Bioenerg. Res.
DOI 10.1007/s12155-014-9486-9

# New Insights into *Clostridia* Through Comparative Analyses of Their 40 Genomes

**Chuan Zhou · Qin Ma · Xizeng Mao · Bingqiang Liu · Yanbin Yin · Ying Xu**



**Abstract** The *Clostridium* genus of bacteria contains the most widely studied biofuel-producing organisms such as *Clostridium thermocellum* and also some human pathogens, plus a few less characterized strains. Here, we present a comparative genomic analysis of 40 fully sequenced clostridial genomes, paying a particular attention to the biomass degradation ones. Our analysis indicates that some of the *Clostridium botulinum* strains may have been incorrectly classified in the current taxonomy and hence should be renamed according to the 16S ribosomal RNA (rRNA) phylogeny. A core-genome analysis suggests that only 169 orthologous gene groups are shared by all the strains, and the strain-specific gene pool consists of 22,668 genes, which is consistent with the fact that these bacteria live in very diverse environments and have evolved a very large number of strain-specific genes to adapt to different environments. Across the 40 genomes, 1.4–5.8 % of genes fall into the carbohydrate active enzyme (CAZyme) families, and 20 out of the 40 genomes may encode cellulosomes with each genome having 1 to 76 genes bearing the cellulosome-related modules such as dockerins and cohesins. A phylogenetic footprinting analysis identified *cis*-regulatory motifs that are enriched in the promoters of the CAZyme genes, giving rise to 32 statistically significant motif candidates.

**Keywords** *Clostridium* · Comparative genomics · Pan-genome · Phylogeny · CAZyme · Motif

Chuan Zhou, Qin Ma, and Xizeng Mao contributed equally to this paper.

**Electronic supplementary material** The online version of this article (doi:10.1007/s12155-014-9486-9) contains supplementary material, which is available to authorized users.

C. Zhou · B. Liu
School of Mathematics, Shandong University, Jinan 250100, Shandong Province, China

Q. Ma · X. Mao · Y. Xu (✉)
Computational Systems Biology Laboratory, Department of Biochemistry and Molecular Biology and Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA
e-mail: xyn@bmb.uga.edu

Q. Ma · X. Mao · Y. Xu
BioEnergy Science Center, Oak Ridge, TN, USA

Y. Yin (✉)
Department of Biological Sciences, Northern Illinois University, Montgomery Hall 325A, DeKalb, IL 60115-2857, USA
e-mail: yyin@niu.edu

Y. Xu
College of Computer Science and Technology, Jilin University, Changchun, Jilin, China

## Introduction

Clostridium is a bacterial genus in the National Center for Biotechnology Information (NCBI)'s taxonomy database. Forty genomes of this genus have been fully sequenced as of April 2012. These bacteria, all being anaerobic *Firmicutes*, live in very different environments using diverse substrates and produce diverse metabolites [1]. Thirty-four of the 40 strains are either pathogens or biomass degraders, highlighting the economic importance of these strains, while the remaining six are less functionally characterized.

We are mainly interested in the biomass degraders among the 40 genomes since some of these bacteria are among the most promising microbes that can be used to produce the second generation biofuels with plant lignocellulosic biomass as the feedstock [2]. For example, *Clostridium thermocellum* (*C. thermocellum*) is a key model organism used for bioenergy research. It is a cellulolytic, thermophilic anaerobic bacterium

Published online: 03 June 2014

that encodes a large number of cellulases and hemicellulases [3]. Its genome also encodes cellulosomes, each being a large protein complex containing a long characteristic scaffold in protein that is bound by multiple glycolsyl hydrolases [4]. A systematic transcriptomic analysis of *C. thermocellum* has shown how the organism changes its gene expressions and reduces the growth rates under low substrate availability [5].As of today, only one study has been reported on a comparative analysis of 11 cellulosome-producing *Clostridia* [6]. Currently, 40 complete clostridial genomes have been sequenced and are publicly available, of which 13 (33 %) and 21 (53 %) have been reported to be biomass degraders [2, 7–11] and pathogens [12–18], respectively.

Here, we report the genomic differences among these 40 *Clostridia*, particularly between the biomass degrader group and the pathogen group derived through a comparative genomic analysis, including a pan-genome analysis [19]. The concept of pan-genome was originally defined as the set of all the (orthologous) genes encoded in any strain of a single species [20], while the core genome is defined as the gene set shared by all the stains of the species. Pan- and core-genome analyses have been found to be informative for understanding the big-picture evolutionary issues [20]. Here, we generalized the definition to the genus level in our comparative study. Our comparative genome analysis also addressed: (a) how are the 40 genomes phylogenetically related, (b) do biomass degraders share any common genomic features and how are they different from the other clostridial genomes, (c) do biomass degraders encode more biomass-related genes such as carbohydrate active enzymes (CAZymes), and (d) do they share any conserved regulatory motifs in their promoter regions? Answers to these questions can help to better understand the molecular mechanisms of biomass degradation and possibly guide the development of more efficient bacterial strains for bioenergy production.

## Methods and Materials

### Data

We retrieved 40 sequenced clostridial genomes from NCBI (http://www.ncbi.nlm.nih.gov/genome/browse/) as of April 2012, shown in Table 1, each given a unique ID, with biomass degraders numbered between 1 and 13, pathogens numbered between 14 and 34, and the remaining six numbered from 35 to 40. The operon information of the genomes needed for orthologous gene prediction is retrieved from the DOOR database [21, 22], which is needed for orthologous gene predictions. The annotated ribosomal RNA (rRNA) information is retrieved from NCBI, which is used in our phylogeny analysis.

### Genomic Information and Annotations of Clostridial Genomes

As part of our comparative genome analyses, the membrane proteins are identified using TMHMM [39], and the signal peptides are predicted using signal P [40], which is used for secretory protein prediction. CAZyme genes are predicted using dbCAN [41], and the CAZy database [42] is used as a benchmark to evaluate the dbCAN program.

### Pan-Genome Analysis

In order to determine if a gene is present in more than one genome, we need to identify the orthologous gene groups across the genomes. Overall the 40 genomes encode 141,710 protein genes. Orthologous gene groups are predicted across these genes using our own tool GOST [43] along with a Markovian clustering algorithm MCL [44]. We did a pan-genome analysis following Tellelin's method [45], based on the predicted orthologous groups. The method requires to randomly select $n$ genomes from the 40 and to estimate the size of the pan-genome based on the selected $n$ genomes and to do this repeatedly using a different combination of $n$ genomes each time to calculate the average of the estimated size. Then, do this for a different $n$ between 1 and 40 to get distributions for the pan-genome size as functions of $n$.

Specifically, we go through all $\binom{40}{n}$ combinations, the number of different combinations in selecting $n$ genomes out of 40, for each $n$ between 1 and 40, to calculate the average size as outlined above. For cases where $\binom{40}{n}$ is too large, we used 50,000 as the upper bound instead. Basically for each $n$ genome combination, we calculated the size of the estimated pan-genome. The average sizes were then calculated across all selected $n$ genome combinations for each $n$. Then, we derive an analytical function using a generic exponential function by fitting the averaged sizes across different $n$'s to derive size distribution functions of $n$, for the pan-genome. Based on whether the function value has a finite upper bound or not as $n$ increases, we predict the pan-genome is open or not.

We have done a similar analysis to estimate the core-genome size of the 40 *Clostridia* to check if the size function converges to zero or not as $n$ increases.

