Indian J Microbiol (Apr–June 2015) 55(2):140–150
DOI 10.1007/s12088-015-0514-z

ORIGINAL ARTICLE

Insights into the Origin of *Clostridium botulinum* Strains: Evolution of Distinct Restriction Endonuclease Sites in *rrs* (16S rRNA gene)

Ashish Bhushan · Tanmoy Mukherjee · Jayadev Joshi · Pratap Shankar · Vipin Chandra Kalia

Received: 2 January 2015 / Accepted: 12 January 2015 / Published online: 18 January 2015
© Association of Microbiologists of India 2015

Abstract Diversity analysis of *Clostridium botulinum* strains is complicated by high microheterogeneity caused by the presence of 9–22 copies of *rrs* (16S rRNA gene). The need is to mine genetic markers to identify very closely related strains. Multiple alignments of the nucleotide sequences of the 212 *rrs* of 13 *C. botulinum* strains revealed intra- and inter-genomic heterogeneity. Low intragenomic heterogeneity in *rrs* was evident in strains 230613, Alaska E43, Okra, Eklund 17B, Langeland, 657, Kyoto, BKT015925, and Loch Maree. The most heterogenous *rrs* sequences were those of *C. botulinum* strains ATCC 19397, Hall, H04402065, and ATCC 3502. *In silico* restriction mapping of these *rrs* sequences was observable with 137 type II Restriction endonucleases (REs). Nucleotide changes (NC) at these RE sites resulted in appearance of distinct and additional sites, and loss in certain others. De novo appearances of RE sites due to NC were recorded at different positions in *rrs* gene. A nucleotide transition A>G in *rrs* of *C. botulinum* Loch Maree and 657 resulted in the generation of 4 and 10 distinct RE sites, respectively. Transitions A>G, G>A, and T>C led to the loss of RE sites. A perusal of the entire NC and in silico RE mapping of *rrs* of *C. botulinum* strains provided insights into their evolution. Segregation of strains on the basis of RE digestion patterns of *rrs* was validated by the cladistic analysis involving six house keeping genes: *dnaN, gyrB, metG, prfA, pyrG,* and *Rho*.

Keywords *Clostridium* · Evolution · Microheterogeneity · Phylogeny · Restriction endonuclease

Introduction

Clostridium botulinum strains have been classified by the Center for Disease Control and Prevention as "Category A agents" with highest—risk threat especially for bioterrorism [1, 2]. Botulinum neurotoxins (BoNTs) produced by *C. botulinum* are extremely lethal, such that 3 g are sufficient to kill the entire population of United Kingdom, and 400 g to wipe out the whole mankind [1, 3]. The need is to identify molecular markers for distinguishing closely related strains [2, 4]. Bacterial identification through sequence analysis of 16S rRNA gene (*rrs*) has been widely exploited [5, 6].

Challenges in identification of *C. botulinum* are the variations arising largely due to the different types (A–G) of neurotoxins [7, 8]. Phylogenetic lineages are of 4 types: (1) Group I—proteolytic *C. botulinum* types A, B and F, and *C. sporogenes,* (2) Group II—nonproteolytic types B, E and F, (3) Group III—types C and D and *C. novyi* type A, and (4) *Clostridium argentinense* (*C. botulinum* type G), related to *C. subterminale* [7, 9, 10]. BoNT toxin is encoded by the *bont* gene. Another feature of genetic variability among *C. botulinum*, is the multiple copies of *rrs* in each genome (http://rrndb.umms.med.umich.edu/search/). Recent studies have shown certain unique features within *rrs,* such as 30–50 nts signatures and Restriction endonucleases (RE) digestion patterns. These can be exploited to identify organisms [5, 6, 11, 12]. *In silico* RE

Electronic supplementary material The online version of this article (doi:10.1007/s12088-015-0514-z) contains supplementary material, which is available to authorized users.

A. Bhushan (✉) · T. Mukherjee · J. Joshi · P. Shankar · V. C. Kalia
Microbial Biotechnology and Genomics, CSIR-Institute of Genomics and Integrative Biology (IGIB), Delhi University Campus, Mall Road, Delhi 110007, India
e-mail: bhushana1422@gmail.com; ashish_bhushan84@yahoo.com

digestion patterns with an array of these enzymes revealed variations in these sites yielding unknown profiles. Of the various REs used in our previous study involving 128 *rrs* sequences of *C. botulinum*, RE-BfaI digestion pattern—5′ 195-20-163-29-146-185-331 (nts) 3′ was reported to be unique. It was thus proposed for usage as a molecular marker for species level identification [6]. However, certain naturally occurring genetic variations in RE sites were observed in closely related strains. These modifications were observed due to changes in the nucleotides of the RE recognition sites. Two questions were raised (1) Can we use these genetic changes as marker to distinguish closely related strains? and (2) What is the evolutionary significance of these nucleotide changes (NC) within RE sites.

Materials and Methods

rrs-Genome Sequence Data Collection

Information on the *rrs* of 13 completed genome sequences of *C. botulinum* used in this study was downloaded from Ribosomal Database Project II (http://rdp.cme.msu.edu/genome/) (Table S1). The genome sizes of different strains of *C. botulinum* were obtained from REBASE Genomes (http://tools.neb.com/~vincze/genomes/), their GC contents (% mol) were retrieved from NCBI database (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html) and the copy numbers of *rrs* within the complete genomes were obtained from rrndb database (http://rrndb.umms.med.umich.edu/search/) [13, 14]. The GC % of each *rrs* was calculated for all *C. botulinum* genomes using BioEdit [17] (Table S1).

Analysis of Intra- and Inter-genomic Heterogeneity

A comparative study of intra- and inter-genomic heterogeneity in *rrs* of 13 completely sequenced genomes of *C. botulinum* was performed by multiple sequence alignment using Clustal X version 2.0.12 followed by Data Analysis in Molecular Biology and Evolution (DAMBE) software Package [16, 17]. It enabled us to bring down the redundancy among the 212 copies of *rrs* of 13 different strains (Tables S1 and S2) to 130 representatives: (1) 130 copies of *rrs* could be reduced to 49 representatives due to 100 % similarity among them, and (2) 81 *rrs* copies showed distinct heterogeneity. Hence, all subsequent analyses were based on these 131 representative copies of *rrs*.

Heterogeneity Analysis of *rrs*

Intragenomic heterogeneity among all the *rrs* sequences within each of the 13 genomes of *C. botulinum* strains was analysed by multiple sequence alignment (Clustal X version 2.0.12): (1) between non substituted *rrs* sequence (Okra, S001014409) and others and by counting the nucleotide changes (NC) with the help of BioEdit, and (2) pairwise alignment between two *rrs* within the strain and by counting NC using BioEdit (Table S1) [15, 16].

