INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY

METHOD

Gupta and Kanter-Eivin, *Int. J. Syst. Evol. Microbiol.* 2023;73:005844

DOI 10.1099/ijsem.0.005844

MICROBIOLOGY SOCIETY

OPEN ACCESS

AppIndels.com server: a web-based tool for the identification of known taxon-specific conserved signature indels in genome sequences. Validation of its usefulness by predicting the taxonomic affiliation of >700 unclassified strains of *Bacillus* species

Radhey S. Gupta* and David A. Kanter-Eivin

Abstract

Taxon-specific conserved signature indels (CSIs) in genes/proteins provide reliable molecular markers (synapomorphies) for unambiguous demarcation of taxa of different ranks in molecular terms and for genetic, biochemical and diagnostic studies. Because of their predictive abilities, the shared presence of known taxon-specific CSIs in genome sequences has proven useful for taxonomic purposes. However, the lack of a convenient method for identifying the presence of known CSIs in genome sequences has limited their utility for taxonomic and other studies. We describe here a web-based tool/server (AppIndels.com) that identifies the presence of known and validated CSIs in genome sequences and uses this information for predicting taxonomic affiliation. The utility of this server was tested by using a database of 585 validated CSIs, which included 350 CSIs specific for ≈45 *Bacillales* genera, with the remaining CSIs being specific for members of the orders *Neisseriales*, *Legionellales* and *Chlorobiales*, family *Borreliaceae*, and some *Pseudomonadaceae* species/genera. Using this server, genome sequences were analysed for 721 *Bacillus* strains of unknown taxonomic affiliation. Results obtained showed that 651 of these genomes contained significant numbers of CSIs specific for the following *Bacillales* genera/families: *Alkalicoccus*, *'Alkalihalobacillaceae'*, *Alteribacter*, *Bacillus* Cereus clade, *Bacillus* Subtilis clade, *Caldalkalibacillus*, *Caldibacillus*, *Cytobacillus*, *Ferdinandcohnia*, *Gottfriedia*, *Heyndrickxia*, *Lederbergia*, *Litchfieldia*, *Margalitia*, *Mesobacillus*, *Metabacillus*, *Neobacillus*, *Niallia*, *Peribacillus*, *Priestia*, *Pseudalkalibacillus*, *Robertmurraya*, *Rossellomorea*, *Schinkia*, *Siminovitchia*, *Sporosarcina*, *Sutcliffiella*, *Weizmannia* and *Caryophanaceae*. Validity of the taxon assignment made by the server was examined by reconstructing phylogenomic trees. In these trees, all *Bacillus* strains for which taxonomic predictions were made correctly branched with the indicated taxa. The unassigned strains likely correspond to taxa for which CSIs are lacking in our database. Results presented here show that the AppIndels server provides a useful new tool for predicting taxonomic affiliation based on shared presence of the taxon-specific CSIs. Some caveats in using this server are discussed.

INTRODUCTION

Genome sequences are now a prerequisite for the description of new prokaryotic species [1–3], and they provide an all-encompassing resource for developing a stable and evolutionarily coherent classification of organisms [3–10]. Based on genome sequences, standardized and readily measurable criteria have now been developed for delineation of species boundaries, based upon average nucleotide identity (ANI; >95–96%) [11–13], average amino acid identity (AAI) [14], digital DNA–DNA hybridization (DDH; >70%) [15, 16] and 16S rRNA similarity (>98.7%) values [12, 17, 18]. Genome sequences also allow for the reconstruction of robust phylogenetic trees based on large datasets of core genes/proteins exhibiting high-resolution at different

Author affiliations: [1]Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Ontario CA L8N 3Z5, Canada.
***Correspondence:** Radhey S. Gupta, gupta@mcmaster.ca
Keywords: *Leuconostocaceae*; *Borreliaceae*; *Neisseriales*; *Caryophanaceae*; conserved signature indels (CSIs); CSIs specific for the taxa within the *Bacillaceae*; *Legionellales* and *Chlorobiales*; taxonomic affiliation of uncharacterized *Bacillus* species; usefulness of CSIs for predicting taxonomic affiliation.
Abbreviations: AAI, average amino acid identity; ANI, average nucleotide identity; CSI, conserved signature indel; GTDB, genomic taxonomy database; NCBI, National Center for Biotechnology Information; OGRI, overall genomic relatedness indices.
Six supplementary tables are available with the online version of this article.

taxonomic levels [4, 19–22]. However, based on genome sequence information, it is also important to develop reliable means for the circumscription of prokaryotic taxa of higher taxonomic ranks (*viz*. genus, family, order, *etc*.) [4, 18]. Of these, the genus-level taxonomic rank is of particular importance in the fields of microbial systematics and microbiology [3, 9, 23–25]. As all species names include genus name, different species within a given genus should be more closely related to each other than to any other taxa [9, 24, 25]. Furthermore, it is expected that all of these species should all commonly share several genotypic, phenotypic or other properties (e.g. pathogenicity profile or potential) [8, 9, 24, 25]. Additionally, as most studies on microorganisms involve examining the properties of the organisms at species and genus levels, the reliable demarcation of genus-level taxa is of great importance in understanding the biological and functional properties of different groups of micro-organisms [3, 8, 9, 24, 25].

Despite the central role of genus-level clades in systematics and microbiology, currently there are no reliable means for the demarcation of genus-level groupings [18, 26, 27]. Most of the known microbial genera are presently delineated based on the clustering of species in phylogenetic trees based on 16S rRNA or other genes/proteins sequences [20–22]. However, branching patterns and groupings of species in phylogenetic trees are affected by a large number of variables including, but not limited to, the number of species included in the dataset as well as the heterogeneity among them, the selected evolutionary models, and specific genes/proteins sequences that are used for phylogenetic analysis [28–32]. Although cut-off values for genus demarcation have been suggested based on 16S rRNA sequence similarity [18], AAI [33], percentage of conserved proteins (POCP) [34], and genome relatedness indices based on ANI and genome alignment fraction [35], these methods generally show poor correlation in distinguishing between different genera [36], and thus far they have proven of limited usefulness in the demarcation of different genera. Parks *et al*. [4] recently created a genome taxonomy database (GTDB) based on phylogenetic analysis of 120 ubiquitous single-copy proteins. The GTDB server uses relative evolutionary divergence of different taxa in their core protein tree as the basis for the delineation of different genera and other higher taxonomic groupings [4]. Although GTDB taxonomy is now an important resource for microbial systematics [3], delineation of different taxa in it is also based on branching of species in a phylogenetic tree. As the branching of species in trees represents a continuum, it does not permit unambiguous demarcation of the boundaries between different genera or other taxonomic clades [4, 7, 37].

The analyses of genome sequences also allow for the identification of numerous highly specific molecular markers, such as conserved signature indels (CSIs) and conserved signature proteins, which are unique characteristics of specific (monophyletic) groups of organisms [38–46]. Of these markers, CSIs in genes/proteins, which are specific for a given group of organisms, have provided particularly useful markers for evolutionary and taxonomic studies [47–49]. The CSIs, which are useful for evolutionary/taxonomic inferences, are of fixed lengths and they are present at specific positions in particular genes/proteins. These indels are also flanked on both sides by conserved regions to ensure that the genetic changes they represent constitute reliable molecular characteristics [48, 50]. The CSIs in genes/proteins result from rare genetic changes, hence their shared presence within a specific group of organisms indicate that the genetic change giving rise to a specific CSI occurred in a common ancestor of the indicated group [7, 48]. In view of the clade specificities of the genetic changes represented by the CSIs, they constitute molecular synapomorphies, which provide dependable evidence, independently of the phylogenetic trees, supporting the common ancestry and relatedness of the species from specific clades [7, 48]. Earlier work provides evidence that these molecular markers exhibit a high degree of predictive ability to be found in other members (newly identified or sequenced) of a given clade [51–57]. These characteristics of CSIs, in conjunction with their rare and discrete molecular nature, indicate that they provide useful means for the circumscription of clades of different ranks in molecular terms [36, 38, 47, 58–63].

In recent years, CSIs have been identified for numerous prokaryotic taxa of different ranks. Their usage, in conjunction with phylogenomic analyses, has led to important changes in the taxonomy of several groups of prokaryotic organisms [38, 47, 58–60, 63–65]. Recently, the use of these molecular markers in conjunction with phylogenomic analysis has led to the division of species from the highly polyphyletic genus *Bacillus* into >30 genera [52, 58, 66]. In the new classification scheme for *Bacillus* species, all newly described genera were reliably distinguished from each other (and other prokaryotic taxa) based on multiple identified CSIs which were specific for each of these genera [52, 58]. Because of the predictive ability of the CSIs to be present in other members of these genera, the described CSIs have also been used by other investigators to assign newly described *Bacillaceae* species, as well as other species, into different genera [54–56, 67–74]. Although the CSIs provide a very useful means for the circumscription of taxa of different ranks (specifically the genus-level clades), and for the assignment of new species into these genera based on the shared presence of CSIs, there is no convenient method available for determining the presence or absence of known taxon-specific CSIs in genome sequences.

To facilitate the identification and use of previously described and validated CSIs in genome sequences, we describe here a web-based tool or server, AppIndels.com, which uses a database of previously identified CSIs (validated to confirm their specificities) to aid in the assignment of species from a given genome into specific groups/taxa for which CSIs are present in the database. We report here the usefulness and reliability of this webserver by examining the results obtained from it for predicting the taxonomic affiliation of 721 unnamed genome sequenced *Bacillus* species into specific *Bacillales* genera. Based on the CSIs currently in the database, the AppIndels server predicted the taxonomic affiliation of 651 of these unnamed *Bacillus* strains into 29 *Bacillales* genera/families. The accuracy of the taxon prediction made by the server was examined by phylogenetic analyses and the results from these studies showed perfect concordance with the taxon affiliation indicated by the server. Thus, the AppIndels server

provides a useful new tool for identifying known CSIs in a genome sequence and using this information for predicting taxonomic affiliation of the analysed species/strain.

METHODS

Characteristics of the database for CSIs

The core of the AppIndels server is a proprietary database of sequences for previously described CSIs, which are uniquely found in specific taxa/clades of prokaryotic organisms. The CSI database used in this study contains information for 585 CSIs specific for several groups of prokaryotic organisms. These include ≈350 CSIs specific for ≈45 genera belonging to three different families within the order *Bacillales* (*viz*. *Bacillaceae*, *Caryophanaceae* and *Leuconostocaceae*) [52, 58, 66, 75]. Of these, >210 CSIs are specific for >30 *Bacillaceae* genera into which species from the genus *Bacillus* were recently reclassified [52, 58, 66], whereas 138 CSIs are specific for the families *Caryophanaceae* and *Leuconostocaceae* and multiple genera and clades within them [66, 75]. In addition, our database also includes information for >230 CSIs that are specific for members of several other prokaryotic taxa including members of the order *Legionellales* [64], *Neisseriales* [76], *Chlorobiales* [77], family *Borreliaceae* [36, 53, 78], as well as some species/genera belonging to the family *Pseudomonadaceae* [79, 80]. Presently, our database includes CSIs for only those clades/taxa whose specificities was examined within the past 2 years [36, 53, 64, 76–78, 81]. Of the CSIs present in our database, the genus *Alkalihalobacillus* [52] was subsequently split into eight genera (*viz*. *Alkalihalobacterium*, *Halalkalibacterium*, *Halalkalibacter*, *Shouchella*, *Pseudalkalibacillus*, *Alkalicoccobacillus*, *Alkalihalophilus* and the emended genus *Alkalihalobacillus*) [80]. Based on their presence in different species, the CSIs for the genus *Alkalihalobacillus* in our database are specific for a family-level taxon, referred to as '*Alkalihalobacillaceae*' in this manuscript and in the server's database. This clade encompasses all of the above noted genera except *Pseudalkalibacillus* and *Alkalicoccobacillus*.

The database of CSIs, which is in SQL format, contains all necessary information regarding CSIs. This information includes: (i) Taxon specificity of each CSI (e.g. *Neobacillus*, *Cytobacillus*, *etc*.). Two additional columns in the database contain information for different CSIs regarding their immediate parent taxon, and its next higher order taxon or clade. This information is useful in the presentation of results for different CSIs in the form of a hierarchical tree. As the circumscriptions of different clades/taxa in our work is based on genome sequence characteristics, the nomenclature for different groups/taxa in our database is most like that used by the NCBI, which is based on genome sequence data [82]. However, the NCBI taxonomy does not list several taxa with validly published names (*viz*. *Caryophanaceae*, *Chitinibacteraceae*, *Aquaspirillaceae*) [83] for which CSIs are present in our database [66, 76]. Such taxa are listed in our database by their LPSN names. All other novel species clades comprising non-validly published names, or those not conforming to either the NCBI or LPSN nomenclature, are identified in our database by their placement within quotation marks. (ii) Amino acid (aa) sequence information for the protein region where the indicated CSI is found (*viz*. DVVAFTRAVSET|PA|LGEERKWVHYGL) and an additional column indicating whether the specific CSI is an insertion or a deletion. The aa sequence information for the CSI also shows the location of the CSI within the sequence. If a specific CSI is an insertion (as in the sequence shown above), then the position of the insertion is marked by vertical lines |aa|. In the sequence shown, the CSI consists of a two aa insertion, where |PA| is found. On the other hand, if a CSI consists of a deletion, then a gap (space) is placed in the input CSI sequence where the deletion is found. The numbers of gaps or spaces in the sequence correspond to the length of the deletion. (iii) E value cut-off for BLASTp searches. As all CSIs in our database are flanked by minimally five conserved/identical aa residues, the E value cut-off we presently use is 1.0 e^{-4}. This cut-off value permits detection of the described CSIs in different orthologous sequences in BLASTp searches. (iv) Name of the protein in which a given CSI is found. (v) Literature references to the publication(s) where a particular CSI was first described and where additional information regarding the specificity of the CSI could be found. (vi) A column indicating the relative weight of a given CSI. Depending upon the number of CSIs that have been identified for different taxa, the weight assigned to an individual CSI varies between 0.15 (for taxa with ≥10 CSIs) to 0.5 for taxon where only two CSIs have been identified. The rationale for weight assignment is discussed in the Results section, and it serves to enhance the specificity of the taxon identification made by the server. Additionally, for the published CSIs that contained long stretches of conserved flanking sequences, in our CSI database,the lengths of the flanking sequences have been reduced to between 6–10 conserved residues on either side. This measure reduces the probability of matching to other similar sequence regions in unrelated genomic sequences.