### *Cis*-Regulatory Motif Prediction and Analysis for CAZyme Genes

For each orthologous group of the CAZyme genes, we have retrieved their promoter sequences of up to 300 bps long in their original genomes on which we did de novo motif finding using BoBro [46] with parameters set at (*k*-3, *c*-1.00, *o*-10, *l*-14, *u*-0.70, *e*-3, *w*-2.00, *b*-0.95, and *N*-6), a tool that our lab developed. We then clustered all the predicted motifs into 32 groups, representing 32 motifs, based on their sequence

**Table 1** Forty clostridial strains with their unique IDs and isolation information

| | Species | Isolation |
|---|---|---|
| 1 | *Clostridium acetobutylicum* ATCC 824 | Garden soil in Connecticut [7, 23] |
| 2 | *Clostridium acetobutylicum* DSM 1731 | Industrial fermentations; German Collection of Microorganisms and Cell Cultures [24] |
| 3 | *Clostridium acetobutylicum* EA 2018 | Soil [25] |
| 4 | *Clostridium beijerinckii* NCIMB 8052 | Soil [26] |
| 5 | *Clostridium cellulolyticum* H10 | Decayed grass compost [27] |
| 6 | *Clostridium cellulovorans* 743B | Wood chip pile [28] |
| 7 | *Clostridium clariflavum* DSM 19732 | Anaerobic sludge of a cellulose-degrading methanogenic bioreactor [29] |
| 8 | *Clostridium phytofermentans* ISDg | Forest soil near the Quabbin Reservoir in Massachusetts |
| 9 | *Clostridium saccharolyticum* WM1 | Methanogenic cellulose-enriched sewage sludge [30] |
| 10 | *Clostridium* sp. BNL1100 | Corn stover enrichment culture [31] |
| 11 | *Clostridium thermocellum* ATCC 27405 | Compost soil [32] |
| 12 | *Clostridium thermocellum* DSM 1313 | Compost soil [32] |
| 13 | *Clostridium lentocellum* DSM 5427 | Estuarine sediment of a river that received both domestic and paper mill waste [9] |
| 14 | *Clostridium botulinum* A str. ATCC *19397* | Laboratory strain probably from foodborne botulism cases in the Western USA [33] |
| 15 | *Clostridium botulinum* A str. ATCC 3502 | Canned peas, California [34] |
| 16 | *Clostridium botulinum* A str. Hall | Canned peas, California [34] |
| 17 | *Clostridium botulinum* A2 str. Kyoto | Infant botulism in Kyoto, Japan, in 1978 |
| 18 | *Clostridium botulinum* A3 str. Loch Maree | Duck liver paste during a botulism outbreak at a hotel in the Scottish highlands in 1922 |
| 19 | *Clostridium botulinum* B str. Eklund 17B | Marine sediments taken off the coast of Washington, USA |
| 20 | *Clostridium botulinum* B1 str. Okra | Foodborne botulism incident in the USA |
| 21 | *Clostridium botulinum* Ba4 str. 657 | Infant botulism case in Texas in 1976 [35] |
| 22 | *Clostridium botulinum* BKT015925 | Outbreak in a Swedish poultry farm |
| 23 | *Clostridium botulinum* E3 str. Alaska E43 | Salmon eggs associated with a foodborne case of botulism in Alaska |
| 24 | *Clostridium botulinum* F str. 230613 | Information not available |
| 25 | *Clostridium botulinum* F str. Langeland | Home-prepared liver paste involved in an outbreak of foodborne botulism on the island of Langeland in Denmark in 1958 |
| 26 | *Clostridium botulinum* H04402 065 | Botulism patient in the UK in 2004 [36] |
| 27 | *Clostridium difficile* 630 | Clinical isolate Switzerland [34] |
| 28 | *Clostridium difficile* BI1 | Human strain isolated in the USA in 1988 [37] |
| 29 | *Clostridium difficile* CD196 | A nonepidemic strain isolated from a patient with PMC in Paris, 1985 [37] |
| 30 | *Clostridium difficile* R20291 | Stoke Mandeville Hospital, UK, in 2006 [37] |
| 31 | *Clostridium tetani* E88 | Natural habitat is soil, dust, and intestinal tracts of various animals; a variant of strain Massachusetts [18] |
| 32 | *Clostridium perfringens* ATCC 13124 | Commonly found in soil, sediments, and the human gastrointestinal tract [17] |
| 33 | *Clostridium perfringens* SM101 | Commonly found in soil, sediments, and the human gastrointestinal tract [17] |
| 34 | *Clostridium perfringens* str. 13 | Commonly found in soil, sediments, and the human gastrointestinal tract [17] |
| 35 | *Clostridium* sp. SY8519 | Isolated from adult human intestine [12] |
| 36 | *Clostridium sticklandii* DSM 519 | Mud water [34] |
| 37 | *Clostridium kluyveri* DSM 555 | Mud, fresh water [34] |
| 38 | *Clostridium kluyveri* NBRC 12016 | Mud, fresh water [34] |
| 39 | *Clostridium novyi* NT | Soil and feces [38] |
| 40 | *Clostridium ljungdahlii* DSM 13528 | Chicken yard waste |

similarities, using a similarity score cutoff =0.55 [47]. To scan a motif across the 40 clostridial genomes for finding additional motif instances, we used the motif scanning tool BBS, which is part of the BoBro package [47] with parameters (*t*-0.95, *s*-5, *e*-1, *c*-0, *u*-0.95, and *n*-0.50). Specifically, consider a set of promoter sequences $P$ for motif scanning in a genome $G$. We calculated (i) the $p$ value for each predicted motif using a null hypothesis that it appears in $\boldsymbol{P}$ by chance (so the smaller a $p$ value, the less likely it is found by chance). Specifically, a motif is considered statistically significant if its $p$ value <3.3e–

5, which has been corrected for multiple testing by the Bonferroni correction based on the estimated (average) number, 300, of transcription factors in each clostridial genome and (ii) an enrichment score of the motif occurring in *P* against in *G* (so the higher the score, the more enriched the motif is in *P* than in the general background *G*) using BBS. This can help to refine the motif prediction.

### Pathway Enrichment Analysis

We have carried out pathway enrichment analyses on the identified orthologous gene groups against pathways in the KEGG database using the KOBAS 2 server [48, 49] on both the biomass degrader group and the pathogen group. The two sets of enriched pathways for the two groups are then statistically compared using a Fisher's exact test. In addition, the FDR control given in the *multtest* R package was applied to correct for the *p* values derived through multiple comparisons [50].

## Results and Discussion

### Phylogenetic Analysis of 40 Clostridial Genomes

We built a 16S rRNA tree over the 40 genomes (see Table 1) and labeled the biomass degraders (red), pathogens (blue), and the remaining (black) in Fig. 1. It is clear from the figure that the biomass degraders and pathogens are not strictly separable in the tree. Eight biomass degraders are clustered together with four strains of *Clostridium difficile* and two other bacteria (cluster I), while five other biomass degraders are clustered together with all the other 17 pathogens and four other *Clostridia* (cluster II). In addition, three *C. botulinum* strains (labeled with blue stars) are not clustered with other *C. botulinum* strains, hence suggesting that these three *C. botulinum* strains should be renamed according to the 16S rRNA phylogeny, which has been observed by other researchers [16, 51].

Meanwhile, we have also labeled the biomass degraders predicted to encode cellulosomes (red stars) if dockerins and cohesins were found in their genomes. Particularly, the scaffold in protein (CipA) in *Clostridium acetobutylicum* [6] is encoded by a pseudo gene, and thus, it may not necessarily make cellulosomes. The cellulosome-encoding bacteria are present in both clusters, defined above. In cluster I, seven out of the eight biomass degraders encode cellulosomes, while in cluster II, four out of five biomass degraders encode cellulosomes. This suggests that the cellulosomal structure may have developed multiple times during evolution, possibly by independent gene transfers into different bacteria.