For the intergenomic heterogeneity, non substituted *rrs* sequences of 13 different *C. botulinum* strains were aligned by multiple sequence alignment (Clustal X version 2.0.12) and a completely non-substituted representative *rrs* sequence (Okra, S001014409) of *C. botulinum* was identified using BioEdit (Table S1) [16, 17]. The intergenomic heterogeneity was calculated by counting NC in the reference sequence which was chosen from representative sequences with the help of BioEdit [16, 17].

Restriction Endonuclease Analysis

A total of 241 Type II REs consisting of 4–6 nts cutters and larger recognition sites (>6 nts) listed in BioEdit [15] were considered for these analyses. Since 104 REs proved to be non-cutters, only 137 REs were used for further analyses (Table S3). Then we concentrated on those RE sites which were common to all the strains and especially those where NCs led to changes in RE sites. Thus only those nucleotides which affected the RE sites were considered for data matrix. Consensus RE patterns—frequency of occurrence of RE sites and the pattern of fragments (nts) were determined for each *rrs* by employing: (1) 4 nts cutters—BstUI(CG'CG), HpyCH4V(TG'CA), TaqI(T'CGA), and Tsp509I('AATT), (2) 5 nts cutters—HpyCH4III(ACn'GT), Hpy188I(TCn'GA), and Tsp45I('GTsAC);and (3) 6 nts cutters—NlaIV(GGn'nCC) and Hpy188III(TC'nnGA) [5] (Tables S4, S5, S6).

Analysis of Nucleotide Changes

NCs leading to changes in RE sites in *rrs* genes of different *C. botulinum* genomes were detected by comparing them with *C. botulinum* strain Okra *rrs*: S001014409 as reference (Table S1). NCs in RE sites within *rrs* were categorized as: (1) appearance of distinct ones, (2) creation of additional RE sites, (3) loss of RE sites, and (4) no evident change in RE sites (Tables 1, 2, S7, S8, S9). In addition, certain NCs were also observed in the non-RE sites in all the strains.

Phylogenetic Analysis

For phylogenetic analysis, *rrs* sequences from all the 13 *C. botulinum* strains and their close phylogenetic associates as reported in literature [2, 8, 18] like *C. argentinense, C. beijerinckii, C. butyricum, C. haemolyticum, C. novyi, C.*

Table 1 Occurrence of distinct restriction endonuclease sites in *rrs* of *Clostridium botulinum* strain

Clostridium botulinum strain	*rrs*		Restriction endonuclease and site[b]	
	Accession No.	NC[a]	Pre-substitution	Post-substitution
657	S001416095	222A>G	AvaII–G'GwCC, NlaIV–GGn'nCC, Sau96I–G'GnCC	ApaI–GGGCC'C, BanII–GrGCy'C, Bme1580I–GkGCm'C, Bsp1286I–GdGCh'C, Cac8I–GCn'nGC, CviJI–rG'Cy, HaeIII–GG'CC, NlaIV–GGn'nCC, PspOMI–G'GGCCC, Sau96I–G'GnCC
	S002289967	229A>G		
Loch Maree	S001014501	1319A>G	HinfI–G'AnTC, Hpy188III–TC'nnGA, NruI–TCG'CGA, TfiI–G'AwTC	AlwI–GATCnn, BstKTI–GAT'C, DpnI–GA'TC, MboI–'GATC
	S002290692	1326A>G		
657	S001416092	1420A>G	NlaIV–GGn'nCC	Cac8I–GCn'nGC, CviJI–rG'Cy, NlaIV–GGn'nCC
	S002289183	1427A>G		
Eklund 17B	S001094720	77G>A	Hpy188I–TCn'GA	BstBI–TT'CGAA, TaqI–T'CGA
	S002288156	83G>A		
ATCC 3502	S000858488	118C>T	Hpy188I–TCn'GA	BstBI–TT'CGAA, TaqI- T'CGA
	S002290042	152C>T		
ATCC 3502	S000858488	363C>T	AciI–C'CGC, Fnu4HI–GC'nGC, TauI–GCsG'C	Hpy99I–CGwCG', HpyCH4IV–A'CGT, TaiI–ACGT'
	S002290042	397C>T		
ATCC 19397	S000891587[c]	1220C>T	HgaI–GACGCnnnnn'nnnnn	HpyCH4 V–TG'CA
	S002289359[c]	1227C>T		
Eklund 17B	S001094707	1232C>T	AciI–C'CGC, BstUI–CG'CG	HpyCH4III–ACn'GT
	S001094714[d]			
	S001094720			
	S002289462	1238C>T		
	S002290368[d]			
	S002288156			
ATCC 3502	S000858482	427T>C	BfaI–C'TAG	KpnI–GGTAC'C, NlaIV–GGn'nCC
	S000858488			
	S000858491			
	S002287580	461T>C		
	S002290042			
	S002287460			
ATCC 19397	S000891578[c]	454T>C	BfaI–C'TAG	Acc65I–G'GTACC, BanI–G'GyrCC, KpnI–GGTAC'C, NlaIV–GGn'nCC
	S000891580[c]			
	S000891587[c]			
	S002287857[c]	461T>C		
	S002288115[c]			
	S002289359[c]			

Table 1 continued

Clostridium botulinum strain	*rrs*		Restriction endonuclease and site[b]	
	Accession No.	NC[a]	Pre-substitution	Post-substitution
Kyoto	S001350515	454T>C		
	S002290406	461T>C		
Langeland	S000891621[d]	454T>C		
	S002288829[d]	461T>C		
230613	S002165522[d]	454T>C		
H04402 065	S002408098	454T>C		
	S002408093	861G>A	HpyCH4III -AC_n'GT BsiEI-CG_ry'CG	MboI-'GATC_, DpnI-GA'TC, BsiEI-CG_ry'CG, PvuI-CG_AT'CG, BstKTI-G_AT'C
BKT015925	S002441243	581C>T	ApoI-r'AATT_y, Tsp509I-'AATT	ApoI-r'AATT_y, Tsp509I-'AATT, EcoRI-G'AATT_C

[a] Nucleotide changes are designated by a ">"-character, 222A > G denotes that at nucleotide position 222 a A is changed to a G

[b] y = C or T, r = G or A, w = T or A, k = T or G, s = G or C, m = C or A, d = A or G or T, h = A or C or T

[c] Intergenomic sequences (see Table S2) with changes at similar position

[d] Intragenomic sequences (see Table S2) with changes at similar position

proteolyticus, C. schirmacherense, C. sporogenes and *C. subterminale* were assembled and aligned using the multiple alignment program Clustal X version 2.0.12 [16]. In order to estimate evolutionary distance, pairwise distances between all species were calculated with the DNADIST of the PHYLIP 3.69 package [5, 19]. The resultant distance matrix was then used to draw a neighbour joining tree with the program NEIGHBOR [19]. The program SEQBOOT [19] was used for statistical testing of the trees by resampling the dataset 1,000 times. The trees were viewed through TreeView Version 1.6.6 [20]. A phylogenetic tree of 130 *rrs* sequences of *C. botulinum* and their close associates was drawn to check if these 131 *rrs* sequences show closeness exclusively to BoNT producers—*C. argentinense,* and *C. butyricum,* or not. We did not observe any discrepancy in the phylogenetic relationships of *C. botulinum* strains with those reported in literature (Fig. S1).

