Indian J Microbiol (July–Sept 2015) 55(3):250–257
DOI 10.1007/s12088-015-0535-7

ORIGINAL ARTICLE

Genome Wide Analysis for Searching Novel Markers to Rapidly Identify *Clostridium* Strains

Anay Kekre[1] · Ashish Bhushan[1] · Prasun Kumar[1] · Vipin Chandra Kalia[1]

Received: 23 April 2015 / Accepted: 8 May 2015 / Published online: 14 May 2015

Abstract Microbial classification is based largely on the 16S rRNA (*rrs*) gene sequence, which is conserved throughout the prokaryotic domain. The Ribosomal Database Project (RDP) has become a reference point for almost all practical purposes. The use of this gene is limited by the fact that it can be used to identify only to the extent to what has been known and is available in the RDP. In order to identify an organism whose *rrs* is not present in the RDP database, we need to generate novel markers to place the unknown on the evolutionary map. Here, sequenced genomes of 27 *Clostridium* strains belonging to 9 species have been used to identify two sets of genes: (1) common to most of the species, and (2) unique to a species. Combinations of genes (*recN*, *dnaJ*, *secA*, *mutS*, and/or *grpE*) and their unique restriction endonuclease digestion (*AluI*, *BfaI* and/or *Tru9I*) patterns have been established to rapidly identify *Clostridium* species. This strategy for identifying novel markers can be extended to all other organisms and diagnostic applications.

Keywords Bacteria · Markers · *Clostridium* · Diagnosis · Restriction endonuclease

Electronic supplementary material The online version of this article (doi:10.1007/s12088-015-0535-7) contains supplementary material, which is available to authorized users.

✉ Vipin Chandra Kalia
vckalia@igib.res.in; vc_kalia@yahoo.co.in

[1] Microbial Biotechnology and Genomics, CSIR - Institute of Genomics and Integrative Biology (IGIB), Delhi University Campus, Mall Road, Delhi 110007, India

Introduction

The mysterious microbial world encompasses organisms having a wide diversity in their metabolic, phenotypic, genomic characteristics. The pursuit to identify microbes has seen a shift from relying upon their morphological and biochemical characteristics to genomic features. The advent of molecular biology and bioinformatic techniques has almost completely revolutionized the concept of bacterial taxonomy and their evolutionary pathways. The transition from single gene sequence to whole genome sequence has given confidence of identifying even the bacteria which are yet to be cultured. In fact, bacterial genomic limit can be extended through metagenomic explorations [1]. Phylogenetic trees provide an evolutionary scale for distinguishing organisms which are distantly placed. However, the output of these tools is complicated by too much of heterogeneity on one extreme to virtually nil variability among the strains. It thus becomes a tough task to identify them in an unambiguous manner [2, 3].

The modern taxonomic classification of microbes is based largely on the gene, which is conserved throughout the prokaryotic domain: the 16S rRNA (*rrs*). The microbial taxonomy was given a new look and the nucleotide sequence of this gene has been so widely adapted that it has become a reference point for almost all practical purposes. The Ribosomal Database Project (RDP) (https://rdp.cme.msu.edu/), which was initiated as a small depository of a few hundred *rrs* sequences, has more than 3.0 million entries (RDP Release 11, Update 3::September 17, 2014:: has 3,019,928 16S rRNAs:: 102,901 Fungal 28S rRNAs entries), at present. The rapidly increasing magnitude of this database is a clear reflection on the influence of the findings of Prof. Carl R. Woese [4, 5]. At times, the *rrs* gene sequence is not able to differentiate very closely related taxa.

In such a scenario, one needs to resort to gene sequences which code for features such as heat shock proteins, ATPase-ß-subunit, RNA polymerases or recombinase etc. In certain cases, additional genes have been identified, which can be used exclusively for distinguishing members within a genus: (1) *rpoB* for *Mycobacterium*; (2) *gyrB* for *Acinetobacter*, *Mycobacterium*, *Pseudomonas*, and *Shewanella*, (3) *gyrA* gene for *Bacillus subtilis*, etc. A few methods generally used for identifying bacterial strains are: Amplified fragment length polymorphism (AFLP), DNA–DNA re-association, Microarray, PCR-ribotyping, multilocus sequence analysis, Randomly amplified polymorphic DNA, and restriction endonuclease (RE) digestion [5, 6].