BLAST analysis with genome sequences of uncharacterized *Bacillus* strains

Genome sequences for the *Bacillus* species/strains analysed in this study were downloaded from the NCBI genome database [82]. The AppIndels server currently uses annotated protein sequences and genome sequence files should have.faa (or.fasta) extension. The quality of the submitted genome sequence is not checked by the server. However, as the genome sequence quality can impact the results, it is recommended that the submitted genomes should be checked for their quality (i.e. completeness or contamination) by using the CheckM server [84]. The AppIndels server uses local BLASTp program for analysis [85]. To analyse a given genome for the presence of matching CSIs, BLASTp searches are carried out with the sequences of all CSIs in the database against the uploaded genome. The observed BLAST hits are examined to check that the indels within the sequences are present in the same location as indicated in the database and they are also flanked by five or more identical/conserved residues on both sides.

For the sequences matching these criteria, information regarding the taxon-specificities of different matching CSIs is gathered. If the total weight of all identified CSIs specific for a taxon exceeds the threshold value of 1.0, a positive identification is made for that taxon. The server then shows the result that a positive match has been found for a specific taxon and displays the number of CSIs identified for that taxon. Using an embedded script, the server also displays the amino acid sequence alignments of the database CSI(s) (i.e. labelled 'Query' in the results) with the sequence in the input genome (labelled 'Subject' in the results). If the matching CSIs are present at multiple taxonomic levels, these results are displayed in the form of a taxonomic hierarchical tree, indicating the number of CSIs present for taxa of different ranks.

In the present work, genome sequences were initially analysed for 394 uncharacterized *Bacillus* species, whose assembly levels were indicated as chromosome (12), complete (72) and contig (312). Subsequently, genomes for 327 additional uncharacterized *Bacillus* species with scaffold-level assembly were also analysed.

Phylogenetic analysis

Phylogenetic trees were reconstructed using genome sequences for specific strains of *Bacillus* species (as indicated in the results section) and generally two species (type species and in most cases one additional species) from different *Bacillaceae* genera for which CSIs are present in our database. Genome sequences for *Lactococcus lactis* and *L. piscium* were used for rooting the tree. All reconstructed phylogenetic trees for the analysed species/strains are based on concatenated sequences for 87 conserved proteins that are part of the phyloeco marker set for the phylum *Bacillota* [86]. Phylogenetic analysis was carried out, using an internally developed pipeline, as described in our earlier work [38, 52, 64]. A maximum-likelihood tree based on the sequence alignment was reconstructed using FastTree 2 [87] and optimized using RAxML 8 [88]. RAxML 8 was also used to calculate the SH-like statistical support values for different nodes and the trees were drawn using MEGA7 [89].

RESULTS

Overall schematic and rationale of the server

Fig. 1 shows an overall schematic of the working of the AppIndels server. The central core of this server is a database of previously identified CSIs, which are specific for different groups (taxa) of micro-organisms. As noted in the Methods section, the server's database presently contains sequence information for >585 CSIs specific for several prokaryotic taxa. Of these, >350 CSIs are specific for ≈45 *Bacillales* genera, which are the focus of this study [66, 75]. Using local BLASTp searches, the server examines the presence of these CSIs in the input genome sequence, and based on this information determines/predicts whether the given genome corresponds to a specific taxon. The working of this server relies on the predictive ability of the CSIs to be present in other members of the indicated taxa (clades). Although earlier work on CSIs provides strong evidence regarding their predictive ability [7, 47, 51–54, 90], to further demonstrate this aspect, in Fig. 2 we present updated sequence information for a previously identified CSI specific for the genus *Neobacillus* [52]. This CSI consists of a four aa deletion in a conserved region of the protein imidazole glycerol phosphate synthase subunit HisH [52]. When this CSI was reported, it was present in 12 different *Bacillus* species, 10 of which were transferred to the genus *Neobacillus* [52]. Two other species sharing the CSIs *viz*. '*B. dielmoensis*' and '*B. rubiinfantis*', were not validly published, hence their transfer to *Neobacillus* was not carried out [83]. Upon updating its sequence information, this CSI in addition to the previously described species, is also shared by seven new *Neobacillus* species and two non-validly published *Bacillus* species (*viz*. '*B. renqingensis*' and '*B. salipaludis*'). This taxonomic inference is congruent with the GTDB taxonomy [4], which also assigns these species to the genus *Neobacillus*. In addition, this CSI is also present in a non-type strain (S-E5) of *Rhodococcus erythropolis*. However, as the type strain of *R. erythropolis* (JCM 25477) does not contain this or any other *Neobacillus*-specific CSIs, the presence of this and other *Neobacillus*-specific (not shown) CSIs in the strain S-E5 is likely due to the mislabeling of its genome. Besides these named species, this CSI is also shared by several unnamed strains of *Bacillus* and *Neobacillus* species (as discussed later). With the exception of these species/strains, this CSI was not present in any other *Bacillota* or other bacteria in the top 500 hits. The observed results demonstrate the specificity of this CSI for *Neobacillus* and provide further evidence of the predictive ability of the CSIs to be present in other members of a specific taxon [67, 68].

Rationale for the weight assignment to different CSIs in the database

Different CSIs which are part of our database were all considered specific or useful for the identification of indicated taxa in our published work [36, 52, 58, 64, 66, 75–77, 79, 80]. However, in these studies, the taxa specificities of the CSIs were generally assessed by examining the top 300–500 BLASTp hits [48, 52, 58, 64, 75]. While most of the described CSIs were specific for the indicated group/taxon by this criterion, in some cases, a CSI was also considered useful and retained, if it was shared by 1–2 other species, which were unrelated to the taxon of interest [7, 52, 58, 75]. As discussed in earlier work [7, 48], the shared presence of a similar CSI in phylogenetically unrelated species occur in isolated cases which can result from several factors including lateral gene transfers, the occurrence of an analogous genetic mutation due to convergent evolution, and also by non-specific factors such as contamination, misclassification or mislabeling of genome sequences. Furthermore, as genome sequence information is rapidly expanding, the possibility of finding a similar CSI in unrelated species also increases. When examining the result for a specific CSI, the significance (or the lack thereof) of isolated exceptions can be readily assessed. Thus, in the example shown

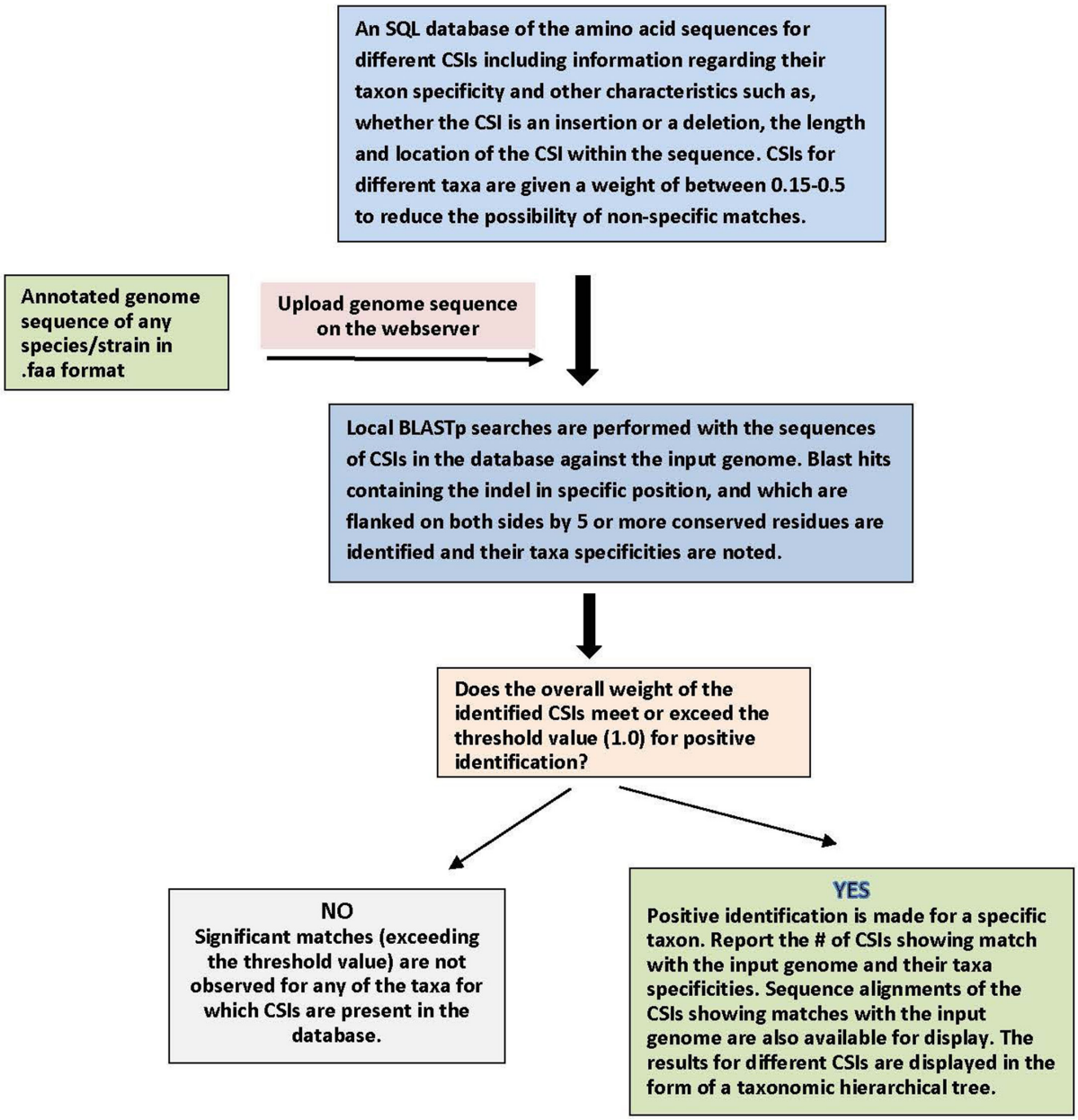

Fig. 1. Schematic diagram showing the working of the AppIndels server using the presence of known taxon-specific CSIs in a genome sequence for taxonomic prediction.

in Fig. 2, the presence of a similar CSI in several non-validly published *Bacillus* species is due to their relatedness to the genus *Neobacillus* (awaiting reclassification), whereas the presence of this CSI in a non-type strain of *R. erythropolis* is very likely due to the mislabeling of its genome. However, when large numbers of CSIs specific for different taxa are used for making predictions of taxonomic affiliation (as in the AppIndels server), these isolated exceptions can lead to misleading inferences. This is of particular concern when no CSIs are present in the database for a taxon to which the analysed genome may be affiliated. In such cases, the only observed matches will be for non-specific taxa. For example, for the *Bacillus* sp. APMAM (Table S3, available in the online version of this article), the prediction made by the server was 'None', as significant numbers of matches were not observed for any taxa for which CSIs are present in the database. However, upon examining the raw results from BLASTp searches, this genome was found to contain 1 CSI each matching the following taxa, *viz*. *Chitinibacteraceae*, *Chromobacteriaceae*, *Bacillus* Subtilis clade, *Ferdinandcohnia*, *Margalitia* and *Chlorobiales*. Based on these results it is difficult to determine whether this genome is affiliated to any of these taxa, and these results, if presented, would be confusing and misleading.

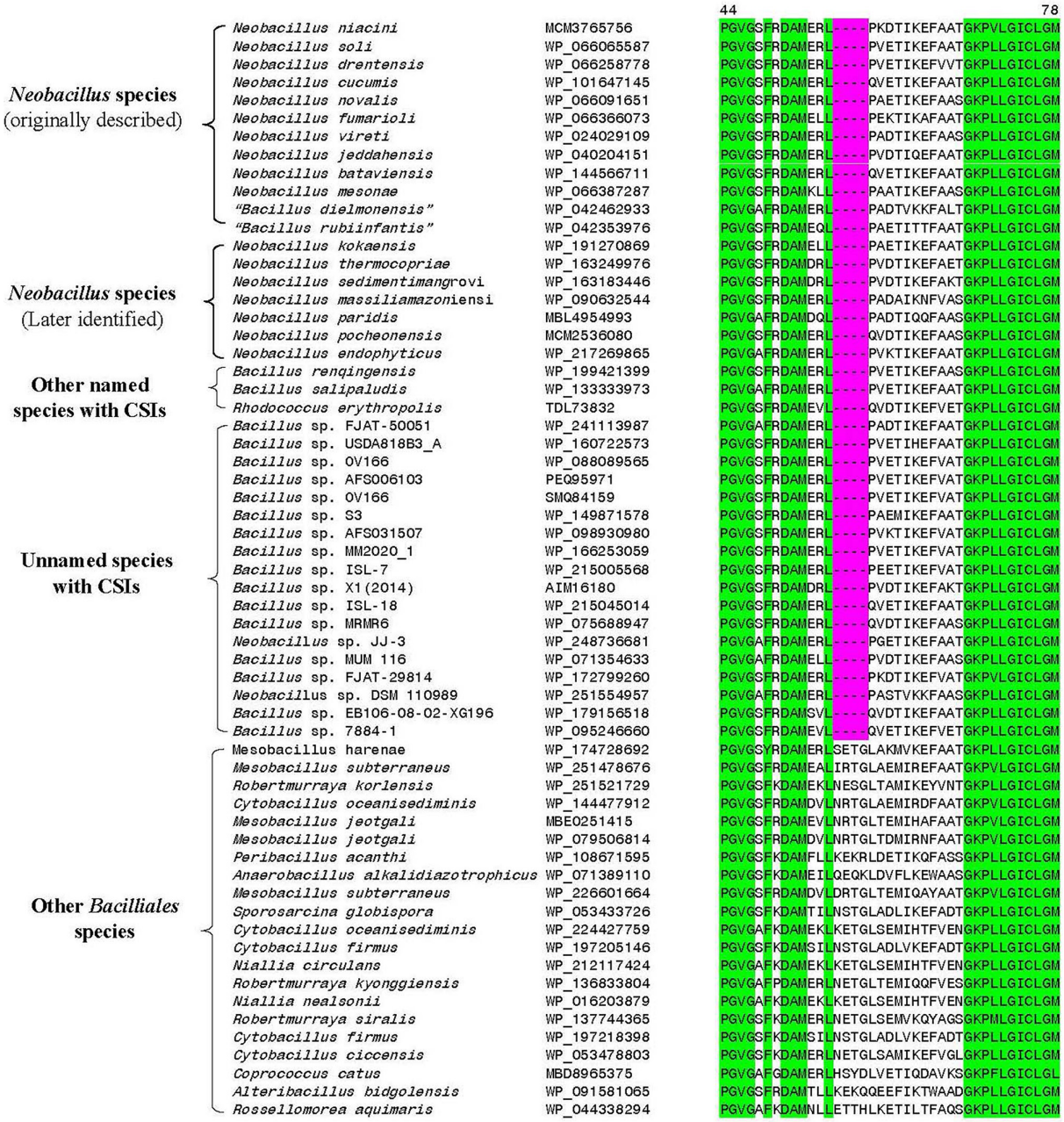

Fig. 2. Updated sequence information for a *Neobacillus*-specific CSI consisting of a four aa deletion (highlighted in pink) in the protein imidazole glycerol phosphate synthase subunit HisH [52]. In addition to the *Neobacillus* species that were known earlier, this CSI is also present in different newly identified *Neobacillus* species and two "*Bacillus* species", which are related to (part of) this genus. This CSI is also present in several unnamed *Bacillus* species and *Neobacillus* species, which based upon the results presented here are affiliated to *Neobacillus*. The results presented here demonstrate the predictive ability of the CSIs to be present in other members of a given group.The flanking aa residues, which are conserved in different species, are highlighted in green.

