Research Article

FN-Identify: Novel Restriction Enzymes-Based Method for Bacterial Identification in Absence of Genome Sequencing

Mohamed Awad,1 Osama Ouda,2 Ali El-Refy,1 Fawzy A. El-Feky,1 Kareem A. Mosa,1,3 and Mohamed Helmy4

1Department of Biotechnology, Faculty of Agriculture, Al-Azhar University, Cairo 11651, Egypt
2Department of Information Technology, Faculty of Computer and Information Sciences, Mansoura University, Mansoura 35516, Egypt
3Department of Applied Biology, College of Sciences, University of Sharjah, P.O. Box 27272, Sharjah, UAE
4Donnelly Centre for Cellular and Biomedical Research, University of Toronto, Toronto, ON, Canada M5S 3E1

Correspondence should be addressed to Mohamed Helmy; mohamed.attiashata@utoronto.ca

Received 31 July 2015; Revised 25 November 2015; Accepted 29 November 2015

Academic Editor: Paul Harrison

Copyright © 2015 Mohamed Awad et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Sequencing and restriction analysis of genes like 16S rRNA and HSP60 are intensively used for molecular identification in the microbial communities. With aid of the rapid progress in bioinformatics, genome sequencing became the method of choice for bacterial identification. However, the genome sequencing technology is still out of reach in the developing countries. In this paper, we propose FN-Identify, a sequencing-free method for bacterial identification. FN-Identify exploits the gene sequences data available in GenBank and other databases and the two algorithms that we developed, CreateScheme and GenelIdentify, to create a restriction enzyme-based identification scheme. FN-Identify was tested using three different and diverse bacterial populations (members of Lactobacillus, Pseudomonas, and Mycobacterium groups) in an in silico analysis using restriction enzymes and sequences of 16S rRNA gene. The analysis of the restriction maps of the members of three groups using the fragment numbers information only or along with fragments sizes successfully identified all of the members of the three groups using a minimum of four and maximum of eight restriction enzymes. Our results demonstrate the utility and accuracy of FN-Identify method and its two algorithms as an alternative method that uses the standard microbiology laboratories techniques when the genome sequencing is not available.

1. Introduction

Bacterial identification is an important routine in the clinical and industrial microbiology laboratories. Microbiologists and researchers stepped up their efforts to improve and facilitate the rapid characterization of various microbial communities. Traditional bacterial identification strategies are mainly based on morphological, biochemical, enzymatic, antigenic, staining, and antibiogram characterization [1]. However, these strategies are time consuming and sometimes fail to identify the bacteria accurately [2]. Many other strategies appear to have improved bacterial identification accuracy, such as automated cellular fatty acid (CFA) analysis, yet these strategies require expensive system and standardized culture condition. Moreover, it cannot differentiate closely related species such as Escherichia coli and Shigella [2]. Protein analysis and phage analysis are also used as methods for bacterial identification [3]. With the presentation and rapid progress of molecular biology and molecular markers, several new and enhanced bacterial identification methods were developed. These methods include plasmid analysis [4], restriction fragment length polymorphism (RFLP) [5], pulsed-field gel electrophoresis (PFGE) [6], random amplified polymorphism DNA (RAPD) [7], fluorescent in situ hybridization (FISH) [8], and DNA Props [9].

In the early 1980s, polymerase chain reaction (PCR) provided novel approaches for bacterial identification through amplification of specific sequences/gene from the bacterial genome. Several ribosomal RNA (rRNA) genes and Internal Transcribed Spacers (ITSs) had been utilized for PCR-based
bacterial identification such as 16S rRNA, 23S rRNA, 5S rRNA, and SSU rRNA [8, 10]. The PCR-based identification uses the ribosomal genes, since ribosomal genes play an important role in living organisms and have functional stability over evolution ages due to rare variation in its sequences through millions of years, which makes them suitable to be used for identification and taxonomical purposes.

Numerous ribosomal RNA genes and ITSs such as Hsp65, rpoB, gyrB, groEL, and recA have been tested as a genetic marker in bacterial identification [11]. However, 16S rRNA is the most widely used ribosomal RNA genes in bacterial identification due to several reasons: (1) the 16S rRNA gene presents in almost all bacterial families; (2) it has functional and evaluation stability; (3) in many cases, multiple copies of the 16S rRNA gene presented in the genome and sometimes differences in sequences present as well, which can be used to distinguish closely related species; (4) the sequence length is about 1500:1550 bp, which is enough for taxonomical purpose and suitable for amplification; (5) the 16S rRNA gene sequence contains conserved regions and variable regions; therefore, it is possible to design a universal primer on these conserved regions for gene sequence amplification [1, 12]. Therefore, several methods for 16S rRNA amplification and analysis were developed: ribotyping [13], denaturing gradient gel electrophoresis (DGGE) [14, 15], temperature gradient gel electrophoresis (TGGE) [15], amplified ribosomal DNA restriction analysis (ARDRA) [16], and terminal restriction fragment length polymorphism (T-RFLP) [17, 18].

With the rapid progress in DNA and RNA sequencing technology, sequencing of 16S rRNA gene and several other genes became a popular method for bacterial identification and phylogenetic reconstruction. Furthermore, it is employed in nucleic acid-based detection, quantification of microbial diversity, and discovery of novel bacterial isolates in different microbiology laboratories [19–22].

Despite the outstanding advancements in speed and accuracy and the remarkable decrease in cost of the sequencing technologies in the recent years, sequencing technologies in developing countries are out of reach for the majority of clinical and research laboratories. This is mainly due to the high cost of establishing sequencing facility and high cost of reagents and maintenance [23–25]. Furthermore, the lack of trained personnel and the limited access to up-to-date scientific information play an important role in constraining the use of such indispensable technology in many clinical and industrial microbiology laboratories in these countries [26, 27]. Most labs depend on outsourcing the DNA/RNA sequencing through using commercial services. Typically, the sample is prepared and sent to a local company that sends it to companies in the EU or China to be sequenced and the results are sent back. Based on our observations, this process is expensive and time consuming (up to several months) and can fail at any point.

In this work, we present a FN-Identify, an efficient and sequencing-free bacterial identification method, as a proposed alternative that can be employed when genome sequencing is inaccessible. FN-Identify, which stands for fragment number-identify, is based on techniques that are available in most of the standard microbiological laboratories.

Our new method depends on sequences available in GenBank and other public databases, such as RDP-II [28], Silva [29], and Greengenes [30], restriction enzymes, and the two FN-Identify algorithms that we developed (Figure 1). We used bacterial population of 33 members (species and strains) of Lactobacillus genus to develop the method and used two other bacterial populations of 33 and 22 members (species and strains) of Pseudomonas and Mycobacterium, respectively, to test the method. FN-Identify successfully identified and differentiated all the species/strains using two different genes 16S rRNA and HSP60, in two independent analyses. The identification scheme and the utilized restriction enzymes, created by FN-Identify, demonstrate its efficiency as a rapid, accurate, and affordable alternative method for bacterial identification in the absence of the sequencing technologies.

2. Materials and Methods

2.1. Bacterial Genomes. We downloaded the 33, 33, and 22 Lactobacillus, Pseudomonas, and Mycobacterium members, respectively, with full genome sequences and annotations from Genome Database of the National Center for Biotechnology Information (NCBI) (September 2013). Table 1 shows the names and GenBank accession number of the Lactobacillus members and Tables S1 and S3 (see Supplementary Material available online at http://dx.doi.org/10.1155/2015/303605) show details of Pseudomonas and Mycobacterium members used in this study.

