Reannotation of the genome sequence of Clostridium difficile strain 630.

Marc Monot, Caroline Boursaux-Eude, Marie Thibonnier, David Vallenet, Ivan Moszer, Claudine Medigue, Isabelle Martin-Verstraete, Bruno Dupuy

Re-annotation of the genome sequence of *Clostridium difficile* strain 630

Marc Monot[1], Caroline Boursaux-Eude[2], Marie Thibonnier[1], David Vallenet[3], Ivan Moszer[2], Claudine Medigue[3], Isabelle Martin-Verstraete[1,4] and Bruno Dupuy[1]

[1] *Institut Pasteur, Laboratoire Pathogenèse des Bactéries Anaérobies, 28 rue du Dr Roux, 75724 Paris Cedex 15, France*

[2] *Institut Pasteur, Intégration et Analyse Génomiques, 28 rue du Dr Roux, 75724 Paris Cedex 15, France*

[3] *CNRS UMR 8030, Université d'Evry & CEA, IG, Genoscope - LABGeM - 2 rue Gaston Crémieux, CP5706, F-91057 Evry cedex, France*

[4] *Université Paris Diderot-Paris 7 – 7 boulevard Diderot, 75012 Paris, France*

Correspondence

Marc Monot

mmonot@pasteur.fr

Tel. +331 45 68 83 16 / Fax. +331 40 61 31 83

A regular update of genome annotations is a prerequisite step to help maintain the accuracy and relevance of the information they contain. Five years after the first publication of the complete genome sequence of *C. difficile* strain 630, we manually re-annotated each of the coding sequences (CDS), using a high-level annotation platform. The function of more than 500 genes annotated previously with putative functions, were re-annotated based on updated sequence similarities of proteins whose functions have been recently identified by experimental data from the literature. We also modified 222 CDS starts, detected 127 new CDS and added the enzyme commission numbers, which were not supplemented in the original annotation. In addition, an intensive project was undertaken to standardise the names of genes and gene products and thus harmonising as much as possible with the HAMAP project. The re-annotation is stored in a relational database that will be available on the MicroScope web-based platform, "http://www.genoscope.cns.fr/agc/microscope/ClostridioScope". The original submission stored in the INSDC nucleotide sequence databases was also updated.

Scope : *Clostridium difficile* special issue of the *Journal of Medical Microbiology*

**The EMBL accession number for the re-annotation of C. difficile strain 630 is AM180355 and its plasmid pCD630 is AM180356.*

36 # INTRODUCTION

37

38 The re-annotation of several model genomes has been recently performed,
39 among these there are *Escherichia coli* for the gram negative bacteria (Luo *et al.*, 2009)
40 and *Bacillus subtilis* for the firmicutes (Barbe *et al.*, 2009). This provided new
41 information about genomic structure and organisation as well as gene function and
42 plays an essential role in defining reference knowledge. In addition the re-annotation of
43 the *B. subtilis* genome also benefits the other firmicutes such as *Clostridium difficile.*

44

45 *C. difficile* is one of the major enteropathogenic clostridia and *C. difficile*-
46 associated diarrhea (CDAD) is currently the most frequently occurring nosocomial
47 infection in many European hospitals. Although toxins are generally recognised as the
48 main virulence factors, *C. difficile* pathogenesis remains poorly understood. The global
49 genetic analysis of *C. difficile* appeared to be an useful approach to find potential
50 mechanisms involved in the bacterial virulence for which an updated of the gene list and
51 corresponding annotations is tremendously important.

52

53 The first complete genome sequence of a *C. difficile* strain (630) was sequenced in
54 2006 (Sebaihia *et al.*, 2006). It led to the development of high throughput projects such
55 as comparative genomic, transcriptomic and proteomic studies (Jain *et al.*, 2010;
56 Janvilisri *et al.*, 2010; Marsden *et al.*, 2010), which were recently reinforced with an
57 increase of multiple genomic projects (Stabler *et al.*, 2009). However, the relevance of all
58 these experiments greatly depends on the information available for the genes
59 particularly their functions experimentally identified or predicted *in silico*. Thus, it is
60 critical that the information is accurate, relevant and useful. This is why we undertook
61 the re-annotation of the *C. difficile* strain 630 genome.

62

63 The advances in second-generation sequencing technologies combined with their
64 relative low cost has led to the increased need for a rapid genome annotation system
65 (Petty, 2010). However the fastest way to obtain an accurate annotation remains to
66 transfer annotation from a reference strain. This requires to have access to a closely
67 related genome for each species annotated to a high standard and regularly updated.

