Journal of Medical Microbiology (2011), **60**, 1193–1199 **DOI** 10.1099/jmm.0.030452-0

Reannotation 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]

Correspondence
Marc Monot
mmonot@pasteur.fr

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

[2]Intégration et Analyse Génomiques, Institut Pasteur, 28 Rue du Dr Roux, F-75724 Paris Cedex 15, France

[3]CNRS UMR 8030, Université d'Evry & CEA, IG, Genoscope – LABGeM, 2 Rue Gaston Crémieux, CP 5706, F-91057 Evry Cedex, France

[4]Université Paris 7-Denis Diderot, 75205 Paris, France

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 *Clostridium difficile* strain 630, we manually reannotated each of the coding sequences (CDSs), using a high-level annotation platform. The functions of more than 500 genes annotated previously with putative functions were reannotated based on updated sequence similarities to proteins whose functions have been recently identified by experimental data from the literature. We also modified 222 CDS starts, detected 127 new CDSs and added the enzyme commission numbers, which were not supplied in the original annotation. In addition, an intensive project was undertaken to standardize the names of genes and gene products and thus harmonize as much as possible with the HAMAP project. The reannotation is stored in a relational database that will be available on the MicroScope web-based platform (https://www.genoscope.cns.fr/agc/microscope/mage/viewer.php?S_id=752&wwwpkgdb=a78e3466ad5db29aa8fe49e8812de8a7). The original submission stored in the (International Nucleotide Sequence Database Collaboration) INSDC nucleotide sequence databases was also updated.

Received 18 January 2011
Accepted 21 February 2011

INTRODUCTION

The reannotation of several model genomes has been recently performed, among these are *Escherichia coli* for the Gram-negative bacteria (Luo *et al.*, 2009) and *Bacillus subtilis* for the *Firmicutes* (Barbe *et al.*, 2009). This provided new information about genomic structure and organization as well as gene function, and plays an essential role in defining reference knowledge. In addition, the reannotation of the *B. subtilis* genome also benefits the study of other *Firmicutes*, such as *Clostridium difficile*.

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

The first complete genome sequence of a *C. difficile* strain (630) was sequenced in 2006 (Sebaihia *et al.*, 2006). It led to the development of high throughput projects such as comparative genomic, transcriptomic and proteomic studies (Jain *et al.*, 2010; Janvilisri *et al.*, 2010; Marsden *et al.*, 2010), which were recently reinforced with an increase of multiple genomic projects (Stabler *et al.*, 2009).

Abbreviations: CDS, coding sequence; HAMAP, high-quality automated and manual annotation of microbial proteomes; PMID, PubMed identifier; PTS, phosphoenolpyruvate-dependent phosphotransferase system; UPC, UniProt Consortium.

The GenBank/EMBL/DDBJ accession numbers for the reannotated sequence of *C. difficile* strain 630 and its plasmid pCD630 are AM180355 and AM180356, respectively.

Tables showing 2006 and 2010 annotations, and keywords relating to functional families, and a figure showing the MicroScope MaGe interface are available as supplementary material with the online version of this paper.

However, the relevance of all these experiments greatly depends on the information available for the genes, particularly their functions as experimentally identified or predicted *in silico*. Thus, it is critical that the information is accurate, relevant and useful. This is why we undertook the reannotation of the *C. difficile* strain 630 genome.

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

We describe in this paper the manual reannotation of all coding sequences (CDSs) 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), and *B. subtilis*, whose genome has been resequenced and reannotated (Barbe *et al.*, 2009). The reannotation resulted in the more accurate definition of the functions 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 through the MicroScope platform to maintain this up-to-date annotation.