It should be noted that a 16S rRNA may have multiple copies in a genome, which may differ in their sequences [43]. To take this into consideration, we have constructed a tree containing all the 16S rRNAs in each of the 40 genomes to evaluate the impact of the 16S rRNA gene as a molecular

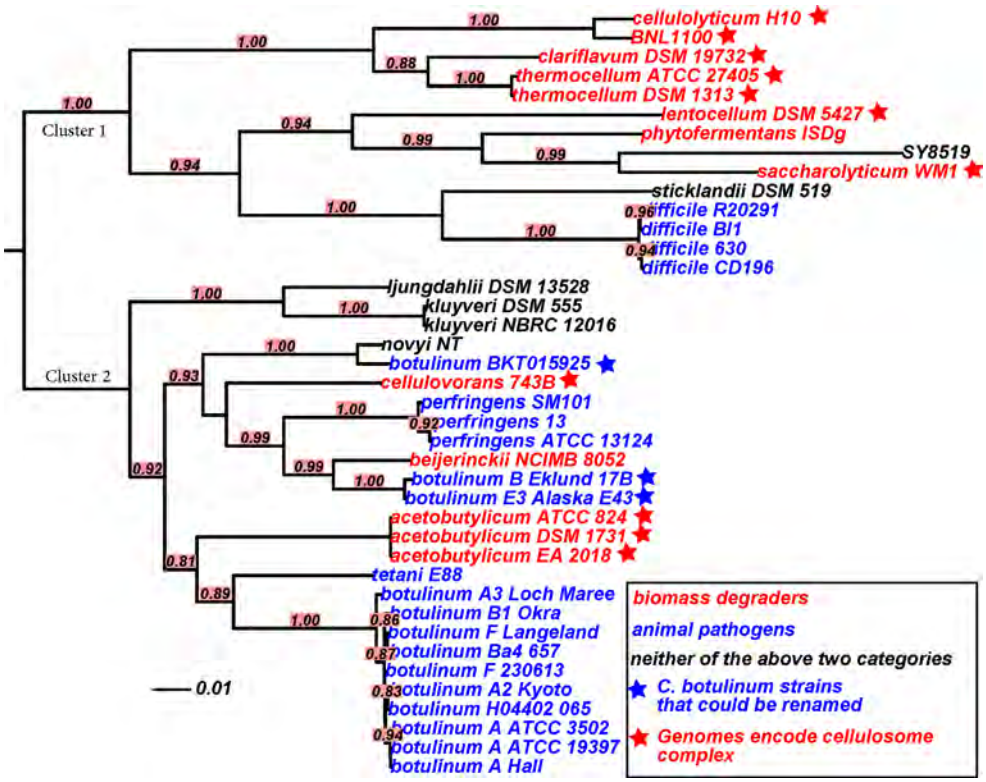



**Fig. 1** A 16S rRNA tree of the 40 clostridial genomes. The biomass degraders are highlighted in *red* and the animal pathogens are in *blue*. The *three blue genomes followed by a star* should be considered for renaming based on this species tree

marker. The new phylogeny is largely in agreement with the one in Fig. 1, except that five 16S rRNAs from two genomes are stray nodes (see Supplementary Fig. S1). We have also looked into the phylogenetic relationship among the 40 genomes based on 31 conserved protein families [52] (Supplementary Fig. S2). The topologies in Fig. 1 and Fig. S2 are basically in agreement.

## Comparison of General Features Between Biomass Degraders and Pathogens

We have compared some general features between the biomass degraders and the pathogens (see Supplementary Table S1). The average genome size of the biomass degraders is 4.5 Mbps (with median of 4.6 Mbps and interquartile range (IQR) of 0.69 Mbps), which is larger than that of the pathogens at 3.8 Mbps (with median of 3.92 Mbps and IQR of 0.5 Mbps). Similar was observed between the number of genes, 4,095 (with median of 4,085 and IQR of 342) versus 3,593 (with median of 3,699 and IQR of 574) on average, the G + C content at 34.9 % (with median of 34.8 % and IQR of 6.43 %) versus 28.3 % (with median of 28.2 % and IQR of 0.36 %), and the coding density at 84.2 % (with median of 84.8 % and IQR of 4.06 %) versus 81.5 % (with median of 81.4 % and IQR of 1.1 %), between these two classes, as shown in Fig. 2. The *p* values of a nonparametric Wilcoxon rank-sum test for the above comparisons are all less than 0.01; hence, the difference is statistically significant between these two classes. These make sense as bacteria tend to lose genes and DNAs (so called *genome reduction*), reduce GC content, and increase the content of noncoding repetitive elements when transforming from free living to parasitic [53–55] like the pathogen group here. In addition, the biomass degraders, overall, have significantly higher percentages of secretory proteins than the pathogens. This is not surprising as biomass degraders encode a large number of biomass-degrading enzymes that need be transported out of the cells.

Knowing that the 34 pathogens and biomass degrader genomes fall into 14 species, ten for biomass degraders, and four for pathogens, we have carried out a similar study to the above but focused on differences between the two groups of *Clostridia* within each species, basically to examine the variations of each group across different species. Table 2 summarizes the analysis results across the 14 species.

## The Core- and Pan-Genome Analysis of the 40 Clostridial Genomes

The 141,710 protein genes encoded in the 40 genomes fall into 41,757 orthologous gene groups using our prediction method. Based on the predicted orthologous genes groups, we have carried out core- and pan-genome analyses. They can be estimated based on the currently available genomes and how the estimated pan- and core-genome sizes change as more genomes are included in the estimation, essentially to check if the estimated pan-genome size converges and the core-genome size converges to a positive value [45].

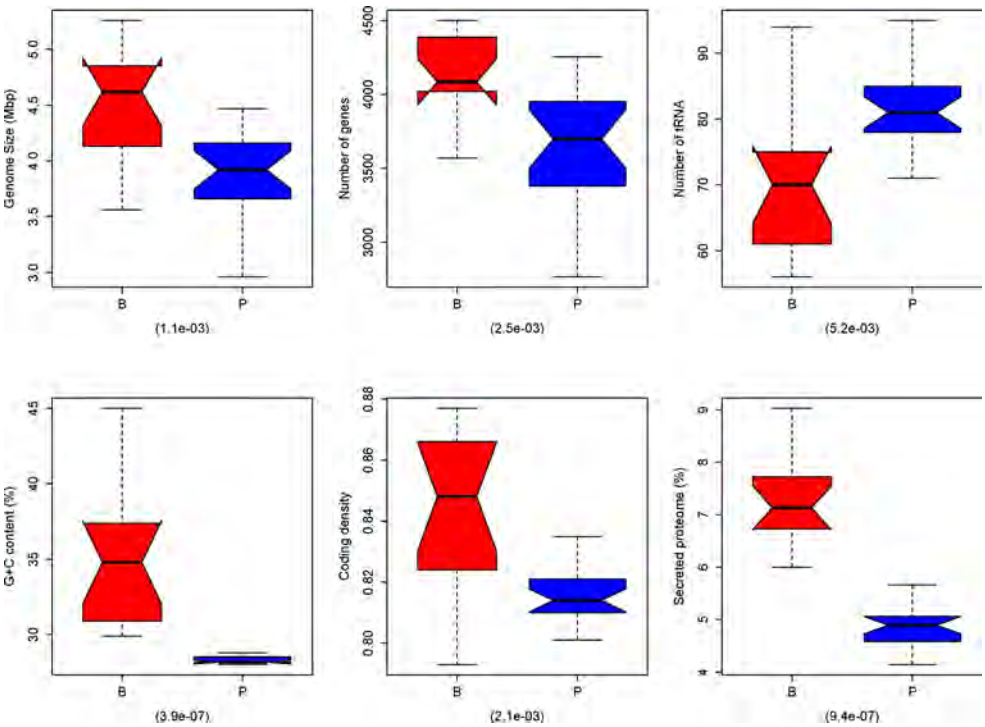



**Fig. 2** A comparison of general features between biomass degraders (*B*) and animal pathogens (*P*). The *p* values of nonparametric Wilcox tests are shown in *brackets* (all *p* values <0.01)