The Latent Features of 16S rDNA

The RDP database, used as reference to identify the newly sequenced 16S rDNA, is limited by the fact that it can be used identify the extent of what has been known and is available. In order to identify the gene sequence which is yet to be seen by the database, it is difficult to visualize how to place the unknown on the evolutionary map. Efforts to resolve the potential problems existing among the different species of (1) *Bacillus*, (2) *Clostridium*, (3) *Pseudomonas*, and (4) *Streptococcus*, revealed the presence of certain latent features in their 16S rDNA gene. The first step involved in the generation of molecular makers was to develop a Phylogenetic Framework, which was composed of sequences, which delineated one species from another i.e. those sequences, which could be used to demarcate the phylogenetic limits of all the known sequences within a species. The second step was to identify motifs (signatures sequences, 30–50 nucleotides (nts) in length), which were unique to a particular species and completely absent from all other species. The third feature, which validates the true identity of the 16S rDNA was the identification of RE, which gives a unique digestion pattern: fragment lengths (nts) and the order of their occurrence. These efforts helped in identification of organisms which were identified initially (by the inventor) only up to genus level [6–9]. This humble beginning in identifying the latent features of those organisms which have been already well identified will help in future to identify and place them on the phylogenetic tree. In fact, these tools have been used to a small extent in certain studies; however, a complete study has been undertaken successfully by others to identify clinically important members of the genus *Streptococcus* [8, 10].

The Mysterious *Clostridium*

Clostridium is a phenotypically and phylogenetically heterogeneous group of strains, which may or may not produce spores and/or toxins, and may give gram-negative or gram-positive reaction [7, 9]. It is tedious to identify them, since their GC content varies from 24 to 58 mol % in *Clostridium perfringens* and *C. barkeri*, respectively. Another major hurdle in identifying *Clostridium* with high precision is the high heterogeneity caused by the presence of multiple copies of *rrs* gene. The need is to look for novel makers for their rapid identification. A novel approach to distinguish very closely related strains of *Clostridium botulinum* was developed recently [11]. However, the method though effective, could be applied to a limited set of strains. In order to identify *Clostridium* present in a mixture of unrelated bacteria, we have identified two sets of genes in *Clostridium* which are: (1) common to most of the species, and (2) unique to a species. A combination of a particular gene or gene set and its (unique) digestion pattern obtained with a specific RE can be exploited to rapidly identify *Clostridium* species.

Materials and Methods

Sequence Data and Comparative Genome Analysis

Completely sequenced genomes of 27 strains of 9 species belonging to genus *Clostridium* were retrieved (http://www.ncbi.nlm.nih.gov/), of which 13 strains belonged to *C. botulinum*, three strains each belonged to *C. acetobutylicum* and *C. perfringens*, 2 strains each were of *C. kluyveri*, and *C. tetani.* The rest of the genomes were of *C. beijerinckii*, *C. cellulovorans*, *C. ljungdahlii*, and *C. novyi* (Table S1). Information of the *Clostridium* genomes for the following parameters such as Accession number, GC percentage, size, and number of genes has been presented (Table S1). Pairwise comparisons among the *Clostridium* genomes were done to identify common (Table 1) and unique genes (Table S2).