METHODS

Identification of new or modified CDSs 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 reannotation 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. The MicroScope MaGe interface is also linked to numerous well-known biological databases and systems. The original gene prediction was systematically checked using 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 was interleaved, 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 three characters long. In addition the locus tag prefix must be separated from the tag value by an underscore ending with a number. So we assigned for all CDS a new locus tag code: 'CD630_00000'. The four-digit number after the underscore is still the original CDS position in the genome. The capital letter of the original identifier was converted to a number, which has been added at the end of each gene: 1 to 9 for gene codes previously ending with capital letter A to I, and 0 for all others e.g. CD0001 was converted into CD630_00010 and CD0163B was converted into CD630_01632. Finally, because the genomic position of the non-coding CDS was defined with only three-digit numbers, we replaced the first number after the locus tag prefix with a 't' or 'r' for tRNA and rRNA, respectively, e.g. CDt001 was changed to CD630_t0010 and CDr001 was changed to CD630_r0010. We used the same coding method for the 11 CDSs encoded by the plasmid pCD630, adding the letter 'p', after the locus tag e.g. CDP01 was changed to CD630_p010.

During the reannotation process using the AMIGene predictions, we identified new CDSs and we assigned them the locus tag of the previous CDS with the last number incremented by one, e.g. a new gene detected after CD630_02670 (previously named CD0267) was coded as CD630_02671. The original locus tag will be kept in the EMBL file using the '/Old_locus_tag' identifier.

Reannotation of the complete *C. difficile* strain 630. The predicted proteins were subjected to a wide range of bioinformatics tools, which includes conserved synteny computations, alignments against TrEMBL and Swiss-Prot databases (UPC, 2011) and TMHMM (Sonnhammer *et al.*, 1998), SignalP (Bendtsen *et al.*, 2004) and PsortB (Yu *et al.*, 2010) software to predict subcellular localization of proteins, as well as InterProScan (Zdobnov & Apweiler, 2001) to identify possible functions of newly discovered proteins (UPC, 2011). This work flow led to an automatic functional annotation for each CDS as previously described (Vallenet *et al.*, 2006). Finally, these pre-computed results served as the basis for the manual reannotation of each CDS.

To normalize the process of manual annotation among multiple users, we set up several guidelines as follows. (a) The product field is filled with the functional annotation for all genes identified with 'hypothetical protein' or 'conserved hypothetical protein' when the gene was not identified. For all others we added 'putative' prior to the product annotation. Pseudogene and gene remnant have a specific nomenclature: 'fragment of' + function + position (N-terminal, C-terminal or centre of the encoding protein). (b) The name of gene was completed by searching in the literature using PubMed data libraries (http://www.ncbi.nlm.nih.gov/pubmed) and when we changed gene names, old names were indicated in the synonymous field. (c) The start sites were modified according to the combination of the graphical data such as coding probability curves deduced from the AMIGene method (Bocs *et al.*, 2003), as well as alignments with orthologous genes (Altschul *et al.*, 1990). Then, the label '/START=' was added in the comment field followed by a capital letter associated with an informative code (M, modified; C, coding curve; S, sequence similarity; O, overlap; R, RBS). (d) PubMed identifiers (PMIDs) of each gene were classified from the specific references to the articles corresponding to orthologous genes and/or the global reviews concerning its function. (e) Protein families were standardized using the same keywords, PMIDs and global classification, such as CMR roles (http://cmr.jcvi.org/cgi-bin/CMR/RoleIds.cgi).

RESULTS AND DISCUSSION

Evaluation of annotation improvement

The original annotation of the *C. difficile* strain 630, published in 2006 (Sebaihia *et al.*, 2006), identified 3776 predicted CDSs. We have updated the annotation of all CDSs and assigned or defined more accurately their functions. During the reannotation process we attributed a class of function to each reannotated gene: (i) 'known' – when the

function was experimentally demonstrated or when a high level of similarities with characterized genes was found; (ii) 'putative' – based on a conserved motif, structural feature or limited similarities; (iii) 'unknown' – when genes were unidentified; and (iv) 'pseudo' – for pseudogenes or gene remnants. The same classification was applied manually to the 2006 annotation to allow comparison of both annotations (Table 1a).

