Gene 623 (2017) 48–62

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Pan-genome analysis of *Clostridium botulinum* reveals unique targets for drug development

Tulika Bhardwaj, Pallavi Somvanshi*

Department of Biotechnology, TERI University, 10, Industrial area, Vasant Kunj, New Delhi 110070, India

ARTICLE INFO

Keywords:
Phylogenomic
Pan genome
Singletons
COG analysis
Pathogenomic
Synteny
Resistome
Toxin/antitoxin

ABSTRACT

Clostridium botulinum, a formidable pathogen is responsible for the emerging cause of food poisoning cases on the global canvas. The endemicity of bacterium *Clostridium botulinum* is reflected by the sudden hospital outbreaks and increased resistance towards multiple drugs. Therefore, a combined approach of in-silico comparative genomic analysis with statistical analysis was applied to overcome the limitation of bench-top technologies. Owing to the paucity of genomic data available by the advent of third generation sequencing technologies, several '*omics*' technologies were applied to understand the underlying evolutionary pattern and lifestyle of the bacterial pathogen using phylogenomics. The calculation of pan-genome, core genome and singletons provides view of genetic repertoire of the bacterial pathogen lineage at the successive level, orthology shared and specific gene subsets. In addition, assessment of pathogenomic potential, resistome, toxin/antitoxin family in successive pathogenic strains of *Clostridium botulinum* aids in revealing more specific targets for drug design and development.

1. Introduction

The advent of next generation sequencing technologies have revolutionized the understanding of basic cellular physiology, microbial genetic repertoire (Adams et al., 1991) and functional diversity at metagenomic level (Olsvik et al., 1993). In addition, bacterial pathogen's whole genome sequencing technology prioritized researcher's interest towards the understanding of underlying pathogenesis by accurately measuring genetic variation within and between pathogenic groups (Méric et al., 2014; Harris et al., 2010; Katz et al., 2013; Rohde et al., 2011; Sheppard et al., 2013). At bench-top level, genetic variation among multiple genomes is inferred by the cost effective and time consuming identification of variable sites characterized as 'SNPs' (Maiden et al., 2013), whole-genome multilocus sequence typing (MLST) approach (Gutacker et al., 2006) etc. To overcome the potential limitations related to these reference based approaches, a comparative genomics based on the sequence similarity search analysis (Prabha et al., 2016) skewed the global interest towards 'omics' strategies. The availability of sequenced data at public repositories and freely accessible databases laid the foundation of 'omics' strategies and consecutive systems biology principles (Bhardwaj and Somvanshi, 2014).

Comparative microbial genomics strategy based on sequence similarity with statistical analysis helps in identifying the essential genetic content shared among all pathogenic isolates as well as subset of genes encoding virulence and novel functions as variable genome (Zhang et al., 2014; Soares et al., 2013; David et al., 2008). Pan-genome signifies both the core and variable genome content of an organism (Rouli et al., 2015; Sahl et al., 2013) while the supragenome (pan-genome) represents the whole genetic repertoire of the isolates under study. Pan genome aid in taxonomic classifications (phylogenomic analysis), precise determination of genomic contents of a group (calculation of core, pan and variable genome) and organism's lifestyle (allopatric or sympatric) (Rouli et al., 2015). We have used this combined approach to unravel the pathogenic potential of food borne pathogen *C. botulinum*.

Clostridium botulinum is an anaerobic, Gram-positive, spore-forming bacteria (Johnson and Bradshaw, 2001; Lund and Peck, 2000). It produces spores that are heat-resistant and exist widely in the environment, and in the absence of oxygen these germinate, grow and then secrete toxins (Shapiro et al., 1998; Woodruff et al., 1992). Foodborne botulism is an intoxication caused by the ingestion of potent neurotoxins in contaminated foods (SubbaRao, 2007). *C. botulinum* is the most

Abbreviation: C. botulinum, *Clostridium botulinum*; COG, Clusters of Orthologous Groups of Proteins; WHO, World Health Organization; HSP, high scoring proteins; SWG, Smith–Waterman–Gotoh; BPGA, Bacterial pan Genome Analysis; MLST, multi locus sequence typing; PSI-BLAST, Position-Specific Iterated BLAST; PSSM, Position Specific Scoring Matrix; SVM, Support Vector Machine; ORF, open reading frame; ROC, receiver operating curve; ATCC, American Type Culture Collection; NCBI, National Centre of Biotechnology Information; GI, genomic islands; VFDB, Virulence Factor Database; KEGG, Kyoto Encyclopedia of Genes & Genomes

* Corresponding author.
E-mail address: pallavi.somvanshi@teriuniversity.ac.in (P. Somvanshi).

http://dx.doi.org/10.1016/j.gene.2017.04.019
Received 2 November 2016; Received in revised form 29 March 2017; Accepted 12 April 2017
Available online 24 April 2017

potent and third most infectious agent that pose the greater risk to human and animal health worldwide (WHO, Palm et al., 2012). The global distribution of *C. botulinum* reveals it to be endemic in selected locations in India and other developing countries (Lund and Peck, 2000). The ability of a pathogen to damage a host and evade host immune defenses arises from a range of complex host-pathogen interactions and can be expressed as the pathogen's toxicity, invasiveness, colonization, and ability to be transmitted to another host (Humeau et al., 2000; Maksymowych, 1999). Despite the research and improvement in therapies, the mortality rates kept dwindling at 40% due to the disease (Shukla et al., 1997).

In this study, completely sequenced *Clostridium botulinum* isolates were subjected to phylogenomic analysis to understand the underlying evolutionary pattern of successive lineages. Pan-genome analysis was carried out to understand the symptoms related to this pathogen with respect to the broad spectrum of hosts. The successive calculation and characterization of the core and pan-genome subset revealed more specific targets for drug design and vaccine development.

2. Materials and methodology

2.1. Genome sequences

Several publically available databases served as the platform for the mining of genome sequences in the draft or incomplete format. Among all publically available repositories, complete genome sequences of *Clostridium botulinum* strains were retrieved from the genome browser of NCBI (http://www.ncbi.nlm.nih.gov/genbank/). A total number of 13 strains were selected and their genomic sequences were downloaded in FASTA and GenBank (gbk) format. The genome strains of food pathogen *Clostridium botulinum* have conserved genomic size ranges from 3.2–4.2 megabase pairs (Mb). The GC content of selected finished strains ranges between 27–29%, with a mean, standard deviation and variation of 28.08, 0.291 and 0.08474 respectively representing a stable bacterial evolution, adaptation, and population structure (Mira et al., 2010) (Supplementary File 1) i.e. genome size in mega base pairs, chromosome accession number, BioProject ID, the number of proteins, genes, Pseudogenes, RNAs, isolation.

2.2. Synteny prediction and 16s RNA

Sequences retrieved in gbk format were subjected to RNAmmer program for the prediction of full length 16s RNA gene sequences (Lagesen et al., 2007). Absynte (Despalins et al., 2011), a tool for displaying the local syntenies in completely sequenced prokaryotic chromosomes was used to identify the syntenic regions shared among the selected thirteen *C. botulinum* strains. Synteny analysis establishes the orthology prediction among *n* number of genomes and infers the underlying important functional relationship among genes. Accordingly, the sequential procedure for synteny analysis comprise 4 stages (i) reference score generation by comparing query protein sequence against itself using BLASTp (ii) similarity search analysis of the query protein sequence (s) against already available bacterial database using TBLASTN to obtain the maximum 'bit score at default parameters' (iii) normalization of the obtained bit score using reference score obtained and additional ranking on the basis of decreasing score (iv) Further, high scoring proteins (HSPs) were compared with each other using the Smith–Waterman–Gotoh (SWG) algorithm to identify paralogs/potential duplicates (Gotoh, 1982). SynMap (Lyons et al., 2008) was used to visualize the syntenic regions relationships among the multiple genome sequences in dot plot format considering *Clostridium botulinum* ATCC 3502 as the reference genome.

2.3. Phylogenomic analysis of Clostridium botulinum strains

GenBank sequences of complete genomes were mined from NCBI ftp site for the phylogenomic analysis at the whole genome level. Gegenees (version 1.1.4) was used for to obtain the phylogenomic relationship among clostridial genomes. It is based on a 'multi-threaded control' algorithm working on two parameters (i) fragment size (ii) step size. Fragments are the contiguous sequence database of the genome sequences generated by Gegenees. To determine the minimum content shared by all the genomes, an all-against-all blast was performed. Further, the minimum shared content obtained after the subtractive genomic analysis was compared with all other strains for similarity percentage identification. Data obtained in the form of distance matrix file was exported into nexus format and a phylogenetic tree was generated using SplitsTree software (version 1.1.4) (Huson and Bryant, 2006; Kloepper and Huson, 2008) using UPGMA method. Heat maps were also generated using the percentage identity results indicating colour spectrum ranging from low similar (red) to high similar (green).

2.4. Pan-genome, core genome and singleton analyses

To calculate the variation in gene content of different strains, pan genome is calculated. The pan genome size of *Clostridium botulinum* was predicted based on the chromosomes of 13 completely sequenced strains compared in this study. The concept of pan genome is not restricted to gene content but extended to structural variations arised due to genomic rearrangements like recombination events, change in location of mobile elements etc. influencing growth rate or pathogenicity of strain (Rocha, 2004). The calculation of pan genome and core genome of two different strains M and N is calculated as (a) pan genome MN was estimated in an additive manner by combining the of gene sets of M and non-orthologous genes of strain N (b) core genome MN estimated in a reductive manner by identifying the orthologs among both strains. Singletons are referred to strain specific genes having no orthologs in corresponding genomic strains. The relationship between the pan-genome and core genome is represented in (Fig. 1).

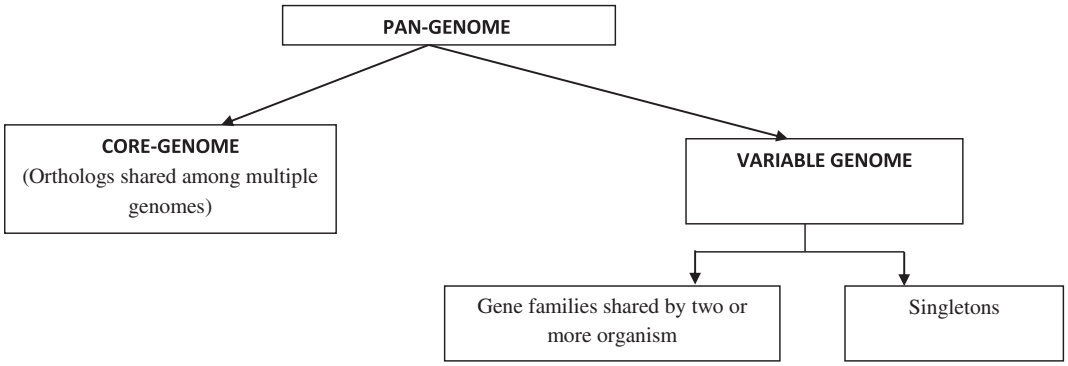


Fig. 1. Pan-genome vs. core-genome.

EDGAR (version 1.2), software for multiple genomic analysis for homology similarity search of query dataset, based on the automatically adjusted specific cut-off was used to calculate pan-genome, core-genome and singletons (Blom et al., 2009). GeneDB (Meyer et al., 2003) annotations were used in EDGAR for functional characterization and Score Ratio Values (SRV) were calculated for homology performance. SRV method estimates the normalized maximal bit score by performing BLAST with itself, with the resulting values ranging between 0 to 1 (Lerat et al., 2003). Finally, the specific cutoff was automatically calculated by a sliding window on the SRV distribution pattern (bimodal or unimodal). Performing SRV analysis aids in the identification of orthologous genes (the pair of genes exhibiting Bidirectional Best Hit where both single hits have an SRV higher than the specific cutoff).

Calculation and representation of core genome and pan genome was based on permutations. Pan genome development trends were calculated though Heaps' Law which follows the power law.

$$n = k*N\gamma \quad (1)$$

where, n = expected number of genes, N = number of genomes, 'k' and 'γ' are proportionality constant and exponent, respectively estimated by using the nonlinear least-squares curve fit to the mean values (Tettelin et al., 2005).

