 and strain 630Δerm. We therefore compared our de novo assembled genome to the reference sequence.

We identified 71 differences between the two strains. These encompass 8 deletions (including the Δerm mutation) [18], 10 insertions (including the duplication in spo0A) [23], 2 insertion-deletions, 50 substitutions and 1 region of complex structural variation (Additional file 1).

Of these, 23 were located intergenically. This includes a 102 bp deletion which likely corresponds to the break-point at 0.68 Mb identified in the short read next generation sequencing (Figure 1B). A complete list of identified structural variants is available as (Additional file 1).

23 of the identified differences are associated with rRNA sequences. We found that strain 630Δerm has acquired an extra ~5 kb rRNA/tRNA cluster that is inserted between CD0011 and CD0012 compared to strain 630 (Table 1, Figure 4). Copy number variations in rRNA operons have previously been noted for C. difficile [39] and may reflect an adaptation to favorable growth conditions in the laboratory. Similar to rRNA operon 6, this operon contains tRNA^{Leu} and tRNA^{Met} genes downstream of the 23S rRNA gene, but the intergenic spacer region (ISR) between the 16S and 23S rRNA genes does not contain a tRNA^{Ala}. A detailed comparison of the ISRs of the different rRNA operons is provided as Additional file 2. A striking number of differences were found in rRNA operon 11 (Figure 4). As observed previously [40], the sequence variations cluster in the 3' region of the 16S rRNA and 5' of the 23S rRNA genes.

We focused our further analysis on the 26 variants that are associated with annotated pseudogenes or open reading frames (Table 1). A 24 bp deletion in CD0632, a conserved protein of unknown function, shortens the arginine-alanine repeat in this protein by 8 amino acids. In two cases, a single basepair insertion restores a pseudogene (CD1388 and CD3156A). This was confirmed by assembling the short read Illumina sequences against both the 630 reference genome and the de novo assembled 630Δerm genome, as a variant was identified in the former but not the latter. CD1388 encodes a putative regulatory protein with a helix-turn-helix motif and CD3156A a conserved protein of unknown function. Interestingly, both proteins encoded by these genes were previously identified in a proteomic analysis [27], indicating that they are expressed in strain 630Δerm. Two in-frame insertions were identified (an extra alanine residue in CD0514 and the published duplication in spo0A/CD1214). Out of 18 identified nucleotide substitutions, 9 were synonymous. These include SNPs in the encoded elongation factor Tu (tufl/CD0058), ribosomal protein L50 (rplC/CD0073) and the putative aminotransferase CD2532. Strikingly, the CD0514 gene, encoding the cell wall protein cwpV [41,42], contains an unusually high density of mutations. In addition to the insertion and 5 synonymous mutations, it contains 2 non-synonymous but conservative mutations.

Other non-synonymous mutations are located in the putative ferric uptake regulator CD0826, the putative acyl-CoA N-acyltransferase CD1190, predicted glyceraldehyde-phosphate dehydrogenase CD1767 (gapB), ethanolamine utilization protein CD1907 (eutG), the hypothetical protein CD2627, the phosphotransferase system protein CD2667 (ptsG-BC) and the transcriptional regulator CD3565. In all these cases, the de novo assembly of the 630Δerm genome was clearly supported by the short read Illumina data.

CTn5 is present in the rumA gene in both 630Δerm (LUMC) and 630Δerm(UCL)

In an attempt to visualize the proposed transposition event (Figure 1B), we generated a dotplot of the genome sequence of our strain versus the reference (Figure 5A). It is immediately evident that the CTn5 element seems to have excised from its original location in CD1844 (encoding a putative cell wall adhesin) and has inserted in an inverted manner in rumA (CD3393) in our isolate of 630Δerm, for clarity hereafter referred to as 630Δerm (LUMC).