2.2. 16S rRNA and HSP60 Extraction. The files that contain the Lactobacillus bacterial genome sequences were processed using Python script to extract each 16S rRNA and HSP60 sequence according to the Lactobacillus genome annotations. Table 2 shows the copy numbers and sequence positions (start-end) of the 16S rRNA and HSP60 sequences in the Lactobacillus members and Tables S2 and S4 show the same details of Pseudomonas and Mycobacterium members used in this study. In one case, Lactobacillus kefiranofaciens ZW3, we
Table 1: Names and GenBank accession number of Lactobacillus species used in this study.

| Strain ID | Organism                                      | GenBank accession number |
|-----------|-----------------------------------------------|--------------------------|
| 1         | Lactobacillus acidophilus 30SC                 | CP002559                 |
| 2         | Lactobacillus acidophilus NCFM                | CP000033                 |
| 3         | Lactobacillus amylovorus GRL II2              | CP002338                 |
| 4         | Lactobacillus amylovorus GRL II18             | CP002609                 |
| 5         | Lactobacillus brevis ATCC 367                 | CP000416                 |
| 6         | Lactobacillus buchneri NRRL B-30929            | CP002652                 |
| 7         | Lactobacillus casei ATCC 334                  | CP000423                 |
| 8         | Lactobacillus crispatus ST1                   | FN692037                 |
| 9         | Lactobacillus delbrueckii subsp. bulgaricus 2038 | CP000156               |
| 10        | Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842 | CR954253            |
| 11        | Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365 | CP000412           |
| 12        | Lactobacillus fermentum CECT 5716             | CP000203                 |
| 13        | Lactobacillus fermentum IFO 3956              | AP008937                 |
| 14        | Lactobacillus gasseri ATCC 33323              | CP000413                 |
| 15        | Lactobacillus helveticus DPC 4571             | CP000517                 |
| 16        | Lactobacillus helveticus H10                  | CP002429                 |
| 17        | Lactobacillus johnsonii DPC 6026              | CP002464                 |
| 18        | Lactobacillus johnsonii FI9785                | FN298497                 |
| 19        | Lactobacillus johnsonii NCC 533               | AE017198                 |
| 20        | Lactobacillus plantarum JDM1                  | CP001617                 |
| 21        | Lactobacillus plantarum subsp. plantarum ST-III | CP002222             |
| 22        | Lactobacillus reuteri DSM 20016               | CP000705                 |
| 23        | Lactobacillus reuteri JCM II2                 | AP007281                 |
| 24        | Lactobacillus rhamnosus ATCC 53103            | AP011548                 |
| 25        | Lactobacillus rhamnosus GG                    | FM119322                 |
| 26        | Lactobacillus rhamnosus Lc 705                | FM119323                 |
| 27        | Lactobacillus sakei 23K                       | CR936503                 |
| 28        | Lactobacillus kefiranofaciens ZW3            | CP002764                 |
| 29        | Lactobacillus Paracasei 8700.2               | CP002391                 |
| 30        | Lactobacillus ruminis ATCC 27782             | CP003032                 |
| 31        | Lactobacillus salivarius CECT 5713           | CP002034                 |
| 32        | Lactobacillus salivarius UCC118              | CP000233                 |
| 33        | Lactobacillus sanfranciscensis TMW 1.1304    | CP002461                 |

*This ID will be used to refer to the species/strains in the text. This table lists the studied Lactobacillus species/strains and their GenBank accession numbers.

had to annotate the 16S rRNA sequences, as its annotation was unavailable in the database. We picked up the 16S rRNA sequences from L. kefiranofaciens ZW3 genome using the same primers successfully used with all other Lactobacillus members. The two primers picked up four copies of 16S rRNA sequences (Table 2 strain ID 28).

2.3. 16S rRNA Primer Selection. We tested 13 different primer sequences obtained from 8 published studies (Table 3). We used Python script to test the primers and compare the sequence positions we got using each primer with 16S rRNA position in the genome annotation in (NCBI), to confirm that the primer would pick the 16S rRNA sequence. Based on this testing, we selected two primers (Table 3, 8F and 1541R) from [31]. The two selected primers appear in all Lactobacillus genomes in this study and with the largest product length (1550 pb).

2.4. HSP60 Primer Design. A universal degenerate primer for picking up HSP60 sequences was designed based on the conserved regions in the HSP60 extracted sequences. We identified the conserved regions by performing multiple sequence alignment (MSA) using CLC Sequence Viewer software (CLC Bio, Swansea, UK). Table 3 shows the sequences of the designed and forward and reverse primers.

2.5. Restriction Enzymes and Restriction Map. We collected the information about restriction enzymes and restriction sites from the database of restriction enzymes (REBASE), Roberts 1980 and Roberts et al., 2010 [39, 40], and the restriction enzyme database attached to the DNA Star software (DNASTAR Inc., Madison, WI, USA). Prediction of the in silico restriction map was performed using the restriction sites information and the seqBuilder tool of Lasergene software tool (DNASTAR Inc., Madison, WI, USA).
Table 2: 16S rRNA and HSP60 copy numbers and genomics positions.

| Strain ID | 16S rRNA copies number | 16S rRNA position   | HSP60 position   |
|-----------|------------------------|---------------------|------------------|
| 1         | 4                      | 57091...58665       | 407805...409506  |
| 2         | 4                      | 59255...60826       | 379688...381333  |
| 3         | 4                      | 66295...67869       | 403452...405083  |
| 4         | 4                      | 55901...57475       | 376234...377865  |
| 5         | 5                      | 706262...707824     | 644545...647079  |
| 6         | 5                      | 259510...261077     | 1429276...1430898|
| 7         | 5                      | 2233684...225318    |                 |
| 8         | 4                      | 2233684...225318    |                 |
| 9         | 9                      | 35825...37395       |                 |
| 10        | 9                      | 45160...46720       |                 |
| 11        | 9                      | 43705...45265       |                 |
| 12        | 5                      | 169391...170958     |                 |
| 13        | 5                      | 477570...479418     |                 |
| 14        | 6                      | 76215...77787       |                 |
| 15        | 4                      | 85110...86682       |                 |
| 16        | 4                      | 546957...548607     |                 |
| 17        | 4                      | 455618...457268     |                 |
| 18        | 4                      | 558550...560200     |                 |
| 19        | 6                      | 115508...115668     |                 |
| 20        | 5                      | 484838...486408     |                 |

Table 2: Continued.

| Strain ID | 16S rRNA copies number | 16S rRNA position   | HSP60 position   |
|-----------|------------------------|---------------------|------------------|
| 11        | 9                      | 43705...45265       |                 |
| 12        | 5                      | 169391...170958     |                 |
| 13        | 5                      | 477570...479418     |                 |
| 14        | 6                      | 76215...77787       |                 |
| 15        | 4                      | 85110...86682       |                 |
| 16        | 4                      | 546957...548607     |                 |
| 17        | 4                      | 455618...457268     |                 |
| 18        | 4                      | 558550...560200     |                 |
| 19        | 6                      | 115508...115668     |                 |
| 20        | 5                      | 484838...486408     |                 |
technologies. Standard research tasks until the availability of sequencing developing alternative methods that can be utilized in doing choice, that is outsourcing. This situation raises the need of running and maintaining the mas well as hiring and training institutions for genomics and provide funds to facilitate to the updated scientific data, literature, and training [26, 27].

remarking speed, continuously improved accuracy, and sequencing is the technology-of-choice for several research, remarkable speed, continuously improved accuracy, and affordable sample processing cost. However, in several developing countries, the genome sequencing technologies are still out of reach for most of researchers and scientists due to several reasons which constrain employing such indispensable technology. Firstly, the high cost of establishing sequencing facility and high cost maintaining the facility in poor-resources countries. Secondly, the lack of well-trained personnel to run the facility. Thirdly, the weak power, Internet, and computational infrastructures. Finally, the limited access to the updated scientific data, literature, and training [26, 27].

