Indian J Microbiol (Oct–Dec 2015) 55(4):366–374
DOI 10.1007/s12088-015-0552-6

ORIGINAL ARTICLE

Genome Wide Search for Biomarkers to Diagnose *Yersinia* Infections

Vipin Chandra Kalia[1] · **Prasun Kumar**[1]

Received: 13 August 2015 / Accepted: 1 September 2015 / Published online: 9 September 2015

Abstract Bacterial identification on the basis of the highly conserved 16S rRNA (*rrs*) gene is limited by its presence in multiple copies and a very high level of similarity among them. The need is to look for other genes with unique characteristics to be used as biomarkers. Fifty-one sequenced genomes belonging to 10 different *Yersinia* species were used for searching genes common to all the genomes. Out of 304 common genes, 34 genes of sizes varying from 0.11 to 4.42 kb, were selected and subjected to in silico digestion with 10 different Restriction endonucleases (RE) (4–6 base cutters). *Yersinia* species have 6–7 copies of *rrs* per genome, which are difficult to distinguish by multiple sequence alignments or their RE digestion patterns. However, certain unique combinations of other common gene sequences—*carB*, *fadJ*, *gluM*, *gltX*, *ileS*, *malE*, *nusA*, *ribD*, and *rlmL* and their RE digestion patterns can be used as markers for identifying 21 strains belonging to 10 *Yersinia* species: *Y. aldovae*, *Y. enterocolitica*, *Y. frederiksenii*, *Y. intermedia*, *Y. kristensenii*, *Y. pestis*, *Y. pseudotuberculosis*, *Y. rohdei*, *Y. ruckeri*, and *Y. similis*. This approach can be applied for rapid diagnostic applications.

Keywords Biomarkers · Diagnosis · Genome · In silico · Restriction endonuclease · *Yersinia*

Electronic supplementary material The online version of this article (doi:10.1007/s12088-015-0552-6) 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

Yersinia as a bacterial pathogen spreads through contaminated food or water and blood transfusion [1–3]. In spite of the availability of a large number of phenotypic and genotypic methods, reliable detection of this pathogenic organism continues to be a challenge [4]. The major hurdle crops up due to their high similarities to many enteric bacteria and their slow-growing nature. Genomes of different *Yersinia* spp. show high level heterogeneity and possess genes responsible for virulence and pathogenesis, which may be located on the plasmids and chromosomes. The need is to develop a sensitive, rapid, and economical method to detect this bacterium. Review of literature reveals that cultural, immunological, and molecular methods are available, but each has its own limitations.

Cultural Methods

The microbiological culturing technique continues to be the gold standard for the detecting pathogens. The culture-based methods, though effective in detecting *Yersinia*, however, are time consuming. Commercially available kits like Biolog and API 20E systems involve expensive strips and equipments, which thus limits their usage on a routine basis [2].

Immunological Methods

Immunoassays permit detection and identification of microbes without culturing them. Latex agglutination is the simplest immunoassay, where latex beads coated with pathogen-specific antibodies agglutinate antigens, and the precipitate is easy to visualize. In Immunomagnetic separation (IMS) assay, magnetic beads coated with antibodies

separate the target organism, which is then confirmed by polymerase chain reaction (PCR) or enzyme-linked immunosorbent assay (ELISA) [5, 6]. A few variants of immunological assays used for detecting *Yersinia* are: ELISA using swine antibodies against lipopolysaccharide, multiplexed sandwich chemiluminescent ELISA and commercial kits to detect the O-antigens [7]. A more sensitive method being employed is the surface plasmon resonance (SPR)-based immunosensor method for assessing antigen–antibody interactions [8]. Due to its ability to detect very low cell density, this method is likely to prove effective for detecting *Yersinia enterocolitica* in food products. This method is highly specific, needs antibodies for each strain, before it can be exploited. As the assay is extremely dependent upon in vitro testing conditions, there is a high likely hood of misinterpretation and false identification.

Molecular Methods

Detection of microbes based on their DNA has led to a range of molecular tools, which have made the methods to be rapid, economical, and precise. Methods used for identifying bacteria include colony hybridization, PCR, microarray, and loop-mediated isothermal amplification (LAMP) among other variants of these basic approaches, including Restriction Endonuclease digestion (RE) [1, 9–12]. Most studies employ 16S rRNA (*rrs*), the most conserved gene, for identifying bacteria. The Ribosomal Database Project (RDP) (https://rdp.cme.msu.edu/) has more than 3.0 million entries as of now. This gene (*rrs*) may not be sufficient to distinguish very closely related taxa or bacteria possessing multiple copies of the gene per genome [13]. In cases where *rrs* alone does not prove effective in distinguishing closely related species, one has to resort other housekeeping genes (HKGs): heat shock proteins, ATPase-ß-subunit, RNA polymerases or recombinase, etc. [14, 15]. For distinguishing members within a genus, a few specific genes have been identified: (1) *gyrA* gene for *Bacillus subtilis*, (ii) *gyrB* for *Acinetobacter*, *Mycobacterium*, *Pseudomonas*, and *Shewanella*, and (iii) *rpoB* for *Mycobacterium*; etc. [12]. The assays commonly used for identifying *Yersinia*, involve genes responsible for pathogenesis: (i) *ail* (attachment and invasion locus, 454 nts), (ii) *inv* (invasion, 570 bp), or (iii) *yst* (*Yersinia* stable toxin, 145 nts), (iv) *myf* (adhesin), and (v) *yop* (yersinia outer protein), (vi) *vir* (transcriptional regulator, 700 nts) genes [16–19]. The process is hindered by the presence of DNA from closely related competing microflora. It is, however, more sensitive than conventionally employed cultural methods. In comparison, to single gene based PCR methods, multiplex PCRs are being preferred. The detection of food-borne pathogens by PCR based methods are being monitored by European Committee for Standardization (CEN) [20]. A few major limitations of these PCR based protocols are (i) high rate of false-positive results, (ii) inability to differentiate viable and non-viable cells. It is thus imperative to include sufficient numbers of negative and internal positive controls [21]. Microarray is another method, which is quite sensitive and effective for identifying the target microbe [9, 22–24]. A more recent DNA based diagnostic tool for identifying food borne pathogens method is the Loop-mediated isothermal amplification (LAMP). It is relatively quite simple and does not involve reagents or any specialized equipment for visualization [10, 25, 26]. The genes used have been *gyrB* (gyrase B), *phoP* (transcriptional response regulator) [26, 27]. A few highly sensitive methods, which have been introduced for detecting *Yersinia* include: (i) a silicon-based optical thin-film biosensor chip [28], (ii) Fourier transform infrared spectroscopy and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry [29–31]. These are beneficial for high-throughput analysis.