We described in this paper the manual re-annotation of all CDS of the *C. difficile* strain 630 genome. For this purpose we used improved methods in bioinformatics, literature surveys and genome data from closely related species such as *Clostridium sticklandii*, which has recently been sequenced (Fonknechten *et al.*, 2010) or *B. subtilis* whose genome has been re-sequenced and re-annotated (Barbe *et al.*, 2009). The re-annotation resulted in the function precision of more than 500 genes and the addition of new CDSs as well as the correction of the start sites of 222 CDSs. All information from laboratory research publications could be continuously integrated though the MicroScope platform to maintain this up-to-date annotation.

METHODS

Identification of new or modified CDS in the *C. difficile* genome

The sequence and the original annotation of the published *C. difficile* 630 genome (Sebaihia *et al.*, 2006) was integrated into the Microscope platform (Vallenet *et al.*, 2009). MicroScope is a web-based framework for the systematic and efficient revision of microbial genome annotation and comparative analysis. Its main features are (i) integration of annotation data from bacterial genomes enhanced by a gene coding re-annotation process using accurate gene models, (ii) integration of results obtained with a wide range of bioinformatics methods, among which exploration of gene context by searching for conserved synteny and reconstruction of metabolic pathways, (iii) an advanced web interface allowing multiple users to refine the automatic assignment of gene product functions. MaGe is also linked to numerous well-known biological databases and systems. The original gene prediction was systematically checked using the AMIGene software (Bocs *et al.*, 2003) and the MICHeck strategy (Cruveiller *et al.*, 2005). The initial identifier of genes 'CD0000(A)' used a prefix of two letters, 'CD', followed by a four-digit number corresponding to the position of CDS in the genome. Whenever a new gene interleave, a capital letter was added in alphabetical order. Since 2006, the locus tag usage has evolved (Cochrane *et al.*, 2008). The prefix now has to contain only alpha-numeric characters and it must be at least 3 characters long. In addition the locus tag prefix must be separated from the tag value by an underscore

100 ending with a number. So we assigned for all CDS a new locus tag code: 'CD630_00000'.
101 The four-digit number after the underscore is still the original CDS position in the
102 genome. The capital letter of the original identifier was converted to a number which
103 has been added at the end of each gene : 1 to 9 for genes previously ended with capital
104 letter A to I, and 0 for all others e.g. CD0001 into CD630_00010 and CD0163B into
105 CD630_01632. Finally, because the genomic position of the non coding CDS was defined
106 with only three-digit numbers, we replaced the first number after the locus tag prefix
107 with a 't' or 'r' respectively for transfer RNA and ribosomal RNA respectively, e.g. CDt001
108 into CD630_t0010 and CDr001 into CD630_r0010. We used the same coding method for
109 the 11 CDSs encoded by the plasmid pCD630, adding the letter 'p', after the locus tag e.g.
110 CDP01 into CD630_p010.
111
112 During the re-annotation process using the AMIGene predictions, we identified
113 new CDS and we assigned them the locus tag of the previous CDS with the last number
114 incremented by 1 e.g. a new gene detected after CD630_02670 (previously named
115 CD0267) was coded CD630_02671. The original locus tag will be kept in the EMBL file
116 using the /Old_locus_tag identifier.
117
118 Re-annotation of the complete *C. difficile* strain 630
119 The predicted proteins were subjected to a wide range of bioinformatics tools,
120 which includes conserved synteny computations, alignments against TrEMBL and
121 SWISS-PROT databases (Apweiler R, 2011) and TMHMM (Sonnhammer *et al.*, 1998),
122 SignalP (Bendtsen *et al.*, 2004) and PsortB (Yu *et al.*, 2010) software to predict
123 subcellular localization of proteins as well as INTERPROSCAN (Zdobnov & Apweiler,
124 2001) to identify possible functions of newly discovered proteins (Apweiler R, 2011).
125 This work flow led to an automatic functional annotation for each CDS as previously
126 described (Vallenet *et al.*, 2006). Finally, these pre-computed results served as basis for
127 the manual re-annotation of each CDS proceeding by inference.
128
129 To normalise the process of manual annotation among multiple users, we set up
130 several guidelines: (a) The product field is filled with the functional annotation for all
131 genes identified with 'Hypothetical protein' or 'Conserved Hypothetical protein' when
132 the gene was not identified. For all others we added 'Putative' prior to the product