The scientific community expected this problem over a decade ago with the rising of the next-generation sequencing technologies [25]. In the following years, many developing countries took steps to utilize these technologies by establishing institutions for genomics and provide funds to facilitate running and maintaining them as well as hiring and training personnel. Reports about case studies in several developing countries including Mexico, Thailand, South Africa, and India show the efforts made to import these technologies and the expected impact on research, public health, and economic development in these countries [41]. Despite these improvements, the problem seems to be still far from being solved, especially in Africa [23, 26], letting the researchers with one choice, that is outsourcing. This situation raises the need of developing alternative methods that can be utilized in doing standard research tasks until the availability of sequencing technologies.

Table 2: Continued.

| Strain ID | 16S rRNA copies number | 16S rRNA position | HSP60 position |
|-----------|------------------------|-------------------|----------------|
| 21        | 5                      | 487643 ·· 489213  | 591466 ·· 593091 |
| 22        | 6                      | 177728 ·· 179296  | 401807 ·· 403435 |
| 23        | 6                      | 177347 ·· 178880  | 401630 ·· 403258 |
| 24        | 5                      | 306772 ·· 308345  | 2303140 ·· 2304732 |
| 25        | 5                      | 307756 ·· 309313  | 2308734 ·· 2310368 |
| 26        | 5                      | 289782 ·· 291339  | 2265733 ·· 2267367 |
| 27        | 7                      | 306178 ·· 307748  | 358686 ·· 360625 |
| 28        | 41                    | 125303 ·· 126858  | 82036 ·· 83667 |
| 29        | 5                      | 274946 ·· 276503  | 224006 ·· 2241640 |
| 30        | 6                      | 274311 ·· 275837  | 650103 ·· 651714 |

Table 2: Continued.

| Strain ID | 16S rRNA copies number | 16S rRNA position | HSP60 position |
|-----------|------------------------|-------------------|----------------|
| 31        | 7                      | 749955 ·· 76521   | 1247027 ·· 1248649 |
| 32        | 7                      | 74540 ·· 76056    | 1246385 ·· 1248007 |
| 33        | 7                      | 485966 ·· 487585  | 481909 ·· 483380 |

1 Our Annotation for 16S rRNA sequences in L. kefiranofaciens ZW3.

3. Results and Discussion

3.1 Genomics in the Developing Countries. Currently, genome sequencing is the technology-of-choice for several research and clinical applications due to its rapid development, remarkable speed, continuously improved accuracy, and affordable sample processing cost. However, in several developing countries, the genome sequencing technologies are still out of reach for most of researchers and scientists due to several reasons which constrain employing such indispensable technology. Firstly, the high cost of establishing sequencing facility and high cost maintaining the facility in poor-resources countries. Secondly, the lack of well-trained personnel to run the facility. Thirdly, the weak power, Internet, and computational infrastructures. Finally, the limited access to the updated scientific data, literature, and training [26, 27].

The scientific community expected this problem over a decade ago with the rising of the next-generation sequencing technologies [25]. In the following years, many developing countries took steps to utilize these technologies by establishing institutions for genomics and provide funds to facilitate running and maintaining them as well as hiring and training personnel. Reports about case studies in several developing countries including Mexico, Thailand, South Africa, and India show the efforts made to import these technologies and the expected impact on research, public health, and economic development in these countries [41]. Despite these improvements, the problem seems to be still far from being solved, especially in Africa [23, 26], letting the researchers with one choice, that is outsourcing. This situation raises the need of developing alternative methods that can be utilized in doing standard research tasks until the availability of sequencing technologies.
3.2. Obtaining Standard Dataset of Bacterial Genomes and Genes. The identification of the family of certain bacteria is usually based on the morphological and other characteristics of the colony, while the identification of the species and strains requires molecular and more sophisticated methods [2, 16, 42, 43]. Therefore, we selected the Lactobacillaceae family as a representative of bacterial population with several industrial and health importance [44–47] to be used in developing FN-Identify method and algorithms. In addition, Lactobacillus members have different important genes used in bacterial identification and barcoding such as 16S rRNA and HSP60 with several differences in sequences and copy numbers. This makes Lactobacillus members ideal for developing and testing new methods for bacterial identification based on the analysis of the restriction patterns of its genes.

We downloaded the 33 complete Lactobacillus genome sequences and annotations available in the NCBI (Table 1). According to the genome annotations, all Lactobacillus genomes have one copy from HSP60 and between four and nine copies of 16S rRNA, except for Lactobacillus kefiranofaciens ZW3 (strain ID: 28, Table 1), where its genome annotation shows absence of 16S rRNA (Table 2). For Lactobacillus kefiranofaciens ZW3 we annotated the 16S rRNA gene using the selected 16S rRNA universal primers (see below). At least two of 16S rRNA copies are in the reverse direction. Strains under the same species have the same number of 16S rRNA copies except Lactobacillus johnsonii strains (strain IDs: 17 and 19, Table 1) since one of them has four copies and the other has six. Tables 1 and 2 list Lactobacillus species/strains used in this study as well as the copy numbers, start and end of each copy, and an ID that we gave to each species/strain that we will use hereafter.

3.3. Primer Selection and Design. In order to select standard universal primer(s) for 16S rRNA sequences from all Lactobacillus genomes, we tested several primers from published literature (Table 3). We performed the in silico screening for each primer using the separated gene sequences as well as the whole genome sequences. Our primers in silico screening show that (8F) and (1541R) primers present in most of the separated 16S rRNA gene sequences with largest product length (see Table 3 for primer sequences). Therefore, we keep the sequences between both primers and exclude all other sequences, including the primers sequences.

In some cases, these two primers are not present in 16S rRNA separated sequences. For instance, the two primers failed with the separated 16S rRNA genes of the strain Lactobacillus salivarius UCC118 (strain ID: 32, Table 1). However, when we used them with the whole genome of the same strain we found 8F and 1541R beginning from nucleotides 74,520 and 76,053, in agreement with the genome annotation of the first 16S rRNA copy (from 74,540 to 76,056). Similarly, Lactobacillus salivarius CECT 5713 (strain ID: 31, Table 1) has the same difference.

In some cases, there was a difference in length between the 16S rRNA returned in silico sequence and the length of the 16S rRNA in the genome annotations. For instance, Lactobacillus johnsonii (strain IDs: 17 and 19, Table 1) returned a 1555 bp sequence when using the two selected primers, while the gene length in the genome annotation was 1650 bp. However, it is within the start and end of the annotated gene, so we accept it. Apart from these few cases, the selected 16S rRNA primers 8F and 1541R performed perfectly with all Lactobacillus genomes. This guarantees that the returned in silico sequences will agree with the isolated sequences in lab.