To overcome this problem, we have introduced a weight element to different CSIs in the server's database to filter out non-specific results from the output. The rationale for the weight assignment to different CSIs is based on our observation that whenever multiple CSIs are known for a taxon, isolated exceptions present in a specific CSI are rarely observed in the other CSIs specific for that group [7, 48, 52, 58]. Based on this information, by assigning a weight element to different CSIs, one can ensure that a positive identification for any taxon is only made when several CSIs specific for that taxon are present in the analysed genome. To implement this, based upon the total number of CSIs known for different taxa, the CSIs have been assigned a weight value of between 0.15 and 0.5. The weight given to an individual CSI is generally less when three or more CSIs are known for a taxon,

and it reaches a value of 0.5 when only two CSIs are known for the taxon. In our CSI database, except for a limited number of genera for which only two CSIs have been identified, for all other taxa three or more CSIs are present in the database. Based on this information, a weight threshold for the identified CSIs has been set at 1.0 for positive identification of any taxon. Thus, a positive identification by the server is only made only when the total weight of the matching CSIs for a specific taxon and its descendants is equal to or exceeds the threshold value of 1.0. This greatly increases the specificity of the taxon prediction made by the server by filtering out results from any non-specific matches.

User interface of the AppIndels server and testing its utility and the reliability of the results obtained

The AppIndels server has a very simple user interface, as shown in Fig. 3(a). After a brief description indicating how the server works, and also indicating some important considerations in its usage and interpreting the result provided by it, the server provides the user options to either submit a query or view the results from earlier queries. The opening page of the server also provides a link where one can search or browse whether the CSIs specific for the group(s) of interest are present in the server's database or not. Upon choosing the 'Submit a Query' option, it takes the user to another page where they are prompted to upload a genome sequence. Genome sequence files to be analysed (in.faa format) can be either dragged and dropped in the provided box, or one can browse for the file, and then select it (Fig. 3b). Upon pressing the 'Start Analysis' button, the server begins the analysis, and results are displayed within a relatively short time (generally <15 s). If the input genome is found to contain CSIs specific for a particular taxon (and their combined weight meets the threshold value of 1.0), the results displayed will indicate: 'CSIs present for [Name of the Taxon]' (Fig. 3c). In Fig. 3c, the results are shown for an uncharacterized *Bacillus* sp. OV166, which the server identified as affiliated to the genus *Neobacillus*. This prediction was based on the presence of 10 CSIs specific for the genus *Neobacillus* in its genome sequence. Information for the sequences matching with the *Neobacillus*-specific CSIs, and the proteins in which these CSIs are present, are also indicated in the provided results. If the input genome sequence contains CSIs matching at more than one taxonomic level, these results are displayed in the form of a hierarchical tree indicating the numbers of CSIs matching at different taxonomic levels, with the name of the most specific taxon indicated at the top. An example of this is shown in Fig. 3(d), where the results are shown for the genome of another uncharacterized *Bacillus* sp. OxB-1. This strain was identified by the server as belonging to the genus *Sporosarcina*. In this case, in addition to the six CSIs which are specific for the genus *Sporosarcina*, its genome also showed 11 matches with the CSIs specific for the family *Caryophanaceae* [66]. References to the key publications for different identified CSIs are also shown on the result page. Furthermore, upon clicking the tab for the CSIs found, information regarding the sequence alignments of the reference CSIs with the sequences from the input genome sequences are displayed. If the input genome sequence showed no match for the CSIs in the database, or if the total weight of the matching CSIs was less than the threshold value of 1.0, the result displayed will be 'None'.

We tested the server initially using the genome sequences for the type species of different genera for which CSIs are present in our database. As expected, the server correctly identified all input genomes to the specific genera (results not shown). Subsequently, we examined the utility of this server for predicting taxonomic affiliations of genome sequences for uncharacterized *Bacillus* species/strains. These studies were initially performed on genomes of 394 uncharacterized *Bacillus* strains with chromosomal, complete and contig-level assemblies as indicated in the Methods. Results from these analyses for different strains, along with the accession numbers of the analysed genomes, and the number of matching CSIs present in them and their taxon-specificities, are summarized in Table 1 and Tables S1–S3. Of the analysed genomes, 107 genomes for whom information is present in Table 1, were identified as belonging to the following 27 *Bacillota* genera/families: *Alkalicoccus* (1), '*Alkalihalobacillaceae*' (12), *Alteribacter* (2), *Caldalkalibacillus* (1), *Caldibacillus* (2), *Cytobacillus* (3), *Ferdinandcohnia* (1), *Gottfriedia* (2), *Heyndrickxia* (1), *Lederbergia* (3), *Litchfieldia* (1), *Margalitia* (2), *Mesobacillus* (2), *Metabacillus* (2), *Neobacillus* (12), *Niallia* (2), *Peribacillus* (10), *Priestia* (14), '*Pseudalkalibacillus*' (5), *Robertmurraya* (3), *Rossellomorea* (11), *Schinkia* (1), *Siminovitchia* (1), *Sutcliffiella* (6), *Weizmannia* (4) and the family *Caryophanaceae* (2). In addition, 244 other genomes were identified as belonging to the genus *Bacillus* (Tables S1 and S2), which consists of two different clades, i.e. the *Bacillus* Subtilis clade, and *Bacillus* Cereus clade [58]. Of the 244 genomes corresponding to genus *Bacillus*, 117 were identified as belonging to the *Bacillus* Subtilis clade (Table S1), whereas the remaining 127 genomes were predicted to be part of the *Bacillus* Cereus clade (Table S2). The remaining 41 genomes from our dataset are listed in Table S3 and were indicated as 'None' to denote that significant numbers of CSIs were not identified in these cases for any taxa for which CSIs are present in our database. Lastly, the results obtained for two genomes (Strain V59.32b, accession number GCA_003429085.1; strain LL01, accession number GCA_001037965.1) were ambiguous. In both these cases, the server indicated the presence of significant numbers of CSIs for two different genera (for the strain LL01 seven CSIs specific for '*Alkalihalobacillaceae*' and nine CSIs specific for *Sutcliffiella*; and for V59.32b, three *Peribacillus*-specific CSIs and two *Niallia*-specific CSIs). The CSIs for these genera present in our database are specific for them (specificities checked and confirmed) and they are not shared by the species from the other indicated genus. To account for these anomalous results, genome sequences for these strains were submitted to the CheckM server [84] for quality and contamination check. The CheckM analysis showed that the genome of strain LL01 is highly contaminated (55.98%), which explains the presence of CSIs specific for two different taxa in this genome. However, the genome for strain V59.32b showed only minimal contamination (1.73%), and it is unclear why it contains CSIs matching two different genera. However, since only one genome out of >700 had an ambiguous result, it is not significant. Due to the ambiguous nature of results from these two genomes, their results were not further considered.

Fig. 3. User interface and the results displayed by the server. (a) The main (starting) user interface of the serve providing a brief introduction to the server; (b) User interface obtained upon clicking "Submit a Query" on the main page. The sequence file to be analysed can either be browsed and selected or dragged and dropped in the provided box. (c) Results displayed by the server for the genome sequence of *Bacillus* sp. OV166, which was identified by the server as affiliated to the genus *Neobacillus*. Total number of *Neobacillus*-specific CSIs identified in this genome is also shown. (d) Results displayed by the server for the *Bacillus* sp. OxB-1. This strain was identified as belonging to the genus *Sporosarcina*, for which both genus-specific as well as family-specific CSIs were identified in the genome. As shown for one of the CSIs, sequence alignments for all identified CSIs are also displayed by the server.

Table 1. Information for *Bacillus* species genomes (chromosomal, complete and contig categories) predicted to belong to specific genera based on shared presence of taxon-specific CSIs

Bacillus **strain**	**Accession no.**	**Identification (No. of CSIs)**	***Bacillus*** **strain**	**Accession no.**	**Identification (No. of CSIs)**
7520-S	GCA_002271655.1	'*Alkalihalobacillaceae*' (10)	CFBP13597	GCA_014839555.1	*Peribacillus* (3)
A134	GCA_009746525.1	'*Alkalihalobacillaceae*' (8)	M6-12	GCA_002860285.1	*Peribacillus* (3)
JCM 19034	GCA_001310635.1	'*Alkalihalobacillaceae*' (6)	mrc49	GCA_002803435.1	*Peribacillus* (3)
JCM 19035	GCA_001310595.1	'*Alkalihalobacillaceae*' (6)	MUM 13	GCA_001866725.1	*Peribacillus* (3)
JCM 19041	GCA_001310375.1	'*Alkalihalobacillaceae*' (5)	RHFB	GCA_016757415.1	*Peribacillus* (3)
JCM 19045	GCA_000576325.1	'*Alkalihalobacillaceae*' (7)	SD075	GCA_017355165.1	*Peribacillus* (3)
JCM 19046	GCA_000576345.1	'*Alkalihalobacillaceae*' (4)	V5-8f	GCA_002863565.1	*Peribacillus* (3)
JCM 19059	GCA_001310655.1	'*Alkalihalobacillaceae*' (7)	AP8	GCA_000321185.1	*Peribacillus* (3)
Marseille-P3800	GCA_900197585.1	'*Alkalihalobacillaceae*' (9)	B4EP4a	GCA_006680105.1	*Peribacillus* (3)
P16(2019)	GCA_007293315.1	'*Alkalihalobacillaceae*' (9)	BA3	GCA_002835675.1	*Peribacillus* (3)
TS-2	GCA_000586495.1	'*Alkalihalobacillaceae*' (9)	ALD	GCA_003400165.1	*Priestia* (2)
YZJH907-2	GCA_017939705.1	'*Alkalihalobacillaceae*' (10)	Aph1	GCA_000409525.1	*Priestia* (2)
FJAT-22090	GCA_001278755.1	*Caryophanaceae* (11)	CMAA 1363	GCA_001757375.1	*Priestia* (2)
N3536	GCA_009676825.1	*Caryophanaceae* (10)	IHB B 7164	GCA_001648135.1	*Priestia* (2)
M4U3P1	GCA_006965545.2	*Alkalicoccus* (4)	MB95	GCA_011319735.1	*Priestia* (2)
FJAT-45348	GCA_002813025.1	*Alteribacter* (5)	ME40	GCA_903970935.1	*Priestia* (2)
KQ-3	GCA_003710255.1	*Alteribacter* (5)	ME75	GCA_903970885.1	*Priestia* (2)
YIM B00319	GCA_016745835.1	*Caldalkalibacillus* (4)	RC	GCA_003400185.1	*Priestia* (2)
DFI.2.34	GCA_020564485.1	*Caldibacillus* (8)	REN51N	GCA_000815345.1	*Priestia* (2)
KG3	GCA_013393595.1	*Caldibacillus* (8)	RP1137	GCA_000500145.1	*Priestia* (2)
7894–2	GCA_002272225.1	*Cytobacillus* (3)	S34	GCA_016624505.1	*Priestia* (2)
NTK034	GCA_017303275.1	*Cytobacillus* (3)	S35	GCA_016624555.1	*Priestia* (2)
ZZV12-4809	GCA_009849585.1	*Cytobacillus* (3)	VT 712	GCA_001614195.1	*Priestia* (2)
HNG	GCA_003400205.1	*Ferdinandcohnia* (8)	Y-01	GCA_003047225.1	*Priestia* (2)
AFS002410	GCA_002556365.1	*Gottfriedia* (14)	7504–2	GCA_002272205.1	*Robertmurraya* (3)
EAC	GCA_002156875.1	*Gottfriedia* (14)	Y1	GCA_014489805.1	*Robertmurraya* (3)
Gen3	GCA_013421425.1	*Heyndrickxia* (4)	Y1	GCA_003586445.1	*Robertmurraya* (3)
IITD106	GCA_019459225.1	*Lederbergia* (3)	BHET2	GCA_005938125.1	*Rossellomorea* (3)
J14TS2	GCA_018333015.1	*Lederbergia* (4)	CH30_1T	GCA_008364765.1	*Rossellomorea* (3)
SD088	GCA_017355205.1	*Lederbergia* (4)	es.034	GCA_002563655.1	*Rossellomorea* (3)
PS06	GCA_014837155.1	*Litchfieldia* (5)	JRC01	GCA_019192665.1	*Rossellomorea* (3)
SAJ1	GCA_003971115.1	*Margalitia* (2)	KH172YL63	GCA_011398925.1	*Rossellomorea* (3)
THG-D6.12	GCA_009766225.1	*Margalitia* (2)	Leaf406	GCA_001426105.1	*Rossellomorea* (3)
SBJS01	GCA_014856545.1	*Mesobacillus* (3)	Marseille-Q1617	GCA_903645295.1	*Rossellomorea* (3)
T33-2	GCA_002860205.1	*Mesobacillus* (3)	MKU004	GCA_001655735.1	*Rossellomorea* (3)