133 annotation. Pseudogene and gene remnant have a specific nomenclature : ”Fragment of”
134 + function + position (N-terminal, C-terminal or center of the encoding protein). (b) The
135 name of gene was completed by searching in the literature using PubMed data libraries
136 (http://www.ncbi.nlm.nih.gov/pubmed) and when we changed gene names, old names
137 were indicated in the synonymous field. (c) The start sites were modified according to
138 the combination of the graphical data such as codage probability curves deduced from
139 the AMIGene method (Bocs *et al.*, 2003), as well as alignments with orthologous genes
140 (Altschul *et al.*, 1990). Then, the label ‘/START=’ was added in the comment field
141 followed by a capital letter associated to an informative code (M: modified, C: coding
142 curve, S: sequence similarity, O : overlap, R : RBS). (d) PubMed identifiers (PMIDs) of
143 each gene were classified from the specific references to the articles corresponding to
144 orthologuous genes and/or the global reviews concerning its function. (e) Protein
145 families were standardized using the same keywords, PMIDs and global classification,
146 such as CMR roles (http://cmr.jcvi.org/cgi-bin/CMR/RoleIds.cgi).
147

148 # RESULTS & DISCUSSION

149

150 Evaluation of annotation improvement

151 The original annotation of the *C. difficile* strain 630, published in 2006 (Sebaihia
152 *et al.*, 2006), identified 3776 predicted coding sequences (CDSs). We have updated
153 annotation of all CDSs and assigned or precised their functions. During the re-annotation
154 process we attributed a class of function to each gene re-annotated: (i) “known”: when
155 function was experimentally demonstrated or when high level of similarities with
156 characterised genes were found (ii) “putative”: based on conserved motif, structural
157 feature or limited similarities, (iii) “unknown”: when genes were unidentified and (iv)
158 “pseudo”: for pseudogenes or gene remnants. The same classification was applied
159 manually to the 2006 annotation to allow comparison of both annotations (Table 1A).
160

161 Thus, 518 and 18 genes whose encoding function was previously described as
162 putative and unknown respectively have now a functional annotation identified by the
163 experimental data from the literature (Table 1A). For example, CD630_26030
164 (previously named CD2603), recognised as a putative response regulator, is now

165 designated cdtR, since it was shown that it controls the binary toxin expression in *C.*
166 *difficile* (Carter *et al.*, 2007). In addition, 117 genes of unknown function have now a
167 putative function. For instance, 12 conserved hypothetical proteins which contain a
168 CRISPR-associated domain (clustered regularly interspaced short palindromic repeats)
169 are annotated "Putative CRISPR-associated family protein". Furthermore, we showed
170 that the ATP synthase epsilon chain, CD630_34670 (CD3467), which was defined as a
171 gene remnant (pseudo class) because of a lack of amino-acid in the C-terminus relative
172 to database matches, actually belongs to the class of "known function". This enzyme
173 usually combines ATP synthesis and hydrolysis but the hydrolysis function is still active
174 in the truncated version (Ferguson *et al.*, 2006).
175
176 Following the re-annotation we included 127 new CDSs and defined 222 new CDS
177 start sites. The majority of the new CDS are divided into putative (25), unknown (86) or
178 pseudogene (15) classes (Table 1B). Only one gene, *CD630_15951* has an orthologue,
179 whose function was experimentally demonstrated. This gene, detected during the
180 proteomic analysis recently performed in *C. difficile* (Lawley *et al.*, 2009), is highly
181 homologous (~60%) to a ferredoxin gene of *Clostridium thermoaceticum (Elliott et al.,*
182 *1982)*.
183
184 We were looking for papers corresponding to each gene, and particularly those
185 published after the original annotation. We added at least one PMID reference number
186 to 64% of the *C. difficile* genes. Like many other genome-wide updates, several
187 specificities were added to the original product function. When possible, we attached
188 new motifs and enzymatic domains identified by INTERPROSCAN, allowing a more
189 accurate description of the original function. For example, putative peptidase enzymes
190 have now family information according to the classification scheme of the MEROPS
191 database (http://merops.sanger.ac.uk). The revised nomenclature of the pathogenicity
192 locus region (Rupnik *et al.*, 2005) has been introduced during re-annotation process as
193 well as genes involved in the *C. difficile* motility and flagellar glycosylation since they
194 were recently published (Twine *et al.*, 2009). A locus tag, product annotation and class
195 comparison between the two annotations performed in 2006 and 2010 were
196 summarised in the Table S1. All information of the CDS re-annotated (Fig. S1), are
197 currently available on the MicroScope platform:

198 http://www.genoscope.cns.fr/agc/microscope/ClostridioScope. We also updated the *C.*
199 *difficile* 630 genomic entry in genomic databases: EMBL, GenBank and DDBJ.
200
201 Deciphering the annotation origin
202 Several pieces of information appeared when we evaluated the source used
203 during the functional annotation of known, putative or unknown genes (Fig. 1). In the
204 known category, 1% of the gene function came from *C. difficile* strain 630 publications,
205 1,5% from other *C. difficile* strains, 4% from other clostridia and 93,5% from other
206 species (Fig. 1A.). The putative category was defined according to the enzymatic domain
207 (40%), homology to mobile elements (20%) or cell localisation (15%) (Fig. 1B). As an
208 example a gene will be annotated "Putative membrane protein" when 3 or more
209 transmembrane helix was detected by TMHMM (Sonnhammer *et al.*, 1998). Finally, we
210 classified the unknown genes from the alignement results with TREMBL (Boeckmann *et*
211 *al.*, 2003). Although 45% were orphan of the *C. difficile* strains, 20% were also found in
212 the genus Clostridium, 15% in the firmicutes phylum and 20% in other bacteria (Fig.
213 1C).
214
215 Concerning genes annotated as known, we noted that only few of them came
216 from a published clostridial experiments (Fig. 1A). This was mainly due to the lack of
217 effective tools to mutate clostridial genes. However gene inactivation method and
218 random mutagenesis system recently developed in *C. difficile* (Cartman & Minton, 2010),
219 should greatly improve the number of publications on *C. difficile* gene functions. Half of
220 the genes with an unidentified function, orphan, are found only in *C. difficile* 630 (Fig.
221 1C.). However, most orphans are present in the *C. difficile* strains already sequenced
222 such as strains 027, CD196 and R20291 (Stabler *et al.*, 2009). This may constitute a
223 source of gene targets that could be used both in research, diagnosis or treatment of the
224 CDAD.
225
226 Miscellaneous improvements
227 To re-annotate the *C. difficile* genome of strain 630 we used the MaGe interface,
228 which contains classic database fields (type, position, name, product, EC numbers) and
229 several specific fields such as gene synonymous (synonyms), authors notes (comments),
230 pubmed identifiers (PMID), product type, localisation and functional classification (Fig.

231 S1). All information found during the re-annotation process that did not fit in the classic
232 fields were added in the specific MaGe field or in the comments. For example, the novel
233 virulence factor called Srl for « Sensitivity regulation of *C. difficile* toxins », (Miura *et al.*,
234 2010) was presented during the third international clostridium difficile symposium.
235 This information was only indicated in the comment field of the CD630_22980 (CD2298)
236 gene until further validation.
237
238 The names of gene products were harmonizing as much as possible with the
239 HAMAP project (Lima *et al.*, 2009). On the other hand, all gene products have now been
240 named with a specific keyword related to their functional family (Fig. S1). Thus,
241 CD630_05310 (CD0531), previously annotated «DeoR-like regulator of transcription» (a
242 regulator of sugar and nucleoside metabolic systems) was re-annotated
243 « Transcriptional regulator (keyword), DeoR family ». We also normalised the
244 annotation of genes that share the same characteristics. As an example, proteins that
245 were only determined according to their membrane localization were annotated:
246 « Putative membrane proteins ». The annotation standardization we used will facilitate
247 the mining of the data using bioinformatics as well as manual search (Fig. S1).
248
249 Membrane Transport
250 The *C. difficile* genome contains a lot of proteins encoding several membrane
251 transport systems: ATP-binding cassette (ABC) transporters, phosphoenolpyruvate-
252 dependent phosphotransferase systems (PTS), charged substrate transporters
253 (antiporters, symporters) and facilitators. The general function of the genes encoding
254 such proteins can be easily determined from bioinformatic approaches, like those used
255 for the protein domain analysis in InterProScan (Zdobnov & Apweiler, 2001). However
256 it is quite difficult to distinguish the exact metabolite they transport, especially when the
257 transport systems have a wide specificity. We reannotated most of the transporter
258 systems by inference including clues about targets using specialized databases such as
259 TransportDB (http://www.membranetransport.org/)(Ren *et al.*, 2007) which compile
260 all information on cytoplasmic membrane transporters. We added a suffix in the
261 classification which indicate, from a global trend to the expected target, the motif
262 (family), the high sequence homology (like) and the evidence of a target metabolite