Cladistic Analysis

For cladistic analysis, nt sequences of 38 HKGs (Table S10) (*abrB, arcA1, atpA, atpC, bofA, ccpA, cphA, cphB, dltB, dltD, dnaA, dnaN, gyrA, gyrB, ispE, ispF, ksgA, mdeA, metG, mnaA, nrdD, nrdG, prfA, psd1, psd2, pyrG, recF, recR, Rho, rpiB, rpmE, serS, sfsA, spmA, spmB, tagO, tdK,* and *Upp*) of 13 *C. botulinum* strains were retrieved from NCBI (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html) in FASTA format. For cladistic analysis, initially trees for all the 38 genes were drawn. Six (*dnaN, gyrB, metG, prfA, pyrG,* and *Rho*) out of these 38 genes were selected on the basis of the high frequency of the close relationship between 2 and 3 strains and compared them with the relationship recorded with *rrs*.

For the final presentation of Cladistic analysis, seven concatenated HKGs including *rrs* of *C. botulinum* strains (Table S1 and S10) were assembled and aligned using the multiple alignment program Clustal X version 2.0.12 [16], and saved in NEXUS format. The maximum parsimony tree was constructed with from the PAUP* (Phylogenetic Analysis Using Parsimony) ver. 4.10b [21] using heuristic search methods and bootstrap test with 100 replicates (Fig. 1).

Results

The availability of thirteen completely sequenced genomes of *C. botulinum* of diverse geographic origins with high microheterogeneity (9–22 copies of *rrs* per genome) (Table S1) [2] prompted us to use them as a model system for exploring unique molecular markers in their *rrs* gene sequences. Comparisons of *rrs* sequences within and between genomes revealed that the most heterogenous *rrs*

Table 2 Evolution of additional restriction endonuclease sites due to nucleotide changes in *rrs* of *Clostridium botulinum* strains

Clostridium botulinum strain	*rrs*		Restriction endonuclease and site[b]
	RDP Accession No.	NC[a]	
ATCC3502	S000858480, S000858482	52G>A	Hpy188III–TC'nnGA
	S000858488, S000858491		
	S002290179, S002287580	86G>**A**	
	S002290042, S002287460		
Loch Maree	S001014494, S001014499	67A>G	Alu I-AG'CT, CviJI-rG'Cy, HindIII–A'AGCTT
H04402 0 65	S002408093		
BKT015925	S002441251, S002441255, S002441247	73A>**G**	
Loch Maree	S002287648, S002290134	74A>**G**	
ATCC 19397	S000891578[c]	79G>A	Hpy188I–TCn'GA
	S002287857	86G>**A**	
	S000891584[c]	78G>A	
	S002290093[C]	85G>**A**	
657	S001416088	80A>G	EcoNI–CCTnn'nnnAGG
	S002287409	87G>**A**	BslI–CCnnnnn'nnGG
Loch Maree	S001014501	82T>C	AciI–C'CGC[d]
H04402 0 65	S002408095	82T>C	AciI–G'CGG
Loch Maree	S002290692	89T>**C**	
657	S001416082[e]	84G>**A**	Tsp509I–'AATT
	S002290186[C]	91G>**A**	
Kyoto	S001350513	84G>**A**	
	S002288993	91G>**A**	
Okra	S001014404	171T>**C**, 174A>G	FatI–'CATG, NlaIII–CATG'[f]
	S002288021	178T>**C**, 181A>G	
Alaska E43	S001094753[e], S001094755, S001094758, S001094762	993G>**A**	MaeIII–'GTnAC, Tsp45I–'GTsAC, AleI–CACnn'nnGTG, MslI–CAynn'nnrTG
	S002289948[C], S002288105, S002290927, S002287743	999G>**A**	
	S001094753[e], S001094755, S001094758, S001094762	1000C>T	HphI –GGTGAnnnnnnnn'
	S002289948[C], S002288105, S002290927, S002287743	1006C>**T**	
ATCC3502	S000858491	1193C>T	HpyCH4 V–TG'CA
Langeland	S000891610	1220C>T	
Loch Maree	S001014484[e]	1220C>**T**	
Okra	S001014404, S001014407	1220C>T	
H04402 0 65	S002408089, S002408093, S002408095	1220C>T	
Langeland	S002287184	1227C>**T**	
ATCC3502	S002287460	1227C>**T**	
Loch Maree	S002290255[C]	1227C>**T**	
Okra	S002288021, S002288411	1227C>**T**	

[a] Nucleotide changess are designated by a ">"-character, 52G>A denotes that at nucleotide 52 a G is changed to a A

[b] y = C or T, r = G or A, s = G or C

[c] Intergenomic sequences (Table S2) with changes at similar position

[d] AciI made 21/23 cuts at C'CGC and 2/23 cuts at G'CGG in Loch Maree strain only