Continued

Table 1. Continued

Bacillus **strain**	**Accession no.**	**Identification (No. of CSIs)**	***Bacillus*** **strain**	**Accession no.**	**Identification (No. of CSIs)**
7586 K	GCA_002271785.1	*Metabacillus* (6)	NTK074B	GCA_017303375.1	*Rossellomorea* (3)
SA1-12	GCA_000981385.1	*Metabacillus* (6)	RO3	GCA_012911895.1	*Rossellomorea* (3)
7884–1	GCA_002272245.1	*Neobacillus* (9)	V3	GCA_015999585.1	*Rossellomorea* (3)
B-jedd	GCA_000821085.1	*Neobacillus* (5)	CHD6a	GCA_001293645.1	*Sutcliffiella* (8)
C11	GCA_011393025.1	*Neobacillus* (5)	DG-18	GCA_020037475.1	*Sutcliffiella* (8)
FJAT-18017	GCA_001278805.1	*Neobacillus* (5)	m3-13	GCA_000175075.1	*Sutcliffiella* (9)
FJAT-27225	GCA_001685025.1	*Neobacillus* (4)	RO1	GCA_012524115.1	*Sutcliffiella* (9)
MM2020_1	GCA_011250595.1	*Neobacillus* (11)	RO2	GCA_012911845.1	*Sutcliffiella* (9)
MRMR6	GCA_001940785.1	*Neobacillus* (8)	THAF10	GCA_009363695.1	*Sutcliffiella* (9)
MUM 116	GCA_001866655.1	*Neobacillus* (8)	JC-4	GCA_008974185.1	*Weizmannia* (2)
OV166	GCA_900177675.1	*Neobacillus* (10)	JC-7	GCA_008974205.1	*Weizmannia* (2)
S3	GCA_005154805.1	*Neobacillus* (11)	KG1	GCA_013393645.1	*Weizmannia* (2)
sid0103	GCA_019219025.1	*Neobacillus* (10)	PP-18	GCA_008974225.1	*Weizmannia* (2)
X1	GCA_000747345.1	*Neobacillus* (11)	Cs-700	GCA_011082085.1	*Pseudalkalibacillus* Clade-2 (6)
RGIG2558	GCA_017431595.1	*Niallia* (2)	es.036	GCA_002563635.1	*Pseudalkalibacillus* Clade-2 (6)
UniB3	GCA_012933555.1	*Niallia* (2)	N1-1	GCA_009818105.1	*Pseudalkalibacillus* Clade-2 (6)
AS04akNAM_105	GCA_012842745.1	*Schinkia* (11)	NTK071	GCA_017303315.1	*Pseudalkalibacillus* Clade-2 (7)
VT-16–64	GCA_001989355.1	*Siminovitchia* (4)	RAR_GA_16	GCA_020165985.1	*Pseudalkalibacillus* Clade-2 (6)
OxB-1	GCA_000829195.1	*Sporosarcina* (17)			

The reliability of the taxonomic predictions made by the server for the analysed genomes was examined by reconstructing a phylogenomic tree based on genome sequence of different *Bacillus* strains which were assigned to specific genera along with genome sequences of the type species (and generally one additional species) from these genera. For the *Bacillus* Subtilis and *Bacillus* Cereus clades, sequences for only a limited number of strains, chosen at random, were included. The resulting tree, which is based on concatenated sequences of 87 conserved proteins (see Methods), is shown in Fig. 4. The clades corresponding to different *Bacillaceae* genera (type species identified with T) and the different unnamed *Bacillus* strains grouping with them are marked in the tree. In this tree, the clades corresponding to different *Bacillaceae* genera are clearly distinguished, and they are generally separated from each other by long branches (Fig. 4). Upon analysing the strains which are grouping with different *Bacillaceae* genera, it was observed that there is 100% concordance between the branching of strains with specific genera and their predicted generic assignment by the server. The observed results provide strong evidence that the predictions made by the AppIndels server regarding taxonomic assignment of different strains are correct and that it provides a useful tool for the assignment of genome-sequenced species/strains to specific genera for which CSIs are present in its database.

In addition to the above genomes, we also analysed 327 additional genome sequences for *Bacillus* species, which consisted of scaffold-level assembly. Results from the server regarding the taxonomic assignment of these genomes/strains are presented in Table 2 and Tables S4–S6. Of these strains, positive assignments were made for 136 strains to the following 19 genera/families (Table 2): *Alteribacter* (1), '*Alkalihalobacillaceae*' (4), *Cytobacillus* (10), *Ferdinandcohnia* (1), *Gottfriedia* (12), *Lederbergia* (6), *Margalitia* (1), *Mesobacillus* (11), *Metabacillus* (3), *Neobacillus* (26), *Niallia* (6), *Peribacillus* (22), *Priestia* (15), *Rossellomorea* (7), *Schinkia* (1), *Siminovitchia* (6), *Sutcliffiella* (1), *Weizmannia* (1) and family *Caryophanaceae* (2). Moreover, 61 strains corresponded to the *Bacillus* Subtilis clade (Table S4) and 103 strains were predicted to be affiliated with the *Bacillus* Cereus clade (Table S5). For the remaining 27 strains from this group (Table S6), no positive assignment was made by the server. A phylogenetic tree was also reconstructed based on genome sequences of *Bacillus* strains from this group which were assigned to specific genera. In

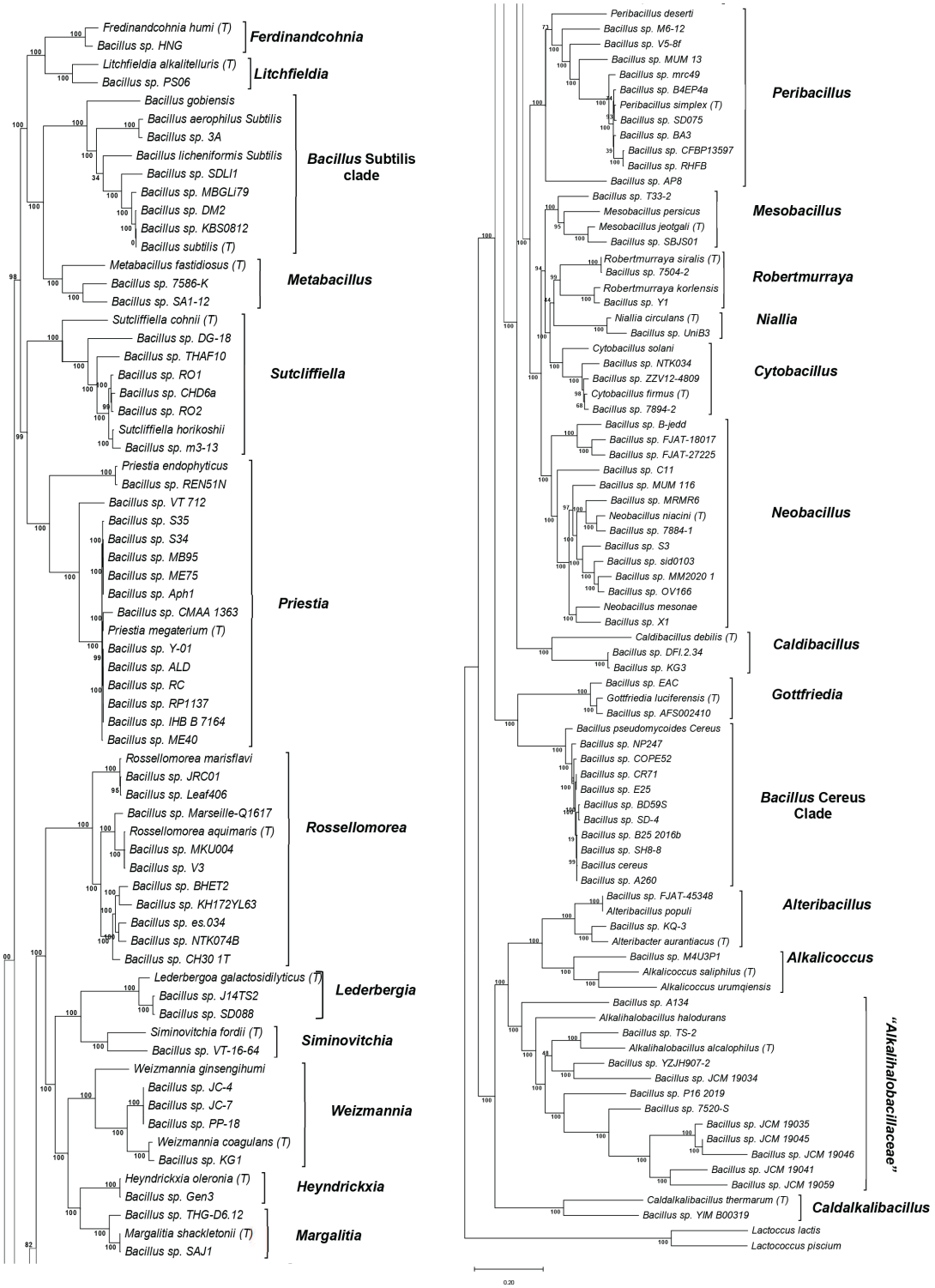

Fig. 4. Phylogenetic tree based on genome sequences for the type species of different *Bacillaceae* genera and genomes of different *Bacillus* species for which positive identification to specific genera was made by the server (Table 1). The results shown here for the initial sets of 394 genomes with chromosome, complete, and contig-level assembly. Of the *Bacillus* strains assigned to *Bacillus* Subtilis and *Bacillus* Cereus clades, only a limited number of strains were included in the analysis. The clades corresponding to different *Bacillaceae* genera and the *Bacillus* strains grouping with them are marked in the tree. The grouping of different *Bacillus* species (strains) with the indicated genera showed perfect correlation with the predicted taxon-assignment by the server. The results for the strains branching with '*Pseudalkalibacilllus*' and *Caryophanaceae* are presented in another figure.

Table 2. Information for *Bacillus* species (scaffold genomes) predicted to belong to specific genera based on shared presence of taxon-specific CSIs

Bacillus **strain**	**Accession no.**	**Identification (No. of CSIs)**	***Bacillus*** **strain**	**Accession no.**	**Identification (No. of CSIs)**
FJAT-45350	GCA_002335805.1	'*Alkalihalobacillaceae*' (6)	LOB377	GCA_014656545.1	*Neobacillus* (11)
FJAT-45037	GCA_002797325.1	'*Alkalihalobacillaceae*' (10)	AFS076308	GCA_002565135.1	*Neobacillus* (10)
C1-1	GCA_003856305.1	'*Alkalihalobacillaceae*' (10)	AFS037270	GCA_002585305.1	*Neobacillus* (10)
BO	GCA_900188535.1	'*Alkalihalobacillaceae*' (10)	ISL-7	GCA_018613065.1	*Neobacillus* (10)
DE0243	GCA_007677425.1	Caryophanaceae (10)	ISL-46	GCA_018613205.1	*Neobacillus* (8)
DE0292	GCA_007676755.1	*Caryophanaceae* (10)	ISL-40	GCA_018613245.1	*Neobacillus* (10)
H-16	GCA_016901055.1	*Alteribacter* (5)	ISL-77	GCA_018613105.1	*Neobacillus* (8)
FJAT-49705	GCA_018343665.1	*Cytobacillus* (2)	MM2020_4	GCA_011250555.1	*Neobacillus* (11)
FJAT-29937	GCA_001509555.1	*Cytobacillus* (2)	ISL-75	GCA_018613095.1	*Neobacillus* (9)
FJAT-21945	GCA_001275655.1	*Cytobacillus* (3)	DE0587	GCA_007665625.1	*Niallia* (2)
S/N-304-OC-R1	GCA_019749375.1	*Cytobacillus* (2)	B1-b2	GCA_009183415.1	*Niallia* (2)
2_A_57_CT2	GCA_000186145.1	*Cytobacillus* (3)	DE0237	GCA_007677495.1	*Niallia* (2)
22–7	GCA_010628845.1	*Cytobacillus* (3)	34–1	GCA_010628745.1	*Niallia* (2)
ISL-47	GCA_018613175.1	*Cytobacillus* (3)	MB2021	GCA_000701605.1	*Niallia* (2)
FJAT-29790	GCA_019039215.1	*Cytobacillus* (2)	522_BSPC	GCA_001076885.1	*Niallia* (2)
J33	GCA_000518885.1	*Cytobacillus* (3)	FJAT-20673	GCA_001636485.1	*Peribacillus* (3)
CRN 9	GCA_013267855.1	*Cytobacillus* (3)	FJAT-21352	GCA_001277355.1	*Peribacillus* (3)
REN16	GCA_020731355.1	*Ferdinandcohnia* (8)	FJAT-22058	GCA_001277335.1	*Peribacillus* (3)
FJAT-25509	GCA_001420605.1	*Gottfriedia* (14)	FJAT-28573	GCA_001541085.1	*Peribacillus* (3)
RG28	GCA_017814275.1	*Gottfriedia* (9)	OV322	GCA_900112495.1	*Peribacillus* (3)
UNCCL81	GCA_900112535.1	*Gottfriedia* (14)	ISL-57	GCA_018613135.1	*Peribacillus* (3)
AFS088145	GCA_002564465.1	*Gottfriedia* (14)	AFS017274	GCA_002561875.1	*Peribacillus* (3)
AFS096315	GCA_002552175.1	*Gottfriedia* (14)	Soil745	GCA_001429835.1	*Peribacillus* (3)
AFS077874	GCA_002566505.1	*Gottfriedia* (14)	ISL-78	GCA_018613455.1	*Peribacillus* (3)
AFS053548	GCA_002575105.1	*Gottfriedia* (14)	RJGP41	GCA_002998155.1	*Peribacillus* (3)
AFS017336	GCA_002555445.1	*Gottfriedia* (14)	ISL-101	GCA_018613475.1	*Peribacillus* (3)
AFS029533	GCA_002585125.1	*Gottfriedia* (14)	ISL-34	GCA_018613325.1	*Peribacillus* (3)
AFS055030	GCA_002574545.1	*Gottfriedia* (13)	Leaf13	GCA_001426005.1	*Peribacillus* (3)
AFS041924	GCA_002577195.1	*Gottfriedia* (13)	ISL-4	GCA_018613275.1	*Peribacillus* (3)
AFS001701	GCA_002559405.1	*Gottfriedia* (14)	Soil768D1	GCA_001429855.1	*Peribacillus* (3)
FJAT-49731	GCA_018343555.1	*Lederbergia* (3)	AFS094228	GCA_002553175.1	*Peribacillus* (3)
FJAT-49754	GCA_018343525.1	*Lederbergia* (3)	OK838	GCA_900188495.1	*Peribacillus* (3)
FJAT-49732	GCA_018343695.1	*Lederbergia* (3)	AFS043905	GCA_002581115.1	*Peribacillus* (3)
FJAT-49682	GCA_018343635.1	*Lederbergia* (3)	AFS026049	GCA_002557105.1	*Peribacillus* (3)
FJAT-49711	GCA_018343715.1	*Lederbergia* (3)	ISL-53	GCA_018613485.1	*Peribacillus* (3)
FJAT-49870	GCA_018343625.1	*Lederbergia* (3)	445_BSPC	GCA_001077295.1	*Peribacillus* (3)
FJAT-49736	GCA_018343615.1	*Margalitia* (2)	220_BSPC	GCA_001054655.1	*Peribacillus* (3)