Thus, 518 and 18 genes whose encoding function was previously described as putative and unknown, respectively, now have a functional annotation identified by experimental data from the literature (Table 1a). For example, CD630_26030 (previously named CD2603), recognized as a putative response regulator, is now designated CdtR, since it was shown that it controls the binary toxin expression in *C. difficile* (Carter *et al.*, 2007). In addition, 117 genes of unknown function have now a putative function. For instance, 12 conserved hypothetical proteins that contain a (clustered regularly interspaced short palindromic repeats) CRISPR-associated domain are annotated 'putative CRISPR-associated family protein'. Furthermore, we showed that the ATP synthase epsilon chain, CD630_34670 (CD3467), which was defined as a gene remnant (pseudo class) because of a lack of an amino-acid in the C terminus relative to database matches, actually belongs to the class of 'known function'. This enzyme usually combines ATP synthesis and hydrolysis but the hydrolysis function is still active in the truncated version (Ferguson *et al.*, 2006).

Following the reannotation we included 127 new CDSs and defined 222 new CDS start sites. The majority of the new CDSs are divided into putative (25), unknown (86) or pseudogene (15) classes (Table 1b). Only one gene, *CD630_15951* has an orthologue, whose function was experimentally demonstrated. This gene, detected during the proteomic analysis recently performed in *C. difficile* (Lawley *et al.*, 2009), is highly homologous (sequence identity) (~60 %) to a ferredoxin-encoding gene of *Clostridium thermoaceticum* (Elliott *et al.*, 1982).

We looked for papers corresponding to each gene, and particularly those published after the original annotation. We added at least one PMID reference number to 64 % of the *C. difficile* genes. Like many other genome-wide updates, several specificities were added to the original product function. When possible, we attached new motifs and enzymic domains identified by InterProScan, allowing a more accurate description of the original function. For example, putative peptidase enzymes now have family information according to the classification scheme of the MEROPS database (http://merops.sanger.ac.uk). The revised nomenclature of the pathogenicity locus region (Rupnik *et al.*, 2005) has been introduced during the reannotation process, as well as genes involved in *C. difficile* motility and flagellar glycosylation since they were recently published (Twine *et al.*, 2009). A locus tag, product annotation and class comparison between the two annotations performed in 2006 and 2010 are summarized in Supplementary Table S1 (available with the online journal). All information about the reannotated CDSs (Supplementary Fig. S1 available with the online journal), is currently available on the MicroScope platform: https://www.genoscope.cns.fr/agc/microscope/mage/viewer.php?S_id=752&wwwpkgdb=a78e3466ad5db29aa8fe49e8812de8a7. We also updated the *C. difficile* 630 genomic entry in the genomic databases EMBL, GenBank and DDBJ.

Table 1. Review of the 2006 annotation update

(a) CDSs were identified and separated according to the four major annotation classes in both 2006 and 2010 annotations: known, when function was experimentally demonstrated; putative, based on conserved motif, structural feature or limited homology; unknown, when function was unidentified; and pseudo, for pseudogenes. Numbers in bold indicate no change, and '+' and '−' correspond to a change between the classes of annotation between the 2006 annotation and the 2010 reannotation. (b) Annotation of the new CDS detected and referenced as known, putative, unknown and pseudo classes.

	2010 annotation			
(a) 2006 annotation	**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

Deciphering the annotation origin

Several pieces of information appeared when we evaluated the source used during the functional annotation of known, putative or unknown genes (Fig. 1). In the known category, 1 % of the gene function came from *C. difficile* strain 630 publications, 1.5 % from other *C. difficile* strains, 4 % from other clostridia and 93.5 % from other species (Fig. 1a). The putative category was defined according to the enzymic domain (40 %), homology to mobile elements (20 %) or cell localization (15 %) (Fig. 1b). As an example a gene will be annotated 'putative membrane protein' when three or more transmembrane helices were detected by TMHMM (Sonnhammer *et al.*, 1998). Finally, we classified the unknown genes from the alignment results with TrEMBL (Boeckmann *et al.*, 2003). Although 45 % were orphans of the *C. difficile* strains, 20 % were found in the genus *Clostridium*, 15 % in the phylum *Firmicutes* and 20 % in other bacteria (Fig. 1c).