After selecting the 8F and 1541R primers as universal primers for 16S rRNA, we used them to annotate the 16S rRNA gene in the Lactobacillus kefiranofaciens ZW3 (strain ID: 28, Table 1) genome. The result shows that the Lactobacillus kefiranofaciens ZW3 genome contains four copies of 16S rRNA sequences, from nucleotide 125,303 to 126,858 (1555 bp), from 142,446 to 144,001 (1555 bp), from 1,350,707 to 1,352,262 (1555 bp), and from 1,818,440 to 1,819,995 (1555 bp) (Table 2).

| ID | Gene name | Name | Sequence | Reference |
|----|-----------|------|----------|-----------|
| 1  | 16S rRNA  | 8F   | 5′AGAGTTTGTATCCCTGGCTCAG3′ | [31] |
| 2  | 16S rRNA  | U1492R | 5′GGTTACCTTGTACGACTT3′ | [32] |
| 3  | 16S rRNA  | 928F | 5′TAAACTYAAAKGATTGACGGG3′ | [33] |
| 4  | 16S rRNA  | 336R | 5′ACTGCTGCSYCCCGTAGAGCTCT3′ | [33] |
| 5  | 16S rRNA  | 1100F | 5′YAAACGAGCGCAACC3′ | [34] |
| 6  | 16S rRNA  | 1100R | 5′AGGTGGCCTCCTTGGTG3′ | [34] |
| 7  | 16S rRNA  | 907R | 5′CCGTCAATTCTTTRAGTTT3′ | [34] |
| 8  | 16S rRNA  | 785F | 5′GGATTAGATACCCTGGTA3′ | [35] |
| 9  | 16S rRNA  | 805R | 5′GACTACCGGAGCTACATTGCT3′ | [36] |
| 10 | 16S rRNA | 515F | 5′GTGCCAGCMCGCGGTAAAA3′ | [34] |
| 11 | 16S rRNA | 518R | 5′GTATTACCGCGCTGCTGG3′ | [37] |
| 12 | 16S rRNA | 27F | 5′AGATTTGATCTGGCTACG3′ | [38] |
| 13 | 16S rRNA | 1541R* | 5′AAGGAGGTGATCCAGCGGGA3′ | [31] |
| 14 | HSP60    | HSP60-F | 5′ATGCCAARGANNTTATTT3′ | Designed |
| 15 | HSP60    | HSP60-R | 5′TCGGCVACNACNGCTTCNGA3′ | Designed |

* 16S rRNA selected primers.
For HSP60 gene, we could not find a universal primer in the published literature. Therefore, we design a universal primer based on the conserved nucleotide sequences of HSP60. The conserved nucleotide sequences were identified by multiple sequence alignment (MSA) using CLC Sequence Viewer software (CLC Bio, Swansea, UK). Based on the alignment results, we were able to design two degenerate primers for HSP60 (HSP60-F and HSP60-R, Table 3): the forward primer (HSP60-F) 5’ATGGCWAARGANTHARTT3’ and the reverse primer (HSP60-R) 5’TCDGCVACNACNGCTTCNGA3’ yielded in 1560 bp for all species while the annotated HSP60 is 1600 bp. Again, we take the sequences between both primers and exclude all other sequences, including the primers sequences.

3.4. In Silico Restriction Map. In order to perform an in silico enzymatic restriction for the 16S rRNA and HSP60 genes, we selected 12 commercially available restriction enzymes from hundred of enzymes that we collected their data. To select these 12 enzymes, we scanned all enzymes using Python script and the information of the restriction site that we collected from the database of restriction enzymes (REBASE) [40], the restriction enzyme database attached to the DNA Star software (DNASTAR Inc., Madison, WI, USA), and other resources [39], against the 16S rRNA and HSP60 sequences. The selected enzymes have different restriction sites, which will help us differentiate the Lactobacillus species through the differences in restriction maps of the selected gene sequences. Next, we performed an in silico enzymatic restriction for the 16S rRNA and HSP60 gene sequences using seqBuilder tool of Lasergene software tool (DNASTAR Inc., Madison, WI, USA).

The in silico enzymatic digest results in DNA fragment lengths ranges approximately from 10 bp to 1570 bp. Since the very short fragments are unobservable in the experiments, we excluded the fragments length less than 30 bp [48]. Although it is expected that the number of return DNA fragments = the number of restriction sites + 1, the results are different from the expected ones and this is mainly due to two reasons: firstly some fragments being equal in length or the difference in lengths being too small to be observed in the gel separation and secondly our exclusion of the very short fragments.

The exclusion of the short fragments was observed in several species and strains from those we used in this study. For instance, Lactobacillus delbrueckii subsp. bulgaricus 2038 (strain ID: 9, Table 1) has six restriction sites for HinfI enzyme but the number of the return DNA fragments was four only. This is because one of the fragments was of length 9 bp, two other fragments are with length of 119, and two other fragments are with very close length (difference is less than 10 bp) [49]. The same strain has five restriction sites for TfiI but the return DNA fragments contain one fragment of length 9 bp. Therefore, it returns five fragments only. Table S5 contains the details of the return DNA fragments for each restriction enzyme.

Other sources of differences in ribotyping between the Lactobacillus genomes are the variation in the 16S rRNA copy numbers between different species and the differences in sequences between the multiple copies within the same genome (Table 2). This leads to difference in restriction sites and number of restriction fragments. One noticeable example for this phenomenon is the Lactobacillus brevis ATCC 367 (strain ID: 5, Table 1), which contains five copies of 16S rRNA genes with three different sequences (Table 2). The restriction of these three different sequences with HinfI enzyme results in four, five, and six DNA fragments since they have three, four, and five HinfI restriction sites, respectively (Table S5). The same three different sequences of 16S rRNA contain two, three, and three restriction sites for TfiI enzyme, respectively. Another example is Lactobacillus fermentum (strain IDs: 12 and 13) that shows similar behavior with the HinfI enzyme (Table 2 and Table S5).

To determine the number of returned DNA fragment from a particular species/strain that contains several copies of 16S rRNA sequences, we compare the lengths of the fragments and exclude the duplicated equal fragments length. This is how the restriction will be done actually in the lab, as the fragments with the same length will be in the same band in the gel. For instance, for Lactobacillus brevis ATCC 367 (strain ID: 5) there are five different copies of 16S rRNA with three different sequences (see above) (Table 2). Restriction with HinfI enzyme returned five fragments for two copies and four for the other copy. After excluding the duplicated fragment lengths, we have seven fragments only in the gel (976 bp, 891 bp, 379 bp, 243 bp, 136 bp, 117 bp, and 86 bp). Supplementary Figure 1 shows comparison of two cases where he fragments number is equal to the expected and where it is not.

For HSP60 gene, the construction of the restriction map was straightforward. Each Lactobacillus species or strain contains one single copy of the gene (Table 2). Therefore, the differentiation between them will be based on differences in restriction patterns between species/strains (Table S7).

4. FN-Identify Method

This section describes our proposed sequencing-free bacterial identification method in detail. The proposed method identifies bacterial species/strains based on the number of fragments and/or fragment lengths that result from the restriction of certain genes using a given set of restriction enzymes. Therefore, we refer to it as the fragment number-identification method or FN-Identify. The main goal of FN-Identify is to establish an identification scheme for bacterial species utilizing fragments patterns of enzymatic restrictions such as the restriction map we built in the above section. The established scheme specifies the set of enzymes that could be employed to identify a given (unknown) gene sequence as well as the order of their application. The identified gene refers to a particular species/strain within the restriction map.