In our endeavour to meet the challenge to provide biomarkers to rapidly identify bacteria, we have developed novel genomic tools to elucidate the latent features of *rrs*, such as: (i) a Phylogenetic Framework, (ii) unique signature sequences, and (iii) unique RE digestion patterns. This approach has enabled identification of organisms up to species level: (i) *Bacillus*, (ii) *Clostridium*, (iii) *Pseudomonas*, and (iv) *Streptococcus* [12, 32–36]. In an attempt to enhance the effectiveness, of this approach, especially, to identify the target bacteria from a mixed population, two sets of genes in *Clostridium* were segregated: (i) common to all species, and (ii) unique to a species. Based on the RE digestion patterns of these genes, unique combination of genes and REs, were suggested to rapidly identify *Clostridium* species [15]. For thorough and effectual surveillance of *Yersinia*, the need is to develop novel and innovative methods, which can prove as powerful tools in the hands of Health Departments to handle any outbreaks. These methods should be easy to use, can detect even very low population densities of the pathogen, culture-independent, in situ and more reliable. In the present work, we have extended the techniques used for distinguishing *Clostridium* strains [15] to develop biomarkers for rapid diagnosis of *Yersinia*.

Materials and Methods

Sequence Data and Comparative Genome Analysis

Completely sequenced genomes of the 51 strains of 10 species belonging to genus *Yersinia* were retrieved (http://www.ncbi.nlm.nih.gov/): *Y. pestis* (22); *Y. pseudotuberculosis* (12); *Y. enterocolitica* (3); *Y. enterocolitica* subsp.

palearctica (2); *Y. ruckeri* (2); and one each of *Y. aldovae*; *Y. enterocolitica* (type O:5); *Y. enterocolitica* subsp. *enterocolitica*; *Y. frederiksenii*; *Y. intermedia*; *Y. kristensenii*; *Y. pestisbiovar Medievalis*; *Y. pestisbiovar Microtus*; *Y. rohdei*; *Y. similis* (Table S1). Information of the *Yersinia* genomes for the following parameters such as Accession number, GC percentage, size, and number of genes has been presented (Table S1). Pair-wise comparisons among the *Yersinia* genomes were done to identify common genes (Table S2). Out of all the genomes, 304 protein encoding common genes (including 6 with two copies each) could be distinguished. Out of these 304 common genes, 34 were selected on the basis of the size of the gene, at an increment of around 100 nucleotides in the range of 114 nucelotides (nts) to 4446 nts (Tables S2 and S3). The most commonly used non-protein coding gene, *rrs* was also taken into consideration, as it is used conventionally for bacterial identification.

Restriction Endonuclease Analysis of Common Genes

A total of 10 Type II REs were considered for digestion on the basis of our previous works [12, 33, 35, 37]. The following REs were used: (i) 4 base cutters *AluI* (AG′CT), *BfaI* (C′TA_G), *BfuCI* (_GATC′), *CviAII* (C_AT′G), *HpyCH4V* (TG′CA), *RsaI* (GT′AC), *TaqI* (T_CG′A), *Tru9I* (T_TA′A), and (ii) 6 base cutters *HaeI* (WGG′CCW), *Hin1I* (GR_CG′YC). All 34 common gene sequences (Table S3) were entered into Cleaver (http://cleaver.sourceforge.net/) to obtain RE digestion patterns. Subsequently, emphasis was laid on Data matrices of those REs, which produced 5–15 fragments. Consensus RE patterns, frequency of occurrence of RE sites and the pattern of nucleotide fragments were determined for each gene by employing the 10 REs listed above. Finally, the study was focused on those RE sites which generated digestion patterns unique to a strain.