263 (specific). However, this classification should be taken with caution since it was mainly
264 deduced from *in silico* analysis rather than from experimental data.
265
266 The table 2 showed annotation of 19 PTS systems with a specific
267 metabolite suggestion. The targeted metabolite was deduced from the INTERPROSCAN
268 motif search but could also be defined by the presence in the same locus of gene
269 encoding enzyme involved in specific sugar assimilation (associated enzyme). As an
270 example *CD630_22690 (CD2269)* is now annotated as "PTS system, fructose-specific
271 IIABC component". This is due to the detection of three motif signatures the mannitol
272 family PTS EII component A, B and C, as well as the presence of the neighbouring gene,
273 the *CD630_22700 (CD2270)*, which encodes an enzyme involved in the utilization of
274 fructose: "Fructose 1-phosphate kinase" as indicated in the gene annotation (Table S1).
275
276 Metabolism update
277 Updating the genome annotation of *C. difficile* led to many changes within the
278 metabolism pathways. The gene cluster involved in the anaerobic oxidative degradation
279 of L-ornithine has been identified in *C. sticklandii* (Fonknechten *et al.*, 2009). From this
280 publication we reannotated genes *CD630_04420 (CD0442)* to *CD630_04480 (CD0448)*
281 whose encoding proteins share high similarities to the ornithine catabolism compounds
282 of *C. sticklandii* e.g. Ord, OrtA, OrtB, OraS, OraE, Or-4 and Orr, respectively (Table S1).
283 This suggested that *C. difficile* could produce acetyl-CoA from the ornithine
284 fermentation. The ability to use a variety of carbohydrates is an important feature for *C.*
285 *difficile* to colonize the host gut. *Enterococcus faecalis* found in the same niche as *C.*
286 *difficile*, provided hints to explore the consistency of a specific pathway required for
287 ethanolamine utilisation, a constituent of an abundant class of phospholipids present in
288 the eucaryotic cell membranes and the host's dietary intake (Del Papa & Perego, 2008)
289 (Fox *et al.*, 2009). Using the *E. faecalis* gene synteny and protein similarities, we were
290 able to reconstruct the whole ethanolamine pathway in *C. difficile*, a cluster of 19 genes,
291 from *CD630_19060 (CD1906) to CD630_19260 (CD1926)* encoding the ethanolamine
292 ammonia-lyase, an alcohol dehydrogenase, a carboxysome associated proteins, the
293 transporter EutH and the two-component system EutV, EutW. (Table S1).
294

295 Interestingly, in *B. subtilis* several enzymes involved in RNA degradation were
296 recently identified (Even *et al.*, 2005) (Shahbabian *et al.*, 2009). In *C. difficile,* a unique
297 Rnase J protein *CD630_12890 (CD1289)* was detected as well as an ortholog of *ymdA*
298 *CD630_13290 (CD1329)*, encoding the Rnase Y protein.
299

300 # CONCLUSION

301

302 Finally, nearly half of the genes of the *C. difficile* strain 630 encode proteins with
303 known function, whereas one-third of the gene products have a putative function and
304 only fifteen percent of proteins with unknown function are encoded by *C. difficile*
305 genome (Table 1A). In addition, 127 new CDSs were discovered (Table 1B) and 222 CDS
306 starts were modified. The re-annotation was performed using a high standard
307 annotation MicroScope platform, which significantly increased the amount of
308 information available for the majority of the CDS, such as literature references, product
309 types, localisation and gene synonymous (Fig. S1).
310

311 Nevertheless, there is still great deal of work to be completed since only 116
312 annotated genes came from a published clostridial experiments. The EMBL entries are
313 now resubmitted and to maintain the annotation up-to-date, all new information would
314 be addressed directly to marc.monot@pasteur.
315

316 # ACKNOWLEDGEMENTS

317

318 Special thanks to Mohamed Sebaihia which allowed us to update the original database
319 entry (AM180355). We thank Richard Stabler who gave several comparison gene files
320 between the reference genome and the 027 strains. Special thanks to Ana Antunes,
321 Sylvie Bouttier, Thomas Candela, Claire Janoir and Johann Peltier who provided helps in
322 specific gene annotations. We are gratful to Marc Griffiths and Erica Porter's help in
323 english corrections.
324

325 # AUTHORS' CONTRIBUTION

326
327 BD, MM, CB-E and IM designed the study. CB-E and MM carried out the major part of the
328 manual re-annotation of the genome together with MT and IM-V. DV and CM were
329 involved in automatic re-annotation and administration of the MicroScope platform.
330 MM, IM-V and BD wrote the manuscript.
331