[e] Intragenomic sequences (see Table S2) with changes at similar position

[f] RE site appeared due to simultaneous changes at positions 171 and 174

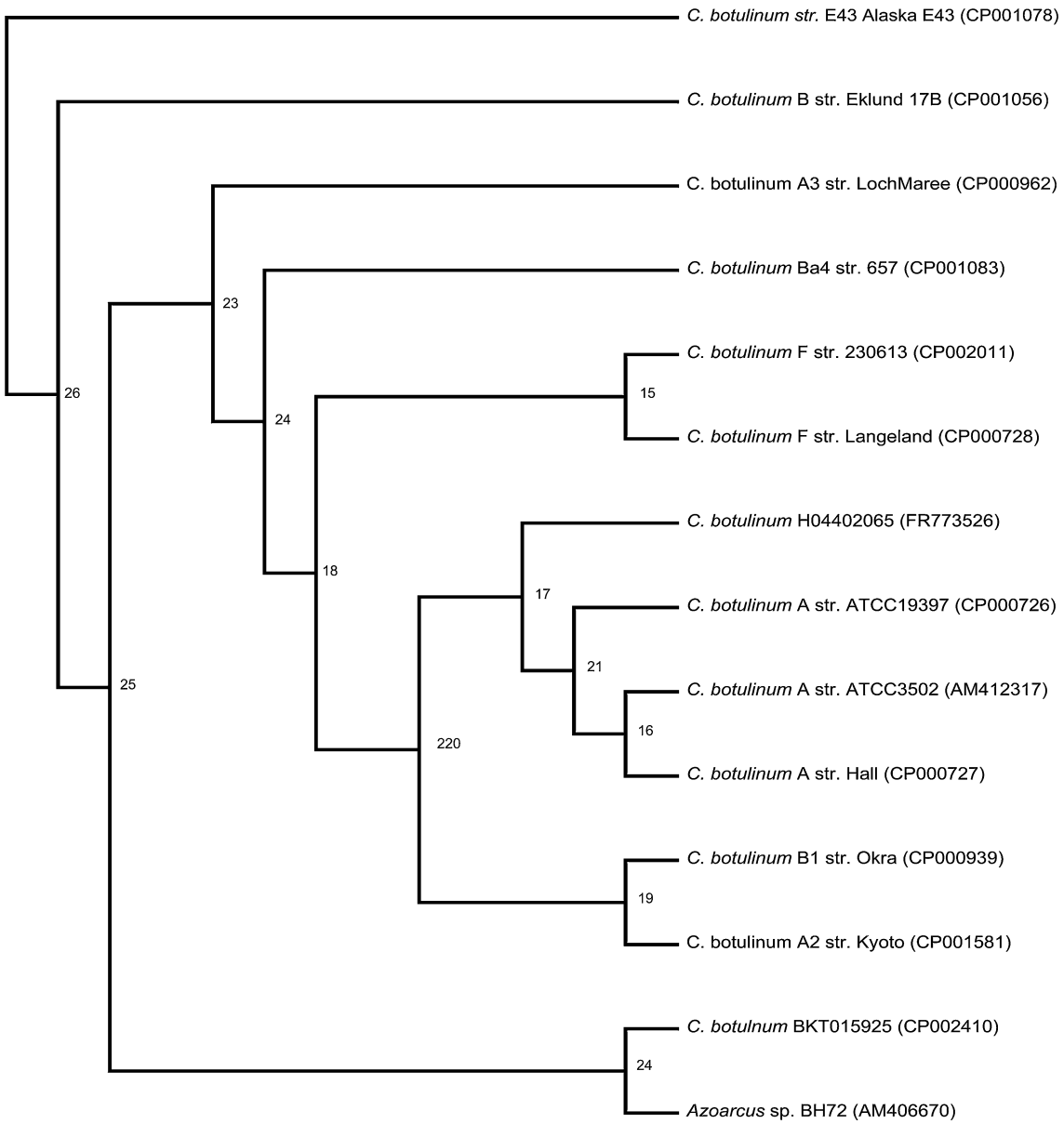

Fig. 1 Maximum parsimony (MP) tree for the concatenated six HKGs with *rrs*. Most-parsimonious tree generated from the concatenated gene sequences of *pyrG, gyrB, dnaN, metG, prfA*, and *RhO* (total length 21113–22616 bp) showing the relationship between the 13 *Clostridium botulinum* strains. MP analysis was performed using Paup* ver. 4.0b10

sequences were observed within *C. botulinum* strains ATCC3502, ATCC19397/Hall, H04402065, Kyoto, Langeland and Okra (Table S2). *In silico* mapping with different type II REs allowed us to establish the extent of variability in digestion patterns in *rrs* due to NC (Tables S4, S5, S6).

Variability in RE Digestion Patterns of *rrs*

Genetic evidence on the origin of the 13 different strains of *C. botulinum* was established through in silico mapping of *rrs* with the following REs: BstUI, Hpy188I, Hpy188III, HpyCH4III, HpyCH4V, NlaIV, TaqI, Tsp45I, Tsp509I (Tables S4, S5, S6). *In silico* RE mapping with Tsp509I resulted in different digestion patterns, each having 3 fragments of distinct lengths (Table S6): (1) 244-40-46 nts, (2) 556-244-40 nts, (3) 461-244-40 nts, and (4) 644-244-40 nts. The appearance of 244-40 nts fragments in all copies of *rrs* of 12/13 strains of *C. botulinum* provides potential evidence of their common origin. This proposal was supported by similar observations with different REs: (1) BstUI (166-100-445), (2) Hpy188I (854-56), (3) Hpy188III (96-88-338-275), (4) NlaIV (1-524), and (5) HpyCH4III

(261-149 nts) (Tables S4, S5, S6). *C. botulinum* strain BKT015925 was quite distinct from the rest 12 strains. Only one copy of *rrs* of strain BKT015925 showed resemblance with respect to digestion pattern obtained with Tsp509I to those from Kyoto (1/18 copies), 657 (2/18 copies) (Table S6). On the other hand, the strain BKT015925 resembled Alaska and Eklund 17B with respect to the digestion patterns obtained with RE BstUI (Table S5).

Variability in the fragments observed at 5′ end of *rrs* sequences with RE -Hpy188I, at 3′ end with RE -BstUI, and at both the ends with RE -HpyCH4III and -TaqI provided evidence for the presence of unique features in these strains. These observations allowed us to segregate the 13 strains of *C. botulinum* into 2 major groups (Tables S4, S5, S6). The first group composed of Alaska E43 and Eklund 17B, showed exactly similar digestion patterns in all their 22 *rrs* copies with three different REs: (1) Hpy188III—5′ 502-96-88-261-76-274-86 (nts) 3′, (2) HpyCH4V—5′ 309-22-217-25-395-204 (nts) 3′, and (3) Tsp509I—5′ 244-40-46 (nts) 3′ (Tables S5, S6).

Evidence of co-evolution of Kyoto and 657 and subsequent independent evolution of Kyoto from the second group of 8 *C. botulinum* strains, can be proposed on the basis of their sharing of RE patterns for—HpyCH4V and Tsp509I. *C. botulinum* strain Kyoto can be separated from 657 due to unique RE digestion patterns in the former: (1) RE-HpyCH4V 5′ 133-415-25-346-49-46-215 (nts) 3′ (Table S5), and (2) RE-Tsp509I 5′ 461-244-40 (nts) 3′ (Table S6). Similarly, unique RE digestion patterns can be used to identify the following: (1) ATCC 19397/Hall by (a) Hpy188I—5′ 70-267-856-56 (nts) 3′ (Table S6), (b) Hpy188III—5′ 308-196-96-88-337-274 (nts) 3′ (Table S5), and (2) ATCC3502 by (a) Hpy188III—5′ 249-196-96-88-337-275 (nts) 3′ (Table S5), and (b) TaqI—5′ 90-784 (nts) 3′ (Table S4).