Restriction Endonuclease Analysis for Common Gene

A total of 22 Type II REs were considered for digestion on the basis of our previous works [6, 7, 9, 11]. Following REs were used: (1) Four base cutters *AluI* (AG′CT), *BfaI* (C′TA_G), *BfuCI* (_GATC′), *Bsp143I* (_GATC′), *BstKTI* (G′AT_C), *BstMBI* (_GATC′), *CviAII* (C_AT′G), *DpnI* (GA′TC), *DpnII* (_GATC′), *FatI* (_CATG′), *FspBI* (C_TA′G), *Hin1II* (′CATG_), *HpyCH4 V* (TG′CA), *Hsp92II* (′CATG_), *MaeI* (C_TA′G), *RsaI* (GT′AC), *TaqI* (T_CG′A), *Tru9I* (T_TA′A), *XspI* (C_TA′G), (2) Five Base cutters *Hsp92I* (GR_C′YC), and (3) Six base cutters *HaeI* (WGG′CCW), *Hin1I* (GR_CG′YC) (Table S3). All 27 common gene sequences (Table 1) were entered into

Table 1 List of genes common among sequenced genomes of *Clostridium* strains (www.ncbi.nlm.nih.gov)

S. No.	Gene	Function/encoded protein	Frequency[a]
Housekeeping genes			
1	*clpB*	ATP-dependent chaperone ClpB	24/27
2	*clpX*	ATP-dependent Clp protease, ATP-binding subunit ClpX	26/27
3	*dnaA*	Chromosomal replication initiator protein DnaA	26/27
4	*dnaJ*	Chaperone protein DnaJ	27/27
5	*ftsA*	Cell division protein FtsA	26/27
6	*ftsY*	Signal recognition particle-docking protein FtsY	25/27
7	*ftsZ*	Cell division protein FtsZ	27/27
8	*galE*	UDP-glucose 4-epimerase	25/27
9	*grpE*	Co-chaperone GrpE	27/27
10	*lepA*	GTP-binding protein LepA	27/27
11	*lexA*	LexA repressor	27/27
12	*minC*	Septum site-determining protein MinC	27/27
13	*minD*	Septum site-determining protein MinD	26/27
14	*mutS*	DNA mismatch repair protein MutS	27/27
15	*nusG*	Transcription termination/antitermination factor NusG	27/27
16	*recA*	Protein RecA	26/27
17	*recJ*	Single-stranded-DNA-specific exonucleaseRecJ	26/27
18	*recN*	DNA repair protein RecN	26/27
19	*recR*	Recombination protein RecR	27/27
20	*ruvA*	Holliday junction DNA helicase RuvA	27/27
21	*ruvB*	Holliday junction DNA helicase RuvB	27/27
22	*secA*	Preproteintranslocase, SecA subunit	27/27
Other genes			
23	*cbiD*	Cobalamin biosynthesis protein CbiD	25/27
24	*cbiM*	Cobalamin biosynthesis protein CbiM	27/27
25	*cbiQ*	Cobalt ABC transporter, permease protein CbiQ	25/27
26	*cbiT*	Precorrin-6Y C5,15-methyltransferase (decarboxylating), CbiT	24/27
27	*hrcA*	Heat-inducible transcription repressor HrcA	27/27

[a] See Table S5–S7

Cleaver (http://cleaver.sourceforge.net/) to obtain RE digestion patterns. Subsequently, emphasis was laid on those REs motifs which were common to all the strains. Data matrices of those REs were taken into consideration which produced 5–15 fragments. Consensus RE patterns, frequency of occurrence of RE sites and the pattern of nucleotide fragments (nts) were determined for each gene by employing: *AluI* (AG′CT), *BfaI* (C′TA_G) and *Tru9I* (T_TA′A).

Restriction Endonuclease Analysis for Unique Gene

A total of 241 Type II REs with recognition sites of ≥4 nucleotides available in BioEdit were used to generate unique RE patterns [12]. Out of these, only 102 REs were used for further analyses (Table S4). Subsequently, the study was focused on those RE sites which were unique to each strain.

Results

The 27 completely sequenced genomes of *Clostridium: C. botulinum* (13), *C. acetobutylicum* and *C. perfringens* (3 each), *C. kluyveri* and *C. tetani* (2 each), *C. beijerinckii*, *C. cellulovorans*, *C. ljungdahlii*, and *C. novyi* (1 each), showed high heterogeneity at genetic level. The number of genes per genome varies from 2427 to 5243 and the overall GC content ranges from 27.4 to 32.02 mol % (Table S1).