To exclude that the finding represents a misassembly in the original 630 genome sequence, and confirm the presence of CTn5 in rumA in 630Δerm (LUMC), we performed various control PCRs (Figure 5B). In strain 630, we found CTn5 inserted in CD1844 and confirmed an intact rumA gene. In contrast, in 630Δerm (LUMC), we detected no product for the left and right junctions...
| AM183055 start | AM183055 end | 630Δerm start | 630Δerm end | Type             | Description                | Region       | Gene name | Function                                      | Details                  |
|----------------|--------------|---------------|--------------|------------------|----------------------------|--------------|-----------|----------------------------------------------|--------------------------|
| 84143          | 89438        | 89522         | 89522        | Substitution     | C > T                     | CD630_00580  | tuf1      | Elongation factor EFTu/EF1A                  | Synonymous               |
| 103225         | 108520       | 108522        | 108522       | Substitution     | G > T                     | CD630_00730  | rplC      | S05 ribosomal protein L3                    | Synonymous               |
| 610336         | 615631       | 615631        | 615631       | Substitution     | G > A                     | CD630_05140  | cwpV      | Cell surface protein Val > Ile              |                         |
| 610480         | 615775       | 615775        | 615775       | Substitution     | C > T                     | CD630_05140  | cwpV      | Cell surface protein Synonymous              |                         |
| 610563         | 615664       | 615664        | 615664       | Insertion        |                            | CD630_05140  | cwpV      | Cell surface protein In frame Ala insertion |                         |
| 610752         | 616050       | 616050        | 616050       | Substitution     | G > A                     | CD630_05140  | cwpV      | Cell surface protein Synonymous              |                         |
| 610840         | 616138       | 616138        | 616138       | Substitution     | C > T                     | CD630_05140  | cwpV      | Cell surface protein Synonymous              |                         |
| 610875         | 616173       | 616173        | 616173       | Substitution     | C > T                     | CD630_05140  | cwpV      | Cell surface protein Synonymous              |                         |
| 755776         | 755800       | 760995        | 760996       | Deletion         |                            | CD630_06320  | Conserved hypothetical protein               | In frame 8aa deletion in repeat region |
| 1000995        | 1006274      | 1006274       | 1006274      | Substitution     | A > G                     | CD630_08260  | Putative ferric-uptake regulator             | Thr > Ala               |
| 1391850        | 1397192      | 1397192       | 1397192      | Substitution     | T > C                     | CD630_11900  | Putative acyl-CoA N-acetyltransferase        | Phe > Leu               |
| 1413060        | 1413077      | 1413077       | 1413077      | Duplication      |                            | CD630_12140  | spo0A2      | Stage 0 sporulation protein A              | 6aa (NVGNE) duplication |
| 1607458        | 1607459      | 1607459       | 1607459      | Insertion        |                            | CD630_13880  | Putative transcriptional regulator          | Restores transcriptional regulator |
| 2044514        | 2049813      | 2049813       | 2049813      | Substitution     | C > G                     | CD630_17670  | gapB        | Glyeraldehyde-3-phosphate dehydrogenase GAPDH | Pro > Ala               |
| 2137467        | 2183040      | 2142764       | 2142764      | Deletion         |                            | CD630_18440  | Putative adhesin                           | Translocation of CTn5; CD1844 restored |
| 2209236        | 2169861      | 2169861       | 2169861      | Substitution     | G > A                     | CD630_19070  | eutG         | Ethanolamine iron-dependent Alcohol dehydrogenase | Gly > Glu               |
| 2924655        | 2881973      | 2881973       | 2881973      | Substitution     | C > T                     | CD630_25320  | Aminotransferase, alanine-glyoxylate transaminase | Synonymous               |
| 3034953        | 2992271      | 2992271       | 2992271      | Substitution     | C > A                     | CD630_26270  | Conserved hypothetical protein               | Gly > Cys               |
| 3080703        | 3038021      | 3038021       | 3038021      | Substitution     | C > T                     | CD630_26670  | ptsG-BC       | PTS system, glucose-specific IIBC component | Val > Ile               |
| 3686534        | 3686535      | 3643756       | 3643756      | Insertion        |                            | CD630_31561  | Conserved hypothetical protein               | Restores conserved hypothetical protein |
| 3967522        | 3967523      | 3924743       | 3970315      | Insertion        |                            | CD630_33930  | rumA          | 23S rRNA (uracil-5-)methyltransferase       | Translocation of CTn5; fuses rumA (CD3393) to CD1844A |||
| Chromosome | Position | Type | Change | Description | Gene | Function |
|------------|----------|------|--------|-------------|------|----------|
| 4166495    | 4169292  | Substitution | G > A | CD630_35650 | Transcriptional regulator, GntR family | Ala > Val |
| 12347      | 12348    | Insertion | 12347_12348ins | multiple | rRNA/tRNA cluster |
| 2317627    | 2320033  | Deletion | 2317633_2320041del | multiple | Loss of erythromycin resistance (Δerm) |