Continued

Table 2. Continued

Bacillus strain	Accession no.	Identification (No. of CSIs)	*Bacillus* strain	Accession no.	Identification (No. of CSIs)
B15-48	GCA_016879755.1	*Mesobacillus* (3)	FJAT-21351	GCA_001275665.1	*Priestia* (2)
ISL-45	GCA_018613195.1	*Mesobacillus* (3)	FJAT-26377	GCA_019039155.1	*Priestia* (2)
ISL-35	GCA_018613355.1	*Mesobacillus* (3)	171095_106	GCA_000514175.1	*Priestia* (2)
REN3	GCA_017912555.1	*Mesobacillus* (3)	AG236	GCA_003337615.1	*Priestia* (2)
ISL-39	GCA_018613295.1	*Mesobacillus* (3)	CBEL-1	GCA_004348885.1	*Priestia* (2)
ISL-41	GCA_018613235.1	*Mesobacillus* (3)	Root147	GCA_001429365.1	*Priestia* (2)
FJAT-27251	GCA_001273925.1	*Mesobacillus* (3)	278922_107	GCA_000514135.1	*Priestia* (2)
ISL-37	GCA_018613315.1	*Mesobacillus* (3)	T_4	GCA_020073875.1	*Priestia* (2)
ISL-55	GCA_018613155.1	*Mesobacillus* (2)	FJAT-14578	GCA_000504165.1	*Priestia* (2)
17 376	GCA_000498695.1	*Mesobacillus* (3)	Root239	GCA_001429385.1	*Priestia* (2)
UBA9944	GCA_003457145.1	*Mesobacillus* (3)	BK006	GCA_004345465.1	*Priestia* (2)
UMB0899	GCA_002871465.1	*Metabacillus* (6)	GbtcB12	GCA_019091355.1	*Priestia* (2)
J37	GCA_000518865.1	*Metabacillus* (6)	Leaf75	GCA_001426025.1	*Priestia* (2)
AFS040349	GCA_002577655.1	*Metabacillus* (6)	Soil531	GCA_001429825.1	*Priestia* (2)
FJAT-49825	GCA_018343535.1	*Neobacillus* (11)	CGMCC 1.16541	GCA_003184905.1	*Priestia* (2)
FJAT-50051	GCA_018343545.1	*Neobacillus* (10)	MCCB 382	GCA_019334305.1	*Rossellomorea* (2)
FJAT-27245	GCA_001273955.1	*Neobacillus* (6)	AFS015802	GCA_002561455.1	*Rossellomorea* (3)
OK048	GCA_900103525.1	*Neobacillus* (11)	DSM 27956	GCA_002237795.1	*Rossellomorea* (3)
UNC41MFS5	GCA_000686805.1	*Neobacillus* (10)	V-88	GCA_003001735.1	*Rossellomorea* (3)
DE0180	GCA_007678255.1	*Neobacillus* (10)	V-88	GCA_900168275.1	*Rossellomorea* (3)
EB106-08-02-XG196	GCA_013396335.1	*Neobacillus* (10)	349Y	GCA_902506085.1	*Rossellomorea* (3)
ISL-18	GCA_018613415.1	*Neobacillus* (11)	ES3	GCA_900106505.1	*Rossellomorea* (2)
AFS031507	GCA_002584635.1	*Neobacillus* (10)	Marseille-P3661	GCA_900240995.1	*Schinkia* (9)
EB01	GCA_000613125.1	*Neobacillus* (5)	Sa1BUA2	GCA_014836955.1	*Siminovitchia* (7)
UNC438CL73TsuS30	GCA_000482325.1	*Neobacillus* (8)	DE0466	GCA_007667595.1	*Siminovitchia* (7)
FJAT-29814	GCA_001510715.1	*Neobacillus* (11)	DE0460	GCA_007667705.1	*Siminovitchia* (7)
AFS073361	GCA_002567005.1	*Neobacillus* (11)	DE0344	GCA_007673645.1	*Siminovitchia* (7)
AFS006103	GCA_002559145.1	*Neobacillus* (11)	HF117_J1_D	GCA_009928435.1	*Siminovitchia* (6)
FJAT-29953	GCA_019039195.1	*Neobacillus* (11)	UBA11286	GCA_003521085.1	*Siminovitchia* (7)
JCA	GCA_000820865.2	*Neobacillus* (11)	FJAT-45066	GCA_002335755.1	*Sutcliffiella* (7)
USDA818B3_A	GCA_009928415.1	*Neobacillus* (10)	SIG189	GCA_015059665.1	*Weizmannia* (2)

this tree (Fig. 5), like the results presented in Fig. 4, the strains assigned to different *Bacillaceae* genera grouped reliably with the type species of these genera, thus providing further evidence regarding the reliability of generic assignment made by the server.

Sequence information for the CSIs specific for *Pseudalkalibacilllus* species [91] and the family *Caryophanaceae* were added to the server's database during revision. Hence, a separate phylogenetic tree was reconstructed for the strains branching with these taxa (Fig. 6). In this tree, 5 *Bacillus* species, which the server predicted as showing affiliation to the family *Caryophanaceae*, branched reliably with (or within) the clade for this family. The CSIs specific for the *Pseudalkalibacilllus*, which are shared by most of the species from this genus, were also shared by five uncharacterized *Bacillus* strains (Table 1). Interestingly, in the phylogenetic tree

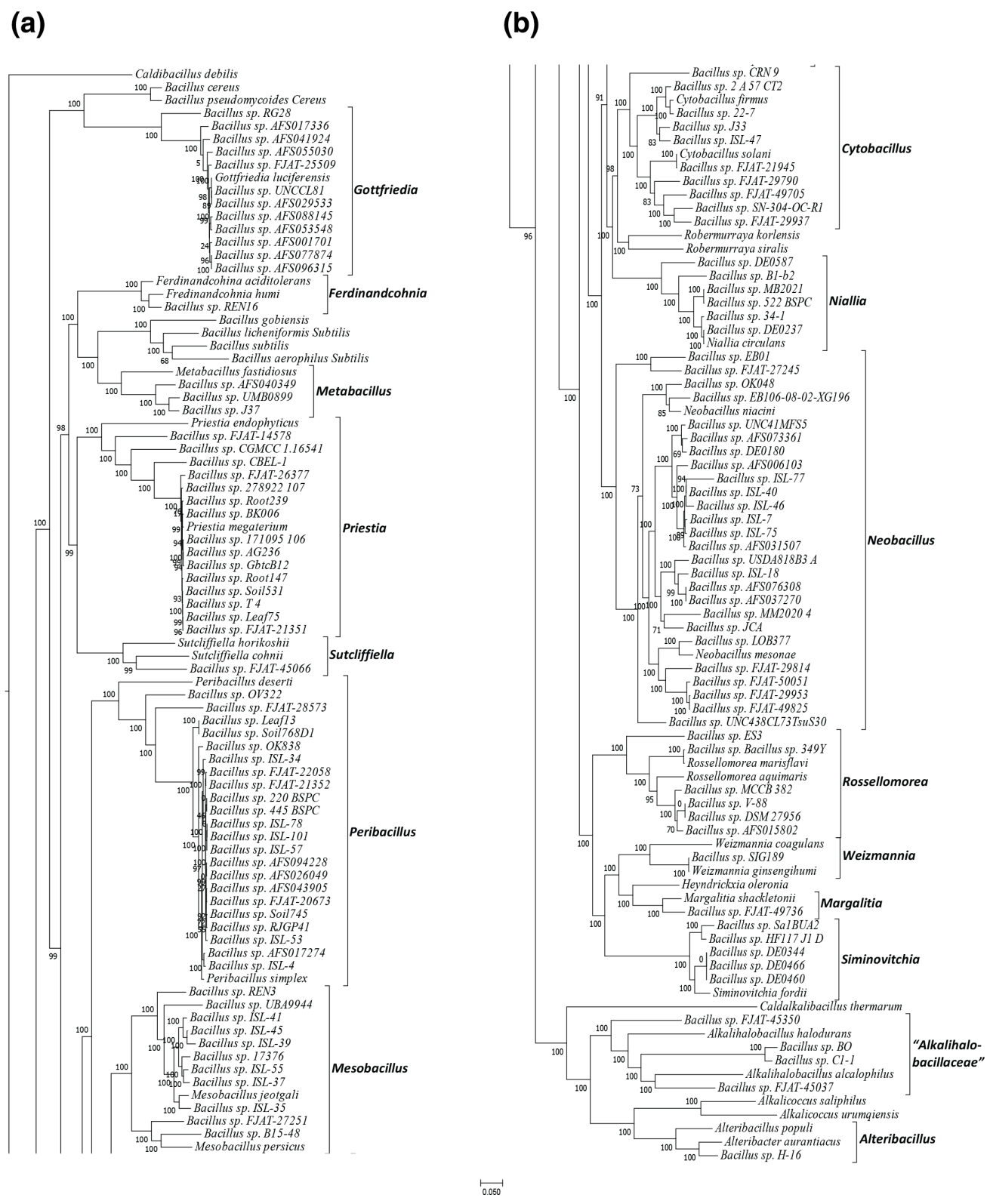

Fig. 5. Phylogenetic tree based on genome sequences for the type species of different *Bacillaceae* genera and the genomes of different *Bacillus* species (strains), with scaffold-level assembly for which positive identification was made by the server to specific genera. The *Bacillus* strains from this group for which taxonomic predictions were made by the server also reliably branched with the indicated genera.

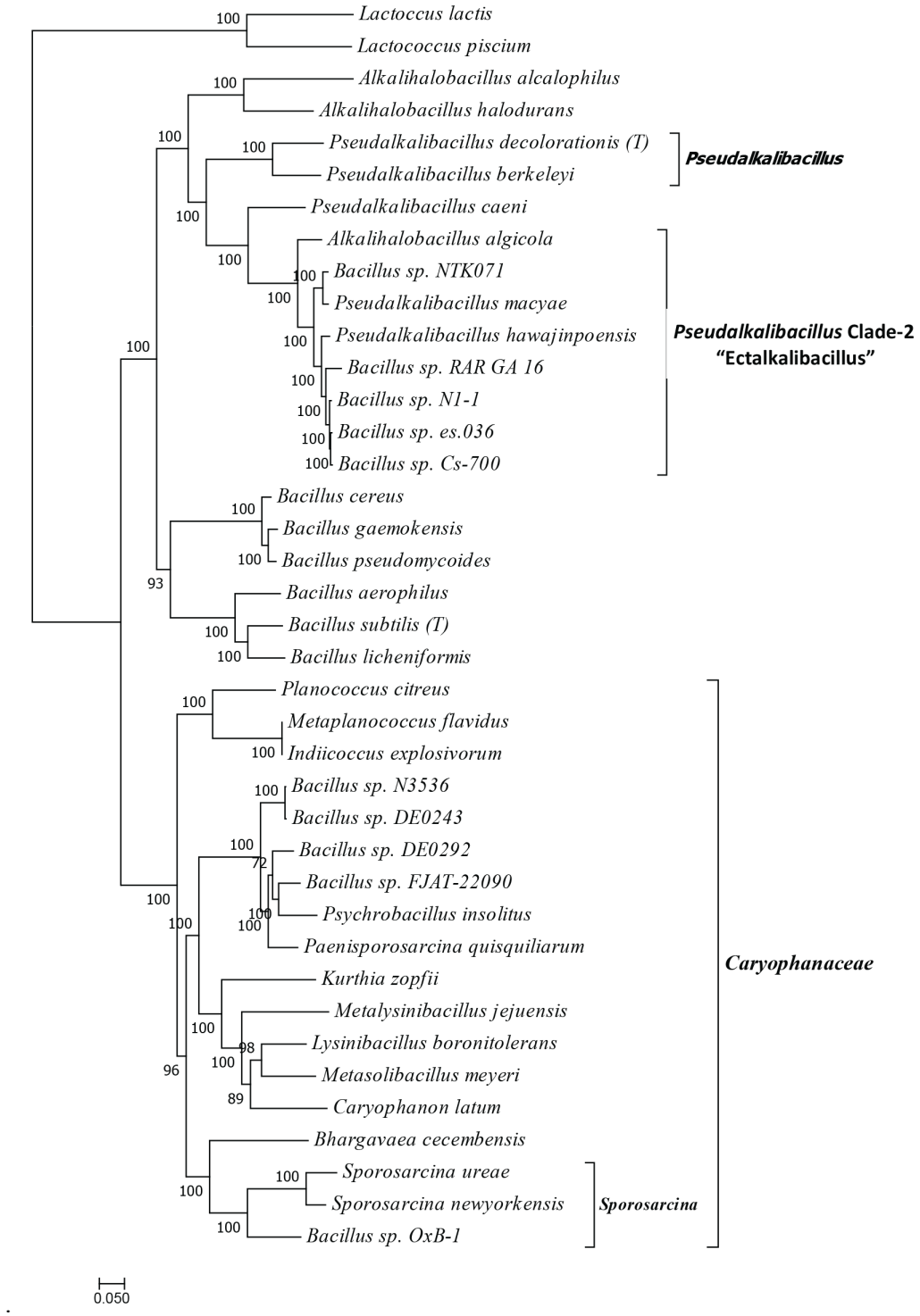

Fig. 6. Phylogenetic tree based on genome sequences for the genus *Pseudalkalibacillus* and representative *Caryophanaceae* species along with the sequences of *Bacillus* species (strains) which were assigned to these taxa. The *Bacillus* strain assigned to these taxa grouped reliably with them. However, the CSIs specific the genus *Pseudalkalibacillus* in our database are specific for a clade marked as *Pseudalkalibacillu*s Clade-2 ('*Ectalkalibacillus*'), which does not include the type species of the genus *Pseudalkalibacillus* (*P. decolorationis*). All *Bacillus* species strains assigned to the *Pseudalkalibacillu*s clade-2, likely constitutes a new genus (tentatively labelled as '*Ectalkalibacillus*'), which will be proposed in future work.

shown in Fig. 6, species from the genus *Pseudalkalibacilllus* [91] formed two distinct clades. All identified CSIs for this genus were specific for the larger of these two clades, which we have labelled as *Pseudalkalibacillus* Clade-2 ('*Ectalkalibacilllus*') in Fig. 6. This clade encompasses most of the species from *Pseudalkalibacilllus* as well as all five uncharacterized *Bacillus* strains which are predicted to be part of this genus. However, *P. decolorationis*, which is type species of *Pseudalkalibacilllus* [91, 92], and another species *P. berkeleyi* formed a separate deeper-branching clade, which is distinct from the '*Ectalkalibacilllus*' clade (Fig. 6). These two species also did not share the CSIs specific for the '*Ectalkalibacilllus*' clade. Hence, it is likely that the species from '*Ectalkalibacilllus*' clade constitute a new genus (to be described in future work after further characterization), which is distinct from the genus *Pseudalkalibacillus* [91].