Novel Markers for Identification of *C. botulinum* Strains

Details of the variability in the fragmnents obtained as a result of digestion of *rrs* of *C. botulinum* with REs—TaqI, Hpy188III, Tsp509I, HpyCH4III, Hpy188I, NlaIV, HPYCH4V and Tsp45I and their use as potential markers has been presented as supplementary material (Tables S4, S5, S6**).** In summary, the various observations confirmed the previous findings that these 13 *C. botulinum* strains have been grouped as (1) proteolytic and (2) non-proteolytic (Alaska E43 and Eklund 17B) [18]. In addition, it may be emphasized that the method employed in this study was able to not only differentiate between group I and group II strains, but could also differentiate individual strains, at least within group II.

The most striking feature which emerged from this in silico RE activity was the variation in the number of fragments in different copies of *rrs* gene of a given *C. botulinum* strain. With RE-BstUI, 22/22 copies of *rrs* of Alaska E43, 14/22 copies of Eklund 17B and 8/10 copies of BKT015925 showed a pattern of 166-100-445-200-94-87-15 nts (Table S5). In the second set of 8/22 copies of Eklund 17B and 2/10 copies of BKT015925, the pattern changed to 166-100-444-200-180-15 nts, which could occur by the merger of 94-86 nts fragments into a single fragment of 180 nts. In the rest of the 8 *C. botulinum* strains, the the merger of the fragments 200-94-86 or 200-180 seem to have given rise to a much bigger fragment of 380 nts (Table S5). Such cases of disappearance of RE sites leading to fewer but larger fragments were observable also with other REs (Tables S4, S5). These observations led us to trace the evolutionary fate of the RE sites which perhaps disappeared primarily through natural genetic mutations.

Nucleotide Changes

Multiple alignments of all copies of *rrs* from different *C. botulinum* strains enabled us to trace the sequence of events of NCs occurring within the *rrs* gene sequences. In each *C. botulinum* genome, 2–6 out of 9–22 *rrs* copies did not show any NC. Further multiple alignments among these *rrs* with no NC enabled us to identify a sequence which could be categorized as a potential candidate for a "common ancestor" to these 13 different strains of *C. botulinum.* This "ancestral" *rrs* (S001014409) sequence belonging to *C. botulinum* strain Okra was devoid of any NC (Table S1). It served as reference *rrs* sequence for elucidating the fate of NC in RE sites, in the following 4 forms: (1) appearance of distinct sites, (2) gain of additional sites, (3) loss of sites, and (4) no evident change. In this study, we did not attempt to elucidate the significance to NC in non-RE sites observed in these *C. botulinum* strains.

Evolution of Distinct *RE* Sites

Changes in nucleotides within RE sites in *rrs* sequences leading to distinct RE sites were observed in 10 *C. botulinum* strains: 657, Loch Maree, Eklund 17B, ATCC 3502, ATCC 19397, Kyoto, Langeland, 230613, H0440265 and BKT015925 (Table 1; Fig. 2). The unique features emerging from the different NC were: (1) the number of *rrs* copies affected within a strain varied up to 6, the most affected being those belonging to 657, Eklund 17B, ATCC 3502, and ATCC 19397, (2) the number of affected sites were 2 per *rrs* copy in cases of Eklund 17B, and ATCC 19397 and 3 in case of ATCC 3502, (3) similar changes at same positions such as 454 in *rrs* copies of different

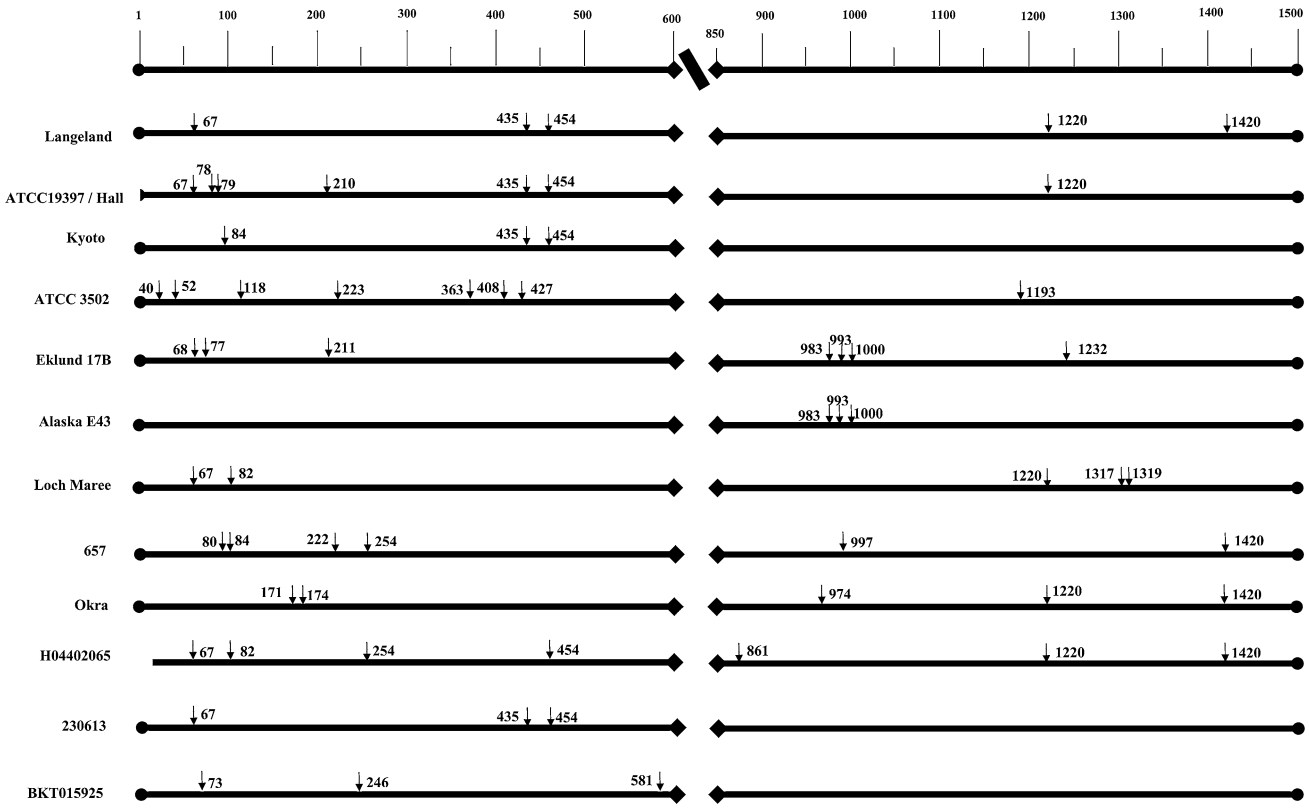

Fig. 2 Mapping of single nucleotide changes (SNC) in restriction endonuclease site(s) in *rrs* of *Clostridium botulinum* strains. No SNCs affecting RE sites were observed in the 500–900 nucleotide region of the *rrs*

strains—ATCC 19397, Kyoto, Langeland, H04402065 and 230613 indicating possible common origin.