Table 1 Structural variants associated with coding sequences (Continued)
of CTn5 in CD1844/CD1878A, indicating that the element is not present at this location. We readily amplified fragments corresponding to the left and right junction of CTn5 when inserted in rumA in C. difficile 630Δerm (LUMC), but not 630, chromosomal DNA. Interestingly, we observed a faint band corresponding to intact rumA even in strain 630Δerm (LUMC). This indicates that a subpopulation of cells does not contain CTn5 at this location, either because it has not inserted yet, or retains the ability to excise spontaneously as previously observed for 630 [8].

The CTn5 insertion site identified here is located immediately downstream of CTn7. A similar tandem arrangement has previously been observed in two clinical PCR ribotype 001 isolates [10,43]. In another clinical isolate (RT027), which lacked a CTn7-like element, a CTn5-like element was found to be integrated at a site homologous to the target site of CTn7 in 630 [43].

The annotation of CD3393 as rumA in C. difficile is based on homology of the predicted protein to E. coli RmRA (also known as RlmD). This enzyme methylates a uracil nucleotide of the ribosomal RNA [44-46]. E. coli rumA mutants perform similarly compared with the wild type strain, in terms of cell growth, antibiotic resistance, and fidelity of translation. However, ΔrumA cells are outcompeted by wild type cells in growth competition assays, which may imply that ribosome function is moderately affected [46].

The translocation of CTn5 to rumA has two major consequences. First, the CD1844 gene, encoding a putative adhesin is restored. Second, the rumA open reading frame is fused to the CD1844A open reading frame resulting into a hybrid protein (CD3393A). CD1844A shows very high similarity (e-value 1e-62, 97% identity) to the C-terminus of an Enterococcus faecalis rumA homolog [EMBL:EOK00135.1]. However, the homology of C. difficile rumA to this gene is limited to the N-terminal TrmA-like domain (COG2265) (Figure 5B). Thus, a link between these open reading frames is also found in other organisms than C. difficile. In order to determine what the phenotypic consequences are of the transposition of CTn5 further experiments are required.

To further our understanding of the origin of the transposition event, we compared the location of CTn5 by PCR in different related strains; a non-passaged isolate of the original 630Δerm [18], hereafter referred to as 630Δerm (UCL), and another erythromycin sensitive derivative of 630, 630E/JIR8094 [47]. We found that in strain 630E the element is present in CD1844/CD1878A, identical to the reference strain, suggesting that the transposition event is not linked to the loss of erythromycin resistance. The 630Δerm (UCL) strain shows prominent bands corresponding to CTn5 at its CD1844/CD1878A location, but also a weak signal for CTn5 at rumA (Figure 5C). Therefore, this isolate likely contains a subpopulation of cells with the transposition identified in this study. It is possible that CTn5 is stable at either location and the stock of the 630Δerm (UCL) is non-clonal, or that CTn5 in 630Δerm (UCL) is highly mobile. During redistribution of the strain, isolates with either insertion could have been selected.

In summary, our data show that integration of CTn5 can occur in at least two different sites in the C. difficile 630Δerm genome, and that the element can switch between these locations during repeated passaging.
Figure 5 CTn5 is present in rumA in 630Δerm but not 630 or 630E. A. Dotplot of the reference sequence for C. difficile 630 (x-axis) versus the de novo assembled 630Δerm sequence (y-axis), indicating the location of CD1844 and rumA (boxed in blue), and the CTn5 element (boxed in red). Note the inverted orientation of the mobile element. B. Schematic representation of the rumA-CD1844A hybrid protein (CD3393A). C. PCR confirmation of the transposition event. For primers used see Methods and Table 2.
Conclusions

The work presented here provides the first reference genome for the widely used *C. difficile* laboratory strain 630Δerm, including the first analysis of major methylation patterns for any *C. difficile* strain. Our work reveals that in addition to insertion, deletions and SNPs, the CTn5 element has moved from its original location within CD1844 to the *runA* gene in our isolate. The observation of such a dramatic rearrangement has important implications for the redistribution of strains with highly mobile genomes and argues for complete resequencing of common lab strains in each laboratory.