Common Gene Analysis

Comparative genomic analyses revealed the presence of genes which were common to all the Clostridial genomes. A total of 27 common genes including 22 housekeeping genes (HKG) were identified on the basis of their high frequency of occurrence (Table 1). A total of 13 genes

Table 2 Unique fragmentation pattern (5′-3′) generated by in silico digestion of common genes present in *Clostridium* strains*: AluI*

Organism	Strain	*recN*	*dnaJ*	*secA*
C. beijerinckii	NCIMB 8052	·240·483·310·50·	·74·33·26·190·491·175·47·	·9·109·435·366·218·40·129·45·35·247·331·95·258·60·
C. botulinum	230613	·162·19·	·90·129·69·	–
	Alaska E43	·575·22·	–	·357·117·120·129·105·362·476·143·150·156·99·
	BKT015925	·99·762·534·	·21·135·63·11·58·309·270·33·76·	·207·312·430·35·45·135·132·15·105·9·129·165·60·150·244·
	Eklund 17B	·13·17·575·121·	–	·357·117·120·129·1086·306·99·
	Kyoto	·162·19·616·285·70·30·213·	·90·129·69·594·33·30·42·15·	–
	657	·162·19·221·275·120·285·70·30·213·	·21·69·129·69·594·33·30·42·15·	·44·6·306·150·87·30·9·102·9·81·39·9·42·438·234·
	Langeland	·162·19·496·120·134·33·118·70·30·213·	·90·129·69·627·30·42·15·	–
	Loch Maree	·162·19·97·54·465·91·76·118·70·30·213·	·15·75·129·69·594·33·72·15·	·44·67·306·150·87·30·9·102·9·81·39·9·42·63·129·246·207·27·
	Okra	·162·19·221·275·120·167·118·70·30·213·	·90·129·69·594·33·72·15·	·44·67·306·150·87·30·9·102·9·81·39·9·42·438·207·27·230·
	H04402 065	·168·19·230·287·126·173·124·73·30·222·	·93·135·72·618·36·30·45·15·	·44·70·321·156·90·30·12·105·9·84·42·9·45·443·13·216·27·
C.cellulovorans	743B	·58·47·920·113·410·	·80·35·278·192·184·	·291·88·7·245·672·42·9·120·9·378·171·73·62·
C. kluyveri	DSM 555	·897·165·	–	–
C.ljungdahlii	DSM 13528	·565·188·286·32·	·282·33·39·85·282·11·	·523·287·9·234·233·486·241·159·134·
C. novyi	NT	·193·132·	·109·342·282·17·181·33·30·57·	·249·63·465·45·153·12·102·37·83·9·354·49·101·99·145·35·
C. perfringens	13	·252·75·9·240·325·150·105·158·246·	·388·120·	·333·36·322·161·387·9·66·54·177·351·165·23·
	ATCC 13124	·4·248·75·9·240·325·150·105·158·246·	·125·157·231·120·	·93·27·306·36·322·161·387·9·66·54·177·351·165·23·
	SM101	·4·248·75·9·240·123·202·150·5·100·158·246·	·305·157·231·120·	·333·36·219·103·161·171·216·75·54·177·351·165·23·
C. tetani	12124569	·88·183·211·113·	·330·156·45·33·	·21·552·56·184·9·314·153·220·372·273·159·
	E88	·121·70·6·159·420·	·124·330·156·45·33·72·	·21·213·339·240·9·314·153·220·372·273·

Symbol (·) indicates RE site in the gene sequences

Table 3 Unique fragmentation pattern (5′-3′) generated by in silico digestion of common genes present in *Clostridium* strains*: BfaI*