**Table 2** Summary of species-dependent comparisons between the two groups

| Species name | Genome size (Mbps) | G + C content (%) | Coding density (%) |
|---|---|---|---|
| ***Clostridium acetobutylicum*** | 4.14 | 30.9 | 86.3 |
| ***Clostridium beijerinckii*** | 6.00 | 29.9 | 79.4 |
| ***Clostridium cellulolyticum*** | 4.07 | 37.4 | 86.6 |
| ***Clostridium cellulovorans*** | 5.26 | 31.2 | 82.4 |
| ***Clostridium clariflavum*** | 4.90 | 35.7 | 79.3 |
| ***Clostridium phytofermentans*** | 4.85 | 35.3 | 81.9 |
| ***Clostridium saccharolyticum*** | 4.66 | 45.0 | 87.0 |
| ***Clostridium* sp. BNL1100** | 4.61 | 37.3 | 87.7 |
| ***Clostridium thermocellum*** | 3.70 | 39.1 | 83.0 |
| ***Clostridium lentocellum*** | 4.71 | 34.3 | 85.2 |
| *Clostridium botulinum* | 3.92 | 28.1 | 80.8 |
| *Clostridium difficile* | 4.27 | 28.8 | 81.8 |
| *Clostridium tetani* | 2.87 | 28.6 | 85.4 |
| *Clostridium perfringens* | 3.10 | 28.4 | 82.6 |

The species highlighted in bold are biomass degraders, and the others are animal pathogens. Each number is calculated as the average value among all strains in the corresponding species

We found that the core genome among the 40 contains 169 orthologous gene groups covering 6,760 genes, and 22,668 out of the remaining 134,950 genes are predicted to be strain-specific genes. Figure 3 shows the occurrence frequencies of all the orthologous gene groups across all the 40 genomes.

We have derived an analytic function, $P(n)$, to estimate the pan-genome size, based on a least-square fit [45] of the size distribution data derived from the 40 genomes (see "Methods and Materials" section). The function is given as follows:

$$P(n) = D + \mathrm{tg}(c)(n-1) + k * \exp[-2/t]\frac{1-\exp[-(n-1)/t]}{1-\exp\frac{-1}{t}}$$

where $k$=1,241.88, $t$=109.11, $D$=3,329.73, and tg(c))= 672.42.

Clearly, $P(n)$ is unbounded as $n$ increases, hence suggesting that the pan-genome is still open. A similar analytic function of the number, $n$, of genomes has been derived for the core genome (see Supplementary Fig. S3 for details).

The pan-genome of the 13 biomass degraders contains 23,538 orthologous gene groups, and the remaining 27 genomes have 23,177 orthologous gene groups, 4,958 of which are shared between the two sets of genomes (see Fig. 4). We noted that the most enriched biological functions/pathways by the 169 core genes are related to RNA binding and ribosomal proteins (see Supplementary Table S2) [56]. The detailed pathway enrichment analysis is given in the following section.

## Pathway Enrichment Analysis

We have analyzed the pathway information for the biomass degrader and the pathogen groups, separately. We found that 179 KEGG pathways are enriched by 18,580 biomass degrader-specific orthologous groups, while 165 pathways are enriched by 18,219 orthologous groups in the other genomes, averaging 13.8 versus 6.1 enriched pathways per genome, suggesting that biomass degraders encode more diverse pathways possibly due to their more diverse living environments. We then carried out pathway-enrichment analysis on the biomass degraders against the pathogen group. We found that the top five pathways enriched in the biomass degrader group are the two-component system, starch and sucrose metabolism, bacterial chemotaxis, cyano amino acid metabolism, and pentose and glucuronate interconversions. This is in agreement with the fact that the biomass degraders need to respond to more complex environments and metabolize more sugar sources. Further details of the enriched pathways can be found in Supplementary Dataset S1. The functions of the strain-specific genes across the 40 genomes, especially the 13 biomass degraders, can be found in Supplementary Table S3.

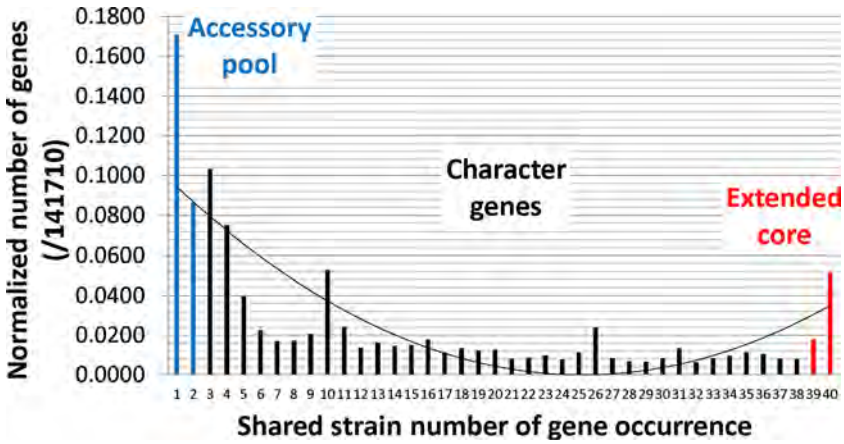



**Fig. 3** Occurrence frequencies of 141,710 genes in the 40 clostridial strains

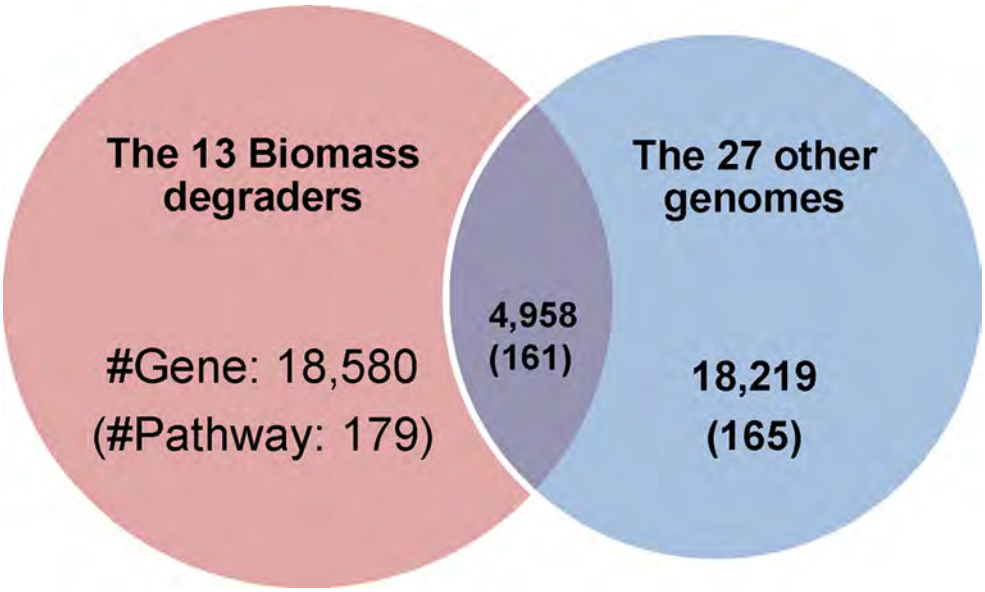



**Fig. 4** A comparison between the numbers of genes in the 13 biomass degraders and the remaining 27 genomes. The pan-genome of biomass degraders is denoted by the *red circle* on the *left*, and the *blue circle* represents the pan-genome of the remaining 27 genomes. The numbers of KEGG pathways are shown in *brackets*

## CAZyme Gene Identification and Analysis

CAZyme genes in each clostridial genome are identified by using the dbCAN server [41], which has an average prediction accuracy at 88.4 % measured against the annotation of the CAZy database [42] (Supplementary Table S4). We found that (i) 1.4–5.8 % of the genes of the 40 genomes fall into CAZyme families and (ii) 20 out of the 40 genomes contain 1 to 76 genes bearing the cellulosome-related modules, dockerins and cohesins, and those genomes known to encode cellulosomes have the highest numbers of proteins having these two modules. It is clear from Fig. 5 that the percentages of the CAZymes and cellulosomal genes are substantially higher in biomass degraders than those in the other genomes. Especially, no or very few genes encoding the cellulosome-related modules are found in the pathogens, suggesting the possibility to use genes bearing the cellulosomal modules as a marker for classification of the clostridial genomes into two groups. However, there are exceptions: genomes 4 and 8 in Fig. 5a have only a few genes. This is because cellulosomes are not required for biomass degraders and many bacteria are known to use free enzymes to degrade biomasses. This is consistent with the data shown in Fig. 5b where genomes 4 and 8 have high percentages of CAZyme genes, like in other genomes of the same group. For example, genome 8 (*Clostridium phytofermentans* ISDg) is well known for its ability to degrade cellulose using free enzymes to produce biofuels.