Methods

Bacterial strains and growth conditions

Our isolate of strain 630Δerm was initially obtained from the Minton lab (University of Nottingham, Nottingham, UK), that in turn received it from the Mullany lab in which it was generated. For the purpose of resequencing the strain was cultured on preseeded CLO plates (Biomerieux), after which it was ented to BHI medium (Oxoid) supplemented with 0.5% yeast extract (Fluka).

Strain 630 was originally obtained from the Mastrantonio lab (Instituto Superiore di Sanità, Rome, Italy) and its use in our lab has been described before [48]. The 630Δerm strain from the Mullany lab (UCL Eastman Dental Institute, London, UK), 630Δerm(UCL), was transported as a glycerol stock on dry ice. Strain 630E was a kind gift of Robert Britton (Michigan State University, East Lansing, MI, USA). All strains were cultured as described for our isolate of strain 630Δerm, which is referred to as 630Δerm (LUMC) where appropriate.

Isolation of chromosomal DNA

For PCR analysis, chromosomal DNA was isolated using the QiaAmp Blood&Tissue kit (Qiagen) according to the manufacturer’s instructions from growth obtained after streaking out the strain directly from the glycerol stock onto CLO plates (Biomerieux). For SMRT sequencing, high molecular weight DNA was isolated from 30 mL of an overnight culture, using the Qiagen GenomicTip 500/ G, according to the manufacturer’s instructions. The quality of the DNA was checked on a Nanodrop ND-200 machine (ThermoFisher), the integrity by agarose gel electrophoresis, and the DNA was quantified on a Qubit instrument (Invitrogen).

Illumina sequencing and analysis

For Illumina sequencing, chromosomal DNA was isolated by Baseclear (Leiden, The Netherlands) from a pellet of bacterial cells derived from 50 mL culture. Data from 50 cycle 500 Mb paired-end read was delivered by Baseclear as 2 fastq files. Sequence reads have been deposited in the ENA Sequence Read Archive (EMBL: ERS550098). A preliminary analysis of the data was performed by aligning the paired-end reads to the reference genome of *C. difficile* strain 630 [GenBank:AM180355] using Geneious R7 (Biomatters, http://www.geneious.com). A more detailed analysis was performed using Stampy [49] and BWA [50]. In a routine quality control (QC) procedure on verifying the alignment, QC metrics including insert-sizes, mapped reads, unmapped reads and reads that align with a deviated pattern (DP; discordant read alignments) were examined. The case where a significant amount of reads cannot align to the reference genome indicates an undefined sequence region in strain 630Δerm or a contamination of the library. In our case, a few regions with discordantly mapped read pairs (DP > 9) were identified (Additional file 3) and validated automatically (Additional file 4). Of the validated breakpoints, the first has matches with the end of the reference assembly and is therefore an artefact of assembling the reads against a linearized genome. This was confirmed by artificially breaking the circular chromosome at a different position and repeating the procedure. Visual inspection in the Integrative Genome Viewer tool [51] on the alignment track (BAM file) was used to determine the nature of the Structural Variations).

Pacific biosciences RSII sequencing and de novo assembly

For single molecule real-time sequencing, a SMRTbell DNA template library with an insert size of ~20 kb was prepared according to the manufacturer’s specification. To this end, chromosomal DNA was fragmented with G-tubes (Covaris). Subsequently, fragmented DNA was end-repaired and ligated to hairpin adapters. SMRT sequencing was carried out on the Pacific Biosciences RSII machine according to standard protocols (Magbead loading, 1x180 min). Sequence reads have been deposited in the ENA Sequence Read Archive (EMBL: ERS550016). Sequencing reads were corrected using the Hgap pipeline [52]. Assembly was performed using Celera Assembler 8.1. We observed unbalanced coverage of two regions of approximately 18.5 kb of the reference genome. These