Core genome development trends follow least square fitting of exponential regression decay to the mean values

$$F(C) = \kappa C*\exp[-n/\tau C] + \Omega \quad (2)$$

where, F(C) = expected core-genome size, n = number of genomes, κC, τC and Ω are constants required for fitness.

For singleton development trends (Tettelin et al., 2008)

$$F(S) = \kappa S*\exp[-n/\tau S] + tg(\theta) \quad (3)$$

where, F (S) = expected singleton number, n = number of genomes, κS, τS and tg(θ) are constants.

2.5. KEGG pathway and COG category analysis

BPGA (Bacterial pan Genome Analysis), a fast and efficient computational pipeline was used in the comparative genomic analysis for the construction of the pan and core genome, in silico multi locus sequence typing (MLST), phylogenetic analysis. In addition, downstream analysis of data subsets under KEGG/COG categories deciphers the platform for further 'omics' approaches. These genomes subset analyses facilitate the identification of genotypic identity related to the serological, ecological or virulent group. BPGA employs the ublast function of USEARCH to identify best hits with respective reference databases and classify them according to KEGG and COG categories. It enables sub grouping of genomes on the basis of habitat, phenotype, morphology, life-style (Chaudhari et al., 2016).

2.6. Pan-genome composition

The complete set of core and variable genome (dispensable genome + singletons) refers to pan-genome. This complete pan-genomic repertoire comprises virulent and non-virulent genes, resistome, and toxin/antitoxin genes.

2.7. Pathogenomic analysis and assessment

Prediction of genomic islands (GI) followed by pathogenic assessment using softwares based on machine learning algorithm aids in the identification of putative virulence factors. IslandViewer 3, a web source for bacterial and archaeal genomic islands prediction is used for *Clostridial* genomes. This open source platform relies on the accuracy of three GI prediction methods: IslandPick (Langille et al., 2008), IslandPath-DIMOB (Hsiao et al., 2003) and SIGI-HMM (Waack et al., 2006). Genomic islands (GI) encode genes involved environmentally adaptations, including antimicrobial resistance and pathogenicity. VirulentPred, a bi-layer cascade Support Vector Machine algorithm for bacterial virulent protein prediction was employed. In first layer analysis, Position-Specific Iterated BLAST (PSI-BLAST) generated Position Specific Scoring Matrix (PSSM) was used for evaluation of protein features like amino acid composition, dipeptide composition, higher order dipeptide composition and optimized remote evolutionary relationships at the individual level. The resultant scores and matrix were then subjected to final SVM based classifier to generate final classifier (virulent or non-virulent). A five-cross validation technique is used for accuracy prediction (Garg and Gupta, 2008).

2.8. Acquired resistance genes prediction

ResFinder, a web service for easy identification of antibiotic resistance genes in prokaryotes was used for *Clostridial* genomes. The analysis was performed by comparing query genomes at individual level against a database of 1862 horizontally acquired resistance genes related to 12 antimicrobial classes at a minimum threshold value of 95%, the minimum length of 50% (Zankari et al., 2012).

2.9. Toxin/antitoxin prediction

To evaluate TA modules in the sequenced bacterial genome, RASTA-Bacteria (Sevin and Barloy-Hubler, 2007) was used. It is a web based annotation tool based on the domain identification by assigning functions to the genes by utilizing new knowledge-based database (TAcddb). Completely sequenced genomes in FASTA format and GenBank files act as input and subjected to the sequential procedure of ORF detection and domain identification to identify TA modules. Identified TA genes were further characterized by their related genomic sequence, locus tag, function, possible partners and ConsDom hits.

2.10. Statistical performance evaluation

To check the accuracy of the adopted pipeline, statistical estimation is required. A web based ROC calculator (Eng, 2013) was considered for ROC analysis over conventional methods. It automatically calculates and generates output in the form of sensitivity, specificity, accuracy and ROC area. The average accuracy of 97.46% indicates the reliable annotation for the functional prediction for further research.

3. Results

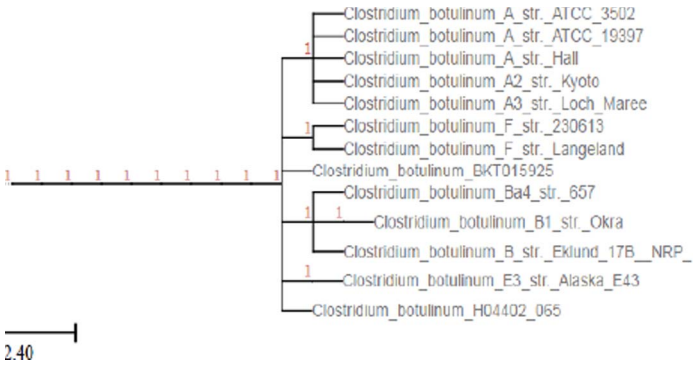
3.1. Genomic features and lifestyle

Complete genome sequences of pathogenic *Clostridium botulinum* ATCC 3502, *Clostridium botulinum* A str. ATCC 19397, *Clostridium botulinum* A2 str. Kyoto, *Clostridium botulinum* A str. Hall, *Clostridium botulinum* A3 str. Loch Maree, *Clostridium botulinum* B1 str. Okra, *Clostridium botulinum* BKT015925, *Clostridium botulinum* B str. Eklund 17B, *Clostridium botulinum* E3 str. Alaska E43, *Clostridium botulinum* Ba4 str. 657, *Clostridium botulinum* F str. 230613, *Clostridium botulinum* F str. Langeland and *Clostridium botulinum* H04402 065 were downloaded from NCBI site (Table 1). Among seven serotypes (A–G) of *Clostridium botulinum*, Type A is considered most potent to cause food-borne botulism globally (Tracy et al., 2011). Although *Clostridium botulinum* genome sequences > 140 are available at NCBI genome database but only the completed ones are considered in this study to avoid misinterpretation and redundancy. The genome size of clostridial strains ranges between 3.2–4.2 megabase pairs (Mb) and have fewer orthologs in phylogenetic members. A strong positive correlation between genome size and GC content ($R^2 = 0.23$, $F_{1,\ 13} = 76.1$, $p < 10^{-6}$) and the trend obtained by performing linear regression (using R statistical package) towards genome reduction represents the

Table 1
Genome statistics of 13 *Clostridium botulinum* genomes.

Genomes	Accession no	Size (Mb)	Proteins	GC%	Genes	Pseudogenes	rRNA	Other RNA
Clostridium botulinum ATCC 3502	NC_009495	3.8	3567	28.2	3767	77	27	7
Clostridium botulinum A str. ATCC 19397	NC_009697	3.86	3552	28.2	3693	34	24	2
Clostridium botulinum A2 str. Kyoto	NC_012563	4.16	3742	28.2	3944	90	27	4
Clostridium botulinum A str. Hall	NC_009698	3.76	3332	28.2	3499	55	24	7
Clostridium botulinum A3 str. Loch Maree	NC_010520	3.99	3540	28.3	3723	73	27	2
Clostridium botulinum B1 str. Okra	NC_010516	3.96	3586	28.3	3751	53	27	3
Clostridium botulinum BKT015925	NC_015425	3.2	2865	28.2	3114	39	30	1
Clostridium botulinum B str. Eklund 17B	NC_010723	3.8	3348	27.5	3505	42	34	4
Clostridium botulinum E3 str. Alaska E43	NC_010674	3.66	3126	27.4	3272	30	34	3
Clostridium botulinum Ba4 str. 657	NC_012658	4.25	3939	28.02	4133	81	27	3
Clostridium botulinum F str. 230613	NC_017297	4.01	3342	28.2	3784	343	27	1
Clostridium botulinum F str. Langeland	NC_009699	4.01	3624	28.2	3786	52	27	2
Clostridium botulinum H04402 065	NC_017299	3.9	3386	28.2	3698	216	27	4

a)



b)

((Clostridium_botulinum_A_str__ATCC_19397_NC_009697:0.000000,Clostridium_botulinum_A_str__Hall_NC_009698:0.000010):0.000200,((((ALL_Clostridium_botulinum_A3_str__Loch_Maree_NC_010520:0.013660,((ALL_Clostridium_botulinum_BKT015925_NC_015425:0.185840,(ALL_Clostridium_botulinum_B_str__Eklund_17B_NC_010674:0.025570,Clostridium_botulinum_E3_str__Alaska_E43_NC_010723:0.024650):0.190280):0.153800,ALL_Clostridium_botulinum_Ba4_str__657_NC_012658:0.012630):0.000320):0.004790,(ALL_Clostridium_botulinum_B1_str__Okra_NC_010516:0.006410,(ALL_Clostridium_botulinum_F_str__230613_NC_017297:0.000130,ALL_Clostridium_botulinum_F_str__Langeland_NC_009699:0.000020):0.005570):0.002140):0.000760,Clostridium_botulinum_A2_str__Kyoto_NC_012563:0.006710):0.000400,Clostridium_botulinum_H04402_065_NC_017299:0.005910):0.006710,ALL_Clostridium_botulinum_A_str__ATCC_3502_NC_009495:0.000150);

Fig. 2. a: A 16 s rRNA tree of 13 *Clostridium botulinum* strains.((Clostridium_botulinum_A_str_ATCC_19397_NC_009697:0.000000, Clostridium_botulinum_A_str_Hall_NC_009698:0.000010):0.000200, ((((ALL_Clostridium_botulinum_A3_str_Loch_Maree_NC_010520:0.013660, ((ALL_Clostridium_botulinum_BKT015925_NC_015425:0.185840, (ALL_Clostridium_botulinum_B_str_Eklund_17B_NC_010674:0.025570, Clostridium_botulinum_E3_str_Alaska_E43_NC_010723:0.024650):0.190280):0.153800, ALL_Clostridium_botulinum_Ba4_str_657_NC_012658:0.012630):0.000320):0.004790, (ALL_Clostridium_botulinum_B1_str_Okra_NC_010516:0.006410, (ALL_Clostridium_botulinum_F_str_230613_NC_017297:0.000130, ALL_Clostridium_botulinum_F_str_Langeland_NC_009699:0.000020):0.005570):0.002140):0.000760, Clostridium_botulinum_A2_str_Kyoto_NC_012563:0.006710):0.000400, Clostridium_botulinum_H04402_065_NC_017299:0.005910):0.006710, ALL_Clostridium_botulinum_A_str_ATCC_3502_NC_009495:0.000150);b: A 16s rRNA tree of 13 *Clostridium botulinum* strains in Newick format.

intracellular lifestyle of *Clostridial* genomes. Lateral gene transfer (LGT) (Tamas et al., 2002; Moran and Plague, 2004), minimal genome selective pressure (Wernegreen, 2005), increased adaptation to host resulting in gene loss (Darby et al., 2007) etc. may contribute to this intrinsic capability of the bacterial lineages. In addition, metabolic reasons (Rocha and Danchin, 2002; Haft et al., 2005) or reparation systems (Toh et al., 2002; Wernegreen et al., 2003) contribute towards AT mutational bias.