Organism	Strain	*recN*	*mutS*
C. beijerinckii	NCIMB 8052	·280·69·774·	·321·101·36·280·17·97·195·588·267·126·
C. botulinum	Loch Maree	·30·1045·	–
	230613	·116·7·23·	–
	Alaska E43	·458·372·567·	·395·625·389·97·297·
	BKT015925	·74·883·177·	·444·42·282·238·1011·258·
	Eklund 17B	·372·171·396·	·165·444·297·486·625·395·
	Kyoto	·116·7·23·1045·	·242·430·225·231·1050·27·6·
	H04402 065	·122·7·23·451·639·	·702·474·1095·30·6·
	657	·116·7·23·547·498·	–
C.cellulovorans	743B	·740·97·16·86·290·97·	·587·80·18·721·74·692·
C.ljungdahlii	DSM 13528	·334·	·882·1287·15·143·
C. novyi	NT	·466·783·	·675·142·317·129·304·129·324·113·229·
C. perfringens	13	·755·469·92·	·297·85·537·207·38·69·277·39·542·192·
	ATCC 13124	·57·755·469·92·	·297·85·537·207·38·346·39·542·195·
	SM101	·57·450·305·469·92·	·297·85·537·207·38·69·277·39·737·
C. tetani	12124569	·11·112·	·916·287·28·26·327·297·133·45·
	E88	·619·	·916·287·28·26·327·430·45·

Symbol (·) RE site in the gene sequences

(including 10 HKGs) were found to be present in 2–4 copies in 21 strains.

In Silico RE Digestion Patterns of Common Genes

In silico RE digestion patterns for all the 27 common genes were obtained with 22 REs, which were selected on the bases of our previous works [6, 7, 9, 11]. The following REs: *AluI* (AG′CT), *BfaI* (C′TA_G) and *Tru9I* (T_TA′A) were generally found to produce 5–15 easily distinguishable fragments, which were thus selected for identifying novel markers (Tables 2, 3, 4, S5–S7).

AluI: RE-*AluI* showed unique digestion patterns in three HKGs: *recN*, *dnaJ* and *secA* among the *Clostridium* strains (Tables 2, S5). On the basis of the digestion of *recN*, with RE-*AluI*, it was possible to distinguish 20 strain out of 27 *Clostridium* strains of 8 species (Table 2) that includes 10 strains of *C. botulinum*, 3 strains of *C. perfringens*, 2 strains of *C. tetani*, one each of *C. beijerinckii* NCIMB 8052, *C. cellulovorans* 743B, *C. kluyveri* DSM 555, *C. ljungdahlii* DSM 13528 and *C. novyi* NT. The interesting unique digestion patterns (nucleotide fragments) was observed with *C. botulinum* 230613 (162·19 nts), *C. botulinum* Alaska E43 (575·22 nts), *C. kluyveri* DSM 555 (897·165 nts) and *C. novyi* NT (193·132 nts), which had only two fragments each. Another set of strains, which have only four unique RE fragments are (1) *C. beijerinckii* NCIMB 8052 (240·483·310·50 nts) (2) *C. botulinum* Eklund 17B (13·17·575·121 nts), and (3) *C. ljungdahlii* DSM 13528 (565·188·286·32). *C. botulinum* strain BKT105925, *C. cellulovorans* strain 743B, *C. tetani* strains 12124569 and E88 were also easily distinguishable on the basis of the unique RE-*AluI* digestion patterns.

Among *C. botulinum* strains Kyoto, 657, Langeland, Loch Maree, Okra and BKT105925, each of them had similar fragments of 162·19 nts at 5′ end and 70·30·213 nts at 3′ end. However, all of them were easily distinguishable on the basis of fragments present between the two ends. Common genes of *C. botulinum* strain H04402 065 had minor similarities with other strains of this species; however, they were still unique and can be used as novel markers. Similarly, the three strains of *C. perfringens* appeared quite close to each other, however, certain fragments were further subdivided to enable easy distinction e.g., 252 nts and 325 nts fragments of strain 13 appeared as 4·248 and as 123·202 nts in strains ATCC 13124 and SM101. Further distinction between *C. perfringens* strains ATCC 13124 and SM101 could be made on the basis of 105 nts fragment being partitioned into 5·100 nts in the later.

Similarly, on the basis of the digestion of *dnaJ* and *secA*, with RE-*AluI*, it was possible to distinguish all the 16 strains listed in Table 2.

BfaI: With RE-*BfaI*, unique digestion patterns of common genes, *recN* and *mutS* of *Clostridium* species (Table 3, S6) could be used as novel markers for 17 and 14 strains, respectively.

