HAL
open science

Introducing SPeDE: High-Throughput Dereplication and Accurate Determination of Microbial Diversity from Matrix-Assisted Laser Desorption–Ionization Time of Flight Mass Spectrometry Data

Charles Dumolin, Maarten Aerts, Bart Verheyde, Simon Schellaert, Tim Vandamme, Felix van Der Jeugt, Evelien de Canck, Margo Cnockaert, Anneleen Wieme, Ilse Cleenwerck, et al.

To cite this version:

Charles Dumolin, Maarten Aerts, Bart Verheyde, Simon Schellaert, Tim Vandamme, et al.. Introducing SPeDE: High-Throughput Dereplication and Accurate Determination of Microbial Diversity from Matrix-Assisted Laser Desorption–Ionization Time of Flight Mass Spectrometry Data. mSystems, 2019, 4 (5), pp.e00437-19. 10.1128/mSystems.00437-19 . hal-02961994

HAL Id: hal-02961994

https://hal.inrae.fr/hal-02961994v1

Submitted on 14 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

METHODS AND PROTOCOLS
Applied and Environmental Science

Introducing SPeDE: High-Throughput Dereplication and Accurate Determination of Microbial Diversity from Matrix-Assisted Laser Desorption–Ionization Time of Flight Mass Spectrometry Data

Charles Dumolin,[a] Maarten Aerts,[a] Bart Verheyde,[a] Simon Schellaert,[b] Tim Vandamme,[a] Felix Van der Jeugt,[b] Evelien De Canck,[a] Margo Cnockaert,[a] Anneleen D. Wieme,[a,c] Ilse Cleenwerck,[a,c] Jindrich Peiren,[a,c] Peter Dawyndt,[b] Peter Vandamme,[a,c] Aurélien Carlier[a]

[a]Laboratory of Microbiology, Department of Biochemistry and Microbiology, Faculty of Sciences, Ghent University, Ghent, Belgium
[b]Computational Biology Laboratory, Department of Applied Mathematics, Computer Science and Statistics, Faculty of Sciences, Ghent University, Ghent, Belgium
[c]BCCM/LMG Bacteria Collection, Department of Biochemistry and Microbiology, Faculty of Sciences, Ghent University, Ghent, Belgium

ABSTRACT The isolation of microorganisms from microbial community samples often yields a large number of conspecific isolates. Increasing the diversity covered by an isolate collection entails the implementation of methods and protocols to minimize the number of redundant isolates. Matrix-assisted laser desorption–ionization time-of-flight (MALDI-TOF) mass spectrometry methods are ideally suited to this dereplication problem because of their low cost and high throughput. However, the available software tools are cumbersome and rely either on the prior development of reference databases or on global similarity analyses, which are inconvenient and offer low taxonomic resolution. We introduce SPeDE, a user-friendly spectral data analysis tool for the dereplication of MALDI-TOF mass spectra. Rather than relying on global similarity approaches to classify spectra, SPeDE determines the number of unique spectral features by a mix of global and local peak comparisons. This approach allows the identification of a set of nonredundant spectra linked to operational isolation units. We evaluated SPeDE on a data set of 5,228 spectra representing 167 bacterial strains belonging to 132 genera across six phyla and on a data set of 312 spectra of 78 strains measured before and after lyophilization and subculturing. SPeDE was able to dereplicate with high efficiency by identifying redundant spectra while retrieving reference spectra for all strains in a sample. SPeDE can identify distinguishing features between spectra, and its performance exceeds that of established methods in speed and precision. SPeDE is open source under the MIT license and is available from https://github.com/LM-UGent/SPeDE.