3.2. Phylogenetic and synteny analysis

To understand the evolutionary relationships among *Clostridial genomes*, a 16S rRNA tree of 13 *Clostridial* genomes was built by performing multiple sequence alignment of genome sequences using MUSCLE. The output of multiple sequence alignment was submitted to PHYLIP to generate the phylogenetic tree at the bootstrap value of 100 using UPGMA algorithm. In Fig. 2, horizontal lines are branches representing the evolutionary lineages changing over time. The scale

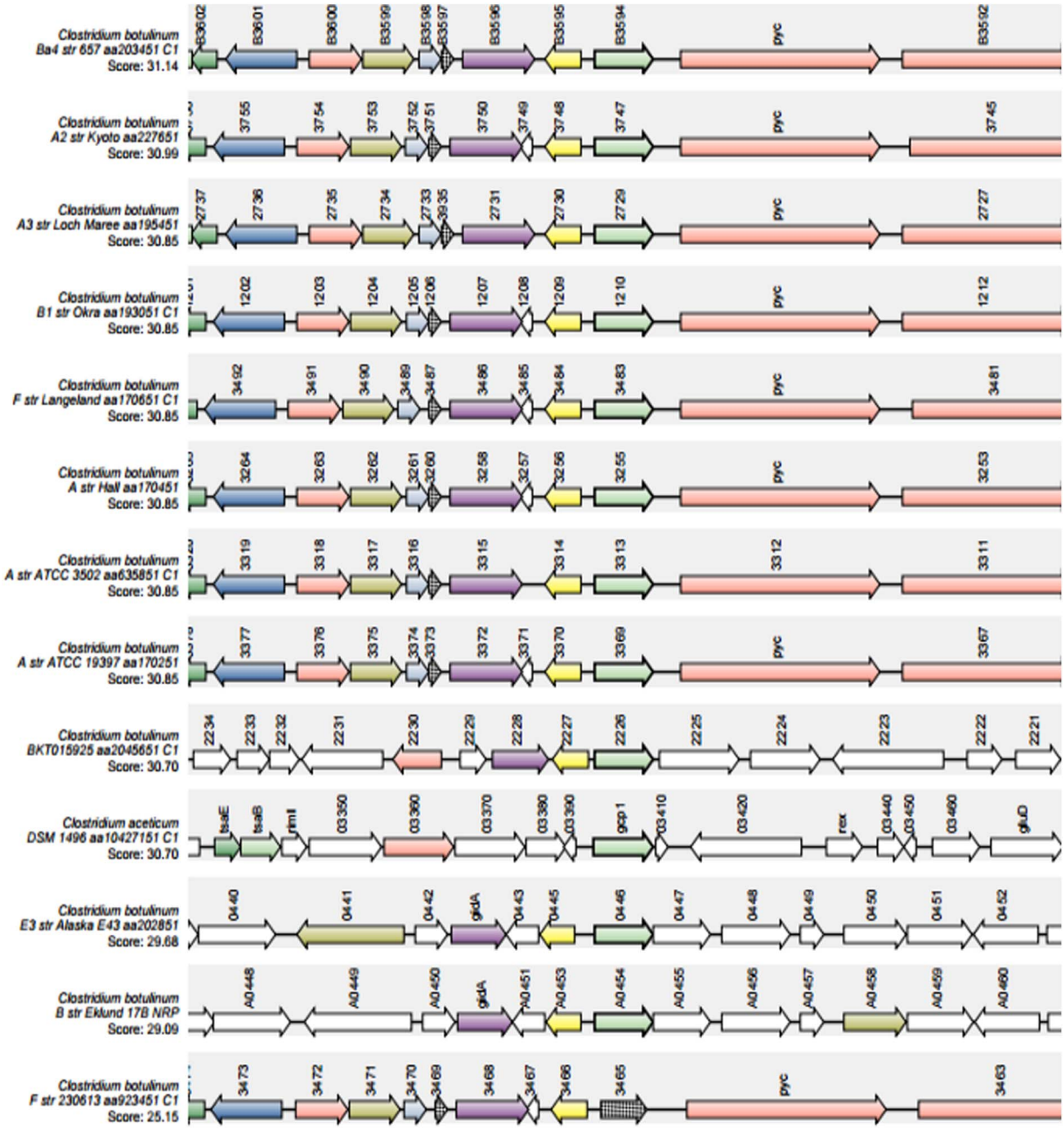


Fig. 3. Representation of local syntenies in completely sequenced 13 *Clostridial* genomes along with the synteny score. Synteny score = maximal bit score by performing blast against all sequences/reference blast score. Genomes are arranged in order of decreasing Synteny score. Colour codes are automatically generated representing a particular coding sequence. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

of 2.40 at the bottom of the tree represents an amount genetic change of '2.40' units. Units referred to nucleotide substitutions per site i.e. units = number of changes or 'substitutions'/length of the sequence. Each node represents the number ranges from 0 to 1 discerning the measure of support to each node, where 1 deciphers maximal support. The overall evolutionary connectivity is represented by the four clades formation from the common ancestor and successive divergence might be the result of gene loss, horizontal gene transfer and host adaptation.

Further, comparative genomic analysis of multiple genomic sequences and identification of orthologous sequences become problematic due to high genomic complexity. Therefore, synteny analysis provides the platform to identify the homologous regions among multiple genomes and aids in phenotypic characterization of unannotated genomic segments. In addition, shared positive synteny (Figs. 2a & b and 3) among *Clostridial* genomes discerns important functional relationships among genomes i.e. adaptation to host, virulence, metabolism, signal transduction, cell division and motility, intracellular trafficking, vesicular transport etc. The syntenic scores range from 31 to 25 represents the high synteny among clostridial genomes (Fig. 3). Visualization of the syntenic relationship among multiple genomes in dot-plot method using SynMap is depicted in Fig. 4.

3.3. Phylogenomic analysis

To understand the evolutionary pattern, the phylogenomic relationships between the *Clostridium botulinum* strains were determined using Gegenees. An all-against-all BLAST program was performed to identify the minimal shared gene content and variable content of individual genome was compared with minimal gene set to percentage identity determination and phylogenomic tree construction. The similarity

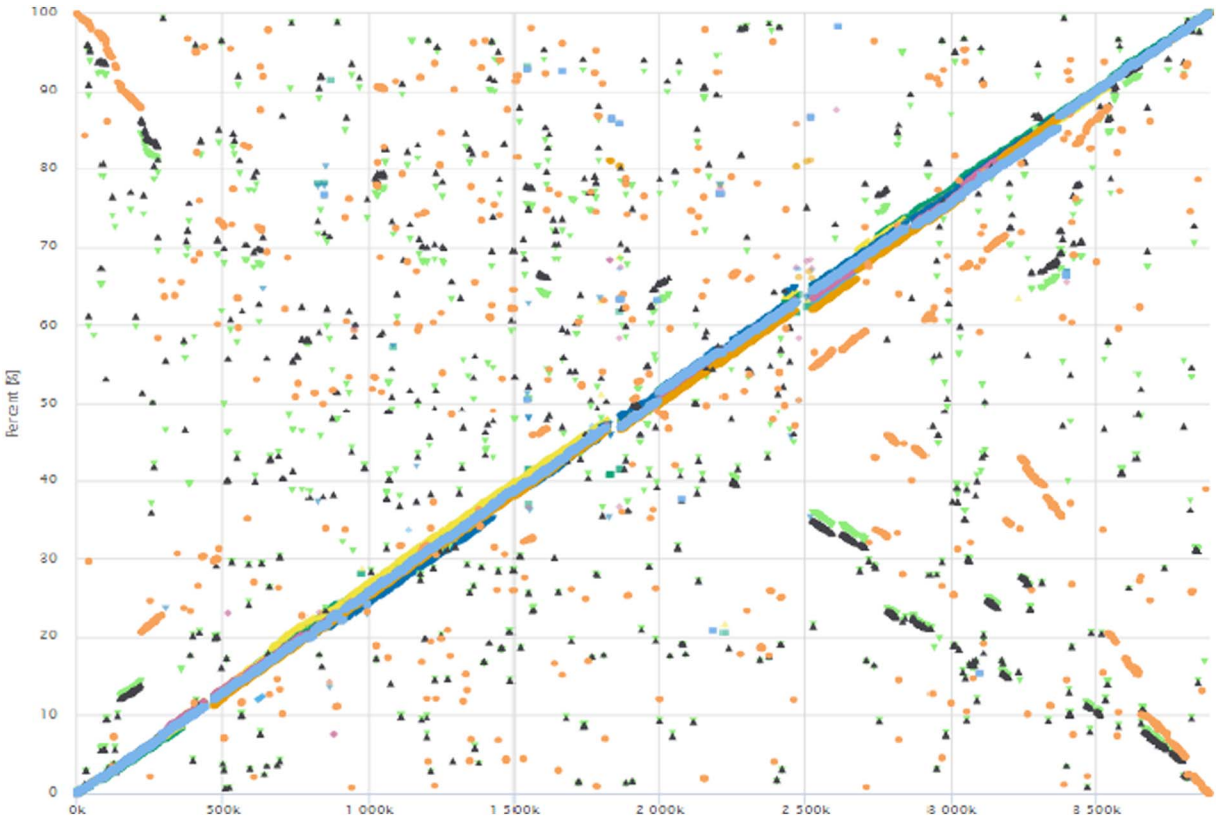


Fig. 4. Visualization of comparative syntenic analysis between 13 Clostridial genomes ● Clostridium botulinum ATCC 19397, ● Clostridium botulinum A Hall, ■ *Clostridium botulinum* A2 str. Kyoto, ● *Clostridium botulinum* H04402 065, ▼ *Clostridium botulinum* Ba4 str. 657, ● *Clostridium botulinum* F str. Langeland, ◆ *Clostridium botulinum* B1 str. Okra, ■ *Clostridium botulinum* F str. 230613, ▲ *Clostridium botulinum* B str. Eklund 17B, ▼ *Clostridium botulinum* E3 str. Alaska E43, ■ *Clostridium botulinum* BKT015925, ◆ *Clostridium botulinum* ATCC 3502, □ *Clostridium botulinum* Type A Hall.

percentage matrix is visualized in the form of heat map with the colour codes of green (region of high similarity) to red (low similarity regions). According to phylogenomic tree generated, type A pathogenic strains *Clostridium botulinum* ATCC 3502, *Clostridium botulinum* A str. ATCC 19397, *Clostridium botulinum* A2 str. Kyoto, *Clostridium botulinum* A str. Hall, *Clostridium botulinum* A3 str. Loch Maree is present in the same cluster. In addition, *Clostridium botulinum* F str. 230613 and *Clostridium botulinum* F str. Langeland are closely related to Type A strains as compared to other. *Clostridium botulinum* Ba4 str. 657, mainly responsible for infant botulism shows the close similarity to *Clostridium botulinum* B1 str. Okra and *Clostridium botulinum* B str. Eklund 17B. The contrasting characteristic was revealed by a high percentage of genetic similarity between *Clostridium botulinum* E3 str. Alaska E43 and *Clostridium botulinum* H04402 065 with a genome-to-genome distance of 96.32. According to a heat map, *Clostridium botulinum* group shares the similarity of > 99% (Fig. 5).

3.4. Pan-genome, core genome and singleton analysis

To overview the genetic repertoire of the *Clostridial* genomes, pan genome was calculated using the aforementioned SRV method used in EDGAR. The sequential addition of the new genome indicates the increase in gene content at each level. As the pan genome includes the minimal shared gene content among the genome group along with non-redundant genes of genomes at the individual level, the resultant contains a total number of 13 1021 genes (approximately 1.6 fold the average number of genes of 13 strains). The pan-genome was curve (Fig. 6a) fitted by the exponential decay model based on Heap's law. Using the formula, the values of the respective constant were calculated using R statistical language. For N (number of genomes) = 13, variables κ and γ were calculated as 3535.74 and 0.41 respectively. A $\alpha \leq 1$ indicates and open pan-genome deciphering the successive addition of accessory genes with each genome introduction. The resultant pan-genome fitted model is represented as $3535.474 * x\hat{}0.414$ ($\alpha = 0.586$). In addition, fitted model is used to evaluate the unique genes for individual genomes using the median value. The model estimates 889 ± 3 genes could be added with every new genome introduction.

In contrast, core-genome analysis converges the total gene content of 13,590 to gene subset of 889.503 and the number of shared genes decreased with new genome introduction. Approximately, only 16% of the genome remains constant with successive genome addition. Core genome development trends (Fig. 6b) follow least square fitting of exponential regression decay to the mean values, $2657.691 * \exp(-x/3.460) + 889.503$ fitted model was observed at confidence interval (95%) from 866.405 to 912.601. The trend predicts (x), the $\kappa * \exp[-x/\tau]$ term will tend towards 0 with the addition of new genome.

Singletons are referred to strain specific genes (having no orthologs in comparative genomes). The singleton development trend (Fig. 6c), $F(S) = \kappa S * \exp[-n/\tau S] + tg(\theta)$ least square decay model was fitted to identify tg (θ). As per calculation and analysis, 287 genes are added to pan genome at every genome introduction.