Results

The 51 completely sequenced genomes of *Yersinia* spp.: *Y. aldovae*, *Y. enterocolitica*, *Y. frederiksenii*, *Y. intermedia*, *Y. kristensenii*, *Y. pestis*, *Y. pseudotuberculosis*, *Y. rohdei*, *Y. ruckeri*, and *Y. similis* (Table S1) showed high heterogeneity at the genetic level. The number of genes within each genome varies from 3219 to 5596 and the overall GC content ranges from 46.96 to 48.05 mol % (Table S1). The initial analysis was based on *rrs* gene, which is conventionally used for identifying bacteria. Equally, we could not trace unique features in *rrs* by which all the genomes could be identified, the focus was switched to other factors, which were common to all the genomes.

rrs Gene Analysis of *Yersinsia* Species

The frequency of occurrence of the *rrs* gene in *Yersinia* strains varied between 6 and 7. Within each genome, these *rrs* genes could be segregated into two groups: (1) 2 copies and (2) 4 or 5 copies. Within each group the copies showed high similarity. Multiple sequence alignments of 335 copies of *rrs* from all the *Yersinia* genomes revealed that these can be represented by 102 sequences, the rest being additional copies. RE digestion patterns of 102 *rrs* sequences representing 51 genomes showed that gene sequences belonging to different genomes clubbed together into groups.

With RE-AluI

102 *rrs* gene sequences from 51 *Yersinia* genomes could be segregated into 6 major groups. RE digestion patterns ·79·42·361·211·207·375· was recorded in both the copies of *rrs* in 35 genomes. This RE digestion pattern was exactly similar to that of one of the two copies of 10 other *Yersinia* genomes. In the latter instance, the second *rrs* copy in all the 10 genomes had a similar RE digestion pattern of ·42·361·211·207·375·. Another two groups of 2 genomes each had the following RE digestion patterns: (i) both the *rrs* copies had ·86·79· 42·328·33·211·207/209·375· (CP009539.1 and CP011078.1), (ii) the two *rrs* copies had the following patterns: (a) ·79·42·361·211·207·375· (CP009759.1) and (b) ·42·361·211·207·375· (CP009757.1). Only two genomes had unique RE-*AluI* digestion patterns in their *rrs*: (i) CP009787.1—·79·42·172·189·210·207·375·; and (ii) CP007448.1—·86·79·42·361·211·207·375·. Thus, out of 51 genomes, RE-*AluI* enabled us to distinguish only two genomes on the basis of their unique *rrs* digestion patterns.

With RE-BfuCI

A total of 7 RE digestion patterns were recorded—(i) 13 ·114·152·424·652·174·56 (in 24 genomes) and (ii) 105·152 ·424·652·174·9 (in 12 genomes), (iii) 18·266·424·652·125 (in 2 genomes), (iv) 105·152·424·652·156 (in 9 genomes), (v) 254·424·652·160 (in 2 genomes). Two genomes (CP010247 and CP011078) had RE-*BfuCI* digestion patterns: (i) 7/13·114·152·424·652/654·174·8/17 (common to both the genomes), and (ii) 10·114·152·424·50·602·174·3 in CP010247 and 17·114·152·424·575·77·174 ·21 in CP011078. In spite of such a variation in the digestion patterns obtained with RE-*BfuCI*, only 2 genomes (CP010247, CP011078), were found to have unique features in one of their *rrs* copies. Hence, unless information on both the *rrs*

copies is available, it will be difficult to distinguish them without any ambiguity.

With RE-CviAII

RE digestion patterns of 51 genomes could be grouped into 9 categories (Table S4). In spite of the availability of a large variation in the RE-*CviAII*, digestion patterns-only 1 genome (CP009539), has certain unique features in all the copies of *rrs*. For identifying CP009935, CP001585 and CP001589, the information on digestion pattern can be exploited only if the nucleotide sequences of both the *rrs* copies are available. RE digestions of *rrs* copies of the rest of the genomes have similarities among themselves, which do not permit their distinction.

With other Restriction Endonucleases

RE digestion of *rrs* with *TaqI* could lead to provide some information, which can be used to distinguish only two *Yersinia* genomes: CP009759 and CP009364. Here, two patterns were recorded within each of the genomes (i) CP009759—(a) 55·760·139·361·216 and (b) 56·760·53·86·361·213, and (ii) CP009364—(a) 55·760·53·86·361·216, and (b) 56·123·636·53·86·361·213.

Even so, with one of the patterns being common, it is difficult to distinguish them in an unambiguous manner. *TaqI* digestion of the rest of the 49 *Yersinia* genomes had an exactly similar digestion pattern: (i) 61·760·53·86·361·264, which could thus not be distinguished by this RE. A slightly better scenario was recorded with RE-*Tru9I* here, it was possible to distinguish only 4 genomes of *Yersinia*: (i) CP007230: 462·9·79·42·26·251·34·52·134·355·98, (ii) CP011078: 2·457·7·88·42·26·251·34·52· 134·355·106, (iii) (a) HF571988: 24·464·88·42·26·251·34·52·134·355·38, and (b) 456·88·42·26 ·251·34·52·134·355·51, and (iv) CP009757: (a) 457·88·42·26·285·52·134·355·90, and (b) 456·88· 42·26·251·34·52·134·355·93. The rest of the 47 genomes had a similar RE *Tru9I* digestion pattern: 462·88·42·26·251·34·52·134·355·141.

The cases with none to few RE digestion sites were as follows: (1) with *HaeI*—no digestion, (ii) with *Hin1I*—single cut site led to two fragments (in 50 genomes) and 3 fragments in CP007448, (iii) with *RsaI*—two cut sites in 50 genomes and 3 cut sites in one copy of *rrs* in genome-CP009367. With RE-*BfaI*—all the copies of *rrs* in all the 51 genomes had a similar digestion pattern: 247·24·14·365·694·241. With RE-*HpyCH4V*—2 RE digestion patterns: (i) 111·110·264·87·333·89·153·92·310·50 (in 50 genomes), and (ii) (a) 109·110·351·333·89·152·22·312·59, and (b) 111·110·264·87·333·89·153·22·362 (in a single genome-CP01 0023).