The idea behind FN-Identify is inspired from two basic observations. First, the number of fragments resulting from each restriction of a DNA sequence (e.g., 16S rRNA gene sequence) would differ based on the employed restriction enzyme. Generally speaking, a given gene sequence G could be split into m_i and m_j fragments if two different enzymes, e_i and e_j, were employed, respectively, where m_i and m_j are likely to be different. Second, some restriction enzymes are more discriminative than other enzymes with respect to

\[ m_i \neq m_j \]
Input: Bacterial family $\mathcal{F}$ and a set of restriction enzymes $\mathcal{E}$.
Output: A tree $T$ representing the created gene-identification scheme

\begin{algorithm}
\begin{algorithmic}
\State Predict the restriction map $\mathcal{R}$ by restricting all species/strains of $\mathcal{F}$ using all enzymes of $\mathcal{E}$.
\If{all enzymes of $\mathcal{E}$ are visited}
\State return
\ElseIf{number of strains/species in $\mathcal{F}$ is 1}
\State return
\Else
\State Search $\mathcal{R}$ to find the restriction enzyme $e_{\max}^x$, that gives the largest number of distinct number-of-fragments when applied to all gene-sequences of $\mathcal{F}$.
\State Group all strains/species having the same number-of-fragments in a distinct group $G$ based on the results produced by $e_{\max}^x$.
\For{each resulting group $G$}
\State createScheme($G$, $\mathcal{E}$);
\State newNode.group $\leftarrow G$;
\State newNodeenzyme $\leftarrow$ nil;
\State newNode.numOfFragments $\leftarrow$ numOfFragments;
\State $T$.addChild(newNode)
\EndFor
\EndIf
\EndIf
\State return $T$
\end{algorithmic}
\caption{CreateScheme($\mathcal{F}$, $\mathcal{E}$).}
\end{algorithm}

different bacterial families. Assume that both $e_i$ and $e_j$ are employed to cut all sequences belonging to a specific bacterial family $\mathcal{F}$. Let $N_{e_i}$ be a set containing the number of fragments resulting from cutting all sequences of $\mathcal{F}$ using $e_i$. Enzyme $e_i$ is said to be more discriminative than $e_j$ if and only if $|N_{e_i}| > |N_{e_j}|$, where $|A|$ denotes the cardinality of set $A$.

For the purpose of illustration, consider an extreme example where all sequences of $\mathcal{F}$ are split into the same number of fragments if $e_i$ is employed, that is, $|N_{e_i}| = 1$, whereas each sequence of $\mathcal{F}$ produces a different number of sequences if $e_j$ is employed; that is, $|N_{e_j}| = n$, where $n$ is the number of sequences of $\mathcal{F}$. Clearly, while $e_i$ can identify all sequences of $\mathcal{F}$ perfectly, $e_j$ does not provide any useful information for discriminating the sequences of $\mathcal{F}$. FN-Identify benefits from the above two observations to create an identification scheme for bacterial genes utilizing only a set of discriminating restriction enzymes. The proposed method consists of two algorithms. The first algorithm, CreateScheme, aims at finding an efficient identification scheme given a bacterial family $\mathcal{F}$ and the adopted set of restriction enzymes $\mathcal{E}$. The second algorithm, GenIdentify, employs the obtained scheme to identify a given unknown gene sequence.

The CreateScheme algorithm (see Algorithm 1) recursively builds a tree $T$ that represents an identification scheme for species/strains of $\mathcal{F}$. Each node of $T$ consists of three components, namely, the processed group of species/strains, the restriction enzyme that produces the largest number of distinct number of fragments when applied to that group, and the number of distinct number of fragments produced. Obviously, the first component of the root node of $T$ (Figure 2) consists of all species/strains of $\mathcal{F}$ and the third component should be 1 since all species/strains of $\mathcal{F}$ consist of only one fragment, that is, the whole sequence. Once the enzyme that produces the largest number of distinct number of fragments, for all members of $\mathcal{F}$, is found, it should be assigned to the second component of root($T$). Algorithm 1 can be described informally as follows.

\begin{itemize}
\item \textbf{Step 1.} Predict the restriction map $\mathcal{R}$ by restricting all species/strains of $\mathcal{F}$ using all enzymes of $\mathcal{E}$.
\item \textbf{Step 2.} Search $\mathcal{R}$ to find the restriction enzyme $e_{\max}^x$ that gives the largest number of distinct number of fragments when applied to all the gene sequences of $\mathcal{F}$.
\item \textbf{Step 3.} Use results obtained from the application of $e_{\max}^x$ to assemble species/strains of $\mathcal{F}$ into different groups according to the resulting number of fragments such that strain sequences that are split into the same number of fragments are grouped together in the same category. As an example, Figure 2 shows that the restriction enzyme $e_{\max}^x$ categorizes the species/strains of $\mathcal{F}$ into three different groups, namely, $G_1^1$, $G_2^1$, and $G_3^1$, where the superscript indicates the tree level (level 1). All species/strains in these groups are fragmented, using $e_{\max}^x$ into 3, 4, and 7 fragments, respectively.
\item \textbf{Step 4.} Apply Step 3 recursively to each resulting group consisting of more than one species/strains. For example, the illustrative example in Figure 2 shows that the first group in level 1, $G_1^1$, is then split into two different groups, $G_1^2$ and $G_2^2$ in level 2, where species/strains in these groups are fragmented into two and three fragments employing the restriction enzyme $e_{\max}^{G_1^2}$, respectively.
\end{itemize}
be possible to identify an unknown gene sequence $\mathcal{G}$ belonging to that is, GeneIdentify node following the restriction enzymes that should be employed first to cut $\mathcal{G}$, that is, $e_{\max}^\mathcal{G}$. Then, the number of fragments of all children (groups) of the current node (root) is retrieved and compared to the number of fragments resulting from cutting $\mathcal{G}$ using $e_{\max}^\mathcal{G}$. The node of the matched group is then visited and its associated restriction enzyme is retrieved and applied to $\mathcal{G}$ in order to decide which node has to be visited in the next level, and so on. This process is continued until a leaf node is met. If such a node is found, the processed gene sequence will be successfully identified as the species/strains at that (leaf) node. Otherwise, the identification process fails. As mentioned earlier, if there are no matching groups at any level of $T$, a different factor such as lengths of fragments could be tried and the identification process will continue.

The GeneIdentify algorithm can be illustrated further using the example shown in Figure 2. Let a strain $\mathcal{G}$ be one of the strains, referred to as strain #3, that belongs to $\mathcal{F}$. In this example (see dashed lines), $\mathcal{G}$ is identified by applying the following sequence of restriction enzymes: $e_{\max}^\mathcal{G}$, $e_{\max}^{G_1}$, and $e_{\max}^{G_2}$. This is because $\mathcal{G}$ is split into three fragments if $e_{\max}^\mathcal{G}$ is employed and two fragments if $e_{\max}^{G_1}$ is employed and no other species/strain is fragmented into the same number of fragments if $e_{\max}^{G_2}$ is employed to cut $\mathcal{F}$.