### Table 2 Oligonucleotides used in this study

| Name           | Sequence (5’ -3’) | Description       |
|----------------|-------------------|-------------------|
| oWKS-1467      | CGCACACTAGGAAGGAAGAG | Left junction CTn5 |
| oWKS-1468      | AGCGGTACGCTGGAAGATG | Left junction CTn5 |
| oWKS-1469      | TAGATGTAGCCCTGCTGAG | Right junction CTn5 |
| oWKS-1470      | AAGGGTTGGATTGCTGAG | Right junction CTn5 |
| oWKS-1471      | CCGTTACGCTGCTGATG  | runA gene         |
| oWKS-1472      | AGGCTTAAATGAGTGC   | runA gene         |

*The repaired junction (CTn5 excised from CD1844) is detected with oWKS-1467 and oWKS-1470. **The insertion of CTn5 into runA is detected by primer combination oWKS-1468/oWKS-1472 and/or oWKS-1469/oWKS-1471.*
regions were found to be nearly identical phages [16], and the unbalanced coverage therefore likely represents an artefact of the unsupervised assembly procedure using the default settings. To correct for this, the assembly was artificially broken into three contigs at these regions and was rejoined using the gap closure software PBJelly [53]. The edited assembly was then validated using reads from a third SMRT cell and polished using Quiver, a consensus algorithm that is part of the SMRT Portal. Subsequently, the consensus sequence was circularized based on the reference sequence of the ancestral 630 strain. We noted that the Pacific Biosciences consensus caller struggles with homopolymeric stretches of adenes and thymines. Therefore a correction was carried out by performing a reference assembly of the short reads from the Illumina sequencing against the reclosed genome, yielding the final genome sequence. This sequence is available from EMBL (EMBL: LN614756).

In silico analysis of the 630\texttt{Δerm} genome sequence

To annotate the de novo assembled genome sequence, we first updated the most recent version of the C. difficile 630 genome sequence [EMBL:AM180355.1] [27] in Artemis [54,55]. Next, we imported the flat genome sequence of strain 630\texttt{Δerm} into Geneious R7 (Biomatters, http://www.geneious.com) and transferred the annotation using the “Live Annotate and Predict” function. The annotation track was manually curated to remove duplicate or missed annotations. The resulting file was saved as a GenBank file, further polished in a text editor and Artemis and submitted to the ENA archive. Genome wheel representations were prepared using Circos [56]. Indels and single nucleotide polymorphisms were identified using the Pacific Biosciences variant caller using the genome of C. difficile strain 630 [8] as a reference and further validated by MUMmer 3.0 [57] and progressiveMAUVE [28]. Subsequently a list of detected structural variants was manually curated (consensus between the alignment of Illumina and PacBio reads to the reference strain and the variants identified by MUMmer and progressiveMAUVE) as concordant description of differences in complex genomic regions could not be achieved by different methods. In addition, for all large structural variants dotplots were generated using Gepard 1.30 [58] using FASTA formatted genome sequences of strains 630 and 630\texttt{Δerm}.

To identify modified bases, kinetic signals were processed for all genomic positions after aligning sequencing reads to the final single chromosome sequence of strain 630\texttt{Δerm}. In order to accurately identify the methylated bases, a threshold of 45 for log-transformed \( P \) values was used after optimizing according to its distribution and minimizing the false positive rate. Genomic positions and identity of the modifications were exported as a GFF file, and imported as a separate track in the genome sequence in Geneious R7. Subsequently, the identification of sequence motifs was performed using the SMRT Portal and sequence logos were prepared using Weblogo (http://weblogo.berkeley.edu/) [59] with 20 bp sequence flanking the modified base.

Analysis of CTn5 translocation

Translocation of CTn5 was confirmed by PCR using primers (Table 2) designed to amplify the left and right junctions of CTn5 as present in the C. difficile strain 630, as well as the \textit{erm} gene (Table 1) using Q5 polymerase (New England Biolabs). Cycling conditions were: initial denaturation 98° C 30 sec, 25 cycles 98°C 10 sec/60°C 30 sec/72°C 1 min 30 sec, and a final extension 72°C for 2 mins. Products were purified (GeneJet PCR purification kit, ThermoScientific) and run on a 0.5× TAE/1.2% agarose gel with a 1 kb + ladder (Fermentas). After staining with ethidium bromide, the DNA bands were visualized on a GelDoc system (Biorad).