DISCUSSION

We describe here the development and validation of a web-based tool that uses sequence information for validated CSIs specific for known prokaryotic taxa for determining the presence of these molecular characteristics in any input genome sequence. If the server identifies that the input genome sequence contains significant numbers of CSIs matching a specific taxon, it predicts that the analysed genome (strain/species) is affiliated to that taxon. The core of this web-based tool (server) is a proprietary database of well-characterized and validated CSIs that are specific for known prokaryotic taxa [7, 48, 49]. Due to their group (taxon)-specificity and predictive ability to be found in other members of the indicated group, the identified CSIs have provided important means for the demarcation of prokaryotic taxa of different ranks in molecular terms, thus aiding in the development of more robust classification schemes for prokaryotic organisms [38, 39, 47, 48, 50, 52, 53, 58]. The CSIs have now been identified for numerous prokaryotic taxa [38, 39, 47, 48, 50, 52, 53, 58, 63, 93, 94] and several investigators have used them for the classification of newly described species into specific taxa [54–56, 62, 67–74]. However, the lack of a convenient method for determining the presence or absence of known CSIs in genome sequences has limited the use of these molecular markers for taxonomic and diagnostic studies.

The AppIndels server fills this gap by providing a simple to use method for determining the presence of previously identified CSIs in a genome sequence and using this information to predict taxonomic affiliation. The server has a simple user interface requiring no expertise to use or interpret the results it provides. In this work, we have demonstrated the utility of this server by analysing the genome sequences of 721 uncharacterized *Bacillus* strains of unknown taxonomic affiliation. The assembly levels of the analysed genomes varied from chromosome to scaffolds and it presented no problem in their analysis by the server. The analyses of these genomes by the AppIndels server showed that 651 of these genomes contained significant numbers of CSIs specific for the 29 *Bacillales* genera/families (Tables 1, 2 and S1–S6). The validity of the taxon assignment/affiliation predicted by the server was examined by the construction of phylogenomic trees. Results from these studies showed that all *Bacillus* strains for which taxonomic predictions were made by the server correctly branched (100% correlation) with the clades corresponding to the predicted genera. Results reported here for the *Bacillus* strains with some additional analyses should be helpful in the classification of some of these strains as novel species in the indicated genera [52, 58].

In our analyses, while >90% of the analysed strains were predicted to be affiliated to specific *Bacillota* genera, for the remaining ≈10% strains, no taxonomic prediction was made by the server. This is not surprising, as the server can only predict taxonomic affiliation to those genera/taxa for which CSIs have been identified for and added to its database. The recent reclassification of *Bacillus* species [52, 58] creatied a number of different genera to which most of the uncharacterized *Bacillus* species are affiliated. However, several *Bacillus* species were not assigned to any genera due to lack of reliable information [52, 58]. These species were placed into the category *Bacillaceae* Incertae Sedis and consequently they have not yet been evaluated for the presence of taxon-specific CSIs [58]. Additionally, no CSIs have been identified for large numbers of other *Bacillaceae* genera [66, 83, 95, 96]. Hence, if the input genome sequence is related to these *Bacillus* species or genera for which no CSIs have been identified, the server will not be able to make any taxonomic prediction. For example, for the five *Bacillus* strains which showed affiliation to the family *Caryophanaceae* [66], the server would not have made any taxonomic prediction if the CSIs for this family were not present in the database.

While the AppIndels server provides a useful tool for predicting the taxonomic affiliation of a genome sequence, there are several important caveats/considerations in using this server for taxonomic inferences or other studies. An important limitation of this server is that it can only make predictions for those taxa for which CSIs are present within the database. Thus, as clearly acknowledged in the title of this paper, this server is not a generalized tool for taxonomic predictions and its utility is limited to only those taxa for which CSIs have been identified and included within the database of AppIndels.com. The server's database at present contains only 585 CSIs covering a limited number of prokaryotic taxa (see Methods). In addition to these CSIs, CSIs specific for several other prokaryotic taxa have been described in earlier work [19, 39, 50, 51, 59, 63, 97–104] and information for them will be added to the server once validation studies have been conducted on them. Information regarding the taxa for which CSIs are present within the AppIndels database, is provided on the server and it can be searched using the provided search box. This information will be updated automatically as new CSIs are added to the server's database. To increase the utility of this server, it is important to identify CSIs for other prokaryotic taxa. However, the identification of CSIs specific for different taxa requires considerable work, as described in an earlier publication [48]. Thus, the progress in

this regard is expected to be slow but it could be expedited by the broader involvement of the scientific community in the identification of CSIs. Guidelines for the submission of well-characterized CSIs for their inclusion to our CSI database will be provided on the website in due course, once the details regarding the reviewing and validation of submitted CSIs have been developed.

As the AppIndels server contains CSIs for a limited number of prokaryotic taxa, when interpreting the results from this server, a positive result showing the presence of significant numbers of CSIs specific for a given group or taxon is very meaningful (significant). In contrast, a negative result is generally not informative or conclusive, and the primary reason for this is the absence of CSIs for the taxon to which the submitted genome may be affiliated. Additionally, if the submitted genome sequence is incomplete, this can also result in a negative result. As elaborated in the Results section, different CSIs in our database have a weight value of between 0.15 and 0.5, and this weight value depends upon the total number of CSIs that are known for a taxon. A positive identification to a specific taxon requires that the total weight of matching taxon-specific CSIs is 1.0 or higher. For most of the taxa in our database four or more specific CSIs are present and there is generally no problem in meeting or exceeding the threshold for positive identification. However, for some taxa, for which only 2–3 CSIs are present in the database, each with a weight value of between 0.4–0.5, matches to all or minimally two of the CSIs must be observed for a positive identification. Consequently, if the gene encoding for one or more of these CSIs is missing from a genome, then the total weight of the matching CSIs will be less than 1.0 and the server will not make a positive identification. Moreover, as mentioned in the Results, while all CSIs in our database were adjudged to be specific for the indicated taxa by the criteria used in our published work [52, 58, 64, 66, 75, 76], there is limited information available at present regarding the long-term stability and specificity of the CSIs within the dynamic framework of prokaryotic taxonomy. Thus, it is possible that the uniqueness and taxon-specificity of some of these CSIs may change over time as sequence information for divergent cultured and uncultured organisms becomes available in public databases. These developments have the potential of diminishing the specificity and consequently predictive ability of some of the identified CSIs for members of the indicated taxa. The impact of such changes in general is expected to be minimal for taxa where multiple specific CSIs have been identified, and where an individual CSI has a low weight (*viz*. <0.3) in the overall threshold for positive identification. However, such changes can have greater impact on the predictive ability of CSIs with higher weight (e.g. 0.4 or higher), thus increasing the possibility of incorrect predictions by the server. To mitigate this possibility, specificity checks will be performed periodically on the CSIs in our database, and the sequence information for any uncertain CSIs will be removed from the database. However, to prevent the possibility of any false-positive or negative results, it is recommended that the predictions from this server be used in conjunction with phylogenetic analyses and other methods, especially when used for taxonomic inferences [4, 6, 105]. In cases where a given species/strain, based on phylogenetic and other analyses, shows reliable association with a specific taxon, but is not predicted by the server, the presence of taxon-specific CSIs in the query genome sequence should be determined by means of BLASTP searches. In such cases, even if a limited number (even one) of taxon-specific CSIs are present in the analysed genome, these results should be considered informative and significant.

The results from this server will also be impacted by the changes in the classification of species for which CSIs are present in its database. Microbial taxonomy is a dynamic discipline and new species related to different taxa are continually being described [3, 8, 25, 83]. The addition of new species to existing taxa, as well as analyses of the interrelationships among given species by different methods, can lead to division of a previously described taxon into multiple taxa, and in some cases, the unification of two or more taxa into a single taxon [3, 8, 25, 83]. In such cases some of the CSIs present in the database, which were earlier specific for a particular taxon, will no longer be specific for the emended taxon bearing the same name. Instead, the originally described CSIs would likely be specific for a clade of either higher or lower phylogenetic depth. Two examples of the effects of species reclassification on the specificities of the CSIs are presented in this work. The genus *Alkalihalobacillus* was described in our work and 10 identified CSIs specific for this genus are present in our database [52]. However, in a later study [91] this genus was split into eight genera including the emended genus *Alkalihalobacillus* containing only a limited number of the original species. As a result of this reclassification, the CSIs present in our database for *Alkalihalobacillus* are no longer specific for only the species from emended genus *Alkalihalobacillus*, but they are now specific for a higher taxonomic clade, encompassing six different genera, which is referred to in our work/database as the family '*Alkalihalobacillaceae*'. In another example discussed here, the genus *Pseudalkalibacillus* described by Joshi *et al*. [91] is shown to comprise two distinct genus-level clades. However, multiple CSIs species for members of this genus that we have identified are specific for only the members of one of these two genus-level clades (Fig. 6). This clade (labelled as *Pseudalkalibacillus* Clade-2), which does not include the type species of the genus *Pseudalkalibacillus*, will be proposed as a new genus (tentatively labelled as '*Ectalkalibacilllus*') in future work. Thus, in using the results from this server, it is important to keep in mind that the changes in microbial classification occur frequently, and some of them may not be reflected in the indicated group-specificity of the described CSIs and in the taxonomic predictions made by the server. Additionally, as indicated in the Methods and Results, the quality of the submitted genome, which is not checked by the server, can also impact the predictions made by the server. Thus, it is recommended that all submitted genomes should be checked prior to their submission for their quality/contamination [84], and these results should be taken into consideration in interpreting the results from the server.

With the above provisos/caveats, the use of AppIndels server should lead to a more definitive demarcation of different prokaryotic genera/taxa. Based on these studies, the placement of species into a specific genus (or other taxonomic groupings) will not only be based on their monophyletic grouping in phylogenetic trees, but also based upon on shared presence of multiple molecular characteristics

(CSIs), which are generally uniquely found in the members of a specific group/taxon [7, 58]. The CSIs in genes/proteins sequences are present at different phylogenetic depths ranging from strains and species specific to those which are specific for higher taxonomic clades (e.g. genus, family, order, class and phylum-specific) [50, 78, 81, 98, 106, 107]. For several taxa (*viz*. families *Borreliaceae*, *Caryophanaceae* and *Leuconostocaceae*; order *Legionellales* [64] and *Chlorobiales* [77]), CSIs are present in the server's database at more than one taxonomic level [47, 50, 62, 78, 81]. In these cases, analysis of a genome should identify CSIs present at different phylogenetic depths providing strong internally consistent evidence supporting the placement of a species/strain into a specific taxon. For example, multiple CSIs are present in the server's database for the family *Borreliaceae* and its two genera, *Borrelia* and *Borreliella* [36, 53, 78, 81]. Based on these CSIs, the server can unambiguously assign any given species or strain related to this family into one of the two genera [36, 53]. Additionally, based on the availability of CSIs, the server is also capable of distinguishing at the species and strain levels. Recently, the available strains of *Pseudomonas aeruginosa* were divided into two separate species, *P. aeruginosa* and *P. paraeruginosa* [79]. These two closely related species cannot be easily distinguished based on the 16S rRNA or other single gene(s) trees [79]. However, multiple CSIs specific for these two species have been identified and they are present in the server database [79]. Based upon them, the server can reliably distinguish the strains belonging to the two species. In earlier work, some CSIs have also been described which are specific for the enterohemorrhagic strains of *Escherichia coli* (O157:H7) [106], or those that can distinguish between very closely related species such as *Bacillus anthracis* from *B. cereus* [107]. Thus, the analysis employed by this server, with the inclusion of CSIs specific for clinically important species/strains, can also serve as a useful diagnostic tool.