Tru9I: In silico digestion pattern analysis of common genes of *Clostridium* species with RE-*Tru9I* (Table 4, S7), revealed that two genes, *mutS* and *grpE* can be used to clearly identify 16 strains. However, from practical point of view, digestion pattern of *mutS* may not be very effective,

Table 4 Unique fragmentation pattern (5′-3′) generated by in silico digestion of common genes present in *Clostridium* strains*: Tru9I*

Organism	Strain	*mutS*	*grpE*
C. beijerinckii	NCIMB 8052	·57·370·82·29·21·70·119·148·63·159·161·133·81·32·25·17·129·79·12·11·60·7· 155·37·32·102·52·77·24·67·146·229·11·	·227·6·83·288·
C. botulinum	657	·43·42·226·62·139·50·70·48·282·30·290·76·57·101·115·17·52·23·55·39· 21·54·75·83·93·9·13·167·322·135·9·	–
	Hall	·43·42·288·139·50·70·48·282·30·290·76·57·101·115·17·52·23·55·12·27· 21·54·75·83·93·9·13·167·206·116·135·9·	–
	H04402 065	·43·45·235·65·145·53·73·48·294·33·302·79·60·104·121·17·55·23·58·12·51· 57·78·86·96·9·13·176·334·141·15·	·27·73·150·179·
	BKT015925	·25·27·41·93·178·29·21·69·49·48·23·211·122·199·125·7·18·63·138·14·10·65· 55·189·35·13·78·20·30·22·6·42·204·65·85·84·144·	·29·20·174·181·33·21·8·30·96·
	Kyoto	·43·42·226·62·139·50·70·48·228·54·30·290·76·57·101·115·17·52·23·55·12· 27·21·54·75·83·93·9·13·167·322·135·9·	–
	Eklund 17B	·55·30·48·206·82·21·8·21·27·43·120·237·77·4·26·79·62·45·133·8·124·23·13·51·24·11· 30·24·55·12·26·22·23·31·168·32·37·65·49·80·9·15·67·234·135·9·	·120·81·15·87·234·46·
	Alaska E43	·135·234·67·15·9·129·65·37·32·168·31·23·22·26·12·55·24·30·11·24·51·13·23· 124·8·133·45·62·79·26·4·77·237·120·43·27·21·8·21·82·117·86·133·59·7·64·	·123·81·15·87·234·46·
C.cellulovorans	743B	–	·56·34·22·12·197·131·11·132·
C.ljungdahlii	DSM 13528	·123·139·110·210·211·174·147·75·56·81·52·153·78·12·177·362·159·15·133·156·9·	·31·164·21·43·57·44·103·20·157·13·
C. novyi	NT	·80·211·23·19·21·50·33·45·48·150·27·90·31·35·226·50·75·219·156·156·24·48· 48·70·50·262·89·4·27·12·55·	·13·16·20·388·21·108·
C. perfringens	ATCC 13124	·210·62·15·51·39·7·24·32·57·139·20·6·22·14·13·80·9·96·57·15·16·20·81·13·120·198·185· 147·72·48·39·81·151·63·36·134·40·17·8·	–
	13	·210·62·15·51·39·7·24·32·57·36·103·20·6·22·14·13·80·9·96·57·15·16·20·81·13·120· 107·91·185·147·72·48·39·52·29·151·63·36·95·36·40·17·8·	·70·87·15·41·22·44·91·11·
	SM 101	·289·210·62·15·51·39·7·24·32·57·36·103·20·6·22·14·13·80·9·96·57·15·16·20·81· 13·120·107·91·185·147·72·48·39·81·151·63·170·40·17·8·	–
C. tetani	12124569	·78·24·41·73·33·60·162·120·99·24·99·105·21·51·306·18·14·43·60·128·316· 27·44·24·78·19·63·123·84·18·	·13·45·132·34·69·87·15·42·12·93·
	E88	·78·24·41·73·33·60·162·120·99·24·204·21·51·15·291·18·14·43·60·128·34·282· 27·44·24·78·19·18·45·123·84·18·	·13·45·120·12·34·69·102·54·93·