Concerning genes annotated as known, we noted that only a few of them came from published clostridial experiments (Fig. 1a). This was mainly due to the lack of effective

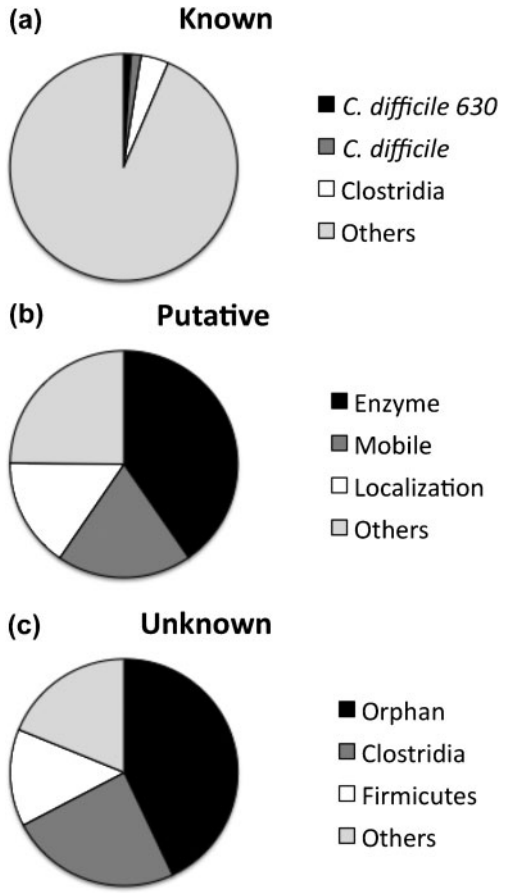

Fig. 1. Distribution of the origins of the functional reannotation. (a) Known functions were identified from the literature references of: *C. difficile* strain 630 (*C. difficile* 630), other *C. difficile* strains (*C. difficile*), *Clostridium* species (clostridia) and others species (others). (b) Putative functions were defined from: enzymic domains (enzyme), homology with mobile elements (mobile), localization in the cell (localization) and the other origins (others). (c) Unknown functions that were found only in: *C. difficile* (orphan), in the *Clostridium* species (clostridia), in the phylum *Firmicutes* (Firmicutes) or in diverse bacteria (others).

tools to mutate clostridial genes. However, the gene inactivation method and random mutagenesis system recently developed in *C. difficile* (Cartman & Minton, 2010) should greatly improve the number of publications on *C. difficile* gene functions. Half of the genes with an unidentified function, orphans, are found only in *C. difficile* 630 (Fig. 1c). However, most orphans are present in the *C. difficile* strains already sequenced such as strains 027, CD196 and R20291 (Stabler *et al.*, 2009). This may constitute a source of gene targets that could be used in research, diagnosis or treatment of *C. difficile*-associated diarrhoea.

Miscellaneous improvements

To reannotate the *C. difficile* genome of strain 630 we used the MaGe interface, which contains classic database fields (type, position, name, product, EC numbers) and several specific fields such as gene synonymous (synonyms), authors notes (comments), PubMed identifiers (PMID), product type, localization and functional classification (Supplementary Fig. S1 available with the online journal). All information found during the reannotation process that did not fit in the classic fields was added in the specific MaGe field or in the comments. For example, the novel virulence factor called Srl, for 'sensitivity regulation of *C. difficile* toxins' (Miura *et al.*, 2010), was presented during the Third International *Clostridium difficile* Symposium (held in Bled, Slovenia, 22–24 September 2010). This information has only been indicated in the comment field of the CD630_22980 (CD2298) gene until further validation.