Lastly, it should be emphasized that the AppIndels server is distinct from other servers as it works by identifying molecular characteristics that are uniquely shared by the members of a specific taxon. There are several other servers available for inferring relatedness of a genome sequence to known taxa based on 16S rRNA similarity values [108], comparison of other readily measurable overall genomic relatedness indices (OGRIs) such as ANI, AAI and DDH [6, 109], as well as branching in phylogenetic trees based on conserved sets of proteins [4]. However, the results provided by these analyses generally represent a continuum from which the placement of any species into a specific taxon is based on empirically suggested/accepted thresholds [6, 13, 14, 16, 33, 105]. These analyses, however, do not identify, or provide information regarding any specific molecular or other characteristic that is commonly and uniquely shared by the members of a given group. In contrast, the AppIndels server identifies a taxon based on highly specific characteristics which are uniquely found in different members of that group/taxon. Although the evolutionary inferences based on the identified CSIs (which AppIndels server uses) regarding taxon assignment in most cases may be similar to the other OGRI-based or phylogenetic approaches, the overall genetic and evolutionary significance of the results provided by the AppIndels server is very different, and it complements the results obtained from other approaches. As indicated earlier, CSIs represent highly specific genetic/molecular changes (synapomorphies), which are exclusively shared by a given group of organisms [7, 28, 48]. As the genotype specifies the phenotype, it is expected that these genetic changes should manifest themselves as novel biochemical or phenotypic properties, which are specific for the indicated groups of organisms [7, 48]. Presently, very few, and in most cases no biochemical or phenotypic characteristics are known that are exclusively found in members of different groups of organisms (genera, family *etc*.). Hence, genetic, and biochemical studies on understanding the cellular functions of the described CSIs provide potential means for identification of novel characteristics that are unique properties of different groups of organisms. Additionally, as most of the identified CSIs are present in conserved regions of genes/proteins, their sequences also provide important means for the development of novel and more specific diagnostic methods for the identification of different organisms by either *in silico* analysis or different commonly used experimental methods [40, 106, 107, 110].

Server availability and data privacy

The use of Appindels.com server is freely available for research purposes to members of the scientific community upon registration. The genome sequence submitted to the server for analysis will not be used or stored for any purpose, and they will be deleted after the analysis is completed. However, the results from these analyses will be retained for a period of at least 30 days and they will be available to the specific user for review/analysis.

Funding information
This work was supported by the research grant (RGPIN-2019–06397) from the Natural Science and Engineering Research Council of Canada to Dr. Radhey S. Gupta and grant support from the Ontario Research Fund.

Acknowledgements
Special thanks and recognition are due to Joseph Manalo, who created an earlier version of this server (Indels.com). We also thank Bashudev Rudra and Sarah Bello for assistance in the formatting of CSI sequences and phylogenetic trees, and for helpful comments on the manuscript.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. Whitman WB. Genome sequences as the type material for taxonomic descriptions of prokaryotes. *Syst Appl Microbiol* 2015;38:217–222.
2. Chun J, Oren A, Ventosa A, Christensen H, Arahal DR, *et al*. Proposed minimal standards for the use of genome data

for the taxonomy of prokaryotes. *Int J Syst Evol Microbiol* 2018;68:461–466.

3. Hugenholtz P, Chuvochina M, Oren A, Parks DH, Soo RM. Prokaryotic taxonomy and nomenclature in the age of big sequence data. *ISME J* 2021;15:1879–1892.
4. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, *et al*. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* 2018;36:996–1004.
5. Chun J, Rainey FA. Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. *Int J Syst Evol Microbiol* 2014;64:316–324.
6. Meier-Kolthoff JP, Göker M. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. *Nat Commun* 2019;10:2182.
7. Gupta RS. Impact of genomics on the understanding of microbial evolution and classification: the importance of Darwin's views on classification. *FEMS Microbiol Rev* 2016;40:520–553.
8. Garrity GM. A new genomics-driven taxonomy of bacteria and archaea: are we there yet? *J Clin Microbiol* 2016;54:1956–1963.
9. Moore ERB, Mihaylova SA, Vandamme P, Krichevsky MI, Dijkshoorn L. Microbial systematics and taxonomy: relevance for a microbial commons. *Res Microbiol* 2010;161:430–438.
10. Sangal V, Goodfellow M, Jones AL, Schwalbe EC, Blom J, *et al*. Next-generation systematics: an innovative approach to resolve the structure of complex prokaryotic taxa. *Sci Rep* 2016;6:38392.
11. Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci* 2005;102:2567–2572.
12. Kim M, Oh HS, Park SC, Chun J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int J Syst Evol Microbiol* 2014;64:346–351.
13. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci* 2009;106:19126–19131.
14. Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, *et al*. Microbial species delineation using whole genome sequences. *Nucleic Acids Res* 2015;43:6761–6771.
15. Meier-Kolthoff JP, Auch AF, Klenk H-P, Göker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 2013;14:60.
16. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, *et al*. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* 2007;57:81–91.
17. Stackebrandt E, Frederiksen W, Garrity GM, Grimont PA, Kampfer P. Report of the *ad hoc* committee for the re-evaluation of the species definition in bacteriology. *Int J Syst Evol Microbiol* 2002;52:1043–1047.
18. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, *et al*. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol* 2014;12:635–645.
19. Gupta RS, Lo B, Son J. Phylogenomics and comparative genomic studies robustly support division of the genus *Mycobacterium* into an emended genus *Mycobacterium* and four novel genera. *Front Microbiol* 2018;9:67.
20. Nouioui I, Carro L, García-López M, Meier-Kolthoff JP, Woyke T, *et al*. Genome-based taxonomic classification of the Phylum *Actinobacteria*. *Front Microbiol* 2018;9:2007.
21. Hördt A, López MG, Meier-Kolthoff JP, Schleuning M, Weinhold L-M, *et al*. Analysis of 1,000+ type-strain genomes substantially improves taxonomic classification of *Alphaproteobacteria*. *Front Microbiol* 2020;11:468.
22. Waite DW, Vanwonterghem I, Rinke C, Parks DH, Zhang Y, *et al*. Comparative genomic analysis of the class *Epsilonproteobacteria* and proposed reclassification to *Epsilonbacteraeota* (phyl. nov.). *Front Microbiol* 2017;8:682.
23. Parker CT, Tindall BJ, Garrity GM. International Code of Nomenclature of Prokaryotes. *Int J Syst Evol Microbiol* 2019;69:S:1–S.
24. Goodfellow M. Microbial systematics: background and uses. In: Priest FG (eds). *Applied Microbial Systematics*. Dordrecht: Kluwer Academic Publishers; 2000. pp. 1–18.
25. Gupta RS. Microbial taxonomy: how and why name changes occur and their significance for (clinical) microbiology. *Clin Chem* 2021;68:134–137.
26. Oren A, Garrity GM. Then and now: a systematic review of the systematics of prokaryotes in the last 80 years. *Antonie van Leeuwenhoek* 2014;106:43–56.
27. Ludwig W, Klenk H-P. Overview: A phylogenetic backbone and taxonomic framework for prokaryotic systamatics. In: Brenner DJ, Krieg NR, Staley JT and Garrity GM (eds). *Bergey's Manual of Systematic Bacteriology*. Berlin: Springer-Verlag; 2005. pp. 49–65.
28. Gupta RS. Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. *Microbiol Mol Biol Rev* 1998;62:1435–1491.
29. Woese CR. How we do, don't and should look at Bacteria and Bacteriology. In: Mea D (eds). *The Prokaryotes: An Evolving Electronic Resource for the Microbiological Community*. New York: Springer-Verlag; 2003.
30. Baldauf SL. Phylogeny for the faint of heart: a tutorial. *Trends Genet* 2003;19:345–351.
31. Felsenstein J. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol* 1996;266:418–427.
32. Lake JA. The order of sequence alignment can bias the selection of tree topology. *Mol Biol Evol* 1991;8:378–385.
33. Konstantinidis KT, Tiedje JM. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr Opin Microbiol* 2007;10:504–509.
34. Qin Q-L, Xie B-B, Zhang X-Y, Chen X-L, Zhou B-C, *et al*. A proposed genus boundary for the prokaryotes based on genomic insights. *J Bacteriol* 2014;196:2210–2215.
35. Barco RA, Garrity GM, Scott JJ, Amend JP, Nealson KH, *et al*. A genus definition for bacteria and archaea based on a standard genome relatedness index. *mBio* 2020;11.
36. Gupta RS. Distinction between *Borrelia* and *Borreliella* is more robustly supported by molecular and phenotypic characteristics than all other neighbouring prokaryotic genera: Response to Margos' et al. "The genus *Borrelia* reloaded" (PLoS One 13(12): e0208432). *PLoS One* 2019;14:e0221397.
37. Puigbò P, Wolf YI, Koonin EV. Seeing the Tree of Life behind the phylogenetic forest. *BMC Biol* 2013;11:46.
38. Adeolu M, Alnajar S, Naushad S, S Gupta R. Genome-based phylogeny and taxonomy of the "*Enterobacteriales*": proposal for *Enterobacterales* ord. nov. divided into the families *Enterobacteriaceae*, *Erwiniaceae* fam. nov., *Pectobacteriaceae* fam. nov., *Yersiniaceae* fam. nov., *Hafniaceae* fam. nov., *Morganellaceae* fam. nov., and *Budviciaceae* fam. nov. *Int J Syst Evol Microbiol* 2016;66:5575–5599.
39. Sawana A, Adeolu M, Gupta RS. Molecular signatures and phylogenomic analysis of the genus *Burkholderia*: proposal for division of this genus into the emended genus *Burkholderia* containing pathogenic organisms and a new genus *Paraburkholderia* gen. nov. harboring environmental species. *Front Genet* 2014;5:429.
40. Gupta RS, Griffiths E. Chlamydiae-specific proteins and indels: novel tools for studies. *Trends Microbiol* 2006;14:527–535.
41. Griffiths E, Gupta RS. Identification of signature proteins that are distinctive of the *Deinococcus-Thermus* phylum. *Int Microbiol* 2007;10:201–208.
42. Griffiths E, Petrich AK, Gupta RS. Conserved indels in essential proteins that are distinctive characteristics of *Chlamydiales* and provide novel means for their identification. *Microbiology* 2005;151:2647–2657.

43. Naushad HS, Gupta RS. Phylogenomics and molecular signatures for species from the plant pathogen-containing order *Xanthomonadales*. *PLoS ONE* 2013;8:e55216.
44. Gupta RS, Chander P, George S. Phylogenetic framework and molecular signatures for the class *Chloroflexi* and its different clades; proposal for division of the class *Chloroflexia* class. nov. [corrected] into the suborder *Chloroflexineae* subord. nov., consisting of the emended family *Oscillochloridaceae* and the family *Chloroflexaceae* fam. nov., and the suborder *Roseiflexineae* subord. nov., containing the family *Roseiflexaceae* fam. nov. *Antonie van Leeuwenhoek* 2013;103:99–119.
45. Bhandari V, Gupta RS. Molecular signatures for the phylum *Synergistetes* and some of its subclades. *Antonie van Leeuwenhoek* 2012;102:517–540.
46. Gao B, Gupta RS. Phylogenetic framework and molecular signatures for the main clades of the phylum *Actinobacteria*. *Microbiol Mol Biol Rev* 2012;76:66–112.
47. Bhandari V, Gupta RS. Phylum *Thermotogae*. In: Rosenberg E, DeLong E, Lory S, Stackebrandt E and Thompson F (eds). *The Prokaryotes- Other Major Lineages of Bacteria and the Archaea*. New York: Springer; 2014. pp. 989–1015.
48. Gupta RS. Identification of conserved indels that are useful for classification and evolutionary studies. In: Goodfellow M, Sutcliffe IC and Chun J (eds). *Bacterial Taxonomy, Methods in Microbiology*, vol. 41. London: Elsevier; 2014. pp. 153–182.
49. Naushad HS, Lee B, Gupta RS. Conserved signature indels and signature proteins as novel tools for understanding microbial phylogeny and systematics: identification of molecular signatures that are specific for the phytopathogenic genera *Dickeya*, *Pectobacterium* and *Brenneria*. *Int J Syst Evol Microbiol* 2014;64:366–383.
50. Bhandari V, Gupta RS. Molecular signatures for the phylum (class) *Thermotogae* and a proposal for its division into three orders (*Thermotogales*, *Kosmotogales* ord. nov. and *Petrotogales* ord. nov.) containing four families (*Thermotogaceae*, *Fervidobacteriaceae* fam. nov., *Kosmotogaceae* fam. nov. and *Petrotogaceae* fam. nov.) and a new genus *Pseudothermotoga* gen. nov. with five new combinations. *Antonie van Leeuwenhoek* 2014;105:143–168.
51. Gupta RS, Naushad S, Chokshi C, Griffiths E, Adeolu M. A phylogenomic and molecular markers based analysis of the phylum *Chlamydiae*: proposal to divide the class *Chlamydiia* into two orders, *Chlamydiales* and *Parachlamydiales* ord. nov., and emended description of the class *Chlamydiia*. *Antonie van Leeuwenhoek* 2015;108:765–781.
52. Patel S, Gupta RS. A phylogenomic and comparative genomic framework for resolving the polyphyly of the genus *Bacillus*: Proposal for six new genera of *Bacillus* species, *Peribacillus* gen. nov., *Cytobacillus* gen. nov., *Mesobacillus* gen. nov., *Neobacillus* gen. nov., *Metabacillus* gen. nov. and *Alkalihalobacillus* gen. nov. *Int J Syst Evol Microbiol* 2020;70:406–438.
53. Barbour AG, Adeolu M, Gupta RS. Division of the genus Borrelia into two genera (corresponding to Lyme disease and relapsing fever groups) reflects their genetic and phenotypic distinctiveness and will lead to a better understanding of these two groups of microbes (Margos *et al*. (2016) There is inadequate evidence to support the division of the genus Borrelia. *Int. J. Syst. Evol. Microbiol.* doi: 10.1099/ijsem.0.001717). *Int J Syst Evol Microbiol* 2017;67:2058–2067.
54. Dobritsa AP, Samadpour M. Reclassification of *Burkholderia insecticola* as *Caballeronia insecticola* comb. nov. and reliability of conserved signature indels as molecular synapomorphies. *Int J Syst Evol Microbiol* 2019;69:2057–2063.
55. Ma Y, Wu X, Li S, Tang L, Chen M, *et al*. Proposal for reunification of the genus *Raoultella* with the genus *Klebsiella* and reclassification of *Raoultella electrica* as *Klebsiella electrica* comb. nov. *Res Microbiol* 2021;172:103851.
56. Jiang L, Wang D, Kim J-S, Lee JH, Kim D-H, *et al*. Reclassification of genus *Izhakiella* into the family *Erwiniaceae* based on phylogenetic and genomic analyses. *Int J Syst Evol Microbiol* 2020;70:3541–3546.
57. Gupta RS. Protein signatures (molecular synapomorphies) that are distinctive characteristics of the major cyanobacterial clades. *Int J Syst Evol Microbiol* 2009;59:2510–2526.
58. Gupta RS, Patel S, Saini N, Chen S. Erratum: Robust demarcation of seventeen distinct *Bacillus* species clades, proposed as novel *Bacillaceae* genera, by phylogenomics and comparative genomic analyses: description of *Robertmurraya kyonggiensis* sp. nov. and proposal for emended genus *Bacillus* limiting it only to the members of the Subtilis and Cereus clades of species. *Int J Syst Evol Microbiol* 2020;70:6531–6533.
59. Campbell C, Adeolu M, Gupta RS. Genome-based taxonomic framework for the class *Negativicutes*: division of the class *Negativicutes* into the orders *Selenomonadales* emend., *Acidaminococcales* ord. nov. and *Veillonellales* ord. nov. *Int J Syst Evol Microbiol* 2015;65:3203–3215.
60. Dobritsa AP, Linardopoulou EV, Samadpour M. Transfer of 13 species of the genus *Burkholderia* to the genus *Caballeronia* and reclassification of *Burkholderia jirisanensis* as *Paraburkholderia jirisanensis* comb. nov. *Int J Syst Evol Microbiol* 2017;67:3846–3853.
61. Cutiño-Jiménez AM, Menck CFM, Cambas YT, Díaz-Pérez JC. Protein signatures to identify the different genera within the *Xanthomonadaceae* family. *Braz J Microbiol* 2020;51:1515–1526.
62. Barbour AG. *Borreliaceae*. In: *Bergey Manual of Systematics of Bacteria and Archaea*. John Wiley & Sons, Inc. Bergey's Trust, 2018.
63. Gupta RS, Sawnani S, Adeolu M, Alnajar S, Oren A. Phylogenetic framework for the phylum *Tenericutes* based on genome sequence data: proposal for the creation of a new order *Mycoplasmoidales* ord. nov., containing two new families *Mycoplasmoidaceae* fam. nov. and *Metamycoplasmataceae* fam. nov. harbouring *Eperythrozoon*, *Ureaplasma* and five novel genera. *Antonie van Leeuwenhoek* 2018;111:1583–1630.
64. Saini N, Gupta RS. A robust phylogenetic framework for members of the order *Legionellales* and its main genera (*Legionella*, *Aquicella*, *Coxiella* and *Rickettsiella*) based on phylogenomic analyses and identification of molecular markers demarcating different clades. *Antonie van Leeuwenhoek* 2021;114:957–982.
65. Gupta RS, Son J, Oren A. A phylogenomic and molecular markers based taxonomic framework for members of the order Entomoplasmatales: proposal for an emended order Mycoplasmatales containing the family Spiroplasmataceae and emended family Mycoplasmataceae comprised of six genera. *Antonie van Leeuwenhoek* 2019;112:561–588.
66. Gupta RS, Patel S. Robust demarcation of the family *Caryophanaceae* (*Planococcaceae*) and its different genera including three novel genera based on phylogenomics and highly specific molecular signatures. *Front Microbiol* 2019;10:2821.
67. Kämpfer P, Glaeser SP, Busse H-J, McInroy JA, Clermont D, *et al*. *Pseudoneobacillus rhizosphaerae* gen. nov., sp. nov., isolated from maize root rhizosphere. *Int J Syst Evol Microbiol* 2022;72.
68. Jiang L, Lee MH, Jeong JC, Kim D-H, Kim CY, *et al*. *Neobacillus endophyticus* sp. nov., an endophytic bacterium isolated from *Selaginella involvens* roots. *Int J Syst Evol Microbiol* 2019;71.
69. Montecillo JAV, Bae H. Reclassification of *Brevibacterium frigoritolerans* as *Peribacillus frigoritolerans* comb. nov. based on phylogenomics and multiple molecular synapomorphies. *Int J Syst Evol Microbiol* 2022;72.
70. Jiang L, Jung WY, Li Z, Lee M-K, Park S-H, *et al*. *Peribacillus faecalis* sp. nov., a moderately halophilic bacterium isolated from the faeces of a cow. *Int J Syst Evol Microbiol* 2019;71.
71. Rai A, Smita N, Shabbir A, Jagadeeshwari U, Keertana T, *et al*. *Mesobacillus aurantius* sp. nov., isolated from an orange-colored pond near a solar saltern. *Arch Microbiol* 2021;203:1499–1507.
72. Jeong JW, Kim YS, Kim SB. *Metabacillus bambusae* sp. nov., isolated from bamboo grove soil. *Int J Syst Evol Microbiol* 2022;72.