Symbol (·) RE site in the gene sequences

Table 5 Potential gene types which can be used for identification of *Clostridium* strains

Organism name	Gene type		
	Common		Unique
	Single copy	Multi-copy	
C. acetobutylicum ATCC 824	N	Y	Y
C. acetobutylicum DSM 1731	N	Y	Y
C. acetobutylicum EA 2018	N	N	Y
C. beijerinckii NCIMB8052	Y	Y	Y
C. botulinum 230613	Y	N	Y
C. botulinum 657	Y	N	Y
C. botulinum Alaska E43	Y	Y	Y
C. botulinum ATCC 19397	N	N	N
C. botulinum ATCC 3502	N	N	Y
C. botulinum BKT015925	Y	N	Y
C. botulinum Eklund 17B	Y	Y	Y
C. botulinum H04402 065	Y	Y	Y
C. botulinum Hall	Y	N	N
C. botulinum Kyoto	Y	N	Y
C. botulinum Langeland	Y	Y	N
C. botulinum Loch Maree	Y	Y	N
C. botulinum Okra	Y	Y	N
C. cellulovorans 743B	Y	Y	N
C. kluyveri DSM 555	Y	N	Y
C. kluyveri NBRC 12016	N	Y	N
C. ljungdahlii DSM 13528	Y	Y	Y
C. novyi NT	Y	N	N
C. perfringens 13	Y	Y	Y
C. perfringens ATCC 13124	Y	N	Y
C. perfringens SM101	Y	N	Y
C. tetani 12124569	Y	Y	Y
C. tetani E88	Y	Y	Y

N No, *Y* Yes

as it generates a large number of small sized fragments (Table 4). *mutS* is the only gene that may be used to differentiate Hall from all other *Clostridium* strains.

Multiple Copies of Common Genes in *Clostridium* Genome

In this study, we found multiple copies of 13 different genes belonging to 22 different strains of *Clostridium.* The number of gene copies varied from 2 to 4, with 2 being the most frequent number (Table S8–S10). In most of the cases, RE digestion patterns varied among the copies as well. By digesting common genes, which were present in multiple copies, we could distinguish an additional 3 strains of *Clostridium*: *C. acetobutylicum* ATCC 824, *C. acetobutylicum* DSM 1731 and *C. kluyveri* NBRC 12016 (Table S8–S10). It may be concluded that using RE—common gene combinations; we could distinguish 24 out of 27 strains used in this study.

Unique Gene Analysis

Pairwise comparison among 27 annotated strains of *Clostridium* species revealed the presence of unique genes. The number of unique genes varied from as low as one in *C. acetobutylicum* strains ATCC 824 and EA 2018, *C. botulinum* strains Alaska E43 and BKT015925 to as high as 31, 35, 40 and 71 in the cases of *C. ljungdahlii* DSM 13528, *C. tetani* 12124569, *C. acetobutylicum* DSM 1731, and *C. kluyveri* DSM555, respectively (Table S2). Out of 27 genomes, only 19 strains were found to have unique genes, which can be exploited for strain level identification.

It indicates that a wide genetic variability is available for distinguishing even very closely related species.

In Silico RE Digestion Patterns of Unique Genes

Unique genes for 19 strains of *Clostridium* and their digestion pattern with REs have been listed in Table S2. These genes can be used either individually or in various combinations to identify organisms up to strain level. In order to increase the validity of the identification, RE patterns of genes with multiple cut sites can be used (Table S2). By combined approach of the RE digestion patterns of common and unique, we can identify 26 out of 27 strains used in this study.