**Algorithm 2: GenIdentify($\mathcal{G}, T$).**

Step 5. The algorithm stops if either (1) the number of species/strains of all groups being processed is one or (2) no further application of any restriction enzyme can discriminate species/strains in groups containing more than one species/strains. The former case indicates that the algorithm can identify all species/strains of $\mathcal{F}$ using the adopted set of restriction enzymes $\mathcal{G}$. The second case, on the other hand, takes place if some species/strains cannot be identified employing $\mathcal{G}$. In this case, another factor, such as the fragment length, can be utilized to break any potential ties among unidentified species/strains.

Once an identification scheme $T$ is created for $\mathcal{F}$, it would be possible to identify an unknown gene sequence $\mathcal{G}$ as belonging to $\mathcal{F}$ or not by traversing $T$ starting from the root node following the GenIdentify algorithm (see Algorithm 2). The first step is to visit the root node of $T$ to specify the restriction enzyme that should be employed first to cut $\mathcal{G}$, that is, $e_{\max}^\mathcal{G}$. Then, the number of fragments of all children (groups) of the current node (root) is retrieved and compared to the number of fragments resulting from cutting $\mathcal{G}$ using $e_{\max}^\mathcal{G}$. The node of the matched group is then visited and its associated restriction enzyme is retrieved and applied to $\mathcal{G}$ in order to decide which node has to be visited in the next level, and so on. This process is continued until a leaf node is met. If such a node is found, the processed gene sequence will be successfully identified as the species/strains at that (leaf) node. Otherwise, the identification process fails. As mentioned earlier, if there are no matching groups at any level of $T$, a different factor such as lengths of fragments could be tried and the identification process will continue.

The GenIdentify algorithm can be illustrated further using the example shown in Figure 2. Let a strain $\mathcal{G}$ be one of the strains, referred to as strain #3, that belongs to $\mathcal{F}$. In this example (see dashed lines), $\mathcal{G}$ is identified by applying the following sequence of restriction enzymes: $e_{\max}^\mathcal{G}$, $e_{\max}^{G_1}$, and $e_{\max}^{G_2}$. This is because $\mathcal{G}$ is split into three fragments if $e_{\max}^\mathcal{G}$ is employed and two fragments if $e_{\max}^{G_1}$ is employed and no other species/strain is fragmented into the same number of fragments if $e_{\max}^{G_2}$ is employed to cut $\mathcal{F}$.

**4.1. Developing FN-Identify Method.** In order to develop our proposed method and algorithms, we used the 16S rRNA sequences of a population of 33 members of *Lactobacillus* (Table 1), an example of bacteria with genes with multiple copies in the genome (Table 2). FN-Identify and the two algorithms were able to identify and differentiate between the 33 species/strains based on the fragment numbers of the 16S rRNA sequences using six restriction enzymes (Figure 3, Supplementary Table 5). For a given species/strains a minimum of one enzyme and maximum of five enzymes were required for the identification (Figure 3 strains ID: 5 and 8, resp.). By adding the fragment length as a second factor, FN-Identify successfully identify and differentiate between the 33 species/strains using five restriction enzymes only. Furthermore, a maximum of three enzymes only was required for the identification of any given species/strains (Figure 4, Supplementary Table 6).

To further improve the identification efficiency of FN-Identify method and algorithms, we used the HSP60 genes as an example for genes with a single copy in the genome (Table 2). Genes represented with a single copy provide less...
variations in sequences (see above). Therefore FN-Identify might require more restriction enzymes to differentiate the 33 species/stains or even may fail in identifying some of them. However, with further tuning of the algorithms, FN-Identify shows comparable performance to what it does in the genes represented with a multiple copies (16S rRNA). It was able to identify the 33 species/stains based on the fragment numbers using six restriction enzymes (Supplementary Figure 2 and Supplementary Table 7). For a given species/stain a minimum of two enzymes and maximum of five enzymes were required for the identification (Supplementary Figure 2 strains ID: 24 and 33, resp.). When we used the fragment length as a second factor, FN-Identify required four restriction enzymes only to identify the 33 species/stains (Supplementary Figure 3 and Supplementary Table 8). Moreover, a maximum of three enzymes only was required for the identification of a given species/stain (Table 4, Supplementary Figure 3, and Supplementary Table 8). In some cases, the gene sequences and copy numbers of two strains are the same. Therefore neither FN-Identify nor the sequencing-based approach can differentiate them, such as strains *Lactobacillus rhamnosus ATCC 53103* and *Lactobacillus rhamnosus GG* (Table 1, strain IDs: 24 and 25) (Figures 3 and 4).

4.2. Testing and Assessment of FN-Identify Method. FN-Identify method and the two algorithms were developed using a training set of 33 members of *Lactobacillus* with two sets of gene sequences (16S rRNA and HSP60). To test FN-Identify method and algorithms performance, we assessed its identification efficiency using two different testing sets from two distinct bacterial groups *Mycobacterium* and *Pseudomonas*. *Mycobacterium* is a Gram-positive bacterial genus from the Mycobacteriaceae family that includes members that cause serious illness such as *Mycobacterium tuberculosis*, which causes tuberculosis. *Pseudomonas* is a Gram-negative bacterial genus from the Pseudomonadaceae that includes important model organisms such as *Pseudomonas aeruginosa*.

We obtained the sequences of the 16S rRNA genes of 22 members of *Mycobacterium* and 33 members of *Pseudomonas* using the same approach that we used with *Lactobacillus* (See Section 2). The variations in the 16S rRNA copy number and differences sequences between the multiple copies within the same genome appear in *Pseudomonas*, whereas the *Mycobacterium* genomes of the 22 members contain only one or two 16S rRNA copies (Supplementary Tables 2 and 4). We applied FN-Identify on the two testing datasets and FN-Identify successfully identified all the members of the two groups using the fragments numbers only and eight and seven enzymes to identify the 33 members of *Pseudomonas* and the 22 members of *Mycobacterium*, respectively (Table 4). Furthermore, for a given species, a maximum of eight and seven enzymes and minimum of seven and five enzymes were required to identify a given member of the *Pseudomonas* and *Mycobacterium* groups, respectively (Table 4, Supplementary Figures 4 and 6, and Supplementary Tables 9 and 11). By adding the fragment length as a second factor, FN-Identify successfully identifies the species/stains of the two groups using seven and four restriction enzymes for *Pseudomonas* and *Mycobacterium* groups, respectively. Furthermore, a maximum of seven and four enzymes and a minimum of four and three enzymes were required for the identification of any given species/stain (Table 4, Supplementary Figures 5 and 7, and Supplementary Tables 10 and 12).

Collectively, these results demonstrate the efficiency and utility of the FN-Identify method and the two developed algorithms in identifying bacterial species/stains within a genus and show that the method is applicable in bacterial groups with distinct properties.
Table 4: Summary of the employed training and testing datasets and FN-Identify performance.

| Bacterial group | Gram\(^1\) | Members | Unique sequences\(^2\) | 16S rRNA Required enzymes\(1\) factor | 2 factors | Max.-Min. Enzymes/species\(^3\) 1 factor | 2 factors | Unique sequences\(^2\) | HSP60 Required enzymes\(1\) factor | 2 factors | Max.-Min. Enzymes/species\(^3\) 1 factor | 2 factors |
|-----------------|-----------|---------|------------------------|--------------------------------------|----------|----------------------------------------|----------|------------------------|--------------------------------------|----------|----------------------------------------|----------|
| Training set    |           |         |                        |                                      |          |                                        |          |                        |                                      |          |                                        |          |
| Lactobacillus   | P         | 33      | 24                     | 6                                    | 5        | 6-6                                    | 5-3      | 23                     | 6                                    | 5        | 4-1                                    | 3-1      |
| Testing sets    |           |         |                        |                                      |          |                                        |          |                        |                                      |          |                                        |          |
| Pseudomonas     | N         | 33      | 32                     | 8                                    | 6        | 8-7                                    | 7-4      | —                      | —                                    | —        | —                                      | —        |
| Mycobacterium   | P         | 22      | 18                     | 7                                    | 4        | 7-5                                    | 4-3      | —                      | —                                    | —        | —                                      | —        |

\(^1\)P: positive and N: negative.