IMPORTANCE Estimation of the operational isolation units present in a MALDI-TOF mass spectral data set involves an essential dereplication step to identify redundant spectra in a rapid manner and without sacrificing biological resolution. We describe SPeDE, a new algorithm which facilitates culture-dependent clinical or environmental studies. SPeDE enables the rapid analysis and dereplication of isolates, a critical feature when long-term storage of cultures is limited or not feasible. We show that SPeDE can efficiently identify sets of similar spectra at the level of the species or strain, exceeding the taxonomic resolution of other methods. The high-throughput capacity, speed, and low cost of MALDI-TOF mass spectrometry and SPeDE dereplication over traditional gene marker-based sequencing approaches should facilitate adoption of the culturomics approach to bacterial isolation campaigns.

KEYWORDS bioinformatics, MALDI-TOF MS, dereplication, microbial ecology

Citation Dumolin C, Aerts M, Verheyde B, Schellaert S, Vandamme T, Van der Jeugt F, De Canck E, Cnockaert M, Wieme AD, Cleenwerck I, Peiren J, Dawyndt P, Vandamme P, Carlier A. 2019. Introducing SPeDE: high-throughput dereplication and accurate determination of microbial diversity from matrix-assisted laser desorption–ionization time of flight mass spectrometry data. mSystems 4:e00437-19. https://doi.org/10.1128/mSystems.00437-19.

Editor Rachel Mackelprang, California State University, Northridge

Copyright © 2019 Dumolin et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to Aurélien Carlier, aurelien.carlier@ugent.be.

C.D. and M.A. contributed equally to this article.

SPeDE: Enabling high-throughput dereplication of MALDI-TOF mass spectra of bacterial isolates @AurelienCarlier @DumolinC

Received 17 July 2019
Accepted 23 August 2019
Published 10 September 2019

Downloaded from http://msystems.asm.org/ on October 14, 2020 by guest

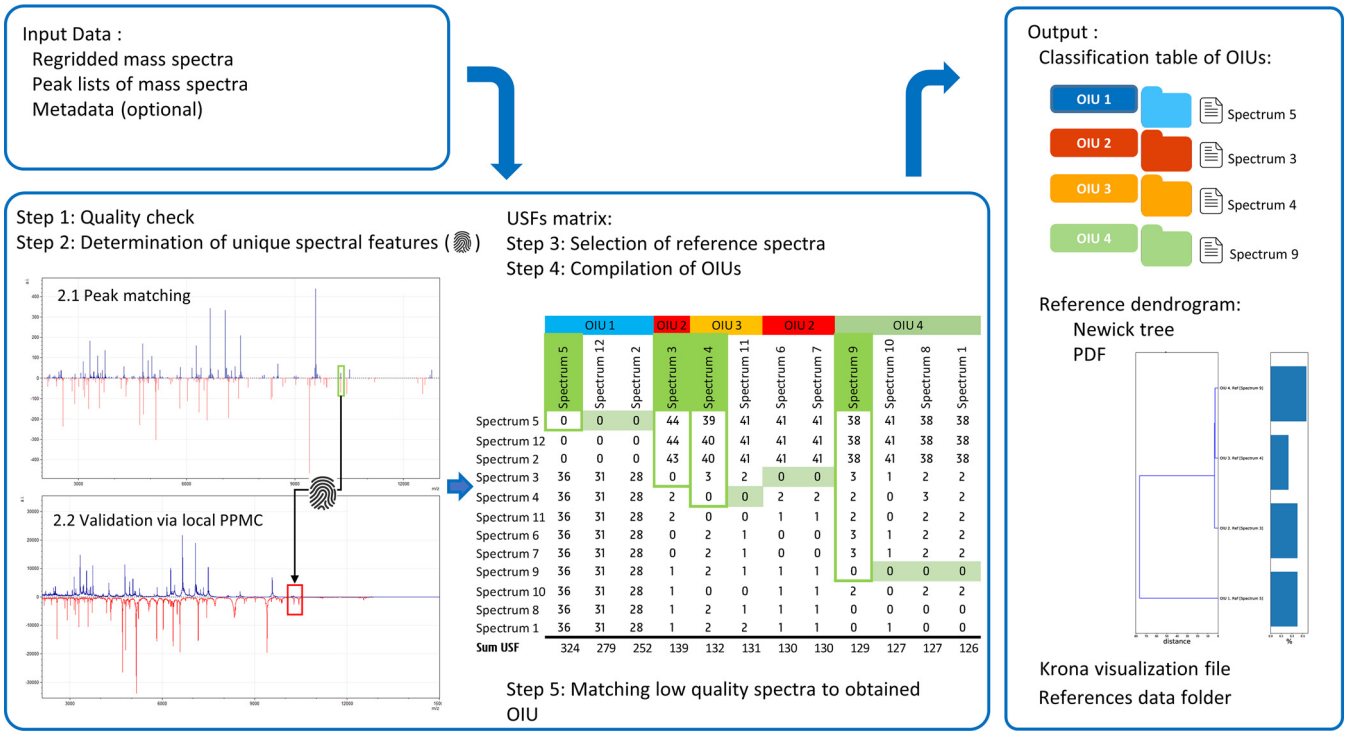
The composition and functioning of the microbiome have been linked to the development of diseases and the performance of industrial processes and have recently emerged as key drivers of agricultural yields and plant health (1–5). Describing microbial diversity is key to understanding niche functioning and is now routinely carried out using culture-independent techniques, such as shotgun or amplicon metagenomics. These techniques allow the determination of the relative abundances of operational taxonomic units (OTUs) and the comparison of large cohorts (6). Metagenomic analyses on environmental samples further confirmed the so-called great plate count anomaly, which is the observation that many microbes seen under the microscope escape cultivation under laboratory conditions (7). Uncultivated bacteria can now be identified and genomes can be reconstructed, allowing the formulation of hypotheses regarding microbe-environment, microbe-microbe, and microbe-host interactions. While these systems biology approaches can be powerful, they are limited in scope and the discovery of fundamentally new processes requires cultures (8). Furthermore, minor microbiota components may have a substantial influence and are not easily captured by culture-independent methods (9). Recently, renewed interest in microbial diversity has led to the development of new methods to overcome the shortcomings of conventional cultivation techniques (10, 11).