Common Gene Analysis

In view of the observations made with the presence of multiple copies of *rrs* gene and high similarity among them in all the sequenced genomes of *Yersinia*, we shifted our focus on other genes. A comparative analysis of all the genes present in the 51 *Yersinia* genomes, we could trace 304 genes, which were common among them. Out of these 304 genes, we selected 34 genes, which varied in size from 114 to 4446 nts, in such a manner that genes of all sizes were represented (Tables S2 and S3).

In silico RE Digestion Patterns

In silico RE digestion patterns of 33genes (in addition to *rrs*), which were common to all the 51 genomes of *Yersinia* with 10 different REs revealed some very interesting features in them (Tables S5-S13 and Tables S14 to S22).

These 33 genes could be categorized into three distinct classes: (i) used for distinguishing most of the genomes—*carB*, *fadJ*, *gluM*, *gltX*, *ileS*, *malE*, *nusA*, *ribD*, and *rlmL* could be (ii) used in combination with certain REs generating supplementary information for identifying those genomes which could not be distinguished on the basis of genes of the category (i)—*aceE*, *aceK*, *cpxP*, *cysJ*, *glpQ*, *gltB*, *gyrB*, *lacZ*, *leuD*, *ligA*, *lolD*, *metH*, *mukB*, *pheT*, *rpoB*, and *secA*, and (iii) genes which could not be used as candidates for distinguishing *Yersinia* species—*rnpA* (114 nts), *mltA* (210 nts), *rplW* (303 nts), *flgC* (405 nts), *nagB* (801 nts), *hisG* (900 nts), *thiP* (1608 nts), and *dnaK* (1902 nts).

In silico Digestion Pattern of Genes as Potential Biomarkers

Out of the 9 genes [category (i)], which could be used for distinguishing genomes, digestion of *fadJ* with different REs proved to be instrumental in distinguishing 11 *Yersinia* genomes: *Y. aldovae* 670-83 (CP009781), *Y. enterocolitica* subsp. enterocolitica 8081 (AM286415), *Y. enterocolitica* LC20 (CP007448), *Y. enterocolitica* WA (CP009367), *Y. enterocolitica* (type O:5) str. YE53/03 (HF571988), *Y. enterocolitica* subsp. palearctica Y11 (FR729477), *Y. intermedia* Y228 (CP009801), *Y. rohdei* YRA (CP009787), *Y. ruckeri* Big Creek 74 (CP011078), *Y. ruckeri* YRB (CP009539), *Y. similis* 228 (CP007230) (Tables 1 and 2) The gene (*fadJ*) gave distinct digestion patterns with REs—*HpyCH4V*, *BfuCI*, *CviAII*, *AluI*, *Hin1I*, *BfaI*, *RsaI*, *TaqI* and *Tru9I*. RE-*HpyCH4V* was the only one which proved effective in distinguishing 11 genomes. In certain combinations *Yersinia* genomes—CP009367, CP009539, CP009787, CP011078, and HF571988, the following REs could not provide distinct digestion patterns: *Hin1I*, *BfaI*,

Table 1 Unique in silico Restriction Endonuclease digestion pattern (5′–3′) of *fadJ* gene of *Yersinia* genomes

Genome	*fadJ* gene digestion with restriction endonucleases			
	HpyCH4V	*BfuCI*	*CviAII*	*AluI*
AM286415	48·294·448·464·52·24·469·301	839·531·13·717	135·149·1019·797	16·763·1321
CP007230	52·179·294·912·608·181·117	210·56·54·556·353·337·569·208	318·109·313·746·257·306·294	37·69·350·358·113·691·565·156·4
CP007448	163·15·8·439·303·464·545·295	831·926·475	273·149·133·886·500·126·165	154·329·388·492·117·447·54·94·157
CP009367	186·294·448·464·52·24·469·301	38·939·531·13·717	422·1019·797	–
CP009539	168·294·603·227·82·343·357·131	92·234·487·452·238·80·381·88·153	112·143·154·1014·782	393·10·1298·184·320
CP009781	186·27·267·313·507·92·343·210·290	275·556·257·433·236·478	273·562·606·518·276	483·376·660·219·240·221·36
CP009787	186·294·484·315·165·111·180·183·19·304	110·721·690·720	67·20·186·25·84·506·553·500·126·96·78	199·666·52·683·360·277·4
CP009801	163·23·27·267·139·345·428·52·768·23	530·27·18·256·48·185·457·131·583	273·109·506·261·292·500·222·72	917·323·240·93·662
CP011078	168·294·830·82·831	92·234·487·452·238·80·381·88·60·93	112·143·109·45·792·222·782	393·1308·184·320
HF571988	186·294·559·353·52·24·469·80·221	977·544·104·613	273·149·1019·797	52·102·711·6·11·35·167·1154
FR729477	480·912·52·24·22·447·80·221	–	–	–

(·): RE site in the gene sequence; (–): no digestion observed

Table 2 Unique in silico Restriction Endonuclease digestion pattern (5′–3′) of *fadJ* gene of *Yersinia* genomes