Additional files

Additional file 1: Table summarizing discordantly mapped read-pairs in the Illumina HiSeq reference alignment of C. difficile 630\texttt{Δerm} (LUMC) versus 630.

Additional file 2: Table summarizing discordantly mapped read-pairs in the Illumina HiSeq reference alignment of C. difficile 630\texttt{Δerm} (LUMC) versus 630.

Additional file 3: Table summarizing validated discordantly mapped read-pairs in the Illumina HiSeq reference alignment of C. difficile 630\texttt{Δerm} (LUMC) versus 630.

Competing interests
The authors declare that they have no competing interests.

Authors’ contributions
EVE, JF, AMS, WKS performed experiments. SYA and APR contributed reagents and tools. EVE, SYA,HPB, JF, AMS, APR, WYL and WKS analyzed data. EVE, SYA, and WKS wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements
We thank Nigel Minton, Robert Britton and Paolo Mastrantonio for strains and Leon Mei of the Sequence Analysis Support Core (LUMC) for facilitating the initial illumina analysis. Furthermore, we want to thank the Geneious team for helpful discussions. This work was supported, in part, by a Veni and a Vidi fellowship from the Netherlands Organization for Scientific Research and a Gisela Thier Fellowship from the Leiden University Medical Center to WKS.

Author details
1Department of Medical Microbiology, Leiden University Medical Center, Leiden, the Netherlands. 2Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands. 3Leiden Genome Technology Center, Human and Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands. 4Wellcome Trust Sanger Institute, Cambridge, Hinxton,UK. 5Sequence Analysis Support Core, Leiden University Medical Center, Leiden, the Netherlands. 6Department of Microbial Diseases, UCL Eastman Dental Institute, London, UK.

Received: 29 September 2014 Accepted: 16 January 2015
Published online: 31 January 2015
References

1. Hall IC, O’Toole E. Intestinal flora in new-born infants: with a description of a new pathogenic anaerobe, *Bacillus difficilis*. Am J Dis Children. 1935;49:390–402.

2. Rupnik M, Wilcox MH, Gerdin DN. *Clostridium difficile* infection: new developments in epidemiology and pathogenesis. Nat Rev Microbiol. 2009;7:526–35.

3. Viswanathan VK, Mallozzi MJ, Vedantam G. *Clostridium difficile* infection: an overview of the disease and its pathogenesis, epidemiology and interventions. Gut Microbes. 2010;1:234–42.

4. Shen A. *Clostridium difficile* toxins: mediators of inflammation. J Innate Immun. 2012;4:19–58.

5. Vedantam G, Clark MA, Miyajima F, Roberts P, Ellison L, Pickard DJ, Martin MJ, et al. Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. Nat Genet. 2013;45:109–13.

6. Smits WK. Hype or hypervirulence: a reflection on problematic *C. difficile* strains. Virulence. 2013;4:592–6.

7. Sebaihia M, Wren BW, Mullany P, Fairweather NF, Minton N, Stabler R, et al. The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. Nat Genet. 2006;38:779–86.

8. Wust J, Sullivan NM, Hardegger U, Wilkins TD. Investigation of an outbreak of antibiotic-resistant *Clostridium difficile* by various typing methods. J Clin Microbiol. 1982;18:1096–101.

9. Brouwer MS, Warburton PJ, Roberts AP, Mullany P, Allan E. Genetic organisation, mobility and predicted functions of genes on integrated, mobile genetic elements in sequenced strains of *Clostridium difficile*. PLoS One. 2011;6:e23014.

10. Brouwer MS, Roberts AP, Mullany P, Allan E. Demonstration of conjugative transfer of clostridial shuttle vectors from *Bacillus subtilis* into toxin producers. Nat Commun. 2013;4:2601.

11. Marinus MG, Casadesus J. Roles of DNA adenine methylation in host-pathogen interactions: mismatch repair, transcriptional regulation, and more. FEBS Microbiol Rev. 2009;33:488–503.