We then put all the identified CAZyme genes and genes containing cellulosomal modules into three pools: 1,045 orthologous groups found only in biomass degraders (A), 478 orthologous groups found only in the other genomes (B), and 130 orthologous groups found in both biomass degraders and the other genomes (C). The detailed distribution of CAZyme domains and cellulosomal modules in these three pools is shown in Table 3. Clearly, the number of orthologous groups in A is larger than those in B, suggesting that CAZyme genes play more important roles in biomass degraders. By comparing A and B, we have found that the % glycoside hydrolases (GH), % carbohydrate-binding modules (CBM), and % polysaccharide lyases (PL) are much higher in A than in B, consistent with the fact that biomass degraders have a greater need to recognize and degrade a larger variety of complex carbohydrates. The higher % glycosyl transferases (GT) in B and C could be due to the fact that GT are used to build polysaccharides, glycoproteins, and other glycol conjugates, which are needed by all bacteria. We have further checked the distribution of different CAZyme domains in the largest five orthologous groups in A, B, and C, respectively, a total of 15. The largest orthologous groups in A are enriched with domains CBM and GH, while the most enriched domain in B is GT. The only one orthologous group enriched with domain CE appears in C (see Supplementary Table S5).

## Prediction of *Cis*-Regulatory Motifs for CAZyme Genes

We have predicted, using BoBro, 378, 273, and 539 *cis*-regulatory motifs enriched in the three pools of orthologous groups: A, B, and C, respectively, given in Supplementary Dataset S2. We found that some motifs are specific to biomass degraders while some others are specific to nonbiomass degraders. All the motifs are grouped into 32 distinct and significant patterns using BBC [47], whose logos are provided in Supplementary Table S6 [57].

Figure 6 shows a heat map for the 32 predicted motifs. Clearly, there are three motif clusters (MCs) in the heat map corresponding to three groups of genomes in the *Clostridium* genus: MC-1, the seven uppermost MCs, encapsulated inside the blue box, which have an overall trend being more

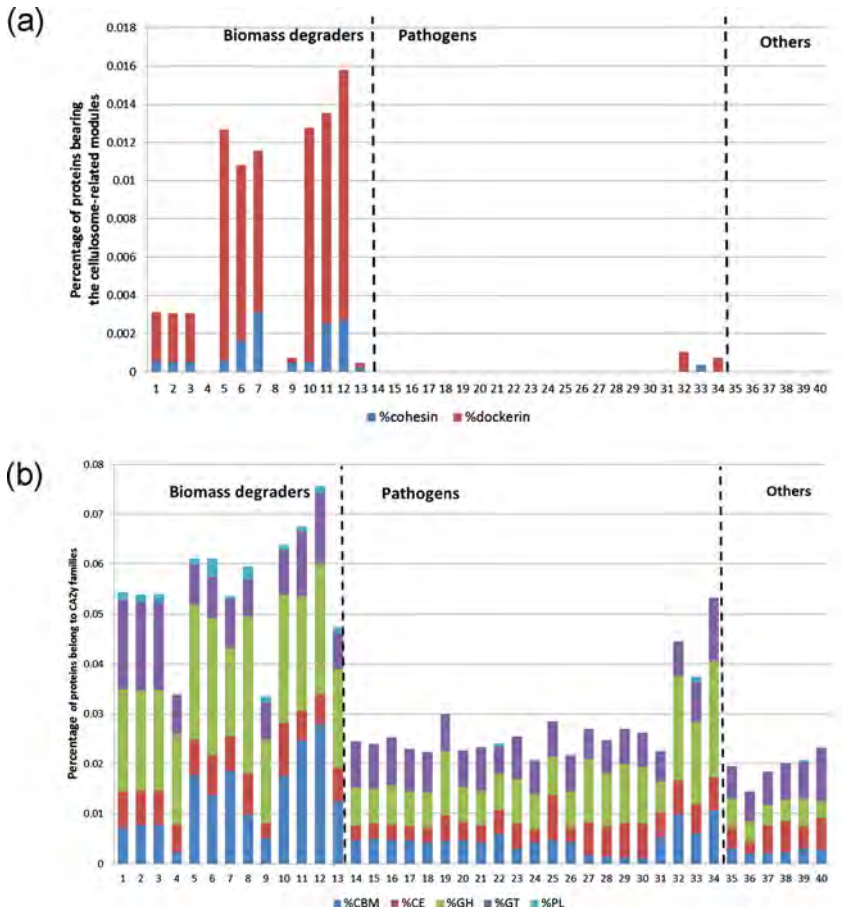



**Fig. 5** **a** The percentage of proteins bearing cellulosome-related modules (cohesin and dockerin). **b** The percentages of proteins in CAZyme families across the 40 genomes, along with five CAZyme domain types: carbohydrate-binding modules (*CBM*), carbohydrate esterases (*CE*), glycoside hydrolases (*GH*), glycosyl transferases (*GT*), and polysaccharide lyases (*PL*)

significant in nonbiomass degraders than in biomass degraders; MC-2, the middle eight MCs inside the red box, which are more significant in biomass degraders than the others (see Table 4 and Supplementary Dataset S2 for downstream CAZyme genes); and MC-3, the bottom 17 MCs, inside the black box, which are significant in all the 40 genomes.

We have compared the 32 predicted motifs to known prokaryotic motif databases, RegTransBase [58] and PRODORIC [59], using MEME [60, 61]. Overall, 29 of the 32 predicted motifs are found in at least one database. Out of these experimentally validated motifs, 7 are in MC-1, 6 in

**Table 3** Distributions of CAZyme domains and cellulosome-related modules in the three gene families, A, B, and C

| Category<br>Total no. of orthologous groups | A<br>1,045 | B<br>478 | C<br>130 |
|---|---|---|---|
| %Cohesin | 3.1 | 0.21 | 0.0 |
| %Dockerin | 10.1 | 0.63 | 0.0 |
| %CBM | 31.3 | 22.0 | 13.1 |
| %GH | 42.7 | 28.9 | 42.3 |
| %GT | 24.5 | 27.0 | 36.9 |
| %CE | 17.6 | 17.0 | 18.4 |
| %PL | 4.3 | 1.5 | 0.0 |