Discussion

In silico mapping of genes with different Type II REs has revealed that digestion patterns vary substantially even between closely related organisms. The variation in RE digestion patterns within a gene originates because of single nucleotide changes, especially those, which fall within the RE recognition motif [11]. Although a large number of REs can be used to digest a gene, however, it has been realized that for driving meaningful conclusions, only a few of them can be employed. Around 22 different REs have been used in this study to identify unique digestion patterns within a gene. It was revealed that out of 2427–5243 genes present in the genomes of *Clostridium* strains, around 27 genes were common to most of them. The presence of these common genes can help in easily identifying the organism at least up to genus level. Now in order to identify the organism up to species level we need another set of markers. It was realized that only three combinations of REs- and HKGs: (1) *AluI-recN*, *dnaJ* and *secA*, (2) *BfaI-recN* and *mutS*, and (3) *Tru9I-mutS* and *grpE*, can be used as novel markers for identifying *Clostridium* strains. In summary, we may conclude that each strain can be identified and further validated by combining the observations made of certain common or unique genes and their RE digestion patterns (Table 5). This study thus provides a unique opportunity to develop diagnostic kits for rapidly identifying strains by amplifying only a very limited number of genes. And perhaps the best part of this study is its potential to be extended to any gene and organism of interest. A few studies have in fact been conducted, where RE digestion patterns of functional genes have been used as markers [8, 10, 13–17].

Acknowledgments We are thankful to the 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. Kalia VC (2010) Extending genomic limits through metagenomic exploration. J Cosmol 13:3625–3627
2. Klappenbach JA, Saxman PR, Cole JR, Schmidt TM (2001) rrndb: the ribosomal RNA operon copy number database. Nucleic Acids Res 29:181–184. doi:10.1093/nar/29.1.181
3. Lal S, Cheema S, Kalia VC (2008) Phylogeny vs genome reshuffling: horizontal gene transfer. Indian J Microbiol 48:228–242. doi:10.1007/s12088-008-0034-1
4. Kalia VC (2013) The Visionary: Prof Carl R. Woese. Indian J Microbiol 53:245–246. doi:10.1007/s12088-013-0417-9
5. Prakash O, Jangid K, Shouche YS (2013) Carl Woese: from Biophysics to evolutionary microbiology. Indian J Microbiol 53:247–252. doi:10.1007/s12088-013-0401-4
6. 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
7. 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
8. Lal D, Verma M, Lal R (2011) Exploring internal features of 16S rRNA gene for identification of clinically relevant species of the genus *Streptococcus*. Ann Clin Microbiol Antimicrob 10:28. doi:10.1186/1476-0711-10-28
9. 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
10. Huma N, Shankar P, Kushwah J, Bhushan A, Joshi J, Mukherjee T, Raju SC, Purohit HJ, Kalia VC (2011) Diversity and polymorphism in AHL-lactonase gene (*aiiA*) of *Bacillus*. J Microbiol Biotechnol 21:1001–1011. doi:10.4014/jmb.1105.05056
11. Bhushan A, Mukherjee T, Joshi J, Shankar P, Kalia VC (2015) Insights into the origin of *Clostridium botulinum* strains: evolution of distinct restriction endonuclease sites in *rrs* (16S rRNA gene). Indian J Microbiol 55:140–150. doi:10.1007/s12088-015-0514-z
12. Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41:95–98
13. Kalia VC, Raju SC, Purohit HJ (2011) Genomic analysis reveals versatile organisms for quorum quenching enzymes: acyl-homoserine lactone-acylase and -lactonase. Open Microbiol J 5:1–13. doi:10.2174/187428580110501000
14. Prakash O, Pandey PK, Kulkarni GJ, Mahale KN, Shouche YS (2014) Technicalities and glitches of terminal restriction fragment length polymorphism (T-RFLP). Indian J Microbiol 54:255–261. doi:10.1007/s12088-014-0461-0
15. 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
16. Verma V, Raju SC, Kapley A, Kalia VC, Kanade GS, Daginawala HF, Purohit HJ (2011) Degradative potential of *Stenotrophomonas* strain HPC383 having genes homologous to *dmp* operon. Bioresour Technol 102:3227–3233. doi:10.1016/j.biortech.2010.11.016
17. Selvakumaran S, Kapley A, Kashyap SM, Daginawala HF, Kalia VC, Purohit HJ (2011) Diversity of aromatic ring-hydroxylating dioxygenase gene in *Citrobacter*. Bioresour Technol 102:4600–4609. doi:10.1016/j.biortech.2011.01.011