The predicted genes (core, unique and accessory) were further classified based on the KEGG and COG category. As per KEGG category analysis (Fig. 7), the genome of all the strains encodes the maximum genes related to 'metabolism' and 'genetic information processing'. Approximately, an equal share of genes was subjected to 'organismal systems' and 'environmental information processing'. Genes related to human diseases almost shared an equal percentage (Table 2).

Surprisingly, classification according to COG category showed the similar results as KEGG analysis up to maximum level. Percentage of genes (Fig. 8) related to metabolism (41.03%) is higher than the others. The percentage of genes in different functional categories was done by combining the sub categories into 4 groups (Information storage and processing, Cellular processes and signaling, Metabolism, poorly char-

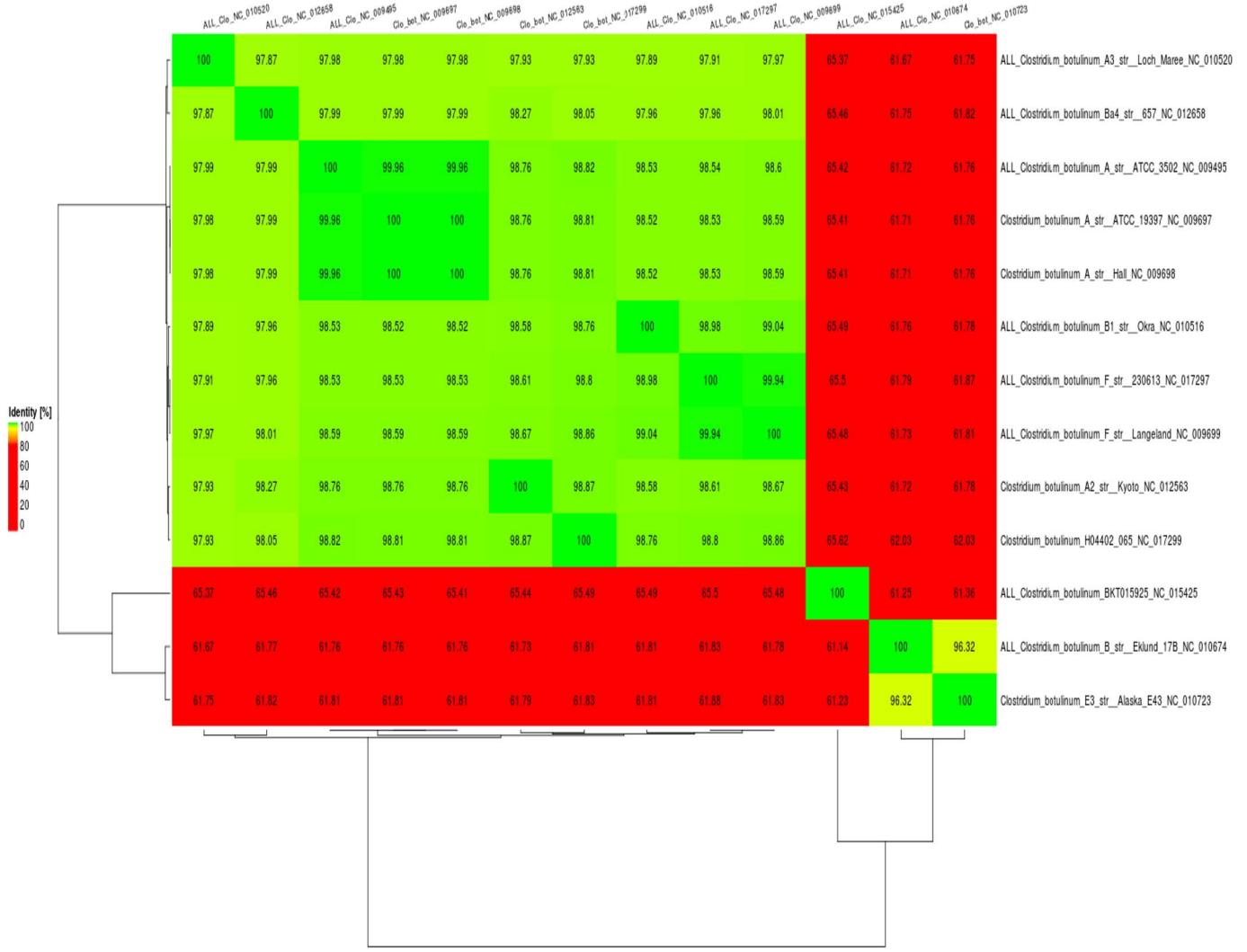


Fig. 5. Phylogenomic tree and heat map visualization of genus *Clostridium*.

acterized). Additionally, the percentage of the 'poorly characterized' genes. In addition, gene count mean of the functional COG categories with specific function code is calculated. The deviation of the gene values from the mean (standard deviation) was also included (Table 3).

3.5. Pathogenomic analysis

To identify the shared Genomic islands (GI) among the 13 *Clostridium botulinum* strains, Island Viewer 3 was analyzed. Comparative identification was done by setting the parameters (Minimum genomic distance = 0.10 units and a maximum distance of 0.40 units). Genomic islands were predicted for each representative strain by employing four algorithms (Integrated, IslandPick, SIGI-HMM and IslandPath-DIMOB). Comparative analysis of the predicted genomic islands among 13 *Clostridial* strains results in 81 core genomic islands among *Clostridial* lineage. Virulence potential was assessed by VirulentPred to estimate the pathogenic potential of the respective genomic islands. Virulent sequences were predicted against the training set including bacterial database (including 512 human pathogens and 332 no-pathogens), MVirDB and VFDB (Virulence Factor Database) at default 0.01 SVM scores. Finally, non-homology search analysis of 79 virulent predicted genomic islands against the human database was performed at default values using BLASTp to identify the potential therapeutic targets. This non-homology search analysis reveals that 35 Genomic Islands (GIs) have potential to cause pathogenicity (Table 4).

4. Discussion

Clostridium botulinum, a gram positive opportunistic pathogen is the major cause of food-borne botulism globally. *C. botulinum* species has 13 complete genomes sequences on NCBI GenBank having a circular chromosome, a plasmid. The total number of 41,382 proteins is encoded by *Clostridial* genomes with the not so variable GC% ranging from 27–28. Pseudogenes, tRNAs, rRNAs are also available in the lineage. Canned peas are the major isolation source for *Clostridium botulinum* A str. ATCC 3502, *Clostridium botulinum* A str. Hall (Pagani et al., 2012) and *Clostridium botulinum* A str. ATCC 19397 (Smith et al., 2007). Infant botulism cases were reported due to *Clostridium botulinum* A2 str. Kyoto and *Clostridium botulinum* Ba4 str. 657 (Edmond et al., 1977). Food-borne botulism cases related to Salmon eggs were reported due to *Clostridium botulinum* BKT015925, *Clostridium botulinum* E3 str. Alaska E43 and *Clostridium botulinum* F str. 230613. In addition, homemade food related botulism cases point to *Clostridium botulinum* H04402 065 36, *Clostridium botulinum* A3 str. Loch Maree (Brazier et al., 2002) and *Clostridium botulinum* B str. Eklund 17B (Table 1).

Phylogenomics integrates the field of evolution and genomics by exploring evolutionary relationships among species through comparative analysis at whole genome level. Following the classical three step procedure of phylogenomic analysis (1. homology clustering 2. multiple sequence alignment and 3. phylogenetic tree) infers tree of life, emergence and spread of antibiotic resistance etc. (Bernard et al.,

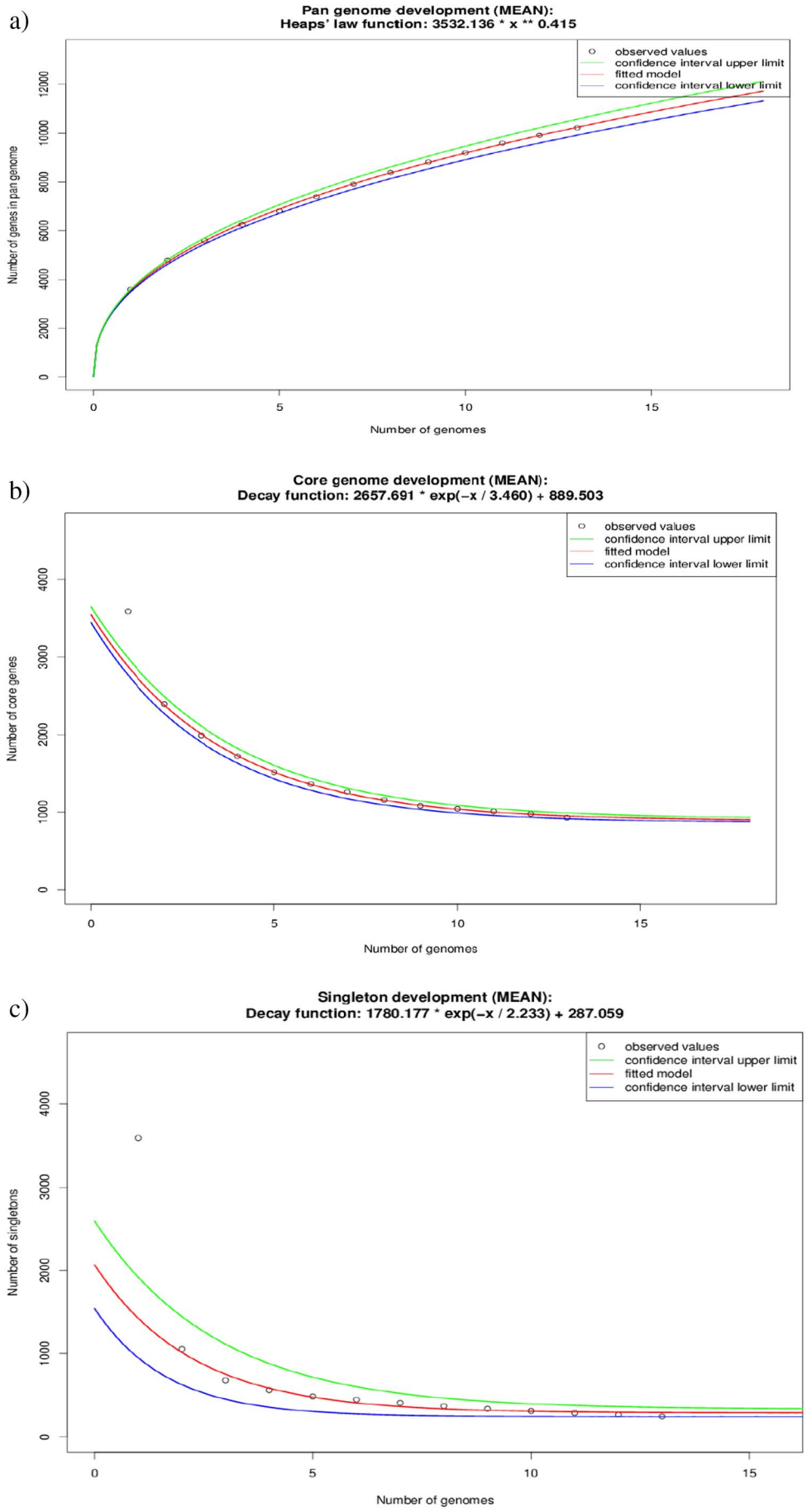


Fig. 6. *C. botulinum* (a) pan, (b) core and accessory genome (c) singleton evolution according to the number of sequenced genomes.