332 # TABLE & FIGURE LEGENDS

333
334 Table 1 : Review of the 2006's annotation update. A) CDSs were identified and separated
335 according to the four major annotation classes in both 2006 and 2010 annotations:
336 (**Known)** when function were experimentally demonstrated, (**Putative**) based on
337 conserved motif, structural feature or limited homology, (**Unknown**) when function are
338 unidentified and (**Pseudo**) for pseudogenes. Dark grey padding numbers indicated no
339 change and "±"and "-" correspond to a change betwenn the classes of annotation
340 between 2006 annotation and 2010 re-annotation. B) Annotation of the new CDS
341 detected and referenced as known, putative, unknown and pseudo classes.
342
343 Table 2 : Re-annotation of the PTS systems according to the metabolite specificity. List of
344 locus tags corresponding to 19 PTS re-annotated. The PTS metabolite was deduced from
345 the motif class detection and/or the presence of associated enzymes involved in a
346 specific sugar metabolism.
347
348 Figure 1 : Distribution of the functional re-annotation origin. A) **Known** functions
349 were identified from the literature references of: *Clostridium difficile* strain 630 (C. diff
350 630), other *Clostridium difficile* strains (C. difficile), *Clostridium* species (Clostridia) and
351 others species (Others). B) **Putative** functions were defined from: enzymatic domains
352 (Enzyme), homology with mobile elements (Mobile), localisation in the cell
353 (Localization) and the remaining origin (Others). C) **Unknown** functions which were
354 found only in: *Clostridium difficile* (Orphan), in the Clostridium species (Clostridia), in
355 the firmicutes plylum (Firmicutes) or in diverse bacteria (Others).

SUPPLEMENTARY DATA

Table S1 : Comparison between 2006 and 2010 annotations. For each CDS, the locus tags, annotation function and classes are compared for both 2006 and 2010 annotations.

Table S2 : Standardization of family product names. The product names were constructed around a keyword specific to the gene's functional family.

Figure S1 : *C. difficile* CDS re-annotation by MaGe. MaGe annotation window for *tcdR* gene. Bold and grey outline focus on information added specifically within this process : mutation, synonyms, comments, PMID, product type, localization, MaGe classification and standard classification (Bioprocess and Roles).

REFERENCES

Altschul, S., Gish, W., Miller, W., Myers, E. & Lipman, D. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **3**.

Apweiler R, M. M., O'Donovan C, Magrane M, Alam-Faruque Y, Antunes R, Barrell D, Bely B, Bingley M, Binns D, Bower L, Browne P, Chan WM, Dimmer E, Eberhardt R, Fazzini F, Fedotov A, Foulger R, Garavelli J, Castro LG, Huntley R, Jacobsen J, Kleen M, Laiho K, Legge D, Lin Q, Liu W, Luo J, Orchard S, Patient S, Pichler K, Poggioli D, Pontikos N, Pruess M, Rosanoff S, Sawford T, Sehra H, Turner E, Corbett M, Donnelly M, van Rensburg P, Xenarios I, Bougueleret L, Auchincloss A, Argoud-Puy G, Axelsen K, Bairoch A, Baratin D, Blatter MC, Boeckmann B, Bolleman J, Bollondi L, Boutet E, Quintaje SB, Breuza L, Bridge A, deCastro E, Coudert E, Cusin I, Doche M, Dornevil D, Duvaud S, Estreicher A, Famiglietti L, Feuermann M, Gehant S, Ferro S, Gasteiger E, Gateau A, Gerritsen V, Gos A, Gruaz-Gumowski N, Hinz U, Hulo C, Hulo N, James J, Jimenez S, Jungo F, Kappler T, Keller G, Lara V, Lemercier P, Lieberherr D, Martin X, Masson P, Moinat M, Morgat A, Paesano S, Pedruzzi I, Pilbout S, Poux S, Pozzato M, Redaschi N, Rivoire C, Roechert B, Schneider M, Sigrist C, Sonesson K, Staehli S, Stanley E, Stutz A, Sundaram S, Tognolli M, Verbregue L, Veuthey AL, Wu CH, Arighi CN, Arminski L, Barker WC, Chen C, Chen Y, Dubey P, Huang H, Mazumder R, McGarvey P, Natale DA, Natarajan TG, Nchoutmboube J, Roberts NV, Suzek BE, Ugochukwu U, Vinayaka CR, Wang Q, Wang Y, Yeh LS, Zhang J. (2011). Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res* **39**, D214-219.

Barbe, V., Cruveiller, S., Kunst, F., Lenoble, P., Meurice, G., Sekowska, A., Vallenet, D., Wang, T., Moszer, I. & other authors (2009). From a consortium sequence to a unified sequence: the Bacillus subtilis 168 reference genome a decade later. *Microbiology* **155**, 1758-1775.

Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**, 783-795.

Bocs, S., Cruveiller, S., Vallenet, D., Nuel, G. & Medigue, C. (2003). AMIGene: Annotation of MIcrobial Genes. *Nucleic Acids Res* **31**, 3723-3726.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C. & other authors (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**, 365-370.