12. Collier J. Epigenetic regulation of the bacterial cell cycle. Curr Opin Microbiol. 2009;12:722–9.

13. Brouwer MS, Roberts AP, Mullany P, Allan E. The *Clostridium difficile* genome: a new era of genome sequencing. Nat Methods. 2010;7:461–9.

14. Madsen CT, Mengel-Jorgensen J, Kirpekar F, Douthwaite S. Identifying the mobile genetic elements conferring antibiotic resistance. FEMS Microbiol Rev. 2011;35:856–71.

15. Jasni AS, Mullany P, Hussain H, Roberts AP. Demonstration of conjugative transfer to and from *Bacillus subtilis*. J Gen Microbiol. 1990;136:1343–9.

16. He M, Sebaihia M, Lawley TD, Stabler R, Dawson LF, Martin MJ, et al. Evolutionary dynamics of *Clostridium difficile* over short and long time scales. Proc Natl Acad Sci U S A. 2010;107:7527–32.

17. Roberts AP, Mullany P. *Tn916*-like genetic elements: a diverse group of modular mobile elements conferring antibiotic resistance. FEMS Microbiol Rev. 2011;35:856–71.

18. Liu Y, Ehsaan M, Philip S, Collery MM, Janoir C, Collignon A, et al. Detecting DNA base modifications using *Clostridium difficile* strains. J Bacteriol. 2005;187:1343–9.

19. Jasiak P, Husain F, Mullany P, Hube B. Three-dimensional structure of the *cloE* gene from *Clostridium difficile* reveals a transposase domain and a large N-terminal variable region. J Mol Biol. 2005;349:529–39.

20. MacPherson CL, Dworkin A, Robert-Voigt A, Lai C, Sarles H, Stanway G, et al. Detection of DNA base modifications using *Clostridium difficile* strains. Nucleic Acids Res. 2005;33:4738–44.

21. Brouwer MS, Mullany P, Allan E. In silico analysis of sequenced strains of *Clostridium difficile* reveals a related set of conjugative transposons carrying a variety of accessory genes. Mol Genet Elements. 2012;28–12.

22. Agerwalla S, Kealey JT, Santi DV, Stroud RM. Characterization of the *C. difficile* PR144 strain. J Bacteriol. 2002;184:6273–8.

23. Brouwer MS, Mullany P, Allan E. In silico analysis of sequenced strains of *Clostridium difficile* reveals a related set of conjugative transposons carrying a variety of accessory genes. Mol Genet Elements. 2012;28–12.

24. Agerwalla S, Kealey JT, Santi DV, Stroud RM. Characterization of the *C. difficile* PR144 strain. J Bacteriol. 2002;184:6273–8.

25. Brouwer MS, Mullany P, Allan E. In silico analysis of sequenced strains of *Clostridium difficile* reveals a related set of conjugative transposons carrying a variety of accessory genes. Mol Genet Elements. 2012;28–12.

26. Agerwalla S, Kealey JT, Santi DV, Stroud RM. Characterization of the *C. difficile* PR144 strain. J Bacteriol. 2002;184:6273–8.

27. Brouwer MS, Mullany P, Allan E. In silico analysis of sequenced strains of *Clostridium difficile* reveals a related set of conjugative transposons carrying a variety of accessory genes. Mol Genet Elements. 2012;28–12.

28. Agerwalla S, Kealey JT, Santi DV, Stroud RM. Characterization of the *C. difficile* PR144 strain. J Bacteriol. 2002;184:6273–8.

29. Brouwer MS, Mullany P, Allan E. In silico analysis of sequenced strains of *Clostridium difficile* reveals a related set of conjugative transposons carrying a variety of accessory genes. Mol Genet Elements. 2012;28–12.

30. Agerwalla S, Kealey JT, Santi DV, Stroud RM. Characterization of the *C. difficile* PR144 strain. J Bacteriol. 2002;184:6273–8.

31. Brouwer MS, Mullany P, Allan E. In silico analysis of sequenced strains of *Clostridium difficile* reveals a related set of conjugative transposons carrying a variety of accessory genes. Mol Genet Elements. 2012;28–12.