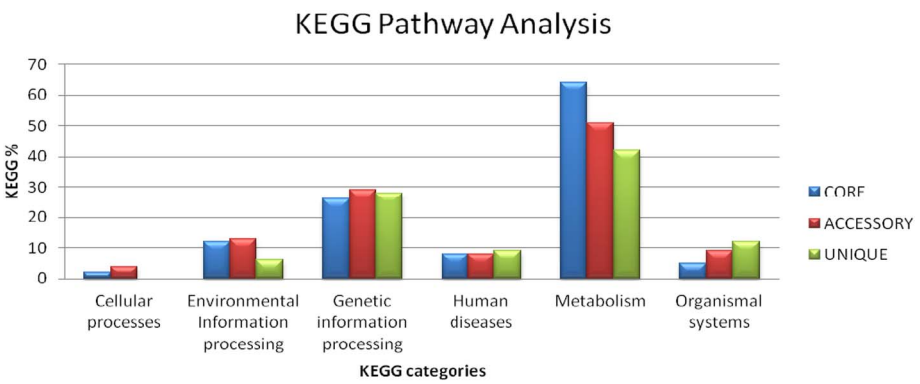


Fig. 7. Classification of genes of *C. botulinum* genes based on KEGG categories.

Table 2
Percentage of distribution of genes in different functional categories (%).

Percentage of genes in different functional categories (%)	
Information storage and processing (B + L + K + A + J)	16.03
Cellular processes and signaling (O + U + T + D + V + M + N + Z)	19.01
Metabolism (F + E + G + I + H + P + Q + C)	35.03
Poorly characterized (R + S)	29.93

2016). At the species level, the phylogenomic analysis reveals the evolutionary pattern in the *Clostridial* lineage. The percentage of sequence similarity about 99% reveals the close proximity to the phylogenetic members indicating the large percentage of share genomics segments. This provides the evidence of increasing susceptibility of the pathogen in host and frequent hospital breakouts.

Pan-genome, core-genome and R_{cp} (ration of core-genome size to that of pan-genome size) were calculated for 13 *Clostridium botulinum* genome sequences as 3535, 2657 and 1.02 respectively. As $R_{cp = 1}$ (approximately) indicating the phylogenetic relationship among the genomes under study. The value of Rcp always lies between 0 and 1 and acts as determinant of genomic diversity and asymmetry in genome sequences available at NCBI (Ghatak et al., 2016). The open genome extrapolation of pan results indicates sympatric lifestyle of the bacteria and its ability for the acquisition of novel species specific genes related to virulence, metabolism, information storage etc. The introduction of genomic islands at individual genome introduction due to the horizon-

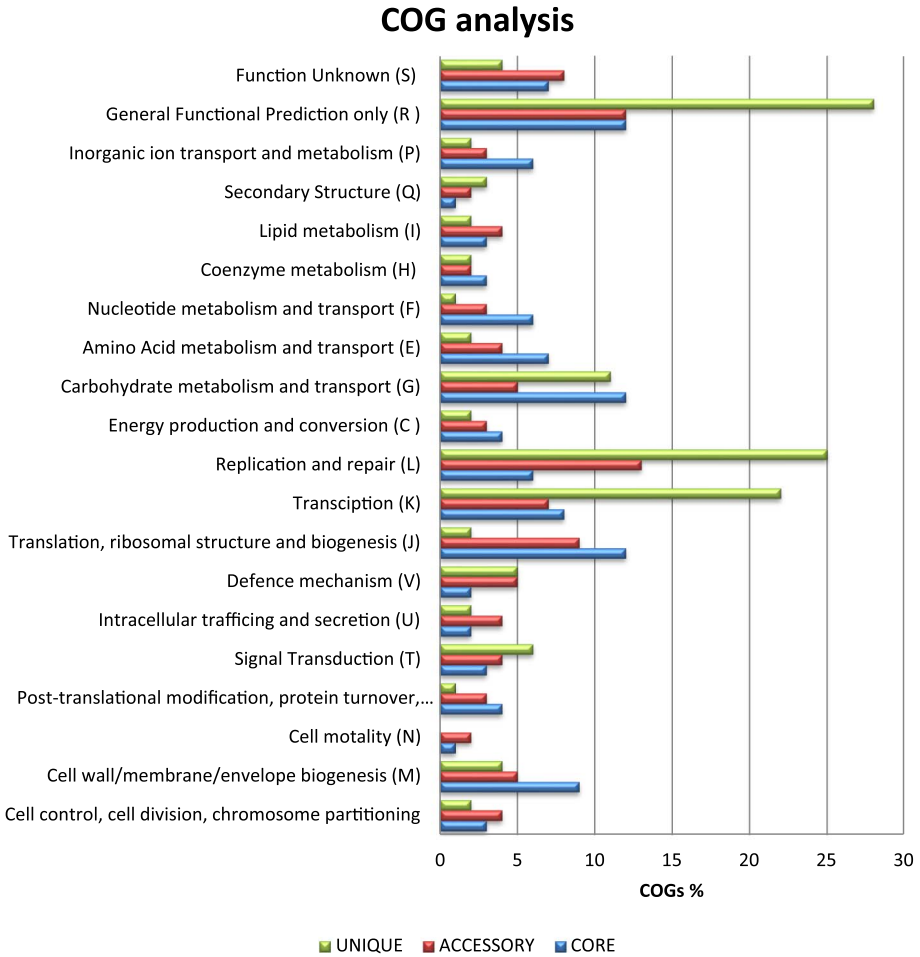


Fig. 8. Classification of genes of *C. botulinum* genes based on COG categories.

Table 3
Representing gene count mean of the functional COG categories with specific function code, gene count mean and standard deviation.

COG categories	Function code	*C. botulinum* (n = 13)
Gene count mean ± standard deviation		
Replication, recombination and repair	L	14.66 ± 9.6
Transcription	K	12.33 ± 8.38
Translation, ribosomal structure and biogenesis	J	7.66 ± 1.53
Posttranslational modification, protein turnover, chaperones	O	2.66 ± 1.52
Intracellular trafficking, secretion, and vesicular transport	U	6 ± 2.64
Signal transduction mechanisms	T	4.33 ± 1.52
Cell cycle control, cell division, chromosome partitioning	D	3 ± 1
Defense mechanisms	V	16.78 ± 9.49
Cell wall/membrane/envelope biogenesis	M	16.78 ± 9.49
Cell motility	N	1 ± 1
Nucleotide transport and metabolism	F	3.33 ± 2.51
Amino acid transport and metabolism	E	4.33 ± 2.51
Carbohydrate transport and metabolism	G	9.33 ± 3.78
Lipid transport and metabolism	I	3 ± 1
Coenzyme transport and metabolism	H	2.33 ± 0.57
Inorganic ion transport and metabolism	P	3.66 ± 2.08
Secondary metabolites biosynthesis, transport and catabolism	Q	2 ± 1
Energy production and conversion	C	3 ± 1
General function prediction only	R	17.33 ± 9.23
Function unknown	S	6.3 ± 2.08
Average		139.81 ± 71.93

tal gene transfer and recombination events adds up to pan genome and related pathogenicity. Core-genome identified infers the lifestyle of bacteria, unusual compositional features or phylogenetic incongruence. This pool of genes encodes heavy-metal and antibiotic resistance, cell-wall components, virulence, metabolic genes, nitrogen fixation and bacteriocins. Singletons, having zero similarity with the closely related strains gained the special attention of the researchers conferring biological individuality and host-specificity and pathogenesis.

As per COG functional category and KEGG category analysis, the majority of the genes (core, accessory and unique) are related to 'Metabolism' and information storage and processing category. Genes involved in carbohydrates, amino acids, nucleotides, coenzymes, lipids, metabolism, inorganic ions and secondary metabolites production, transport and secretion, production and conversion of energy are maximum in number rather than other categories. However, the numbers of genes in category 'poorly characterized' indicates the upcoming possibility of discerning the pathogenesis of bacterial lineage. This aids in suppressing the possibility of disease outbreaks and surveillance.

The pathogenomic analysis provides the obscure pathogenic potential of the opportunistic pathogen by a subtractive analysis involving computational algorithms and machine learning approaches. Functional characterization of the hypothetical proteins (genomic segments) identified indicated the existence of genes involved in type III secretion system, invasion and adhesion. The presence of *SecE*, a preprotein translocation unit causing membrane associated virulence in *Brucella abortis* (DelVecchio et al., 2002) indicates the invasive virulence in *Clostridial* genomes. A propanediol utilization protein (*PduL*) required for the catabolism of propanediol was assessed as pathogenic in causing virulence as in *Salmonella typhimurium* (Bobik et al., 1997).

Table 4
List of 35 Genomic Islands (GIs) having potential to cause pathogenicity representing protein name, IslandViewer prediction method, prediction method (virulent/non-virulent), prediction scores calculated on the basis of VirulentPred algorithm.

S. no.	Protein name	Island Viewer prediction method	Prediction result	Prediction scores
1.	Propanediol utilization protein *Pdu*L-like protein	IslandPath-DIMOB	Virulent	1.0606
2.	Acetaldehyde dehydrogenase	IslandPath-DIMOB	Non-virulent	− 0.468
3.	Propanediol/ethanolamine utilization protein	IslandPath-DIMOB	Virulent	1.0606
4.	Ethanolamine utilization protein *Eut*Q-like protein	IslandPath-DIMOB	Virulent	1.0606
5.	Ethanolamine utilization protein *Eut*J family protein	IslandPath-DIMOB	Virulent	1.0606
6.	GTP-binding protein, *Eut*P/*Pdu*V family	IslandPath-DIMOB	Virulent	1.0606
7.	Hypothetical protein	IslandPath-DIMOB	Virulent	1.0606
8.	Ethanolamine utilization protein *Eut*S-like protein	IslandPath-DIMOB	Virulent	1.0606
9.	Glycyl-radical activating family protein	IslandPath-DIMOB	Virulent	1.0606
10.	Formate acetyltransferase	IslandPath-DIMOB	Non-virulent	1.0606
11.	Aldehyde dehydrogenase	IslandPath-DIMOB	Virulent	1.0606
12.	Microcompartments family protein	IslandPath-DIMOB	Virulent	1.0606
13.	*Mer*R family transcriptional regulator	IslandPath-DIMOB	Virulent	1.0606
14.	Iron-containing alcohol dehydrogenase	IslandPath-DIMOB	Virulent	1.0606
15.	Phage integrase	IslandPath-DIMOB	Virulent	1.0606
16.	Nitroimidazole resistance protein	IslandPath-DIMOB	Virulent	1.0606
17.	tRNA pseudouridine synthase B	IslandPath-DIMOB	Virulent	1.0606
18.	DHH family protein	IslandPath-DIMOB	Virulent	1.0606
19.	Ribosome-binding factor A	IslandPath-DIMOB	Virulent	1.0623
20.	Translation initiation factor IF-2	IslandPath-DIMOB	Virulent	1.0606
21.	Ribosomal protein L7Ae family protein	IslandPath-DIMOB	Virulent	1.0606
22.	Recombinase, phage *Rec*T family	IslandPath-DIMOB	Virulent	1.0600
23.	50S ribosomal protein L30	IslandPath-DIMOB	Virulent	1.0606
24.	30S ribosomal protein S5	IslandPath-DIMOB	Virulent	1.0606
25.	50S ribosomal protein L18	IslandPath-DIMOB	Virulent	1.0606
26.	50S ribosomal protein L6	IslandPath-DIMOB	Virulent	1.0606
27.	30S ribosomal protein S14	IslandPath-DIMOB	Virulent	1.0611
28.	50S ribosomal protein L5	IslandPath-DIMOB	Virulent	1.0606
29.	Elongation factor Tu	IslandPath-DIMOB	Virulent	1.0606
30.	Elongation factor G	IslandPath-DIMOB	Virulent	1.0609
31.	Ribosomal protein L7Ae family protein	IslandPath-DIMOB	Virulent	1.0606
32.	DNA-directed RNA polymerase subunit beta′	IslandPath-DIMOB	Virulent	1.0606
33.	Transcription antitermination protein *Nus*G	IslandPath-DIMOB	Virulent	1.0612
34.	Preprotein translocase subunit *Sec*E	IslandPath-DIMOB	Virulent	1.0606
35.	RNA polymerase factor sigma-70	IslandPath-DIMOB	Virulent	1.0606

Table 5
Listing antibiotic resistance genes of 13 *Clostridial* genomes at individual level representing Locus tag, gene name, product and related EC number.