A NC A>G in *rrs* of *C. Botulinum* 657 and Loch Maree had a dramatic impact on the RE sites. Here, 3 and 10 distinct RE sites were generated at positions 1,420/1,427 and 222/229, respectively, in strain 657. On the other hand, similar transition of A>G resulted in 4 distinct RE sites at position 1,319/1,326 in Loch Maree (Table 1). Similar observations were also made in the rest of the 3 strains of this group. The mutations of G>A in *rrs*: (1) S001094720 of *C. botulinum* Eklund 17B modified Hpy188I–TCn**'G**A to BstBI–TT'CG**A**A and TaqI–T'CG**A**. Transition of C>T affected three *C. botulinum* strains ATCC 3502, ATCC 19397 and Eklund 17B. Here, 6 RE sites were found to have evolved into 6 distinct RE sites. In addition, it was observed that RE-AciI got transformed into HpyCH4IV in ATCC 3502 and into HpyCH4III in strain Eklund 17B. In contrast to C>T, a reversal of T>C was found to influence 6 strains: ATCC 3502, ATCC 19397, Kyoto, 230613, H04402065 and Langeland. In all these 6 strains the same RE—BfaI site C'**T**AG was affected with slightly different out puts. This suggests a common origin of these 6 strains.

Evolution of Additional RE Sites

Appearance of additional RE sites due to NCs were recorded at different positions along the entire length of the *rrs* gene (Table 2; Fig. 2). Most of the mutations were transitions: (1) G>A at positions—52, 78, 79, 84, 85, 86, 87, 91, 993 and 999; (2) A>G at positions 67, 73, 74, 80, 174 and 181; (3) T>C at positions 82. 89, 171 and 178; and (4) C>T at positions 1,000, 1,006, 1,193 and 1,220 and 1,227. These NCs resulted in the generation of additional RE sites: (1) G>A lead to the generation of RE sites for Hpy188I–TCn'G**A**, Hpy188II–ITC'nnG**A,** Hpy188III–TC'nnG**A,** Tsp509I-**'A**ATT, AleI–C**A**Cnn'nnGTG, MaeIII–'GTn**A**C, MslI–C**A**ynn'nnrTG, and Tsp45I–'GTs**A**C; (2) A>G lead to the formation of sites for activities of REs: Alu I-A**G**'CT, CviJI-r**G**'Cy, HindIII–A'A**G**CTT, BslI–CCnnnnn'nn**G**G, and EcoNI–CCTnn'nnn**AG**G; (3) T>C resulted in the evolution of the sites for REs: AciI–C'CG**C** and AciI–G'**C**GG; (4) **C>**T transition resulted in the evolution of the sites for REs: HphI–GG**T**GAnnnnnnnn', and HpyCH4V–**T**G'CA. It may be noted here that in *C. botulinum* strain Okra, distinct RE sites for FatI–'**CAT**G and NlaIII–**CAT**G', appeared due to simultaneous

changes at positions 171T>**C** and 174A>**G**. It may be further remarked that none of these NCs was due to transversions (Purine ↔ Pyrimidine).

The impact of the appearances of additional RE sites was most evident among 4–8 *rrs* copies of *C. botulinum* strains ATCC 3502, ATCC 19397, Alaska E43) and Loch Maree. This group was followed by strains 657 and Okra, where 4 *rrs* copies gained RE sites (Table 2). The least affected strains were Kyoto and Langeland where gain in RE sites was evident only in a single copy each of *rrs*. Eklund 17B was the only strain where no gain in RE sites was recorded.

A detailed scrutiny based on the appearance of additional and distinct RE sites or loss, within *rrs* of different strains provided insights into their origin. The simultaneous appearance of additional RE sites in multiple copies within a strain and among different strains provided further evidence for common origin and of duplication events in these *C. botulinum* strains. An evidence of common origin of ATCC3502, Langeland, Loch Maree, H04402065 and Okra was obtained on the basis of transition 1220/1227C>T taking place simultaneously in 1–3 copies of *rrs* of each of the 4 strains (Table 2). Similarly, a transition 84/91G>A in a copy each of strain 657 and Kyoto indicate their possible common origin. Transition events in multiple copies within a strain were recorded as (1) 52/86G>A in 8 copies of ATCC 3502, (2) 67A>G in 2 copies within Loch Maree and one copies of H04402065, (3) 78/79/86G>A in 8 copies of ATCC 19397, (4) 993/999G>A and 1000/1006C>T in 8 copies of Alaska E43 and (5) 1220C>T in 2 copies of Okra, which provide support to duplication events (Table S7).

A transition G>A at positions 67/68 in 8 *rrs* copies belonging to 5 different strains—ATCC 19397, Eklund 17B, 230613, H04402065 and Langeland resulted in the loss of sites for 3 REs—AluI, CviJI and HindIII (Table S8). Similarly, A>G transition at position 435 was found to result in simultaneous loss of site for RE-BfaI in 6 copies of *rrs* belonging to 4 different strains –ATCC 19397, Kyoto, 230613 and Langeland (Table S7). These evidences support common origin of these 6 strains and subsequent diversification in Eklund 17B and Kyoto. A single transition affecting loss of RE site in multiple copies of *rrs* at the same position within a strain supports duplication events in the following cases: (1) 40G>A in 2 copies in ATCC 3502, (2) 67/68/74G>A in (a) 2 copies of Langeland, (b) 6 copies of ATCC 19397 and Eklund 17B each, (3) 254/261C>T in 4 copies of 657, (4) 408A>G in 3 copies of ATCC 3502, (5) 435/442A>G in 8 copies of ATCC19397 (Table S8).

Further support to the evolutionary events of common origin and duplication events was also recorded with NC in *rrs* leading to the generation of distinct RE sites (Table 1). A single transition event of 454/461T>C was observed in 12 *rrs* sequences belonging to 5 different *C. botulinum* strains ATCC 19397, Kyoto and Langeland, 23063 and H04402065. Here the simultaneous effect on RE-BfaI site CTAG can be explained by a common origin of these *rrs* sequences. Another similar T>C NC affecting RE-BfaI at position 427/461 in 6 copies of *rrs* of ATCC 3502 can be explained by a duplication event.

A summary of all these effects of NCs (Fig. 2) on RE sites lead us to conclude that 230613, Kyoto, Langeland, and H04402065 had a common origin, which later on diversified into Loch Maree on one hand, most probably from Kyoto and on the other hand ATCC 19397/Hall, originated from H04402065. It is also quite evident that Langeland might have given rise to Loch Maree and Okra. Redundancy in different copies of *rrs* within a strain was evident due to simultaneous NC in multiple copies such as (1) 4 NC events at positions 40, 52, 408 and 427 in 2–4 copies within ATCC 3502, (2) 3 NC events at positions 67, 78/79, and 435 in 3–4 copies of ATCC 19397, (3) simultaneous NC at positions 993 and 1,000 in 4 copies of Alaska E43, and (4) in 2–3 *rrs* copies each of strains 657, Eklund 17B, Langeland, Loch Maree and Okra.