High-throughput culture-based methods, also called culturomics, aim to recover the diversity of cultivable microorganisms present in a sample (12). In a landmark study, Lagier and colleagues cultured more than 900,000 isolates from the human gut (13). These included over 77% of all prokaryotes previously identified in the human gut, in addition to organisms corresponding to previously unassigned OTUs (13). However, the application of high-throughput cultivation techniques leads to the isolation of multiple conspecific strains or genetically identical clones, inflating the downstream cost of analyses. To control cost, various dereplication methods can be implemented. These typically involve typing and filtering out conspecifics based on 16S rRNA gene sequence or matrix-assisted laser desorption–ionization time-of-flight (MALDI-TOF) mass spectrometry (MS) profiles (14–16). MALDI-TOF MS has proven to be a fast and cost-effective method for dereplication and identification when coupled to a profile database with extensive taxonomic coverage of a sample, with as little as 0.1% of isolates from the gut microbiota further requiring identification via 16S rRNA gene sequencing (17). Furthermore, recent technical and technological advances have enabled MALDI-TOF MS profiling of bacteria at the strain level (18), indicating a taxonomic resolution potentially exceeding that of 16S rRNA sequencing for some taxa. However, commercial MALDI-TOF MS databases are mostly populated with spectra for isolates of clinical or food safety relevance, which makes identification coupled to dereplication impractical for environmental samples containing a high diversity of rarely sampled bacteria (19). Retroactively updating the database with spectra from newly cultured isolates is possible but time-consuming, compromising the throughput of the culturomics approach. In addition, extending commercial databases may involve lengthy clearance procedures by regulatory agencies, such as the U.S. Food and Drug Administration.