Locus tag	Gene	Product	EC number
Clostridium botulinum ATCC 3502			
CBO2021	catB	Chloramphenicol acetyltransferase	2.3.1.28
CBO0214		Vancomycin B-type resistance protein VanW	
CBO0215		Vancomycin B-type resistance protein VanW	
CBO2769	gyrA	Topoisomerase IV subunit A	5.99.1.-
CBO2770	gyrB	Topoisomerase IV subunit B	5.99.1.-
CBO2857	cfr	23S rRNA (adenine(2503)-C(Chaudhari et al., 2016))-methyltransferase	2.1.1.224
CBO3488	rpoB	DNA-directed RNA polymerase beta subunit	2.7.7.6
CBO0619	cat	Chloramphenicol acetyltransferase	2.3.1.28
CBO0646		MDR-type permease	
CBO0007	gyrA	DNA gyrase subunit A	5.99.1.3
CBO0820		Vancomycin B-type resistance protein VanW	
CBO0828	dhpS	Alternative dihydrofolate reductase 2/dihydropteroate synthase	2.5.1.15
Clostridium botulinum ATCC 19397			
CLB_1961	cat-3	Chloramphenicol acetyltransferase	2.3.1.28
CLB_0255		Vancomycin B-type resistance protein VanW	
CLB_0256		Vancomycin B-type resistance protein VanW	
CLB_2712		Topoisomerase IV subunit A	5.99.1.-
CLB_2713		Topoisomerase IV subunit B	5.99.1.-
CLB_2801		23S rRNA (adenine(2503)-C(Chaudhari et al., 2016))-methyltransferase	2.1.1.224
CLB_3545	rpoB	DNA-directed RNA polymerase beta subunit	2.7.7.6
CLB_0659	cat-1	Chloramphenicol acetyltransferase (EC)	2.3.1.28
CLB_0683		MDR-type permease	
CLB_0007	gyrA	DNA gyrase subunit A (EC)	5.99.1.3
CLB_0861		Vancomycin B-type resistance protein VanW	
CLB_0869	folP	Alternative dihydrofolate reductase 2/dihydropteroate synthase	2.5.1.15
CLB_0439		Undecaprenyl-diphosphatase	3.6.1.27
Clostridium botulinum A2 str. Kyoto			
CLM_0485		Undecaprenyl-diphosphatase	3.6.1.27
CLM_0264		Vancomycin B-type resistance protein VanW	
CLM_0265		Vancomycin B-type resistance protein VanW	
CLM_2238	catB	Chloramphenicol acetyltransferase	2.3.1.28
CLM_0295		Transcriptional regulatory protein	
CLM_3137		Topoisomerase IV subunit A	5.99.1.-
CLM_3138		Topoisomerase IV subunit B	5.99.1.-
CLM_3236		23S rRNA (adenine(2503)-C(Chaudhari et al., 2016))-methyltransferase	2.1.1.224
CLM_3956	rpoB	DNA-directed RNA polymerase beta subunit	2.7.7.6
CLM_0620		Two-component system response regulator	
CLM_0622		ABC transporter, ATP-binding protein	
CLM_0727		Chloramphenicol acetyltransferase	2.3.1.28
CLM_0752		MDR-type permease	
CLM_0007	gyrA	DNA gyrase subunit A	5.99.1.3
CLM_0799	blaI	Beta-lactamase repressor BlaI	
CLM_0827		ABC transporter, ATP-binding protein	
CLM_0962		Vancomycin B-type resistance protein VanW	
CLM_0970	folP	Alternative dihydrofolate reductase 2/dihydropteroate synthase	2.5.1.15
CLM_1026		Chloramphenicol acetyltransferase	2.3.1.28
Clostridium botulinum B str. Eklund			
CB17B1432		MDR-type permease	
CB17B0200	rpoB	DNA-directed RNA polymerase beta subunit	2.7.7.6
CB17B0715	gyrB	Topoisomerase IV subunit B	5.99.1.-
CB17B1362		Chloramphenicol acetyltransferase	2.3.1.28
CB17B2496	upk1	Undecaprenyl-diphosphatase	3.6.1.27
Clostridium botulinum B1 str. Okra			
CLD_0819	gyrA	DNA gyrase subunit A	5.99.1.3
CLD_2604	cat	Chloramphenicol acetyltransferase	2.3.1.28
CLD_1803		Topoisomerase IV subunit A	5.99.1.-
CLD_1802		Topoisomerase IV subunit B	5.99.1.-
CLD_1710		23S rRNA (adenine(2503)-C(Chaudhari et al., 2016))-methyltransferase	2.1.1.224
CLD_1016	rpoB	DNA-directed RNA polymerase beta subunit	2.7.7.6
CLD_0561		Vancomycin B-type resistance protein VanW	
CLD_0560		Vancomycin B-type resistance protein VanW	
CLD_0529		Transcriptional regulatory protein	
CLD_0115		MDR-type permease	
CLD_0056		ABC transporter, ATP-binding protein	
CLD_3752		Vancomycin B-type resistance protein VanW	
CLD_3744	folP	Alternative dihydrofolate reductase 2/dihydropteroate synthase	2.5.1.15
Clostridium botulinum Ba4 str. 657			
CLJ_B0301		Chloramphenicol acetyltransferase	2.3.1.28
CLJ_B0473	uppP_2	Undecaprenyl-diphosphatase	3.6.1.27
CLJ_B0862		Vancomycin B-type resistance protein VanW	

(continued on next page)

Table 5 *(continued)*

Locus tag	Gene	Product	EC number
CLJ_B0870	folP	Alternative dihydrofolate reductase 2/dihydropteroate synthase	2.5.15
CLJ_B2225		Chloramphenicol acetyltransferase	2.3.1.28
CLJ_B0007	gyrA	DNA gyrase subunit A	5.99.1.3
CLJ_B2999		Topoisomerase IV subunit A	5.99.1.-
CLJ_B3000		Topoisomerase IV subunit B	5.99.1.-
CLJ_B3097		23S rRNA (adenine(2503)-C(Chaudhari et al., 2016))-methyltransferase	2.1.1.224
CLJ_B3797	rpoB	DNA-directed RNA polymerase beta subunit	2.7.7.6
CLJ_B0262		Vancomycin B-type resistance protein VanW	
CLJ_B0263		Vancomycin B-type resistance protein VanW	
CLJ_B0701	cat	Chloramphenicol acetyltransferase	2.3.1.28
CLJ_B0716		MDR-type permease	
Clostridium botulinum A3 str. Loch Maree			
CLK_0108		ABC transporter, ATP-binding protein	
CLK_0221		Vancomycin B-type resistance protein VanW	
CLK_0230	folP	Alternative dihydrofolate reductase 2/dihydropteroate synthase	2.5.1.15
CLK_1474		Chloramphenicol acetyltransferase	2.3.1.28
CLK_2156		Topoisomerase IV subunit A	5.99.1.-
CLK_2157		Topoisomerase IV subunit B	5.99.1.-
CLK_2251		23S rRNA (adenine(2503)-C(Chaudhari et al., 2016))-methyltransferase	2.1.1.224
CLK_3139	gyrA	DNA gyrase subunit A	5.99.1.3
CLK_2932	rpoB	DNA-directed RNA polymerase beta subunit	2.7.7.6
CLK_3396		Vancomycin B-type resistance protein VanW	
CLK_3397		Vancomycin B-type resistance protein VanW	
CLK_0019	cat	Chloramphenicol acetyltransferase	2.3.1.28
CLK_0039		MDR-type permease	
CLK_3430		Chloramphenicol acetyltransferase	2.3.1.28
CLK_3605		Undecaprenyl-diphosphatase	3.6.1.27
Clostridium botulinum BKT015925			
CbC4_0220	rpoB	DNA-directed RNA polymerase beta subunit	2.7.7.6
CbC4_1604		Topoisomerase IV subunit A	5.99.1.-
CbC4_1905		Undecaprenyl-diphosphatase	3.6.1.27
Clostridium botulinum E3 str. Alaska			
CLH_0230		DNA-directed RNA polymerase beta subunit	2.7.7.6
CLH_0747		Topoisomerase IV subunit A	5.99.1.-
CLH_2348	uppP	Undecaprenyl-diphosphatase	3.6.1.27
Clostridium botulinum F str. Langeland			
CLI_2087	cat-2	Chloramphenicol acetyltransferase	2.3.1.28
CLI_0279		Vancomycin B-type resistance protein VanW	
CLI_0280		Vancomycin B-type resistance protein VanW	
CLI_2821		Topoisomerase IV subunit A	5.99.1.-
CLI_2822		Topoisomerase IV subunit B	5.99.1.-
CLI_2911		23S rRNA (adenine(2503)-C(Chaudhari et al., 2016))-methyltransferase	2.1.1.224
CLI_3671	rpoB	DNA-directed RNA polymerase beta subunit	2.7.7.6
CLI_0699	cat-1	Chloramphenicol acetyltransferase	2.3.1.28
CLI_0722		MDR-type permease	
CLI_0007	gyrA	DNA gyrase subunit A	5.99.1.3
CLI_0782		ABC transporter, ATP-binding protein	
CLI_0901		Vancomycin B-type resistance protein VanW	
CLI_0910	folP	Alternative dihydrofolate reductase 2/dihydropteroate synthase	2.5.1.15
Clostridium botulinum F str. 230613			
CBF_2073	catB	Chloramphenicol acetyltransferase (EC)	2.3.1.28
CBF_0248		Vancomycin B-type resistance protein VanW	
CBF_2813		Topoisomerase IV subunit A (EC)	5.99.1.-
CBF_2902		23S rRNA (adenine(2503)-C(Chaudhari et al., 2016))-methyltransferase (EC)	2.1.1.224
CBF_3657	rpoB	DNA-directed RNA polymerase beta subunit (EC 2.7.7.6)	2.7.7.6
CBF_0668		Chloramphenicol acetyltransferase (EC 2.3.1.28)	2.3.1.28
CBF_0690		MDR-type permease	
CBF_0007	gyrA	DNA gyrase subunit A (EC)	5.99.1.3
CBF_0750		ABC transporter, ATP-binding protein	
CBF_0872		Vancomycin B-type resistance protein VanW	
CBF_0881	folP	Alternative dihydrofolate reductase 2/dihydropteroate synthase (EC)	2.5.1.15
Clostridium botulinum H04402065			
H04402_02050		Chloramphenicol acetyltransferase	2.3.1.28
H04402_00203		Vancomycin B-type resistance protein VanW	
H04402_00204		Vancomycin B-type resistance protein VanW	
H04402_02852		Topoisomerase IV subunit A (EC)	5.99.1.-
H04402_02853		Topoisomerase IV subunit B	5.99.1.-
H04402_02943		23S rRNA (adenine(2503)-C(Chaudhari et al., 2016))-methyltransferase	2.1.1.224
H04402_03591		DNA-directed RNA polymerase beta subunit	2.7.7.6
H04402_00694		Chloramphenicol acetyltransferase	2.3.1.28
H04402_00716		MDR-type permease	
H04402_00007		DNA gyrase subunit A	5.99.1.3

(continued on next page)

Table 5 *(continued)*

Locus tag	Gene	Product	EC number
H04402_00755		Beta-lactamase repressor BlaI	
H04402_00880		Vancomycin B-type resistance protein VanW	
H04402_00888		Alternative dihydrofolate reductase 2/dihydropteroate synthase	2.5.1.15
H04402_00956		Chloramphenicol acetyltransferase	2.3.1.28
Clostridium botulinum A str. Hall			
CLC_1966	cat-2	Chloramphenicol acetyltransferase	2.3.1.28
CLC_0270		Vancomycin B-type resistance protein VanW	
CLC_0271		Vancomycin B-type resistance protein VanW	
CLC_2645		Topoisomerase IV subunit A	5.99.1.-
CLC_2646		Topoisomerase IV subunit B	5.99.1.-
CLC_2734		23S rRNA (adenine(2503)-C(Chaudhari et al., 2016))-methyltransferase	2.1.1.224
CLC_3433	rpoB	DNA-directed RNA polymerase beta subunit	2.7.7.6
CLC_0674	cat-1	Chloramphenicol acetyltransferase	2.3.1.28
CLC_0698		MDR-type permease	
CLC_0007	gyrA	DNA gyrase subunit A	5.99.1.3
CLC_0875		Vancomycin B-type resistance protein VanW	
CLC_0883	folP	Alternative dihydrofolate reductase 2/dihydropteroate synthase	2.5.1.25