Discussion

Genetic and functional variation due to NCs (transitions or transversions) within the coding regions may cause amino acid exchange or even lead to stop codons. In the non-conservative region, these may influence structure and/or function of the protein [22]. In general, NCs are detectable due to their association with a wide range of downstream expressions: phenotypic and biochemical—ranging from fitness, antimicrobial susceptibility and virulence [23–27]. In contrast, *rrs* does not code for a protein per se. Here, NCs are likely to provide a very narrow range of information on the variability in its structure and function [27]. On the other hand, an evaluation of upstream "expression" in *rrs* due to NCs is more likely to reflect variations in RE activities, as the available multiple action sites are distributed along the length of the *rrs* [5, 6, 11]. Mutation in *rrs* of *Mycoplasma* species at position 997 (C>T) in RE site of AluI (AĜCT) yielded unknown Amplified rDNA Restriction Analysis profiles [28]. Variations in RE maps in *rrs* with AluI, HaeIII and MspI were reported and selected to distinguish *Pseudomonas* isolates [29, 30]. NCs as markers are gaining importance as tools for molecular genetic analysis [31], strain discrimination [32] and are being used to gain insights into their evolutionary history [33] and as important prognostic markers [34].

A summary based on all the NC and RE digestion patterns of *rrs* of *C. botulinum* strains permit us to propose that ATCC 3502, ATCC 19397/Hall, and Langeland are

the closest to each other and were closely related to Loch Maree and Okra, prior to their divergence. High nt identity among ATCC3502, ATCC 19397 and Hall belonging to BoNT/A1 has in fact illustrated their clonal nature [35]. On the other hand, ATCC 3502 seems to have diverged in a different direction, since it had NCs at positions quite different from those observed in ATCC 19397/Hall, and Langeland (Fig. 2). In the study carried out to establish evolutionary relationships among *C. botulinum* A1 genomes, ATCC19379 has been shown to fall close too Hall A and ATCC3502 [8]. Independent evolution of Eklund 17B appears to have happened after its common origin with Alaska E43, which is supported by the only case of transversion C>A at similar position i.e., 983 in only one copy of *rrs* in both of them (Table S8). The strains 657 and Kyoto might have originated from a common ancestor, with divergence occurring due to genetic changes in Kyoto. Thus, in spite of their reported diverse geographical origins, these closely related strains can be identified on the basis of the unique in silico RE maps of their *rrs* and cladistic analysis (Fig. 1). In fact, *Pseudomonas* spp. originating from diverse biogeographical locations were also found to be genetically quite similar [36].

The evolutionary significance of NCs can be viewed to impart resistance to attacks by REs in these *C. botulinum* strains. Simultaneous evolution of novel and additional RE sites in 230613, Alaska E43, ATCC 19397/HALL, ATCC 3502, H04402065, Kyoto, Langeland and Loch Maree can be interpreted to have higher susceptibility to RE attack in these organisms. In contrast, 657, BKT015925, Eklund 17B and Okra have evolved to become more resistant to RE attack by not allowing the evolution of any novel RE sites and by loosing a large number of existing RE sites (8–65) (Table S9). ATCC 19397/Hall, ATCC 3502 and Langeland seems to have evolved in a unique manner since they lost RE sites with high frequency (36–44),which can optimize the process of evolution of resistance to REs. Strain 657 added 30 RE sites, which made it more susceptible to RE attack. Eklund 17B is a strain, which can be recognized to have become highly resistant to REs attack on account of higher loss in RE sites and no evolution of novel RE sites.

Conclusion

In conclusion, the presence of NCs and multiple RE sites within a single gene provides opportunities to elucidate the sequence of evolutionary events. NC at the same position within multiple copies of *rrs* belonging to different *C. botulinum* strains indicated towards a common origin of ATCC 19397 and Langeland and their subsequent diversifications. Secondly, NCs causing specific changes in RE sites (Fig. 2), and unique RE digestion patterns of *rrs* can be used as molecular markers for distinguishing closely related strains of *C. botulinum.*

Acknowledgments We are thankful to Director of CSIR-Institute of Genomics and Integrative Biology (IGIB), and CSIR project GENESIS (BSC0121) for providing the necessary funds, facilities and moral support.

References

1. Arnon SS, Schechter R, Inglesby TV, Henderson DA, Bartlett JG, Ascher MS, Eitzen E, Fine AD, Hauer J, Layton M, Lillibridge S, Osterholm MT, O'Toole T, Parker G, Perl TM, Russell PK, Swerdlow DL, Tonat K (2001) *Botulinum* toxin as a biological weapon: medical and public health management. JAMA 285:1059–1070. doi:10.1001/jama.285.8.1059
2. 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 189:818–832. doi:10.1128/JB.01180-06
3. Editorial (2011) Microbiology by numbers. Nat Rev Microbiol 9:628
4. Johnson EA, Tepp WH, Bradshaw M, Gilbert RJ, Cook PE, McIntosh ED (2005) Characterization of *Clostridium botulinum* strains associated with an infant botulism case in the United Kingdom. J Clin Microbiol 43:2602–2607. doi:10.1128/JCM.43.6.2602-2607.2005
5. Porwal S, Lal S, Cheema S, Kalia VC (2009) Phylogeny in aid of the present and novel microbial lineages: diversity in *Bacillus*. PLoS ONE 4:e4438. doi:10.1371/journal.pone.0004438
6. Kalia VC, Mukherjee T, Bhushan A, Joshi J, Shankar P, Huma N (2011) Analysis of the unexplored features of *rrs* (16S rDNA) of the genus *Clostridium*. BMC Genom 12:18. doi:10.1186/1471-2164-12-18
7. Peck MW (2009) Biology and genomic analysis of *Clostridium botulinum*. Adv Microb Physiol 55:183–2655. doi:10.1016/S0065-2911(09)05503-9
8. Ng V, Lin WJ (2014) Comparison of assembled *Clostridium botulinum* A1 genomes revealed their evolutionary relationship. Genomics 103:94–106. doi:10.1016/j.ygeno.2013.12.003
9. Hutson RA, Thompson DE, Lawson PA, Schocken-Itturino RP, Böttger EC, Collins (1993) Genetic interrelationships of proteolytic *Clostridium botulinum* type A, B, and F and other members of the *Clostridium botulimun* complex as revealed by small-subunit rRNA gene sequences. Antonie Van Leeuwenhoek 64:273–283. doi:10.1007/BF00873087
10. 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 Genom 12:185. doi:10.1186/1471-2164-12-185
11. Verma V, Raju SC, Kapley A, Kalia VC, Daginawala HF, Purohit HJ (2010) Evaluation of genetic and functional diversity of *Stenotrophomonas* isolates from diverse effluent treatment plants. Bioresour Technol 101:7744–7753. doi:10.1016/j.biortech.2010.05.014
12. Bhushan A, Joshi J, Shankar P, Kushwah J, Raju SC, Purohit HJ, Kalia VC (2013) Development of genomic tools for the identification of certain *Pseudomonas* up to species level. Indian J Microbiol 53:253–263. doi:10.1007/s12088-013-0412-1
13. Klappenbach JA, Saxman PR, Cole JR, Schmidt TM (2001) rrndb: the ribosomal RNA operon copy number database. Nucl Acids Res 29:181–184. doi:10.1093/nar/29.1.181