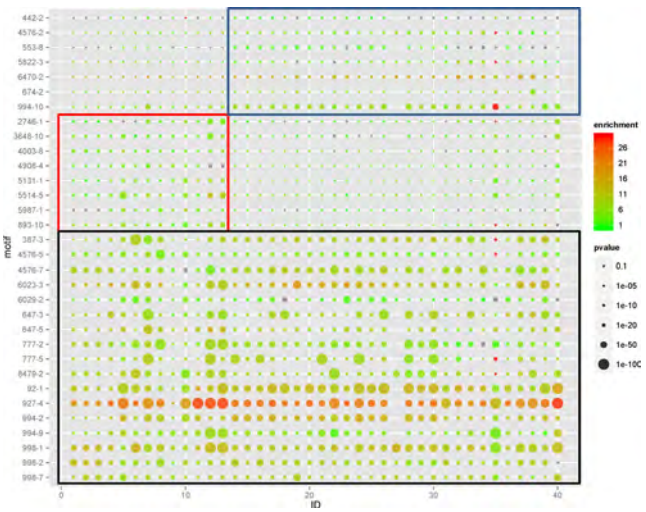
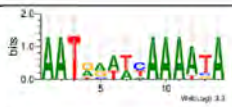
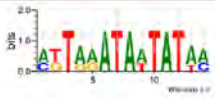
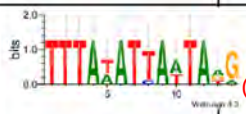
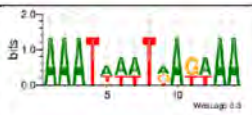
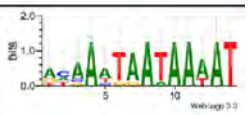
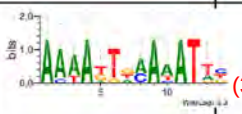



**Fig. 6** A heat-map for predicted motifs of the CAZyme genes in 40 clostridial genomes, with 32 rows (representing 32 motif clusters) and 40 columns (representing 40 genomes). The *size* and *color* of each *solid circle* represent the statistical significance of each predicted motif, measured using the *p* values and enrichment scores, the calculation of which can be found in "Methods and Materials" section. Specifically, the lower a *p* value is, the larger the corresponding circle is, and the enrichment score of a motif increases along with the color changing from green to red

**Table 4** The logos of eight motifs (MC-2) with high statistical significance in biomass degrader genomes, with *p* values under each ID, all less than 1.0e–5

| ID | LOGO | ID | LOGO | ID | LOGO |
|---|---|---|---|---|---|
| 2746-1 (1.6e-15) |  | 3468-10 (1.3e-19) |  | 4003-8 (4.9e-9) |  |
| 4908-4 (7.0e-6) |  | 5131-1 (3.8e-7) |  | 5514-5 (6.9e-15) |  |
| 5987-1 (2.5e-10) |  | 893-10 (3.0e-11) |  | | |

MC-2, and 16 in MC-3. Interestingly, motifs for the nucleoid-associated proteins, ArcA and IHF, are only in MC-2, suggesting that the regulation of the corresponding genes in biomass degraders may be done through chromosomal folding and refolding [62, 63]. ExuR in MC-2 is previously known to be involved in transport and catabolism of galacturonate and glucuronate [64, 65], which is important in biomass degradation. More details can be found in Table 5 and Supplementary Dataset S2.

## Summary of New Findings

We have gained new insights about *Clostridia* through a comparative genomic analysis, particularly in terms of the genomic-level differences between the 13 biomass degraders and the other genomes of 40 sequenced genomes.

The current classification of the *C. botulinum* strains is based on if they encode botulinum toxin-producing genes, but not on 16S rRNA sequence similarities, which is more accurate and used most widely in determining the taxonomical positions of bacterial species. Our 16S rRNA-based phylogenic analyses strongly suggest that three *C. botulinum* strains (genome nos. 19, 22, and 23) should be renamed since they are in different locations from the other *C. botulinum* strains and they also differ from the other *C. botulinum* strains in general genome features such as genome size, GC content, and the number of genes (Supplementary Fig. S4).

Our comparative analyses revealed that biomass degraders tend to have larger genome sizes, more genes, higher GC contents, and higher percentages of secretory proteins than the other clostridial strains. In addition, biomass degraders also have larger pan-genomes and encode more metabolic pathways. We believe that all these reflect the different lifestyles and the complexities of their living environments. Specifically, as shown in Table 1, the 13 biomass degraders are mostly isolated from soils and biomass composts, whereas the 21 pathogenic bacteria are host associated. Soils and biomass composts are environmentally harsher than host-associated niches such as intestinal tracts. Thus, biomass degraders need more genes and metabolic pathways to process their foods: complex carbohydrate molecules. Due to the same reason, biomass degraders tend to encode more CAZyme proteins and cellulosome-related modules than the other genomes. Future studies will be needed to test if this is a general feature of other bacteria that use plant biomass as their main nutrient sources. Lastly, eight potentially important regulatory motifs are found to be conserved in the promoter regions of CAZyme genes that are unique to biomass degraders. These motifs are worth further experimental investigating, as they might be good targets for improving biofuel production. Overall, we believe that our analyses could provide useful information to mechanism studies of biomass degradation and possibly designing more efficient biomass degraders.

**Table 5** Overlapping results between the 32 predicted motifs and known motifs in the RegTransBase and PRODORIC databases, all with *p* values less than 0.01 (calculated by MEME)

| Motif clusters | TFs in RegTransBase | TFs in PRODORIC |
|---|---|---|
| MC-1 | NagC, XylR, and VP2396 | Ada, XylR, and CaiF |
| MC-2 | ArcA, ExuR, CscR, and RL4253 | IHF, SigB, Fur, AhrC, and CodY |
| MC-3 | ModE, Fnr, MtaR, NagC, VP2396, IscR, TnrA, Atu4556, and YP00846 | CovR, LexA, TnrA, Hpr, GlnR, TyrR, YhiX, CodY, NagC, PhoP, GltC, and resD |

**Acknowledgments** This research was supported in part by the National Science Foundation (#NSF DEB-0830024 and NSF MCB-0958172), the US Department of Energy's BioEnergy Science Center (BESC) grant through the Office of Biological and Environmental Research, and National Science Foundation of China (NSFC 61272016 and 61303084). The BioEnergy Science Center is a US Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. Funding for open access charge was provided by the US Department of Energy's BioEnergy Science Center (BESC).

**Author Contribution** Y.Y. and Y.X. conceived the basic idea and planned the project. Q.M. and C.Z. carried out the experiments and analyzed the data. X.M. did the pathway enrichment analysis and proposed good suggestions to interpret the data in the view of biology. All authors edited the manuscript and approved the final manuscript. Q.M., C.Z., and X.M. contributed equally to this paper.