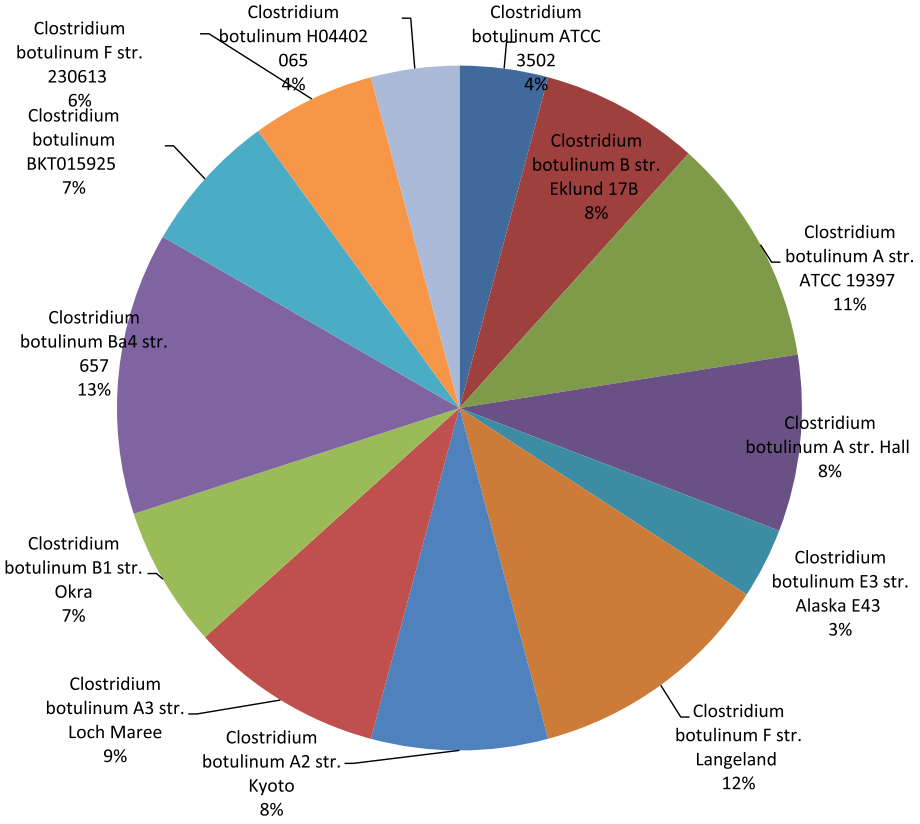


Fig. 9. Pie chart representing total number of genes encoding seven toxin/antitoxin families for 13 *Clostridium* genomes under study in terms of percentage.

Presence of *EutQ* (an ethanolamine utilization protein) indicates the pathogenic potential of *Clostridial* lineage as experimentally validated in *Salmonella* (Roof and Roth, 1988; Chang and Chang, 1975) *Enterococcus* (Del Papa and Perego, 2008), *Erwinia* (Yang et al., 2004), *Flavobacterium* (Castillo et al., 2016), *Klebsiella* (Scarlett and Turner, 1976), *Mycobacterium* (Morth et al., 2004), *Pseudomonas* (Wilson et al., 1966) etc. Ethanolamine acts as a source of carbon and/or nitrogen promoting successful colonization of bacteria into the intestine. In addition, innate immune functions are affected by the breakdown of ethanolamine and disrupt gut functions. *Eut* genes upregulated expression was due to a presence of global virulence regulator *Csr*A regulating SPI-I (*Salmonella* Pathogenicity Island 1) and flagellar genes as in *Salmonella typhimurium* (Garsin, 2010).

Prediction of acquired resistance mechanism in *Clostridial* genomes reflects the universal resistance against vancomycin and chloramphenicol. Resistance mechanisms for glycopeptides, rifampicin, fusidic acid and fosfomycin were not identified. Various topoisomerases and methyltransferases identified indicate antibiotic resistance (Table 5). In addition, evaluating toxin/antitoxin modules in pathogenic bacteria provide platform to researchers for exploring new dimensions of antimicrobial therapies. Toxin/antitoxin module characterized by small operons encoding toxin and cognate antitoxins (Unterholzner et al., 2013). These modules acts as global metabolic stress manager and considered indispensable for bacteria survival in dynamically changing environmental conditions (Pandey and Gerdes, 2005). Either targeting the neutralizing 'antitoxin' or disrupting the inferring interaction between toxin and antitoxin can lead to cell death. Therefore, characterizing the virulent potential of these modules aids in developing novel therapies to suppress

Table 6
Toxin-antitoxin number in each of the seven known toxin-antitoxin families for the 13 *Clostridial* genomes.

Genome	VapI	HipB	MazF	PemK	SpotVB_AbrB	phd/doc	ParE/D	Σ
Clostridium botulinum ATCC 3502	1	1	1	1	1	–	–	5
Clostridium botulinum B str. Eklund 17B	3	3	1	1	1	–	–	9
Clostridium botulinum A str. ATCC 19397	6	7	–	–	–	–	–	13
Clostridium botulinum A str. Hall	5	5	–	–	–	–	–	10
Clostridium botulinum E3 str. Alaska E43	1	1	1	–	–	–	1	4
Clostridium botulinum F str. Langeland	5	7	–	–	1	–	1	14
Clostridium botulinum A2 str. Kyoto	3	5	2	–	–	–	–	10
Clostridium botulinum A3 str. Loch Maree	3	6	1	–	–	–	1	11
Clostridium botulinum B1 str. Okra	4	1	1	–		1	1	8
Clostridium botulinum Ba4 str. 657	6	7	–	2	–	1	–	16
Clostridium botulinum BKT015925	5	3	–	–	–	–	–	8
Clostridium botulinum F str. 230613	3	3	–	1	–	–	–	7
Clostridium botulinum H04402 065	2	1	1	1	–	–	–	5

Table 7
List of accuracy, sensitivity, specificity and ROC area of various bioinformatics tools used for phylogenomic analysis, pan-genome, core-genome and singleton analysis, pathogenomic analysis of 13 *Clostridium botulinum* strains.

S. no.	Software	Accuracy of prediction (%)	Sensitivity	Specificity	ROC area
1.	RNAmmer	97	100%	76.9%	0.893
2.	Absynte	96	100%	68%	0.61
3.	BLASTp	100	100%	n/a	n/a
4.	TBLASTN	100	100%	n/a	n/a
5.	SynMap	94	100%	57.1%	0.903
6.	Gegenees	99	100%	83.3%	0.94
7.	SplitsTree	99	100%	84%	0.997
8.	EDGAR	98	100%	50%	0.601
9.	BPGA	98	100%	76%	0.68
10.	IslandViewer 3	95	100%	44.4%	0.603
11	VirulentPred	97	100%	83.3%	0.893
12.	ResFinder	96	100%	62%	0.66
13.	RASTA	98	100%	81%	0.91
	Average	97.46	100%	57.08%	0.648

the burden of antibiotic resistance (Georgiades and Raoult, 2011). The percentage of genes encoding for seven toxin/antitoxin families were evaluated for *Clostridium* genomes individually and represented in the form of the pie chart (Fig. 9) (Table 6). An average of 97.46% (Table 7) for statistically evaluating the performance of pipeline used to support the study and results obtained.

5. Conclusion

Increased antimicrobial resistance, in hospital environment colonizing the mucous membrane and adaptation in host emerges the need to restrict the evolution of *Clostridium* pathogenesis towards the human population. Increased pathogenic potential of endemic lineages are due to the presence of drug resistant determinants still the contribution of the other genetic elements accounting for intrinsic pathogenic capability is still unclear. This study illustrates the comparative genomic and pan-genomic analysis of multiple *Clostridium botulinum* genomes. Core-genome calculation reveals high genomic similarity among the genomes with not so variable GC content. An open genome endowed by *Clostridial* lineage indicates the possibility of the addition of new gene sets with every genome introduction along with novel strain specific genes (Singletons). The persistence of singletons represents the ability to acquire novel virulence traits, resulting in a threat to the human population. Further, identification and pathogenomic analysis of genomic islands (GIs) using machine learning approaches and bi-layer computational strategies aid in characterizing potential vaccine and therapeutic targets.

Acknowledgement

The authors are thankful to Department of Biotechnology, TERI University, New Delhi for providing the facility and technical support during the preparation of the manuscript.

References

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., et al., 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252 (5013), 1651–1656.
Bernard, G., Chan, C.X., Ragan, M.A., 2016. Alignment-free microbial phylogenomics under scenarios of sequence divergence, genome rearrangement and lateral genetic transfer. Sci. Rep. 6, 28970.
Bhardwaj, T., Somvanshi, P., 2014. Plant Systems Biology: Insights and Advancements. In: Plant Omics: The Omics of Plant Science. Springer Publications, 978-81-322-2171-5, pp. 791–819.
Blom, J., Albaum, S.P., Doppmeier, D., Puhler, A., Vorhölter, F.J., Zakrzewski, M., 2009. EDGAR: a software framework for the comparative analysis of prokaryotic genomes. BMC Bioinforma. 20 (10), 154.
Bobik, T.A., Xu, Y., Jeter, R.M., Otto, K.E., Roth, J.R., 1997. Propanediol utilization genes (*pdu*) of *Salmonella typhimurium*: three genes for the propanediol dehydratase. J. Bacteriol. 179 (21), 6633–6639.
Brazier, J.S., Duerden, B.I., Hall, V., Salmon, J.E., Hood, J., Brett, M.M., McLauchlin, J., George, R.C., 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.
Castillo, D., Chritiansen, R.H., Dalsgaard, I., Madsen, L., Espejo, R., Middelboe, M., 2016. Comparative genome analysis provides insights into the pathogenicity of *Flavobacterium psychrophilum*. PLoS One 11 (4), e0152515.
Chang, G.W., Chang, J.T., 1975. Evidence for the B12-dependent enzyme ethanolamine deaminase in *Salmonella*. Nature 254, 150–151.
Chaudhari, N.M., Gupta, V.K., Dutta, C., 2016. BPGA — an ultra-fast pan-genome analysis pipeline. Sci. Rep. 6, Article number: 24373.
Darby, A.C., Cho, N.H., Fuxelius, H.H., Westberg, J., Andersson, S.G., 2007. Intracellular pathogens go extreme: genome evolution in the *Rickettsiaes*. Trends Genet. 23, 511–520.
David, A.R., Rosovitz, M.J., Myers, M.J., Mongodin, E.F., Frickel, W.F., Gajer, P., et al., 2008. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. J. Bacteriol. 190 (20), 6881–6893.
Del Papa, M.F., Perego, M., 2008. Ethanolamine activates a sensor histidine kinase regulating its utilization in *Enterococcus faecalis*. J. Bacteriol. 9 (1), 43–46.
DelVecchio, V.G., Kapatral, V., Elzer, P., Patra, G., Mujer, C.V., 2002. The genome of *Brucella melitensis*. Vet. Microbiol. 90, 587–592.
Despalins, A., Marsit, S., Oberto, J., 2011. Absynte: a web tool to analyze the evolution of orthologous archaeal and bacterial gene clusters. Sci. Math. Bioinforma. 27 (20), 2905–2906.
Edmond, B.J., Guerra, F.A., Blake, J., Hempler, S., 1977. Case of infant botulism in Texas. Tex. Med. 73 (10), 85–88.
Eng, J., 2013. ROC analysis: web-based calculator for ROC curves. Johns Hopkins University, Baltimore. (Available from: http://www.jrocfit.org).
Garg, A., Gupta, D., 2008. VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. BMC Bioinforma. 9, 62.
Garsin, D.A., 2010. Ethanolamine utilization in bacterial pathogens: roles and regulation. Nat. Rev. Microbiol. 8 (4), 290–295.
Georgiades, K., Raoult, D., 2011. Genomes of the most dangerous epidemic bacteria have a virulence repertoire characterized by fewer genes but more toxin-antitoxin modules. PLoS One 6, e17962.
Ghatak, S., Blom, J., Das, S., Sanjukta, R., Puro, K., 2016. Pan-genome analysis of *Aeromonas hydrophila*, *Aeromonas veronii* and *Aeromonas caviae* indicates phylogenomic diversity and greater pathogenic potential for *Aeromonas hydrophila*. Antonie Van Leeuwenhoek 109 (7), 945–956.