The names of gene products were harmonized as much as possible with the HAMAP project (Lima *et al.*, 2009). However, all gene products have now been named with a specific keyword related to their functional family (Supplementary Table S2 available with the online journal). Thus, CD630_05310 (CD0531), previously annotated 'DeoR-like regulator of transcription' (a regulator of sugar and nucleoside metabolic systems) was reannotated 'transcriptional regulator (keyword), DeoR family'. We also normalized the annotation of genes that share the same characteristics. As an example, proteins that were only determined according to their membrane localization were annotated 'putative membrane protein'. The annotation standardization we used will facilitate the mining of the data using bioinformatics as well as by manual search (Supplementary Fig. S1 available with the online journal).

Membrane transport

The *C. difficile* genome contains a lot of proteins encoding several membrane transport systems: ATP-binding cassette (ABC) transporters, phosphoenolpyruvate-dependent phosphotransferase systems (PTSs), charged substrate transporters (antiporters, symporters) and facilitators. The general function of the genes encoding such proteins can be easily determined from bioinformatic approaches, like those used for the protein domain analysis in InterProScan (Zdobnov & Apweiler, 2001). However, it is quite difficult to distinguish the exact metabolite they transport, especially when the transport systems have a wide specificity. We reannotated most of the transporter systems by inference, including clues about targets using specialized databases such as TransportDB (http://www.membranetransport.org/) (Ren *et al.*, 2007), which compiles all information on cytoplasmic membrane transporters. We added a suffix in the classification that indicated, from a global trend to the expected target: the motif (family), the high sequence homology (like) and the evidence of a target metabolite (specific). However, this classification should be treated with caution since it was mainly deduced from *in silico* analysis rather than from experimental data.

Table 2 shows annotations of 19 PTSs with a specific metabolite suggestion. The targeted metabolite was deduced

Table 2. Reannotation of the PTSs according to the metabolite specificity

List of the locus tags corresponding to the 19 PTSs reannotated. The PTS metabolites were deduced from the motif class detection and/or the presence of associated enzymes involved in a specific sugar metabolism.

Locus tag	Motif class	Associated enzyme	Proposed PTS metabolite
CD630_04690	Glucose	CD630_04680	Sucrose
CD630_03880	Glucose	CD630_03890	β-Glucoside
CD630_30970	Glucose	CD630_30950 / CD630_30960	β-Glucoside
CD630_31160	Glucose	CD630_31150	β-Glucoside
CD630_31250	Glucose	CD630_31240	β-Glucoside
CD630_31370	Glucose	CD630_31360	β-Glucoside
CD630_26660/CD630_26670	Glucose	–	Glucose
CD630_30580/CD630_30610	Glucose	CD630_30600	α-Glucoside
CD630_22690	Mannitol	CD630_22700	Fructose
CD630_30750	Mannitol	CD630_30740	Tagatose
CD630_30860	Mannitol	CD630_30850	2-*O*-α-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_30150	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

from the InterProScan motif search but could also be defined by the presence in the same locus of a gene encoding an enzyme involved in specific sugar assimilation (associated enzyme). For example CD630_22690 (CD2269) is now annotated as 'PTS system, fructose-specific IIABC component'. This is due to the detection of three motif signatures, the mannitol family PTS EII component A, B and C, as well as the presence of the neighbouring gene, *CD630_22700* (*CD2270*), which encodes an enzyme involved in the utilization of fructose – 'fructose 1-phosphate kinase' as indicated in the gene annotation (Supplementary Table S1 available with the online journal).