14. Roberts RJ, Vincze T, Posfai J, Macelis D (2010) REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. Nucl Acids Res 38:D234–D236. doi:10.1093/nar/gkp874
15. Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 1999:95–98
16. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948. doi:10.1093/bioinformatics/btm404
17. Xia X, Xie Z (2001) DAMBE: data analysis in molecular biology and evolution. J Hered 92:371–373. doi:10.1093/jhered/92.4.371
18. Sasaki Y, Takikawa N, Kojima A, Norimatsu M, Suzuki S, Tamura Y (2001) Phylogenetic positions of *Clostridium novyi* and *Clostridium haemolyticum* based on 16S rDNA sequences. Int J Syst Evol Microbiol 51:901–904. doi:10.1099/00207713-51-3-901
19. Felsenstein J, Phylip (Phylogeny Inference Package) version 3.57c. (1993) Department of Genetics, University of Washington, Seattle Distribution: http://evolution.genetics.washington.edu/phylip.html
20. Page RDM (1996) TreeView: an application to display phylogenetic trees on personal computer. Comput Appl Biosci 12:357–358. doi:10.1093/bioinformatics/12.4.357
21. Swofford DL (2003) PAUP*: phylogenetic analysis using parsimony (* and other methods). version 4.0.b10. Sinauer Associates, Sunderland, MA
22. Klockgether J, Munder A, Neugebauer J, Davenport CF, Stanke F, Larbig KD, Heeb S, Schöck U, Pohl TM, Wiehlmann L, Tümmler B (2010) Genome diversity of *Pseudomonas aeruginosa* PAO1 laboratory strains. J Bacteriol 192:1113–1121. doi:10.1128/JB.01515-09
23. Smith EE, Buckley DG, Wu Z, Saenphimmachak C, Hoffman LR, D'Argenio DA, Miller SI, Ramsey BW, Speert DP, Moskowitz SM, Burns JL, Kaul R, Olson MV (2006) Genetic adaptation by *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients. Proc Nat Acad Sci USA 103:8487–8492. doi:10.1073/pnas.0602138103
24. Dethlefsen L, Schmidt TM (2007) Performance of the translational apparatus varies with the ecological strategies of bacteria. J Bacteriol 189:3237–3245. doi:10.1128/JB.01686-06
25. Lauro FM, Chastain RA, Blankenship LE, Yayanos AA, Bartlett DH (2007) The unique 16S rRNA genes of piezophiles reflect both phylogeny and adaptation. Appl Environ Microbiol 73:838–845. doi:10.1128/AEM.01726-06
26. López-López A, Benlloch S, Bonfá M, Rodríguez-Valera F, Mira A (2007) Intragenomic 16S rDNA divergence in *Haloarcula marismortui* is an adaptation to different temperatures. J Mol Evol 65:687–696. doi:10.1007/s00239-007-9047-3
27. Jensen S, Frost P, Torsvi V (2009) The nonrandom microheterogeneity of 16S rRNA genes in *Vibrio splendidus* may reflect adaptation to versatile life styles. FEMS Microbiol Lett 294:207–215. doi:10.1111/j.1574-6968.2009.01567
28. Stakenborg T, Vicca J, Butaye P, Maes D, De Baere T, Verhelst R, Peeters J, de Kruif A, Haesebrouck F, Vaneechoutte M (2005) Evaluation of amplified rDNA restriction analysis (ARDRA) for the identification of *Mycoplasma* species. BMC Inf Dis 5:46. doi:10.1186/1471-2334-5-46
29. Saikia R, Sarma RK, Yadav A, Bora TC (2010) Genetic and functional diversity among the antagonistic potential fluorescent pseudomonads isolated from tea rhizosphere. Curr Microbiol 62:434–444. doi:10.1007/s00284-010-9726-y
30. Mulet M, David Z, Nogales B, Bosch R, Lalucat J, García-Valdés E (2011) *Pseudomonas* diversity in crude-oil-contaminated intertidal sand samples obtained after the prestige oil spill. Appl Environ Microbiol 77:1076–1085. doi:10.1128/AEM.01741-10
31. Suh Y, Vijg J (2005) SNP discovery in associating genetic variation with human disease phenotypes. Mutat Res 573:41–53. doi:10.1016/j.mrfmmm.2005.01.005
32. Dos Vultos T, Mestre O, Rauzier J, Golec M, Rastogi N, Rasolofo V, Tonjum T, Sola C, Matic I, Gicquel B (2008) Evolution and diversity of clonal bacteria: the paradigm of *Mycobacterium tuberculosis*. PLoS ONE 3:1538. doi:10.1371/journal.pone.0001538
33. Mestre O, Luo T, Dos Vultos T, Kremer K, Murray A, Namouchi A, Jackson C, Rauzier J, Bifani P, Warren R, Rasolofo V, Mei J, Gao Q, Gicquel B (2011) Phylogeny of *Mycobacterium tuberculosis* Beijing strains constructed from polymorphisms in genes involved in DNA replication, recombination and repair. Plus One 6:e16020. doi:10.1371/journal.pone.0016020
34. Dötsch A, Pommerenke C, Bredenbruch F, Geffers R, Häussler S (2009) Evaluation of a microarray-hybridization based method applicable for discovery of single nucleotide polymorphisms (SNPs) in the *Pseudomonas aeruginosa* genome. BMC Genom 10:29. doi:10.1186/1471-2164-10-29
35. 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:e1271. doi:10.1371/journal.pone.0001271
36. Morris CE, Sands DC, Vanneste JL, Montarry J, Oakley B, Guilbaud C, Glaux C (2010) Inferring the evolutionary history of the plant pathogen *Pseudomonas syringae* from its biogeography in headwaters of rivers in North America, Europe, and New Zealand. MBio 1:e00107–e00110. doi:10.1128/mBio.00107-10