Methods based on global similarity measures combined with hierarchical clustering of mass spectra have been developed to uncouple dereplication from identification (15, 20). A shortcoming of this strategy is that the identification of redundant profiles relies either on visual inspection of dendrograms or on the use of a predefined distance cutoff value to delineate clusters of similar spectra (21, 22). Predefined cutoff values do not consider the variability of profiles between taxa and thus need to be adjusted according to the taxonomic composition of the samples. Both shortcomings limit the taxonomic resolution and require user intervention, which contribute to making dereplication analysis time-consuming and subject to reproducibility issues. To our knowledge, no fast and user-friendly tool has been developed to dereplicate large sets of spectral data obtained from culturomics studies without relying on prior identification.

We introduce SPeDE, an algorithm and software implementation enabling high-throughput isolate dereplication using comparison of MALDI-TOF MS profiles. SPeDE

Downloaded from http://msystems.asm.org/ on October 14, 2020 by guest



	OIU 1			OIU 2	OIU 3		OIU 2		OIU 4			
	Spectrum 5	Spectrum 12	Spectrum 2	Spectrum 3	Spectrum 4	Spectrum 11	Spectrum 6	Spectrum 7	Spectrum 9	Spectrum 10	Spectrum 8	Spectrum 1
Spectrum 5	0	0	0	44	39	41	41	41	38	41	38	38
Spectrum 12	0	0	0	44	40	41	41	41	38	41	38	38
Spectrum 2	0	0	0	43	40	41	41	41	38	41	38	38
Spectrum 3	36	31	28	0	3	2	0	0	3	1	2	2
Spectrum 4	36	31	28	2	0	0	2	2	2	0	3	2
Spectrum 11	36	31	28	2	0	0	1	1	2	0	2	2
Spectrum 6	36	31	28	0	2	1	0	0	3	1	2	2
Spectrum 7	36	31	28	0	2	1	0	0	3	1	2	2
Spectrum 9	36	31	28	1	2	1	1	1	0	0	0	0
Spectrum 10	36	31	28	1	0	0	1	1	2	0	2	2
Spectrum 8	36	31	28	1	2	1	1	1	0	0	0	0
Spectrum 1	36	31	28	1	2	2	1	1	0	1	0	0
Sum USF	324	279	252	139	132	131	130	130	129	127	127	126

FIG 1 Schematic representation of the SPeDE algorithm. See Materials and Methods for a detailed description of the algorithm. Briefly, all possible pairs of peak lists in a data set are compared (step 2.1). Peaks which are not shared by a pair of spectra are validated by estimating the Pearson product moment correlation (PPMC) between raw spectra in a local area surrounding the peak (step 2.2). Peaks with a PPMC below a threshold value are considered discriminating. The number of discriminating peaks or unique spectral features (USFs) between pairs of spectra is computed and tabulated (step 3). Pairs of spectra for which no USFs are found in at least one of the elements are matched and clustered into operational isolation units (OIUs; step 4). All spectra with a quality too poor to be considered for inclusion as a reference spectrum are matched to an OIU to give a reliable abundance estimate for each OIU. The output of SPeDE includes a table of representative spectra for each OIU and a USF distance matrix between all spectra which can be used to generate a dendrogram or a Krona plot.

discriminates MS spectra through the detection of unique spectral features with adjustable sensitivity. We validated the program with a data set of more than 5,000 spectra obtained from 167 different strains belonging to 132 genera across six phyla, for which we also obtained whole-genome sequences.

RESULTS

Rationale and purpose of SPeDE. We designed SPeDE to identify isolates from recurrent taxa in culture collections in a time- and cost-efficient manner. MALDI-TOF MS spectra are compared in a pairwise manner using local and global measures. SPeDE does not attempt to classify the samples according to a reference database but instead computes the number of exclusive spectral features between pairs of spectra. Pairs of spectra for which all features of one element are shared with the other are considered redundant and grouped in single operational isolation units (OIUs) represented by a reference spectrum. Program parameters affect the sensitivity at which discriminating features are detected, allowing various degrees of taxonomic resolution.