\(^2\)Members with differences in 16S rRNA sequences. In some cases two or more members have 100% similarity in 16S rRNA sequences. Those members are considered as one entry to FN-Identify.

\(^3\)The maximum and minimum number of enzymes required identifying a given member of the group.
### 4.3. Applications and Future Perspective.

The assessment of FN-Identify method and the two developed algorithms shows the potentials of the method, with standard microbiology protocols and instruments. FN-Identify is a computational method that is designed as an aid that helps designing and minimizing the experimental procedures required for bacterial identification. Ideally, FN-Identify interfaces with the experimental and clinical workflows through receiving inputs (expected bacterial group, gene(s) to be used for identification, and list of restriction enzymes) and provides outputs that lead the later bench exterminates (list and order of enzymatic restriction experiments and the identification scheme that is used to interpret the experimental results).

To be fully utilized, FN-Identify needs a software tool that is connected with a database of gene sequence (e.g., 16S rRNA and HSP60) in different bacterial families and database of restriction enzymes. The software should implement the two algorithms and automate the selection of the species and the enzymes as well as automating building the restriction map and the identifying scheme. We are currently building this tool as a webserver that provides these services for free to enable the scientific community in the developing countries to utilize FN-Identify.

### 5. Conclusion

Bacterial identification is an important routine that is required in several microbiological and environmental applications and research. The current techniques are highly dependent on genome sequencing techniques that target certain genes that present almost in all bacterial species. Although the genome sequencing techniques observed outstanding improvements in accuracy and decrease in cost, developing countries remain far from employing these indispensable technologies due to several barriers. Therefore, alternative sequencing-independent methods are required to facilitate the needed tasks with affordable costs and using the available facilities. We developed FN-Identify method, a sequencing-independent method for bacterial identification, using standard microbiological protocols and instruments, restriction enzymes, and two algorithms that we developed (CreateScheme and GeneIdentify). FN-Identify was tested against standard bacterial populations of 22 and 33 bacterial species/strains of the Mycobacterium and Pseudomonas groups, respectively. The method successfully differentiate all the species/strains in two independent analyses using two different genes 16S rRNA and HSP60 for each of the two groups. A webserver is being developed for FN-Identify to automate the scheme building and maximize the utilization of the method. We believe that FN-Identify is a useful alternative to the sequencing methods when they are out of reach.

### Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.
References

[1] X. X. Ning, "Bacterial identification based on 16S ribosomal RNA gene sequence analysis," in Advanced Techniques in Diagnostic Microbiology, pp. 323–332, Springer, Boston, Mass, USA, 2006.

[2] G. M. Garrity, D. J. Brenner, N. R. Krieg, and J. T. Staley, Bergey’s Manual of Systematic Bacteriology—Volume 2: The Proteobacteria, Part B: The Gammaproteobacteria, Springer, 2005.

[3] Y. - W. Tang, G. W. Procop, and D. H. Persing, "Molecular diagnostics of infectious diseases," Clinical Chemistry, vol. 43, no. 11, pp. 2021–2038, 1997.

[4] E. Tietze and H. Tschape, "Plasmid pattern analysis of natural bacterial isolates and its epidemiological implication," Journal of Hygiene, vol. 90, no. 3, pp. 475–488, 1983.

[5] D. Paffett, C. Barberio, E. Casalone et al., "DNA fingerprinting by random amplified polymorphic DNA and restriction fragment length polymorphism is useful for yeast typing," Research in Microbiology, vol. 146, no. 7, pp. 587–594, 1995.

[6] D. C. Schwartz and C. R. Cantor, "Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis," Cell, vol. 37, no. 1, pp. 67–75, 1984.

[7] J. G. K. Williams, A. R. Kubelik, J. I. Livak, J. A. Rafalski, and S. V. Tingey, "DNA polymorphisms amplified by arbitrary primers are useful as genetic markers," Nucleic Acids Research, vol. 18, no. 22, pp. 6531–6535, 1990.

[8] G. J. Olsen, D. J. Lane, S. J. Giovannoni, N. R. Pace, and D. A. Stahl, "Microbial ecology and evolution: a ribosomal RNA approach," Annual Review of Microbiology, vol. 40, pp. 337–365, 1986.

[9] R. Amann and W. Ludwig, "Ribosomal RNA-targeted nucleic acid probes for studies in microbial ecology," FEMS Microbiology Reviews, vol. 24, no. 5, pp. 555–563, 2000.

[10] H. K. Reee and R. A. Zimmermann, "Organization and expression of the 16S, 23S and 5S ribosomal RNA genes from the archaebacterium Thermoplasm acidophilum," Nucleic Acids Research, vol. 18, no. 15, pp. 4471–4478, 1990.

[11] R. J. Case, Y. Boucher, I. Dahllof, C. Holmström, W. F. Doolittle, and S. Kjelleberg, "Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies," Applied and Environmental Microbiology, vol. 73, no. 1, pp. 278–288, 2007.

[12] M. G. Block and A. Ouellette, "Genetic identification of eleven aquatic bacteria using the 16S rDNA gene," Journal of Research Across the Disciplines, pp. 1–46, 2012.

[13] F. Grimont and P. A. D. Grimont, "Ribosomal ribonucleic acid gene restriction patterns as potential taxonomic tools," Annales de l’Institut Pasteur / Microbiologie, vol. 137, no. 1, supplement 2, pp. 165–175, 1986.

[14] G. Muzyer, E. C. de Waal, and A. G. Uitterlinden, "Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA," Applied and Environmental Microbiology, vol. 59, no. 3, pp. 695–700, 1993.

[15] G. Muzyer, "DGGE/TGGE a method for identifying genes from natural ecosystems," Current Opinion in Microbiology, vol. 2, no. 3, pp. 317–322, 1999.

[16] M. Vaneechoutte, H. De Beenhouwer, G. Claeys et al., "Identification of Mycobacterium species by using amplified ribosomal DNA restriction analysis," Journal of Clinical Microbiology, vol. 31, no. 8, pp. 2061–2065, 1993.

[17] W. - T. Liu, T. L. Marsh, H. Cheng, and L. J. Forney, "Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA," Applied and Environmental Microbiology, vol. 63, no. 11, pp. 4516–4522, 1997.

[18] S. D. Brugger, L. Frei, P. M. Frey, S. Aebi, K. Mühlemann, and M. Hilty, "16S rRNA terminal restriction fragment length polymorphism for the characterization of the nasopharyngeal microbiota," PLoS ONE, vol. 7, no. 12, Article ID e52241, 2012.

[19] L. Chen, Y. Cai, G. Zhou et al., "Rapid Sanger sequencing of the 16S rRNA gene for identification of some common pathogens," PLoS ONE, vol. 9, no. 2, Article ID e88866, 2014.