Genome	*fadJ* gene digestion with restriction endonucleases				
	Hin1I	*BfaI*	*RsaI*	*TaqI*	*Tru9I*
AM286415	2100	586·1144·370	323·162·195·243·1177	1490·610	10·414·34·147·324·24·741·403·3
CP007230	347·632·1364	451·1017·875	336·332·574·830·132·139	726·303·239·63·303·15·108·213·276·97	722·888·330·403
CP007448	617·1615	997·1235	291·188·516·66·628·135·408	119·1791·322	20·128·595·324·24·1014·127
CP009367	–	724·1144·370	–	1628·610	148·414·34·147·324·24·741·403·3
CP009539	–	400·306·1499	–	–	–
CP009781	67·469·456·871·372	1588·647	1194·762·279	1286·202·116·9·315·307	148·448·147·324·1168
CP009787	–	1123·1118	623·195·481·390·552	1235·369·9·15·613	20·128·414·34·495·741·409
CP009801	1894·341	220·2015	182·279·600·75·61·731·10·297	331·883·195·195·258·119·254	20·128·448·147·333·15·671·473
CP011078	–	706·1499	–	–	–
HF571988	–	–	362·148·1728	1604·24·282·328	–
FR729477	–	–	–	–	–

(·): RE site in the gene sequence; (–): no digestion observed

RsaI, *TaqI* and *Tru9I*. FR729477 could be distinguished from all others on the basis of only one RE-*HpyCH4V*.

In order to distinguish genomes in addition to those 11 listed with *fadJ*, all gene-RE combinations were searched. It provided unique digestion patterns by which another 10 *Yersinia* genomes can be easily identified: *Y. frederiksenii* Y225 (CP009364.1), *Y. kristensenii* Y231 (CP009997.1), *Y. pestis* Harbin35 (CP009704.1), *Y. pestis* biovar Medievalis str. Harbin 35 (CP001608.1), *Y. pestis* D106004 (CP001585.1), *Y. pestis* D182038 (CP001589.1), *Y. pestis* str. Pestoides B (CP010023.1), *Y. pseudotuberculosis* ATCC 6904 (CP008943.1), *Y. pseudotuberculosis* IP 31758 (CP000720.1), and *Y. pseudotuberculosis* str. PA3606 (CP010067.1) (Table 3). Here, the following genes—*aceE*, *aceK*, *cpxP*, *cysJ*, *glpQ*, *gltB*, *gyrB*, *lacZ*, *leuD*, *ligA*, *lolD*, *metH*, *mukB*, *pheT*, *rpoB*, and *secA*, on digestion with specific REs—*AluI*, *BfaI*, *HpyCH4V*, *CviAII*, *TaqI*, *BfuCI*, *RsaI*, *Hin1I*, *Tru9I* (Table 3) provided information on unique digestion patterns. The patterns obtained with specific RE-gene combination can be used as biomarkers for demarcating *Yersinia* genomes with high specificity.

Discussion

The urgency to identify the disease causing organisms is always there, especially in case of an epidemic outbreak. Morphological and biochemical characteristics prove effective to a certain extent in identifying the organism. However, modern biological methods have narrowed down the search to the use of highly conserved genes such as *rrs* and a few HKGs for bacterial identification. The use of *rrs* has been the most extensive and adapted even by not so well equipped laboratories. In order to extend the usage of *rrs* for distinguishing very closely related strains, and provide biomarkers for rapidly identifying bacteria, certain latent features of *rrs* have been elucidated using molecular approaches. The use of REs to deduce unique *rrs* digestion

Table 3 *Yersinia* genomes distinguished based on unique Gene: Restriction Endonuclease approach

Genome	Gene	RE	Restriction endonuclease digestion Pattern
CP000720	*glpQ*	*AluI*	419·271·222·22·182
CP001608	*ligA*	*AluI*	345·204·153·533·179·570·29
CP010067	*gyrB*	*AluI*	884·266·12·775·216·262
	gltB	*BfaI*	71·3224·483·538·142
	cysJ	*HpyCH4V*	63·330·222·469·26·150·210·303·84
	mukB	*AluI*	81·661·335·70·155·78·192·111·75·15·66·712·152·122·193·46·13·34·12·149·22·402·412·155·78·117
		CviAII	477·54·298·96·104·1231·303·180·63·125·168·748·290·229·6·6·80
		HpyCH4V	112·62·57·87·171·531·204·36·117·168·72·211·68·276·252·393·58·401·123·747·138·118·41·15
		TaqI	68·279·507·195·648·270·12·231·42·1383·15·59·304·30·83·332
FR729477	*gltB*	*AluI*	533·331·321·159·57·226·25·40·340·914·320·103·90·220·151·400·205·5·21
		BfuCI	131·55·365·28·30·714·81·150·180·600·143·16·42·843·168·243·356·106·38·172
		RsaI	7·1804·24·39·257·289·667·192·216·321·233·81·10·321
	metH	*Hin1I*	1400·6·2129·161
	mukB	*BfuCI*	69·83·249·367·204·1317··207·140·157·732·921
	pheT	*AluI*	634·528·174·13·569·98·48·79·245
	secA	*AluI*	111·150·85·141·6·230·181·33·268·348·516·473·173
		BfuCI	42·33··113·456·192·417·328·12·380·418·74·190·60
		TaqI	104·147·372·24·549·120·56·328·216·192·33·574
CP008943	*lacZ*	*BfaI*	919·1899·103·280
CP009364	*aceE*	*BfuCI*	27·5·27·52·189·105·320·207·135·28·171·425·15·61·11·49·44·106·687
	cpxP	*CviAII*	159·108·231
CP009997	*aceE*	*BfuCI*	27·5·27·52·189·105·320·207·135·28·171·425·15·61·11·49·44·106·687
CP001589	*aceK*	*CviAII*	121·48·918·17·79·219·264·62
CP009367	*cpxP*	*Hin1I*	37·461
CP010023	*leuD*	*HpyCH4V*	46·44·234
		Tru9I	128·162··34
CP001585	*lolD*	*Tru9I*	221·218·118·148
CP009704	*rpoB*	*Tru9I*	293·120·1011·384·18·375·6·531·429·69·793