32. Agerwalla S, Kealey JT, Santi DV, Stroud RM. Characterization of the *C. difficile* PR144 strain. J Bacteriol. 2002;184:6273–8.

33. Brouwer MS, Mullany P, Allan E. In silico analysis of sequenced strains of *Clostridium difficile* reveals a related set of conjugative transposons carrying a variety of accessory genes. Mol Genet Elements. 2012;28–12.

34. Brouwer MS, Mullany P, Allan E. In silico analysis of sequenced strains of *Clostridium difficile* reveals a related set of conjugative transposons carrying a variety of accessory genes. Mol Genet Elements. 2012;28–12.

35. Brouwer MS, Mullany P, Allan E. In silico analysis of sequenced strains of *Clostridium difficile* reveals a related set of conjugative transposons carrying a variety of accessory genes. Mol Genet Elements. 2012;28–12.

36. Brouwer MS, Mullany P, Allan E. In silico analysis of sequenced strains of *Clostridium difficile* reveals a related set of conjugative transposons carrying a variety of accessory genes. Mol Genet Elements. 2012;28–12.

37. Brouwer MS, Mullany P, Allan E. In silico analysis of sequenced strains of *Clostridium difficile* reveals a related set of conjugative transposons carrying a variety of accessory genes. Mol Genet Elements. 2012;28–12.

38. Brouwer MS, Mullany P, Allan E. In silico analysis of sequenced strains of *Clostridium difficile* reveals a related set of conjugative transposons carrying a variety of accessory genes. Mol Genet Elements. 2012;28–12.

39. Brouwer MS, Mullany P, Allan E. In silico analysis of sequenced strains of *Clostridium difficile* reveals a related set of conjugative transposons carrying a variety of accessory genes. Mol Genet Elements. 2012;28–12.

40. Brouwer MS, Mullany P, Allan E. In silico analysis of sequenced strains of *Clostridium difficile* reveals a related set of conjugative transposons carrying a variety of accessory genes. Mol Genet Elements. 2012;28–12.

41. Brouwer MS, Mullany P, Allan E. In silico analysis of sequenced strains of *Clostridium difficile* reveals a related set of conjugative transposons carrying a variety of accessory genes. Mol Genet Elements. 2012;28–12.

42. Brouwer MS, Mullany P, Allan E. In silico analysis of sequenced strains of *Clostridium difficile* reveals a related set of conjugative transposons carrying a variety of accessory genes. Mol Genet Elements. 2012;28–12.

43. Brouwer MS, Mullany P, Allan E. In silico analysis of sequenced strains of *Clostridium difficile* reveals a related set of conjugative transposons carrying a variety of accessory genes. Mol Genet Elements. 2012;28–12.

44. Brouwer MS, Mullany P, Allan E. In silico analysis of sequenced strains of *Clostridium difficile* reveals a related set of conjugative transposons carrying a variety of accessory genes. Mol Genet Elements. 2012;28–12.

45. Brouwer MS, Mullany P, Allan E. In silico analysis of sequenced strains of *Clostridium difficile* reveals a related set of conjugative transposons carrying a variety of accessory genes. Mol Genet Elements. 2012;28–12.

46. Brouwer MS, Mullany P, Allan E. In silico analysis of sequenced strains of *Clostridium difficile* reveals a related set of conjugative transposons carrying a variety of accessory genes. Mol Genet Elements. 2012;28–12.
49. Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 2011;21:936–9.

50. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 2009;25:1754–60.

51. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14:178–92.

52. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10:563–9.

53. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the gap: upgrading genomes with pacific biosciences RS long-read sequencing technology. PLoS One. 2012;7:e47768.

54. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, et al. Artemis: sequence visualization and annotation. Bioinformatics. 2000;16:944–5.

55. Carver T, Berriman M, Tivey A, Patel C, Rohme U, Barrell BG, et al. Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. Bioinformatics. 2008;24:2672–6.

56. Krzywinski M, Schein J, Biro I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19:1639–45.

57. Kurtz S, Phillippy A,Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5:R12.

58. Krumsiek J, Arnold R, Rattei T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics. 2007;23:1026–8.

59. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14:1188–90.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at www.biomedcentral.com/submit