## References

1. Tracy BP, Jones SW, Fast AG, Indurthi DC, Papoutsakis ET (2011) Clostridia: the importance of their exceptional substrate and metabolite diversity for biofuel and biorefinery applications. Curr Opin Biotechnol 23(3):364–381
2. Hemme CL, Mouttaki H, Lee YJ, Zhang G, Goodwin L, Lucas S, Copeland A, Lapidus A, Glavina del Rio T, Tice H, Saunders E, Brettin T, Detter JC, Han CS, Pitluck S, Land ML, Hauser LJ, Kyrpides N, Mikhailova N, He Z, Wu L, Van Nostrand JD, Henrissat B, He Q, Lawson PA, Tanner RS, Lynd LR, Wiegel J, Fields MW, Arkin AP, Schadt CW, Stevenson BS, McInerney MJ, Yang Y, Dong H, Xing D, Ren N, Wang A, Huhnke RL, Mielenz JR, Ding SY, Himmel ME, Taghavi S, van der Lelie D, Rubin EM, Zhou J (2010) Sequencing of multiple clostridial genomes related to biomass conversion and biofuel production. J Bacteriol 192(24):6494–6496. doi:10.1128/JB.01064-10
3. Demain AL, Newcomb M, Wu JH (2005) Cellulase, clostridia, and ethanol. Microbiol Mol Biol Rev 69(1):124–154
4. Bayer EA, Lamed R, White BA, Flint HJ (2008) From cellulosomes to cellulosomics. Chem Rec 8(6):364–377. doi:10.1002/tcr.20160
5. Raman B, McKeown CK, Rodriguez M Jr, Brown SD, Mielenz JR (2011) Transcriptomic analysis of *Clostridium thermocellum* ATCC 27405 cellulose fermentation. BMC Microbiol 11:134. doi:10.1186/1471-2180-11-134
6. Tamaru Y, Miyake H, Kuroda K, Nakanishi A, Matsushima C, Doi RH, Ueda M (2011) Comparison of the mesophilic cellulosome-producing *Clostridium cellulovorans* genome with other cellulosome-related clostridial genomes. Microb Biotechnol 4(1):64–73. doi:10.1111/j.1751-7915.2010.00210.x
7. Nolling J, Breton G, Omelchenko MV, Makarova KS, Zeng Q, Gibson R, Lee HM, Dubois J, Qiu D, Hitti J, Wolf YI, Tatusov RL, Sabathe F, Doucette-Stamm L, Soucaille P, Daly MJ, Bennett GN, Koonin EV, Smith DR (2001) Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*. J Bacteriol 183(16):4823–4838. doi:10.1128/JB.183.16.4823-4838.2001
8. Wang Y, Li X, Mao Y, Blaschek HP (2011) Single-nucleotide resolution analysis of the transcriptome structure of *Clostridium beijerinckii* NCIMB 8052 using RNA-Seq. BMC Genomics 12:479. doi:10.1186/1471-2164-12-479
9. Miller DA, Suen G, Bruce D, Copeland A, Cheng JF, Detter C, Goodwin LA, Han CS, Hauser LJ, Land ML, Lapidus A, Lucas S, Meincke L, Pitluck S, Tapia R, Teshima H, Woyke T, Fox BG, Angert ER, Currie CR (2011) Complete genome sequence of the cellulose-degrading bacterium *Cellulosilyticum lentocellum*. J Bacteriol 193(9):2357–2358. doi:10.1128/JB.00239-11
10. Feinberg L, Foden J, Barrett T, Davenport KW, Bruce D, Detter C, Tapia R, Han C, Lapidus A, Lucas S, Cheng JF, Pitluck S, Woyke T, Ivanova N, Mikhailova N, Land M, Hauser L, Argyros DA, Goodwin L, Hogsett D, Caiazza N (2011) Complete genome sequence of the cellulolytic thermophile *Clostridium thermocellum* DSM1313. J Bacteriol 193(11):2906–2907. doi:10.1128/JB.00322-11
11. Kopke M, Held C, Hujer S, Liesegang H, Wiezer A, Wollherr A, Ehrenreich A, Liebl W, Gottschalk G, Durre P (2010) *Clostridium ljungdahlii* represents a microbial production platform based on syngas. Proc Natl Acad Sci U S A 107(29):13087–13092. doi:10.1073/pnas.1004716107
12. Yokoyama S, Oshima K, Nomura I, Hattori M, Suzuki T (2011) Complete genomic sequence of the O-desmethylangolensin-producing bacterium *Clostridium* rRNA cluster XIVa strain SY8519, isolated from adult human intestine. J Bacteriol 193(19):5568–5569. doi:10.1128/JB.05637-11
13. Skarin H, Hafstrom T, Westerberg J, Segerman B (2011) *Clostridium botulinum* group III: a group with dual identity shaped by plasmids, phages and mobile elements. BMC Genomics 12:185. doi:10.1186/1471-2164-12-185
14. Seedorf H, Fricke WF, Veith B, Bruggemann H, Liesegang H, Strittmatter A, Miethke M, Buckel W, Hinderberger J, Li F, Hagemeier C, Thauer RK, Gottschalk G (2008) The genome of *Clostridium kluyveri*, a strict anaerobe with unique metabolic features. Proc Natl Acad Sci U S A 105(6):2128–2133. doi:10.1073/pnas.0711093105
15. Bettegowda C, Huang X, Lin J, Cheong I, Kohli M, Szabo SA, Zhang X, Diaz LA Jr, Velculescu VE, Parmigiani G, Kinzler KW, Vogelstein B, Zhou S (2006) The genome and transcriptomes of the anti-tumor agent *Clostridium* novyi-NT. Nat Biotechnol 24(12):1573–1580. doi:10.1038/nbt1256
16. Hill KK, Smith TJ, Helma CH, Ticknor LO, Foley BT, Svensson RT, Brown JL, Johnson EA, Smith LA, Okinaka RT, Jackson PJ, Marks JD (2007) Genetic diversity among botulinum neurotoxin-producing clostridial strains. J Bacteriol 189(3):818–832. doi:10.1128/JB.01180-06
17. Myers GS, Rasko DA, Cheung JK, Ravel J, Seshadri R, DeBoy RT, Ren Q, Varga J, Awad MM, Brinkac LM, Daugherty SC, Haft DH, Dodson RJ, Madupu R, Nelson WC, Rosovitz MJ, Sullivan SA, Khouri H, Dimitrov GI, Watkins KL, Mulligan S, Benton J, Radune D, Fisher DJ, Atkins HS, Hiscox T, Jost BH, Billington SJ, Songer JG, McClane BA, Titball RW, Rood JI, Melville SB, Paulsen IT (2006) Skewed genomic variability in strains of the toxigenic bacterial pathogen, *Clostridium perfringens*. Genome Res 16(8):1031–1040. doi:10.1101/gr.5238106
18. Bruggemann H, Baumer S, Fricke WF, Wiezer A, Liesegang H, Decker I, Herzberg C, Martinez-Arias R, Merkl R, Henne A, Gottschalk G (2003) The genome sequence of *Clostridium tetani*, the causative agent of tetanus disease. Proc Natl Acad Sci U S A 100(3):1316–1321. doi:10.1073/pnas.0335853100
19. Lapierre P, Gogarten JP (2009) Estimating the size of the bacterial pan-genome. Trends Genet 25(3):107–110
20. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pan-genome. Curr Opini Genet Dev 15(6):589–594. doi:10.1016/j.gde.2005.09.006
21. Mao F, Dam P, Chou J, Olman V, Xu Y (2009) DOOR: a database for prokaryotic operons. Nucleic Acids Res 37(Database issue):D459–D463. doi:10.1093/nar/gkn757
22. Mao X, Ma Q, Zhou C, Chen X, Zhang H, Yang J, Mao F, Lai W, Xu Y (2013) DOOR 2.0: presenting operons and their functions through dynamic and integrated views. Nucleic Acids Res. doi:10.1093/nar/gkt1048
23. Weyer ER, Rettger LF (1927) A Comparative study of six different strains of the organism commonly concerned in large-scale