Symbol filled circle (·) indicates RE site in the gene sequence

Table 4 Segregation of *Yersinia* species with different genes

Segregated by *fadJ*	Segregated by genes other than *fadJ*[a]
Y. aldovae 670-83 CP009781.1	*Y. frederiksenii* Y225 CP009364.1
Y. enterocolitica WA CP009367.1	*Y. kristensenii* Y231 CP009997.1
Y. enterocolitica (*type O:5*) *str. YE53/03* HF571988.1	*Y. pestis* Harbin35 CP009704.1
Y. enterocolitica LC20 CP007448.1	*Y. pestis biovarMedievalis str.* Harbin 35 CP001608.1
Y. enterocolitica subsp. *enterocolitica* 8081 AM286415.1	*Y. pestis* D106004 CP001585.1
Y. enterocolitica subsp. *palearctica* Y11 FR729477.2	*Y. pestis* D182038 CP001589.1
Y. intermedia Y228 CP009801.1	*Y. pestis str.* Pestoides B CP010023.1
Y. rohdei YRA CP009787.1	*Y. pseudotuberculosis* ATCC 6904 CP008943.1
Y. ruckeri Big Creek 74 CP011078.1	*Y. pseudotuberculosis* IP 31758 CP000720.1
Y. ruckeri YRB CP009539.1	*Y. pseudotuberculosis str.* Pa3606 CP010067.1
*Y. similis*228 CP007230.1	
Could not be segregated	
Y. enterocolitica 2516-87 CP009838.1	*Y. pestis KIM10* + AE009952.1
Y. enterocolitica subsp. *palearctica* 105.5R(r) CP002246.1	*Y. pestis* Nepal516 CP000305.1
Y. pestis Dodson CP009844.1	*Y. pestis* Pestoides F CP000668.1
Y. pestis El Dorado CP009785.1	*Y. pestis* Pestoides F CP009715.1
Y. pestis Nicholisk 41 CP009991.1	*Y. pestis* Pestoides G CP010247.1
Y. pestis PBM19 CP009492.1	*Y. pestis* Z176003 CP001593.1
Y. pestis Shasta CP009723.1	*Y. pseudotuberculosis*1 CP009786.1
Y. pestis A1122 CP002956.1	*Y. pseudotuberculosis* EP2/+ CP009759.1
Y. pestis A1122 CP009840.1	*Y. pseudotuberculosis* MD67 CP009757.1
Y. pestis Angola CP000901.1	*Y. pseudotuberculosis* IP 32953 B X936398.1
Y. pestis Angola CP009935.1	*Y. pseudotuberculosis* IP32953 CP009712.1
Y. pestis Antiqua CP000308.1	*Y. pseudotuberculosis* PB1/+ CP001048.1
Y. pestis Antiqua CP009906.1	*Y. pseudotuberculosis* PB1/+ CP009780.1
Y. pestisbiovarMicrotus str. 91001 AE017042.1	*Y. pseudotuberculosis* YPIII CP000950.1
Y. pestis CO92-AL590842.1	*Y. pseudotuberculosis* YPIII CP009792.1

[a] See Table 3 for list of genes

patterns have proved to be quite effective in the cases of: (i) *Bacillus*, (ii) *Clostridium*, (iii) *Pseudomonas*, and (iv) *Streptococcus* [12, 32–36]. In spite of a roaring success observed in the usage of *rrs*, the phylogenetic analysis of *rrs* gene sequences has not proved very useful in an unambiguous manner in the case of *Yersinia.* This limitation in employing *rrs* in distinguishing *Yersinia* species has been assigned to the presence of its multiple copies per genome and high sequence similarity among themselves [38]. A similar scenario was observed in *Clostridium* spp., which also possesses multiple copies of *rrs*. These *rrs* copies also show high similarity [33, 35]. In order to circumvent this problem, we have employed a genome wide analysis for searching novel biomarkers in *Clostridium* [15]. Compared to only 22 genes being common to all the 27 genomes of *Clostridium* spp., we could observe 304 genes to be common to 51 sequenced genomes of *Yersinia*.