Metabolism update

Updating the genome annotation of *C. difficile* led to many changes relating to the metabolism pathways. The gene cluster involved in the anaerobic oxidative degradation of L-ornithine has been identified in *C. sticklandii* (Fonknechten *et al.*, 2009). From this publication we reannotated genes *CD630_04420* (*CD0442*) to *CD630_04480* (*CD0448*) whose encoding proteins share high similarities to the ornithine catabolism compounds of *C. sticklandii*, Ord, OrtA, OrtB, OraS, OraE, Or-4 and Orr, respectively (Supplementary Table S1 available with the online journal). This suggested that *C. difficile* could produce acetyl-CoA from ornithine fermentation. The ability to use a variety of carbohydrates is an important feature of *C. difficile* for colonization of the host gut. *Enterococcus faecalis*, found in the same niche as *C. difficile*, provided hints to explore the consistency of a specific pathway required for ethanolamine utilization, a constituent of an abundant class of phospholipids present in eukaryotic cell membranes and the host's dietary intake (Del Papa & Perego, 2008; Fox *et al.*, 2009). Using the *E. faecalis* gene synteny and protein similarities, we were able to reconstruct the whole ethanolamine pathway in *C. difficile*, a cluster of 19 genes, from *CD630_19070* (*CD1907*) to *CD630_19250* (*CD1925*) inclusive, encoding the ethanolamine ammonia-lyase, an alcohol dehydrogenase, carboxysome associated proteins, the transporter EutH and the two-component system EutV/EutW (Supplementary Table S1 available with the online journal).

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

Conclusion

Finally, nearly half of the genes of the *C. difficile* strain 630 encode proteins with known function, whereas one-third of the gene products have a putative function and only 15 % of the proteins are of unknown function (Table 1a). In addition, 127 new CDSs were discovered (Table 1b) and 222 CDS starts were modified. The reannotation was performed using a high standard annotation MicroScope platform, which significantly increased the amount of information available for the majority of the CDSs, such as literature references, product types, localization and

synonymous genes (Supplementary Fig. S1 available with the online journal).

Nevertheless, there is still great deal of work to be completed since only 116 annotated genes came from published clostridial experiments. The EMBL entries are now resubmitted and to keep the annotation up-to-date, all new information should be addressed directly to Marc Monot (marc.monot@pasteur).

ACKNOWLEDGEMENTS

Special thanks to Mohamed Sebaihia (Sanger Institute, Cambridge, UK) who allowed us to update the original database entry (AM180355). We thank Richard Stabler (London School of Hygiene and Tropical Medicine, London, UK) who gave several comparison gene files between the reference genome and 027 strains. Special thanks to Ana Antunes (University of Siena, Italy); Sylvie Bouttier, Thomas Candela and Claire Janoir (Faculte de Pharmacie Paris XI, Chatenay Malabri, France); and Johann Peltier (Universite de Rouen, France) who provided help with specific gene annotations. We are grateful to Marc Griffiths and Erica Porter (Paris, France) for their help with English corrections. B. D., M. M., C. B.-E. and I. M. designed the study. C. B.-E. and M. M. carried out the major part of the manual reannotation of the genome together with M. T., and I. M.-V. D. V. and C. M. were involved in automatic reannotation and administration of the MicroScope platform. M. M., I. M.-V. and B. D. wrote the manuscript.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403–410.

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. & Médigue, 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** (database issue), D5–D12.

Cruveiller, S., Le Saux, J., Vallenet, D., Lajus, A., Bocs, S. & Médigue, C. (2005). MICheck: a web tool for fast checking of syntactic annotations of bacterial genomes. *Nucleic Acids Res* **33** (web server issue), W471–W479.

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., Bréchemmier-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 F_1F_0-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. & Dougan, G. (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** (database issue), D471–D478.

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 SRL modulates toxin B sensitivity of intestinal epithelial cells. In *3rd International Clostridium difficile Symposium, Bled, Slovenia, 22–24 September 2010*. Basel: European Society of Clinical Microbiology and Infectious Diseases.

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** (Database issue), D274–D279.

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., Cerdeño-Tárraga, 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.

UPC (2011). Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res* **39** (database issue), D214–D219.

Vallenet, D., Labarre, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., Lajus, A., Pascal, G., Scarpelli, C. & Médigue, C. (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.