production of butyl alcohol and acetone by the biological process. J Bacteriol 14(6):399–424
24. Bao G, Wang R, Zhu Y, Dong H, Mao S, Zhang Y, Chen Z, Li Y, Ma Y (2011) Complete genome sequence of *Clostridium acetobutylicum* DSM 1731, a solvent-producing strain with multireplicon genome architecture. J Bacteriol 193(18):5007–5008. doi:10.1128/JB.05596-11
25. Hu S, Zheng H, Gu Y, Zhao J, Zhang W, Yang Y, Wang S, Zhao G, Yang S, Jiang W (2011) Comparative genomic and transcriptomic analysis revealed genetic characteristics related to solvent formation and xylose utilization in *Clostridium acetobutylicum* EA 2018. BMC Genomics 12:93. doi:10.1186/1471-2164-12-93
26. O'Brien RW, Morris JG (1971) Oxygen and the growth and metabolism of *Clostridium acetobutylicum*. J Gen Microbiol 68(3):307–318
27. Giallo J, Gaudin C, Belaich JP, Petitdemange E, Caillet-Mangin F (1983) Metabolism of glucose and cellobiose by cellulolytic mesophilic *Clostridium* sp. strain H10. Appl Environ Microbiol 45(3):843–849
28. Tamaru Y, Miyake H, Kuroda K, Nakanishi A, Kawade Y, Yamamoto K, Uemura M, Fujita Y, Doi RH, Ueda M (2010) Genome sequence of the cellulosome-producing mesophilic organism *Clostridium cellulovorans* 743B. J Bacteriol 192(3):901–902. doi:10.1128/JB.01450-09
29. Shiratori H, Sasaya K, Ohiwa H, Ikeno H, Ayame S, Kataoka N, Miya A, Beppu T, Ueda K (2009) *Clostridium clariflavum* sp. nov. and *Clostridium caenicola* sp. nov., moderately thermophilic, cellulose-/cellobiose-digesting bacteria isolated from methanogenic sludge. Int J Syst Evol Microbiol 59(Pt 7):1764–1770. doi:10.1099/ijs.0.003483-0
30. William D, Murray Awk, And L, Van Den Berg (1982) C*lostridium saccharolyticurn* sp. nov., a saccharolytic species from sewage sludge. Int J Syst Bacteriol 132–135
31. Li LL, Taghavi S, Izquierdo JA, van der Lelie D (2012) Complete genome sequence of *Clostridium* sp. strain BNL1100, a cellulolytic mesophile isolated from corn stover. J Bacteriol 194(24):6982–6983. doi:10.1128/JB.01908-12
32. Tamaru Y, Miyake H, Kuroda K, Ueda M, Doi RH (2010) Comparative genomics of the mesophilic cellulosome-producing *Clostridium cellulovorans* and its application to biofuel production via consolidated bioprocessing. Environ Technol 31(8-9):889–903. doi:10.1080/09593330.2010.490856
33. Smith TJ, Hill KK, Foley BT, Detter JC, Munk AC, Bruce DC, Doggett NA, Smith LA, Marks JD, Xie G, Brettin TS (2007) Analysis of the neurotoxin complex genes in *Clostridium botulinum* A1-A4 and B1 strains: BoNT/A3, /Ba4 and /B1 clusters are located within plasmids. PLoS One 2(12):e1271. doi:10.1371/journal.pone.0001271
34. Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC (2012) The Genomes OnLine Database (GOLD) v. 4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 40(Database issue): D571–D579. doi:10.1093/nar/gkr1100
35. Edmond BJ, Guerra FA, Blake J, Hempler S (1977) Case of infant botulism in Texas. Tex Med 73(10):85–88
36. Carter AT, Pearson BM, Crossman LC, Drou N, Heavens D, Baker D, Febrer M, Caccamo M, Grant KA, Peck MW (2011) Complete genome sequence of the proteolytic *Clostridium botulinum* type A5 (B3′) strain H04402 065. J Bacteriol 193(9):2351–2352. doi:10.1128/JB.00072-11
37. He M, Sebaihia M, Lawley TD, Stabler RA, Dawson LF, Martin MJ, Holt KE, Seth-Smith HM, Quail MA, Rance R, Brooks K, Churcher C, Harris D, Bentley SD, Burrows C, Clark L, Corton C, Murray V, Rose G, Thurston S, van Tonder A, Walker D, Wren BW, Dougan G, Parkhill J (2010) Evolutionary dynamics of *Clostridium difficile* over short and long time scales. Proc Natl Acad Sci U S A 107(16):7527–7532. doi:10.1073/pnas.0914322107
38. Brazier JS, Duerden BI, Hall V, Salmon JE, Hood J, Brett MM, McLauchlin J, George RC (2002) Isolation and identification of *Clostridium* spp. from infections associated with the injection of drugs: experiences of a microbiological investigation team. J Med Microbiol 51(11):985–989
39. Sonnhammer EL, 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
40. Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8(10):785–786
41. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y (2012) dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res 40(W1):W445–W451
42. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res 37(Database issue):D233–D238. doi:10.1093/nar/gkn663
43. Case RJ, Boucher Y, Dahllof I, Holmstrom C, Doolittle WF, Kjelleberg S (2007) Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Appl Environ Microbiol 73(1):278–288. doi:10.1128/AEM.01177-06
44. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30(7):1575–1584
45. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJ, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A 102(39):13950–13955. doi:10.1073/pnas.0506758102
46. Li G, Liu B, Ma Q, Xu Y (2011) A new framework for identifying cis-regulatory motifs in prokaryotes. Nucleic Acids Res 39(7):e42. doi:10.1093/nar/gkq948
47. Ma Q, Liu B, Zhou C, Yin Y, Li G, Xu Y (2013) An integrated toolkit for accurate prediction and analysis of cis-regulatory motifs at a genome scale. Bioinformatics 29(18):2261–2268. doi:10.1093/bioinformatics/btt397
48. Mao X, Cai T, Olyarchuk JG, Wei L (2005) Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics 21(19):3787–3793. doi:10.1093/bioinformatics/bti430
49. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li CY, Wei L (2011) KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res 39(Web Server issue):W316–W322. doi:10.1093/nar/gkr483
50. Kim KI, van de Wiel MA (2008) Effects of dependence in high-dimensional multiple testing problems. BMC Bioinforma 9:114. doi:10.1186/1471-2105-9-114
51. Hutson RA, Thompson DE, Collins MD (1993) Genetic interrelationships of saccharolytic *Clostridium botulinum* types B, E and F and related clostridia as revealed by small-subunit rRNA gene sequences. FEMS Microbiol Lett 108(1):103–110
52. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. Science 311(5765):1283–1287. doi:10.1126/science.1123061
53. Moran NA (2002) Microbial minimalism: genome reduction in bacterial pathogens. Cell 108(5):583–586

54. Ochman H, Davalos LM (2006) The nature and dynamics of bacterial genomes. Science 311(5768):1730–1733. doi:10.1126/science.1119966
55. Rocha EP, Danchin A (2002) Base composition bias might result from competition for metabolic resources. Trends Genet 18(6):291–294. doi:10.1016/S0168-9525(02)02690-2
56. da Huang W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4(1):44–57. doi:10.1038/nprot.2008.211
57. Ma Q, Zhang H, Mao X, Zhou C, Liu B, Chen X, Xu Y (2014) DMINDA: an integrated web server for DNA motif identification and analyses. Nucleic Acids Res. doi:10.1093/nar/gku315
58. Kazakov AE, Cipriano MJ, Novichkov PS, Minovitsky S, Vinogradov DV, Arkin A, Mironov AA, Gelfand MS, Dubchak I (2007) RegTransBase—a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. Nucleic Acids Res 35(Database issue):D407–D412. doi:10.1093/nar/gkl865
59. Munch R, Hiller K, Barg H, Heldt D, Linz S, Wingender E, Jahn D (2003) PRODORIC: prokaryotic database of gene regulation. Nucleic Acids Res 31(1):266–269
60. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37(Web Server issue):W202–W208. doi:10.1093/nar/gkp335
61. Tanaka E, Bailey T, Grant CE, Noble WS, Keich U (2011) Improved similarity scores for comparing motifs. Bioinformatics 27(12):1603–1609. doi:10.1093/bioinformatics/btr257
62. Dillon SC, Dorman CJ (2010) Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. Nat Rev Microbiol 8(3):185–195. doi:10.1038/nrmicro2261
63. Benza VG, Bassetti B, Dorfman KD, Scolari VF, Bromek K, Cicuta P, Lagomarsino MC (2012) Physical descriptions of the bacterial nucleoid at large scales, and their biological implications. Rep Prog Phys Phys Soc 75(7):076602. doi:10.1088/0034-4885/75/7/076602
64. Hugovieux-Cotte-Pattat N, Robert-Baudouy J (1982) Regulation and transcription direction of exuR, a self-regulated repressor in *Escherichia coli* K-12. J Mol Biol 156(1):221–228
65. Rodionov DA, Mironov AA, Rakhmaninova AB, Gelfand MS (2000) Transcriptional regulation of transport and utilization systems for hexuronides, hexuronates and hexonates in gamma purple bacteria. Mol Microbiol 38(4):673–683