In silico digestion of nucleotide sequences 34 genes with different type II REs has allowed us to deduce unique combinations of genes and REs. A thorough analysis of RE digestion patterns revealed that 21 out of 51 strains representing 10 species of *Yersinia*. These differences in the RE digestion patterns arise because of changes in the nucleotide within the RE recognition motif [37]. By this approach we can use novel biomarkers—*carB*, *fadJ*, *gluM*, *gltX*, *ileS*, *malE*, *nusA*, *ribD*, and *rlmL* and their RE digestion patterns for rapidly identifying *Yersinia* species (Table 4). It may be remarked that these genes have not been reported earlier as biomarkers for identifying *Yersinia*. Previously, the genes involved in pathogenesis: (i) *ail*, *inv*, *yst*, *myf*, *vir*, and *yop* have been used for identifying *Yersinia* [16–19]. Incidentally, these 5 genes are not present in all the strains of *Yersinia*, making it difficult to use them as universal biomarkers [39, 40]. Even *gyrB* gene, which is commonly used for bacterial identification, did not prove very effective in distinguishing *Yersinia* strains being studied here. On the other hand, the biomarkers identified in this study are common to all the *Yersinia* strains, are

highly specific and have a unique RE digestion pattern. Thus by using specific primer sets, we can amplify these genes through polymerase chain reaction. The amplicon can be digested with the specific RE, through which we can detect this organism even among a large group of unrelated bacterial populations e.g., those present in a soil sample, contaminated food or water sample or wounds and even on medical equipments, etc. Our data provides information to detect even non-pathogenic *Yersinia*, which may be present by chance. In spite of the fact that these genes, including *fadJ* are present in genus such as *Pseudomonas*, their RE digestion patterns were remarkably different (Data not shown). These biomarkers can thus be used for developing diagnostic kits specifically for *Yersinia*.

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. 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
2. Gupta V, Gulati P, Bhagat N, Dhar MS, Virdi JS (2015) Detection of *Yersinia enterocolitica* in food: an overview. Eur J Clin Microbiol Infect Dis 34:641–650. doi:10.1007/s10096-014-2276-7
3. Santos-Montañez J, Benavides-Montaño JA, Hinz AK, Vadyvaloo V (2015) *Yersinia pseudotuberculosis* IP32953 survives and replicates in trophozoites and persists in cysts of *Acanthamoeba castellanii*. FEMS Microbiol Lett. doi:10.1093/femsle/fnv091
4. Laukkanen-Ninios R, Fredriksson-Ahomaa M, Korkeala H (2014) Enteropathogenic *Yersinia* in the pork production chain: challenges for control. Compr Rev Food Sci Food Saf 13:1165–1191. doi:10.1111/1541-4337.12108
5. Hochel I, Skvor J (2007) Characterization of rabbit antibodies for immunochemical detection of *Yersinia enterocolitica*. Folia Microbiol (Praha) 52:511–518. doi:10.1007/BF02932112
6. Kapperud G, Vardund T, Skjerve E, Hornes E, Michaelsen TE (1993) Detection of pathogenic *Yersinia enterocolitica* in foods and water by immunomagnetic separation, nested polymerase chain reactions, and colorimetric detection of amplified DNA. Appl Environ Microbiol 59:2938–2944
7. Thibodeau V, Frost EH, Quessy S (2001) Development of an ELISA procedure to detect swine carriers of pathogenic *Yersinia enterocolitica*. Vet Microbiol 82:249–259. doi:10.1016/S0378-1135(01)00356-X
8. Oh BK, Lee W, Chun BS, Bae YM, Lee WH, Choi J-W (2005) Surface plasmon resonance immunosensor for the detection of *Yersinia enterocolitica*. Colloids Surf A Physicochem Eng Asp 257–258:369–374. doi:10.1016/j.colsurfa.2004.10.083
9. Myers KM, Gaba J, Al-Khaldi SF (2006) Molecular identification of *Yersinia enterocolitica* isolated from pasteurized whole milk using DNA microarray chip hybridization. Mol Cell Probes 20:71–80. doi:10.1016/j.mcp.2005.09.006
10. Tomita N, Mori Y, Kanda H, Notomi T (2008) Loop-mediated isothermal amplification (LAMP) of gene sequences and simple visual detection of products. Nat Protoc 3:877–882. doi:10.1038/nprot.2008.57
11. Giannino ML, Aliprandi M, Feligini M, Vanoni L, Brasca M, Fracchetti F (2009) A DNA array based assay for the characterization of microbial community in raw milk. J Microbiol Methods 78:181–188. doi:10.1016/j.mimet.2009.05.015
12. 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
13. Chen J, Miao X, Xu M, He J, Xie J, Wu X, Chen G, Yu L, Zhang W (2015) Intra-genomic heterogeneity in 16S rRNA genes in strictly anaerobic clinical isolates from periodontal abscesses. PLoS One 10:e0130265. doi:10.1371/journal.pone.0130265
14. Pei AY, Oberdorf WE, Nossa CW, Agarwal A, Chokshi P, Gerz EA, Jin Z, Lee P, Yang L, Poles M, Brown SM, Sotero S, DeSantis T, Brodie E, Nelson K, Pei Z (2010) Diversity of 16S rRNA genes within individual prokaryotic genomes. Appl Environ Microbiol 76:3886–3897. doi:10.1128/AEM.02953-09AEM.02953-09
15. Kekre A, Bhushan A, Kumar P, Kalia VC (2015) Genome wide analysis for searching novel markers to rapidly identify *Clostridium* strains. Indian J Microbiol 55:250–257. doi:10.1007/s12088-015-0535-7
16. Khushiramani R, Shukla J, Tuteja U, Batra HV (2006) *Yersinia* outer-membrane protein B (YopB): a tool for identification of *Yersinia pestis* isolates. J Med Microbiol 55:467–469. doi:10.1099/jmm.0.46382-0
17. Paixão R, Moreno LZ, Sena de Gobbi DD, Raimundo DC, Hofer E, Matté MH, Sebastiana T, Ferreira P, de Moura Gomes VT, Costa BLP, Moreno AM (2013) Characterization of *Yersinia enterocolitica* biotype 1A Strains isolated from swine slaughterhouses and markets. Sci World J 2013:769097. doi:10.1155/2013/769097
18. Garzetti D, Susen R, Fruth A, Tietze E, Heesemann J, Rakin A (2014) A molecular scheme for *Yersinia enterocolitica* pathoserotyping derived from genome-wide analysis. Int J Med Microbiol 304:275–283. doi:10.1016/j.ijmm.2013.10.007
19. von Altrock A, Seinige D, Kehrenberg C (2015) *Yersinia enterocolitica* isolates from wild boars hunted in Lower Saxony, Germany. Appl Environ Microbiol 81:4835–4840. doi:10.1128/AEM.00550-15
20. Malorny B, Tassios PT, Rådström P, Cook N, Wagner M, Hoorfar J (2003) Standardization of diagnostic PCR for the detection of foodborne pathogens. Int J Food Microbiol 83:39–48. doi:10.1016/S0168-1605(02)00322-7
21. Mäde D, Reiting R, Strauch E, Ketteritzsch K, Wicke A (2008) A real-time PCR for detection of pathogenic *Yersinia enterocolitica* in food combined with an universal internal amplification control system. J Consum Prot Food Saf 3:141–151. doi:10.1007/s00003-008-0341-9
22. Ikeda M, Yamaguchi N, Tani K, Nasu M (2006) Detection of food poisoning bacteria in fresh vegetables using DNA microarray. J Health Sci 52:36–42. doi:10.1248/jhs.52.36
23. Bang J, Beuchat LR, Gu MB, Chang HI, Ryu JH (2010) Identification of *Yersinia enterocolitica* using a random genomic DNA microarray chip. Lett Appl Microbiol 51:665–670. doi:10.1111/j.1472-765X.2010.02954.x
24. Kim DH, Lee BK, Kim YD, Sk Rhee, Kim YC (2010) Detection of representative enteropathogenic bacteria, *Vibrio* spp., pathogenic *Escherichia coli*, *Salmonella* spp., *Shigella* spp., and *Yersinia enterocolitica*, using a virulence factor gene-based oligonucleotide microarray. J Microbiol 48:682–688. doi:10.1007/s12275-010-0119-5
25. Mori Y, Nagamine K, Tomita N, Notomi T (2001) Detection of loop mediated isothermal amplification reaction by turbidity derived from magnesium pyrophosphate formation. Biochem