Carter, G. P., Lyras, D., Allen, D. L., Mackin, K. E., Howarth, P. M., O'Connor, J. R. & Rood, J. I. (2007). Binary toxin production in Clostridium difficile is regulated by CdtR, a LytTR family response regulator. *J Bacteriol* **189**, 7290-7301.

Cartman, S. T. & Minton, N. P. (2010). A mariner-based transposon system for in vivo random mutagenesis of Clostridium difficile. *Appl Environ Microbiol* **76**, 1103-1109.

Cochrane, G., Akhtar, R., Aldebert, P., Althorpe, N., Baldwin, A., Bates, K., Bhattacharyya, S., Bonfield, J., Bower, L. & other authors (2008). Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Res* **36**, D5-12.

Cruveiller, S., Le Saux, J., Vallenet, D., Lajus, A., Bocs, S. & Medigue, C. (2005). MICheck: a web tool for fast checking of syntactic annotations of bacterial genomes. *Nucleic Acids Res* **33**, W471-479.

Del Papa, M. F. & Perego, M. (2008). Ethanolamine activates a sensor histidine kinase regulating its utilization in Enterococcus faecalis. *J Bacteriol* **190**, 7147-7156.

Elliott, J. I., Yang, S. S., Ljungdahl, L. G., Travis, J. & Reilly, C. F. (1982). Complete amino acid sequence of the 4Fe-4S, thermostable ferredoxin from Clostridium thermoaceticum. *Biochemistry* **21**, 3294-3298.

Even, S., Pellegrini, O., Zig, L., Labas, V., Vinh, J., Brechemmier-Baey, D. & Putzer, H. (2005). Ribonucleases J1 and J2: two novel endoribonucleases in B.subtilis with functional homology to E.coli RNase E. *Nucleic Acids Res* **33**, 2141-2152.

Ferguson, S. A., Keis, S. & Cook, G. M. (2006). Biochemical and molecular characterization of a Na+-translocating F1Fo-ATPase from the thermoalkaliphilic bacterium Clostridium paradoxum. *J Bacteriol* **188**, 5045-5054.

Fonknechten, N., Perret, A., Perchat, N., Tricot, S., Lechaplais, C., Vallenet, D., Vergne, C., Zaparucha, A., Le Paslier, D. & other authors (2009). A conserved gene cluster rules anaerobic oxidative degradation of L-ornithine. *J Bacteriol* **191**, 3162-3167.

Fonknechten, N., Chaussonnerie, S., Tricot, S., Lajus, A., Andreesen, J. R., Perchat, N., Pelletier, E., Gouyvenoux, M., Barbe, V. & other authors (2010). Clostridium sticklandii, a specialist in amino acid degradation:revisiting its metabolism through its genome sequence. *BMC Genomics* **11**, 555.

Fox, K. A., Ramesh, A., Stearns, J. E., Bourgogne, A., Reyes-Jara, A., Winkler, W. C. & Garsin, D. A. (2009). Multiple posttranscriptional regulatory mechanisms partner to control ethanolamine utilization in Enterococcus faecalis. *Proc Natl Acad Sci U S A* **106**, 4435-4440.

Jain, S., Graham, R. L., McMullan, G. & Ternan, N. G. (2010). Proteomic analysis of the insoluble subproteome of Clostridium difficile strain 630. *FEMS Microbiol Lett* **312**, 151-159.

Janvilisri, T., Scaria, J. & Chang, Y. F. (2010). Transcriptional profiling of Clostridium difficile and Caco-2 cells during infection. *J Infect Dis* **202**, 282-290.

Lawley, T. D., Croucher, N. J., Yu, L., Clare, S., Sebaihia, M., Goulding, D., Pickard, D. J., Parkhill, J., Choudhary, J. & other authors (2009). Proteomic and genomic characterization of highly infectious Clostridium difficile 630 spores. *J Bacteriol* **191**, 5377-5386.

Lima, T., Auchincloss, A. H., Coudert, E., Keller, G., Michoud, K., Rivoire, C., Bulliard, V., de Castro, E., Lachaize, C. & other authors (2009). HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Res* **37**, D471-478.

Luo, C., Hu, G. Q. & Zhu, H. (2009). Genome reannotation of Escherichia coli CFT073 with new insights into virulence. *BMC Genomics* **10**, 552.

Marsden, G. L., Davis, I. J., Wright, V. J., Sebaihia, M., Kuijper, E. J. & Minton, N. P. (2010). Array comparative hybridisation reveals a high degree of similarity between UK and European clinical isolates of hypervirulent Clostridium difficile. *BMC Genomics* **11**, 389.