[20] D. Tewari, S. Cieply, and J. Livengood, "Identification of bacteria recovered from animals using the 16S ribosomal RNA gene with pyrosequencing and Sanger sequencing," Journal of Veterinary Diagnostic Investigation, vol. 23, no. 6, pp. 1104–1108, 2011.

[21] P. C. Y. Woo, J. L. L. Teng, J. K. L. Wu et al., "Guidelines for interpretation of 16S rRNA gene sequence-based results for identification of medically important aerobic Gram-positive bacteria," Journal of Medical Microbiology, vol. 58, no. 8, pp. 1030–1036, 2009.

[22] J. G. Caporaso, C. L. Lauber, W. A. Walters et al., "Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample," Proceedings of the National Academy of Sciences of the United States of America, vol. 108, supplement 1, pp. 4516–4522, 2011.

[23] S. El-Metwally, O. M. Ouda, and M. Helmy, Next Generation Sequencing Technologies and Challenges in Sequence Assembly, vol. 7, Springer, New York, NY, USA, 1st edition, 2014.

[24] Human Genetics Programme and Chronic Diseases and Health Promotion, Genetics, Genomics and the Patenting of DNA, World Health Organization, Geneva, Switzerland, 2005.

[25] A. Alwan and B. Modell, "Recommendations for introducing genetics services in developing countries," Nature Reviews Genetics, vol. 4, no. 1, pp. 61–68, 2003.

[26] K. E. Nelson and B. Jones-Nelson, Eds., Genomics Applications for the Developing World, Springer, New York, NY, USA, 1st edition, 2012.

[27] L. M. Parsons, Â. Somoskovi, C. Gutierrez et al., "Laboratory diagnosis of tuberculosis in resource-poor Countries: challenges and opportunities," Clinical Microbiology Reviews, vol. 24, no. 2, pp. 314–350, 2011.

[28] J. R. Cole, B. Chai, R. J. Farris et al., "The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis," Nucleic Acids Research, vol. 33, pp. D294–D296, 2005.

[29] C. Quast, E. Pruesse, P. Yilmaz et al., "The SILVA ribosomal RNA gene database project: improved data processing and web-based tools," Nucleic Acids Research, vol. 41, no. 1, pp. D590–D596, 2013.

[30] T. Z. DeSantis, P. Hugenholtz, N. Larsen et al., "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB," Applied and Environmental Microbiology, vol. 72, no. 7, pp. 5069–5072, 2006.

[31] U. Edwards, T. Rogall, H. Blöcker, M. Emde, and E. C. Böttger, "Isolation and direct complete nucleotide determination of entire genes. Characterization of a gene coding for 16S ribosomal RNA," Nucleic Acids Research, vol. 17, no. 19, pp. 7843–7853, 1989.

[32] A. L. Reysenbach, L. J. Giver, G. S. Wickham, and N. R. Pace, "Differential amplification of rRNA genes by polymerase chain reaction," Applied and Environmental Microbiology, vol. 58, no. 10, pp. 3417–3418, 1992.
[33] S. Weidner, W. Arnold, and A. Pühler, “Diversity of uncultured microorganisms associated with the seagrass Halophila stipulacea estimated by restriction fragment length polymorphism analysis of PCR-amplified 16S rRNA genes,” Applied and Environmental Microbiology, vol. 62, no. 3, pp. 766–771, 1996.

[34] S. Turner, K. M. Pryer, V. P. W. Miao, and J. D. Palmer, “Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis,” The Journal of Eukaryotic Microbiology, vol. 46, no. 4, pp. 327–338, 1999.

[35] C. Vannini, G. Rosati, F. Verni, and G. Petroni, “Identification of the bacterial endosymbionts of the marine ciliate Euplotes magnicirratus (Ciliophora, Hypotrichia) and proposal of ‘Candidatus Devosia euplotis’,” International Journal of Systematic and Evolutionary Microbiology, vol. 54, no. 4, pp. 1151–1156, 2004.

[36] T. D. Leser, J. Z. Amenuvor, T. K. Jensen, R. H. Lindecrona, M. Boye, and K. Møller, “Culture-independent analysis of gut bacteria: the pig gastrointestinal tract microbiota revisited,” Applied and Environmental Microbiology, vol. 68, no. 2, pp. 673–690, 2002.

[37] Y. Sekiguchi, H. Takahashi, Y. Kamagata, A. Ohashi, and H. Harada, “In situ detection, isolation, and physiological properties of a thin filamentous microorganism abundant in methanogenic granular sludges: a novel isolate affiliated with a clone cluster, the green non-sulfur bacteria, subdivision I,” Applied and Environmental Microbiology, vol. 67, no. 12, pp. 5740–5749, 2001.

[38] W. G. Weisburg, S. M. Barns, D. A. Pelletier, and D. J. Lane, “16S Ribosomal DNA amplification for phylogenetic study,” Journal of Bacteriology, vol. 173, no. 2, pp. 697–703, 1991.

[39] R. J. Roberts, “Restriction and modification enzymes and their recognition sequences,” Nucleic Acids Research, vol. 8, no. 1, pp. r63–r80, 1980.

[40] R. J. Roberts, T. Vincze, J. Posfai, and D. Macelis, “REBASE—a database for DNA restriction and modification: enzymes, genes and genomes,” Nucleic Acids Research, vol. 38, supplement 1, pp. D234–D236, 2010.

[41] B. Séguin, B.-J. Hardy, P. A. Singer, and A. S. Daar, “Human genomic variation initiatives in emerging economies and developing countries,” Nature Reviews Genetics, vol. 9, no. 1, pp. S3–S4, 2008.

[42] D. R. Cullimore, Practical Atlas for Bacterial Identification, CRC Press, 2nd edition, 2010.

[43] S. R. Herbel, W. Vahjen, L. H. Wieler, and S. Guenther, “Timely approaches to identify probiotic species of the genus Lactobacillus,” Gut Pathogens, vol. 5, no. 1, article 27, 2013.

[44] E. López-Huertas, “Safety and efficacy of human breast milk Lactobacillus fermentum CECT 5716. A mini-review of studies with infant formulae,” Beneficial Microbes, vol. 6, no. 2, pp. 219–224, 2015.

[45] W. A. Walker and R. S. Iyengar, “Breast milk, microbiota, and intestinal immune homeostasis,” Pediatric Research, vol. 77, pp. 220–228, 2015.

[46] L. Morelli, “Yogurt, living cultures, and gut health,” The American Journal of Clinical Nutrition, vol. 99, no. 5, supplement, pp. 1248S–1250S, 2014.

[47] D. A. Savaiano, “Lactose digestion from yogurt: mechanism and relevance,” American Journal of Clinical Nutrition, vol. 99, no. 5, supplement, pp. 1251S–1255S, 2014.

[48] M. Hannig, G. Braker, J. Dippner, and K. Jürgens, “Linking denitrifier community structure and prevalent biogeochemical parameters in the pelagial of the central Baltic Proper (Baltic Sea),” FEMS Microbiology Ecology, vol. 57, no. 2, pp. 260–271, 2006.

[49] W. Sun, “General procedures,” in Molecular Diagnostics, W. W. Grody, R. M. Nakamura, C. M. Strom, and F. L. Kiechle, Eds., pp. 49–57, Elsevier, 2010.