Biophys Res Commun 289:150–154. doi:10.1006/bbrc.2001.5921
26. Gao H, Lei Z, Jia J, Wang S, Chen Y, Sun M, Liang C (2009) Application of loop-mediated isothermal amplification for detection of *Yersinia enterocolitica* in pork meat. J Microbiol Methods 77:198–201. doi:10.1016/j.mimet.2009.02.001
27. Li Y, Jiang M, Liu W, Zhang L, Zhang S, Zhao X, Xiang R, Liu Y (2010) Loop-mediated isothermal amplification method targets to the *phoP* gene for detection of *Yersinia enterocolitica*. Mol Cell Probes 24:68–71. doi:10.1016/j.mcp.2009.10.003
28. Bai S, Zhao J, Zhang Y, Huang W, Xu S, Chen H, Fan L-M, Chen Y, Deng XW (2010) Rapid and reliable detection of 11 food-borne pathogens using thin-film biosensor chips. Appl Microbiol Biotechnol 86:983–990. doi:10.1007/s00253-009-2417-6
29. Kuhm AE, Suter D, Felleisen R, Rau J (2009) Identification of *Yersinia enterocolitica* at the species and subspecies levels by Fourier transform infrared spectroscopy. Appl Environ Microbiol 75:5809–5813. doi:10.1128/AEM.00206-09
30. Ayyadurai S, Flaudrops C, Raoult D, Drancourt M (2010) Rapid identification and typing of *Yersinia pestis* and other *Yersinia* species by matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry. BMC Microbiol 10:285. doi:10.1186/1471-2180/10/285
31. Stephan R, Cernela N, Ziegler D, Pflüger V, Tonolla M, Ravasi D, Fredriksson-Ahomaa M, Hächler H (2011) Rapid species specific identification and subtyping of *Yersinia enterocolitica* by MALDI-TOF mass spectrometry. J Microbiol Methods 87:150–153. doi:10.1016/j.mimet.2011.08.016
32. 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
33. 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
34. 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
35. 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
36. Kalia VC (2015) Let's explore the latent features of genes to identify bacteria. J Mol Genet Med 9:e105. doi:10.4172/1747-0862.1000E105
37. 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
38. Shivaji S, Bhanu NV, Aggarwal RK (2000) Identification of *Yersinia pestis* as the causative organism of plague in India as determined by 16S rDNA sequencing and RAPD-based genomic fingerprinting. FEMS Microbiol Lett 189:247–252. doi:10.1111/j.1574-6968.2000.tb09238.x
39. Bhagat N, Virdi JS (2007) Distribution of virulence associated genes in *Yersinia enterocolitica* biovar 1A correlates with clonal groups and not the source of isolation. FEMS Microbiol Lett 266:177–183. doi:10.1111/j.1574-6968.2006.00524.x
40. Kishore P, Lalitha KV, Joseph TC, Thampuran N (2012) Biotyping and antibiotic resistance profile of *Yersinia enterocolitica* associated with seafoods from South-west coast of India. Fish Technol 49:64–71