Miura, M., Kato, H. & Matsushita, O. (2010). A novel virulence factor *clostridium difficile* SRL modulates toxin B sensitivity of instestinal epithelial cells. *3rd ICDS abstract book*.

Petty, N. K. (2010). Genome annotation: man versus machine. *Nat Rev Microbiol* **8**, 762.

Ren, Q., Chen, K. & Paulsen, I. T. (2007). TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Res* **35**, D274-279.

Rupnik, M., Dupuy, B., Fairweather, N. F., Gerding, D. N., Johnson, S., Just, I., Lyerly, D. M., Popoff, M. R., Rood, J. I. & other authors (2005). Revised nomenclature of Clostridium difficile toxins and associated genes. *J Med Microbiol* **54**, 113-117.

Sebaihia, M., Wren, B. W., Mullany, P., Fairweather, N. F., Minton, N., Stabler, R., Thomson, N. R., Roberts, A. P., Cerdeno-Tarraga, A. M. & other authors (2006). The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome. *Nat Genet* **38**, 779--786.

Shahbabian, K., Jamalli, A., Zig, L. & Putzer, H. (2009). RNase Y, a novel endoribonuclease, initiates riboswitch turnover in Bacillus subtilis. *EMBO J* **28**, 3523-3533.

Sonnhammer, E. L., von Heijne, G. & Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* **6**, 175-182.

Stabler, R. A., He, M., Dawson, L., Martin, M., Valiente, E., Corton, C., Lawley, T. D., Sebaihia, M., Quail, M. A. & other authors (2009). Comparative genome and phenotypic analysis of Clostridium difficile 027 strains provides insight into the evolution of a hypervirulent bacterium. *Genome Biol* **10**, R102.

Twine, S. M., Reid, C. W., Aubry, A., McMullin, D. R., Fulton, K. M., Austin, J. & Logan, S. M. (2009). Motility and flagellar glycosylation in Clostridium difficile. *J Bacteriol* **191**, 7050-7062.

Vallenet, D., Labarre, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., Lajus, A., Pascal, G., Scarpelli, C. & other authors (2006). MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Res* **34**, 53--65.

Vallenet, D., Engelen, S., Mornico, D., Cruveiller, S., Fleury, L., Lajus, A., Rouy, Z., Roche, D., Salvignol, G. & other authors (2009). MicroScope: a platform for microbial genome annotation and comparative genomics. *Database (Oxford)* **2009**, bap021.

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S. C., Ester, M. & other authors (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* **26**, 1608-1615.

Zdobnov, E. M. & Apweiler, R. (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847-848.

Table 1

A

2006 \ 2010	**Known** 47%	**Putative** 37%	**Unknown** 14%	**Pseudo** 2%
Known 34%	1222	-63	-0	-10
Putative 47%	+518	1163	-115	-12
Unknown 17%	+18	+177	397	-7
Pseudo 2%	+1	+2	+1	63

B

New CDS	**Known**	**Putative**	**Unknown**	**Pseudo**
127	1	25	86	15

Table 2

Locus tag	Motif class	Associated enzyme	Propposed PTS metabolites
CD630_04690	Glucose	CD630_04680	Sucrose
CD630_03880	Glucose	CD630_03890	b-glucoside
CD630_30970	Glucose	CD630_30950 / CD630_30960	b-glucoside
CD630_31160	Glucose	CD630_31150	b-glucoside
CD630_31250	Glucose	CD630_31240	b-glucoside
CD630_31370	Glucose	CD630_31360	b-glucoside
CD630_26660 / CD630_26670	Glucose	-	Glucose
CD630_30580 / CD630_30610	Glucose	CD630_30600	a-glucoside
CD630_22690	Mannitol	CD630_22700	Fructose
CD630_30750	Mannitol	CD630_30740	Tagatose
CD630_30860	Mannitol	CD630_30850	2-O-a-mannosyl-D-glycerate
CD630_23320 / CD630_23330	Mannitol	CD630_23310	Mannitol
CD630_00410 / CD630_00420 / CD630_00430	Mannitol	-	Galactitol
CD630_36450 / CD630_36470 / CD630_36480	Lactose	-	Lichenan
CD630_28800 / CD630_28830 / CD630_28840	Lactose	CD630_28820	Cellobiose
CD630_30130 / CD630_30140 / CD630_3015	Mannose	CD630_30120	Mannose
CD630_25660 / CD630_25670 / CD630_25680	Mannose	CD630_25690	Mannose
CD630_30670 / CD630_30680 / CD630_30690 / CD630_30700	Mannose	CD630_30710	Xyloside
CD630_07640 / CD630_07650 / CD630_07660 / CD630_07670	Sorbitol	CD630_07680	Sorbitol