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).

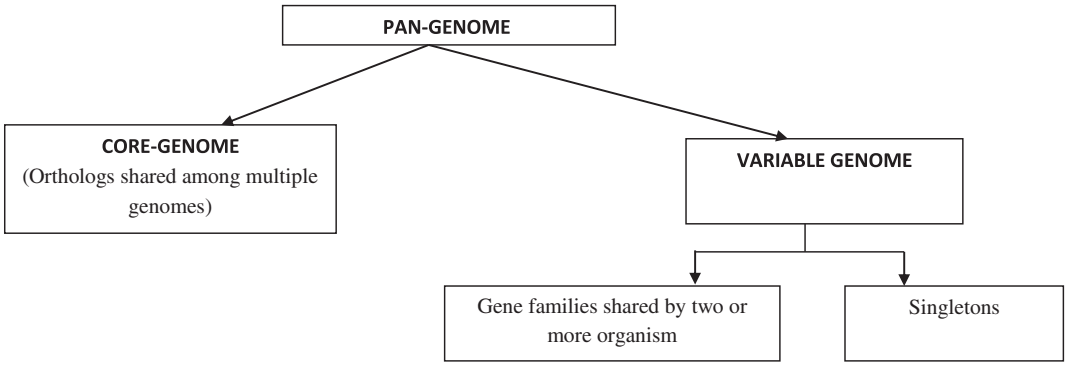

Fig. 1. Pan-genome vs. core-genome.