Gotoh, O., 1982. An improved algorithm for matching biological sequences. J. Mol. Biol. 162, 705–708.
Gutacker, M.M., Mathema, B., Soini, H., Shashkina, E., Kreiswirth, B.N., Graviss, E.A., et al., 2006. Single-nucleotide polymorphism-based population genetic analysis of *Mycobacterium tuberculosis* strains from 4 geographic sites. J. Infect. Dis. 193, 121–128.
Haft, D.H., Selengut, J.D., Brinkac, L.M., Zafar, N., White, O., 2005. Genome properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. Bioinformatics 21, 293–306.
Harris, S.R., Feil, E.J., Holden, M.T., Quail, M.A., Nickerson, E.K., Chantratita, N., et al., 2010. Evolution of MRSA during hospital transmission and intercontinental spread. Science 327, 469–474.
Hsiao, W., Wan, I., Jones, S.J., Brinkman, F.S., 2003. IslandPath: aiding detection of genomic islands in prokaryotes. Bioinformatics 19, 418–420.
Humeau, Y., Doussau, F., Grant, N.J., Poulain, B., 2000. How botulinum and tetanus neurotoxins block neurotransmitter release. Biochimie 82, 427–446.
Huson, D.H., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23, 254–267.
Johnson, E.A., Bradshaw, M., 2001. *Clostridium botulinum* and its neurotoxins: a metabolic and cellular perspective. Toxicon 39 (1703–1), 722.
Katz, L.S., Petkau, A., Beaulaurier, J., Tyler, S., Antonova, E.S., et al., 2013. Evolutionary dynamics of *Vibrio cholerae* O1 following a single-source introduction to Haiti. MBio 4.
Kloepper, T.H., Huson, D.H., 2008. Drawing explicit phylogenetic networks and their integration into SplitsTree. BMC Evol. Biol. 8, 22.
Lagesen, K., Hallin, P., Rødland, E.A., Stærfeldt, H., Rognes, T., Ussery, D.W., 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35 (9), 3100–3108.
Langille, M., Hsiao, W., Brinkman, F., 2008. Evaluation of genomic island predictors using a comparative genomics approach. BMC Bioinforma. 9, 329.
Lerat, E., Daubin, V., Moran, N.A., 2003. From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-proteobacteria. PLoS Biol. 1, E19.
Lund, B.M., Peck, M.W., 2000. Clostridium botulinum. In: Lund, B.M., et al. (Eds.), The Microbiological Safety and Quality of Food. Aspen, Gaithersburg, MD, pp. 1057–1109.
Lyons, E., Pedersen, B., Kane, J., Alam, M., Ming, R., Tang, H., Wang, X., Bowers, J., Paterson, A., Lisch, D., et al., 2008. Finding and comparing systemic regions among Arabidopsis and the out groups papaya, poplar, and grape: CoGe with rosins. Plant Physiol. 148 (4), 1772–1781.
Maiden, M.C., van Rensburg, M.J., Bray, J.E., Earle, S.G., Ford, S.A., et al., 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat. Rev. Microbiol. 11, 728–736.
Maksymowych, A.B., 1999. Pure botulinum neurotoxin is absorbed from the stomach and small intestine and produces peripheral neuromuscular blockade. Infect. Immun. 67, 4708–4712.
Méric, G., Yahara, K., Mageiros, L., Pascoe, B., Maiden, M.C.J., Jolley, K.A., et al., 2014. A reference pan-genome approach to comparative bacterial genomics: identification of novel epidemiological markers in pathogenic *Campylobacter*. PLoS One 9 (3), e92798.
Meyer, F., Goesmann, A., McHardy, A.C., Bartels, D., Bekel, T., et al., 2003. GenDB—an open source genome annotation system for prokaryote genomes. Nucleic Acids Res. 31, 2187–2195.
Mira, A., Martin- Cuadrado, A.B., D' Auria, G., Rodriguez-Valera, F., 2010. The bacterial pan-genome: a new paradigm in microbiology. Int. Microbiol. 13, 45–57.
Moran, N.A., Plague, G.R., 2004. Genomic changes following host restriction in bacteria. Curr. Opin. Genet. Dev. 14, 627–633.
Morth, J.P., Feng, V., Perry, L.J., Svergun, D.I., Tucker, P.A., 2004. The crystal and solution structure of a putative transcriptional antiterminator from *Mycobacterium tuberculosis*. Structure 12, 1595–1605.
Olsvik, O., Wahlberg, J., Petterson, B., Uhlen, M., Popovic, T., Wachsmuth, I.K., Fields, P.I., 1993. Use of automated sequencing of polymerase chain reaction-generated amplicons to identify three types of cholera toxin subunit B in *Vibrio cholerae* O1 strains. J. Clin. Microbiol. 31, 22–25.
Pagani, I., Liolios, K., Jansson, J., Chen, I.M., Smirnova, T., Nosrat, B., Markowitz, V.M., Kyrpides, N.C., 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.
Palm, D., Johansson, K., Ozin, A., Friedrich, A., Grundmann, H., Larsson, J., Struelens, M., 2012. Molecular Epidemiology of Human Pathogens: How to Translate Breakthroughs Into Public Health Practice. Stockholm. (November 2011).
Pandey, D.P., Gerdes, K., 2005. Toxin-antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes. Nucleic Acids Res. 33, 966–976.
Prabha, R., Singh, D.P., Sinha, S., Ahmad, A., Rai, A., 2016. Genome-wide comparative analysis of codon usage bias and codon context patterns among cyanobacterial genomes. Mar. Genomics 1874-7787 (16), 30120–30129.
Rocha, E.P., 2004. The replication-related organization of bacterial genomes. Microbiology 150, 1609–1627.
Rocha, E.P., Danchin, A., 2002. Base composition bias might result from competition for metabolic resources. Trends Genet. 18, 291–294.
Rohde, H., Qin, J., Cui, Y., Li, D., Loman, N.J., Hentschke, M., et al., 2011. Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. N. Engl. J. Med. 365, 718–724.
Roof, D.M., Roth, J.R., 1988. Ethanolamine utilization in *Salmonella typhimurium*. J. Bacteriol. 170, 38556.
Rouli, L., Merhej, V., Fournier, P.E., Raoult, D., 2015. The bacterial pangenome as a new tool for analyzing pathogenic bacteria. New Microbes New Infect. 7, 72–85.
Sahl, J.W., Gillece, J.D., Schupp, J.M., Waddel, L.V.G., Driebe, E.M., Engelthaler, D.M., et al., 2013. Evolution of a pathogen: a comparative genomics analysis identifies a genetic pathway to pathogenesis in *Acinetobacter*. PLoS One 8 (1), e54287.
Scarlett, F.A., Turner, J.M., 1976. Microbial metabolism of amino alcohols. Ethanolamine catabolism mediated by coenzyme B12-dependent ethanolamine ammonia-lyase in *Escherichia coli* and *Klebsiella aerogenes*. J. Gen. Microbiol. 95, 173–176.
Sevin, E.W., Barloy-Hubler, F., 2007. RASTA-bacteria: a web-based tool for identifying toxin-antitoxin loci in prokaryotes. Genome Biol. 8 (8), R155.
Shapiro, R.L., Hatheway, C., Swedlow, D.L., 1998. Botulism in the United States: a clinical and epidemiologic review. Ann. Intern. Med. 129, 221–228.
Sheppard, S.K., Didelot, X., Jolley, K.A., Darling, A.E., Pascoe, B., Meric, G., et al., 2013. Progressive genome-wide introgression in agricultural *Campylobacter coli*. Mol. Ecol. 22, 1051–1064.
Shukla, H.D., Sharma, S.K., Singh, B.R., 1997. Identification of a hemagglutinin present in the neurotoxin complex of *Clostridium botulinum* type A, as a heat shock protein. J. Infect. Dis. 123, 34.
Smith, T.J., Hill, K.K., Foley, B.T., Detter, J.C., Munk, A.C., Bruce, D.C., Doggett, N.A., Smith, L.A., Marks, J.D., Xie, G., Brettin, T.S., 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.
Soares, S.C., Silva, A., Trost, E., Blom, J., Ramos, R., Carneiro, A., et al., 2013. The pan-genome of the animal pathogen *Corynebacterium pseudotuberculosis* reveals differences in genome plasticity between the biovar ovis and equi strains. PLoS One 8 (1), e53818.
SubbaRao, G.M., 2007. Food Safety Knowledge, Attitudes and Practices of Mothers—Findings from Focus Group Studies in South India. National Institute of Nutrition (NIN).
Tamas, I., Klasson, L., Canback, B., Naslund, A.K., Eriksson, A.S., Wernegreen, J.J., et al., 2002. 50 million years of genomic stasis in endosymbiotic bacteria. Science 296, 2376–2379.
Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., et al., 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial pan-genome. Proc. Natl. Acad. Sci. U. S. A. 102, 13950–13955.
Tettelin, H., Riley, D., Cattuto, C., Medini, D., 2008. Comparative genomics: the bacterial pan-genome. Curr. Opin. Microbiol. 11, 472–477.
Toh, T., Martin, W., Nei, M., 2002. Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. Proc. Natl. Acad. Sci. U. S. A. 99, 12944–12948.
Tracy, B.P., Jones, S.W., Fast, A.G., Indurthi, D.C., Papoutsakis, E.T., 2011. *Clostridia*: the importance of their exceptional substrate and metabolite diversity for biofuel and biorefinery applications. Curr. Opin. Biotechnol. 23 (3), 364–381.
Unterholzner, S.J., Poppenberger, B., Rozhon, W., 2013. Toxin–antitoxin systems: biology, identification, and application. Mob. Genet. Elem. 3 (5), e26219.
Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W., Surovcik, K., Meinicke, P., Merkl, R., 2006. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. BMC Bioinforma. 7, 142.
Wernegreen, J.J., 2005. For better or worse: genomic consequences of intracellular mutualism and parasitism. Curr. Opin. Genet. Dev. 15, 572–583.
Wernegreen, J.J., Degnan, P.H., Lazarus, A.B., Palacios, C., Bordenstein, S.R., 2003. Genome evolution in an insect cell: distinct features of an anti-bacterial partnership. Biol. Bull. 204, 221–231.
Wilson, S.A., Wachira, S.J., Norman, R.A., Pearl, L.H., Drew, R.E., 1966. Transcription antitermination regulation of the *Pseudomonas aeruginosa* amidase operon. EMBO J. 15, 5907–5916.
Woodruff, B.A., Griffin, P.M., McCroskey, L.M., Smart, J.F., Wainwright, R.B., Bryant, R.G., Hutwagner, L.C., Hatheway, C.L., 1992. Clinical and laboratory comparison of botulism from toxin types A, B, and E in the United States, 1975–1988. J. Infect. Dis. 166, 1281–1286.
Yang, S., Perna, N.T., Cooksey, D.A., Okinaka, Y., Lindow, S.E., Ibekwe, A.M., Keen, N.T., Yang, C.H., 2004. Genome-wide identification of plant-upregulated genes of *Erwinia chrysanthemum* 3937 using a GFP-based IVET leaf array. Mol. Plant-Microbe Interact. 17, 999–1008.
Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., et al., 2012. Identification of acquired antimicrobial resistance genes. J. Antimicrob. Chemother. 67 (11), 2640–2644.
Zhang, L., Xiao, D., Pang, B., Zhang, Q., Zhou, H., Zhang, L., et al., 2014. The core proteome and pan proteome of *Salmonella* paratyphi A epidemic strains. PLoS One 9 (2), e89197.