73. Lee SY, Son JS, Hwang YJ, Shin JH, Ghim SY. *Metabacillus elymi* sp. nov., isolated from the Rhizosphere of *Elymus tsukushiensis*, a plant native to the Dokdo Islands, Republic of Korea. *Antonie van Leeuwenhoek* 2021;114:1709–1719.
74. Montecillo JAV. Phylogenomics and comparative genomic analyses support the creation of the novel family *Ignatzschineriaceae* fam. nov. comprising the genera *Ignatzschineria* and *Wohlfahrtiimonas* within the order *Cardiobacteriales*. *Res Microbiol* 2023;174:103988.
75. Bello S, Rudra B, Gupta RS. Phylogenomic and comparative genomic analyses of *Leuconostocaceae* species: identification of molecular signatures specific for the genera *Leuconostoc*, *Fructobacillus* and *Oenococcus* and proposal for a novel genus *Periweissella* gen. nov. *Int J Syst Evol Microbiol* 2022;72.
76. Chen S, Rudra B, Gupta RS. Phylogenomics and molecular signatures support division of the order *Neisseriales* into emended families *Neisseriaceae* and *Chromobacteriaceae* and three new families *Aquaspirillaceae* fam. nov., *Chitinibacteraceae* fam. nov., and *Leeiaceae* fam. nov. *Syst Appl Microbiol* 2021;44:126251.
77. Bello S, Howard-Azzeh M, Schellhorn HE, Gupta RS. Phylogenomic analyses and molecular signatures elucidating the evolutionary relationships amongst the *Chlorobia* and *Ignavibacteria* species: Robust demarcation of two family-level clades within the order *Chlorobiales* and proposal for the family *Chloroherpetonaceae* fam. nov. *Microorganisms* 2022;10:1312.
78. Adeolu M, Gupta RS. A phylogenomic and molecular marker based proposal for the division of the genus *Borrelia* into two genera: the emended genus *Borrelia* containing only the members of the relapsing fever *Borrelia*, and the genus *Borreliella* gen. nov. containing the members of the Lyme disease *Borrelia* (*Borrelia burgdorferi* sensu lato complex). *Antonie van Leeuwenhoek* 2014;105:1049–1072.
79. Rudra B, Duncan L, Shah AJ, Shah HN, Gupta RS. Phylogenomic and comparative genomic studies robustly demarcate two distinct clades of *Pseudomonas aeruginosa* strains: proposal to transfer the strains from an outlier clade to a novel species *Pseudomonas paraeruginosa* sp. nov. *Int J Syst Evol Microbiol* 2022;72:11.
80. Rudra B, Gupta RS. Phylogenomic and comparative genomic analyses of species of the family *Pseudomonadaceae*: Proposals for the genera *Halopseudomonas* gen. nov. and *Atopomonas* gen. nov., merger of the genus *Oblitimonas* with the genus *Thiopseudomonas*, and transfer of some misclassified species of the genus *Pseudomonas* into other genera. *Int J Syst Evol Microbiol* 2021;71.
81. Gupta RS, Mahmood S, Adeolu M. A phylogenomic and molecular signature based approach for characterization of the phylum Spirochaetes and its major clades: proposal for a taxonomic revision of the phylum. *Front Microbiol* 2013;4:217.
82. Sayers EW, Agarwala R, Bolton EE, Brister JR, Canese K, *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2019;47:D23–D28.
83. Parte AC. LPSN - the List of Prokaryotic Names with Standing in Nomenclature. *Int J Syst Evol Microbiol* 2018;68:1825–1829.
84. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015;25:1043–1055.
85. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402.
86. Wu D, Jospin G, Eisen JA, Brochier-Armanet C. Systematic identification of gene families for use as "markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. *PLoS One* 2013;8:e77033.
87. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;5:e9490.
88. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30:1312–1313.
89. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* 2013;30:2725–2729.
90. Bhandari V, Ahmod NZ, Shah HN, Gupta RS. Molecular signatures for *Bacillus* species: demarcation of the *Bacillus subtilis* and *Bacillus cereus* clades in molecular terms and proposal to limit the placement of new species into the genus *Bacillus*. *Int J Syst Evol Microbiol* 2013;63:2712–2726.
91. Joshi A, Thite S, Karodi P, Joseph N, Lodha T. *Alkalihalobacterium elongatum* gen. nov. sp. nov.: an antibiotic-producing bacterium isolated from Lonar Lake and reclassification of the genus *Alkalihalobacillus* into seven novel genera. *Front Microbiol* 2021;12:722369.
92. Heyrman J, Balcaen A, Rodriguez-Diaz M, Logan NA, Swings J, *et al*. *Bacillus decolorationis* sp. nov., isolated from biodeteriorated parts of the mural paintings at the Servilia tomb (Roman necropolis of Carmona, Spain) and the Saint-Catherine chapel (Castle Herberstein, Austria). *Int J Syst Evol Microbiol* 2003;53:459–463.
93. Gupta RS, Shami A. Molecular signatures for the *Crenarchaeota* and the *Thaumarchaeota*. *Antonie van Leeuwenhoek* 2011;99:133–157.
94. Gupta RS, Lali R. Molecular signatures for the phylum Aquificae and its different clades: proposal for division of the phylum Aquificae into the emended order *Aquificales*, containing the families *Aquificaceae* and *Hydrogenothermaceae*, and a new order *Desulfurobacteriales* ord. nov., containing the family *Desulfurobacteriaceae*. *Antonie van Leeuwenhoek* 2013;104:349–368.
95. Collins MD, Lawson PA, Willems A, Cordoba JJ, Fernandez-Garayzabal J, *et al*. The phylogeny of the genus *Clostridium*: proposal of five new genera and eleven new species combinations. *Int J Syst Bacteriol* 1994;44:812–826.
96. Collins MD. The genus *Brevibacterium*. In: Dworkin M, Falkow S, Rosenberg E, Schleifer KH and Stackebrandt E (eds). *The Prokaryotes: An Evolving Electronic Resource for the Microbiological Community*. New York: Springer-Verlag; 2006. pp. 1013–1019.
97. Patel S, Gupta RS. Robust demarcation of fourteen different species groups within the genus *Streptococcus* based on genome-based phylogenies and molecular signatures. *Infect Genet Evol* 2018;66:130–151.
98. Alnajar S, Gupta RS. Phylogenomics and comparative genomic studies delineate six main clades within the family *Enterobacteriaceae* and support the reclassification of several polyphyletic members of the family. *Infect Genet Evol* 2017;54:108–127.
99. Gupta RS, Naushad S, Fabros R, Adeolu M. Erratum to: a phylogenomic reappraisal of family-level divisions within the class *Halobacteria*: proposal to divide the order *Halobacteriales* into the families *Halobacteriaceae*, *Haloarculaceae* fam. nov., and *Halococcaceae* fam. nov., and the order *Haloferacales* into the families, *Haloferacaceae* and *Halorubraceae* fam nov. *Antonie van Leeuwenhoek* 2016;109:1521–1523.
100. Zhang G, Gao B, Adeolu M, Khadka B, Gupta RS. Phylogenomic analyses and comparative studies on genomes of the *Bifidobacteriales*: identification of molecular signatures specific for the order *Bifidobacteriales* and its different subclades. *Front Microbiol* 2016;7:978.
101. Ho J, Adeolu M, Khadka B, Gupta RS. Identification of distinctive molecular traits that are characteristic of the phylum "Deinococcus-Thermus" and distinguish its main constituent groups. *Syst Appl Microbiol* 2016;39:453–463.
102. Gupta RS, Naushad S, Baker S. Phylogenomic analyses and molecular signatures for the class *Halobacteria* and its two major clades: a proposal for division of the class *Halobacteria* into an emended order *Halobacteriales* and two new orders, *Haloferacales* ord. nov. and *Natrialbales* ord. nov., containing the novel families *Haloferacaceae* fam. nov. and *Natrialbaceae* fam. nov. *Int J Syst Evol Microbiol* 2015;65:1050–1069.

103. Naushad S, Adeolu M, Wong S, Sohail M, Schellhorn HE, *et al*. A phylogenomic and molecular marker based taxonomic framework for the order *Xanthomonadales*: proposal to transfer the families *Algiphilaceae* and *Solimonadaceae* to the order *Nevskiales* ord. nov. and to create a new family within the order *Xanthomonadales*, the family *Rhodanobacteraceae* fam. nov., containing the genus *Rhodanobacter* and its closest relatives. *Antonie van Leeuwenhoek* 2015;107:467–485.
104. Ravinesan DA, Gupta RS. Molecular signatures for members of the genus *Dehalococcoides* and the class *Dehalococcoidia*. *Int J Syst Evol Microbiol* 2014;64:2176–2181.
105. Kim O-S, Cho Y-J, Lee K, Yoon S-H, Kim M, *et al*. Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. *Int J Syst Evol Microbiol* 2012;62:716–721.
106. Wong SY, Paschos A, Gupta RS, Schellhorn HE. Insertion/deletion-based approach for the detection of *Escherichia coli* O157:H7 in freshwater environments. *Environ Sci Technol* 2014;48:11462–11470.
107. Ahmod NZ, Gupta RS, Shah HN. Identification of a *Bacillus anthracis* specific indel in the yeaC gene and development of a rapid pyrosequencing assay for distinguishing *B. anthracis* from the *B. cereus* group. *J Microbiol Methods* 2011;87:278–285.
108. Chun J, Lee J-H, Jung Y, Kim M, Kim S, *et al*. EzTaxon: a web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences. *Int J Syst Evol Microbiol* 2007;57:2259–2261.
109. Meier-Kolthoff JP, Carbasse JS, Peinado-Olarte RL, Göker M. TYGS and LPSN: a database tandem for fast and reliable genome-based classification and nomenclature of prokaryotes. *Nucleic Acids Res* 2022;50:D801–D807.
110. Gao B, Gupta RS. Conserved indels in protein sequences that are characteristic of the phylum *Actinobacteria*. *Int J Syst Evol Microbiol* 2005;55:2401–2412.