Optimization of similarity threshold parameters for the identification of nonredundant spectra. The SPeDE algorithm relies on peak matching coupled to spectrum similarity in a focused area around the peaks to determine unique spectral features (USFs). For details of the algorithm, see the Materials and Methods section and Fig. 1. For each pair of spectra, peaks are matched if they fall within a predetermined peak accuracy window, calculated by the *m/z* difference between peak centroids. Peaks matched and peaks considered to be unique for one of the spectra are validated by calculating the Pearson product moment correlation (PPMC) between raw spectra in a local area around each peak. Spectral features are considered unique if the PPMC value is below a preset threshold. This step allows us to detect peaks which may have been

Downloaded from http://msystems.asm.org/ on October 14, 2020 by guest

missed by the peak-calling algorithm with high sensitivity, reducing the risk of erroneously identifying discriminating features. Both the size of the peak accuracy window and the PPMC value threshold are expected to affect the outcome of peak matching and, by extension, the outcome of the dereplication process.

To determine optimal values for the PPMC threshold and peak accuracy window, we applied SPeDE to a benchmark data set of 167 strains covering 132 genera and 143 species. These included several species of *Burkholderia* and *Lactobacillus*, which are notoriously difficult to discriminate by mass spectrometry (23–25). To account for closely related strains in the estimation of true- and false-positive spectral matches, we divided the data set into 149 OTUs, which we defined as groups of strains with an intragroup pairwise genome-wide average nucleotide identity (ANI) of >98%. This ANI threshold was empirically determined to give the widest precision range in response to SPeDE parameter changes. We obtained 19 to 32 spectra for each strain, for a total of 5,228 spectra. A quality assessment routine built in the SPeDE algorithm rejected 28 spectra as poor quality, as they contained less than 5 peaks with a signal-to-noise ratio (S/N) of >30. The total number of spectra used for benchmarking of the algorithm was thus 5,200.

Overall, varying the size of the peak accuracy window from 500 to 1,000 ppm had a negligible impact on the dereplication ratio (taken as the ratio of the number of OTUs/number of reference spectra) or on the precision, i.e., the ability of the algorithm to correctly match the spectra to an OTU (see Fig. S1 in the supplemental material). Varying the local PPMC threshold value had the most impact on overall performance. Increasing the local PPMC threshold values resulted in higher precision, but this came at the cost of lower dereplication ratios (Fig. S1). Gains in precision were incremental only for PPMC threshold values of >50%, rising from 95.3% to a maximum of 99.8%. In contrast, the dereplication ratio dipped rapidly from 70.5% to below 50% at local PPMC threshold values of 50% and 66%, respectively (Fig. S1A and B).

Therefore, we selected a local accuracy window of 700 ppm and a local PPMC threshold of 50% as the default parameter settings which offered the best compromise between precision (95.3%) and dereplication (70.5%).

We were also interested in determining optimal parameter values for the clustering of conspecific strains in a data set. We repeated our parameter search, this time defining an $OTU_{95\%}$ as a group of strains with an intragroup pairwise genome-wide average nucleotide identity (ANI) of >95%, a commonly accepted threshold for the circumscription of taxonomic species (26, 27) (Fig. S1C and D). With this definition, our benchmarking data set contained 150 $OTU_{95\%}$ values, each corresponding to a biological species. Precision values reached a plateau for local PPMC threshold values of 50 and above (precision > 99.5%), corresponding to dereplication ratios of <68.5%. In contrast, the dereplication ratio decreased rapidly at local PPMC threshold values of >20, with precision remaining >97.8%. We thus recommend setting the local PPMC threshold parameter value at 20 if dereplication at the species level is desired.

Performance and taxonomic resolution of SPeDE. Our benchmark data set contained 5,200 MALDI-TOF mass spectra passing the SPeDE quality control. These mass spectra represented 167 strains. Dereplication analysis of these spectra with a PPMC threshold of 50% and an accuracy window width of 700 ppm resulted in 210 distinct clusters of spectra, or OIUs. A representative spectrum was picked for each cluster (see Materials and Methods for details), which we refer to here as the reference spectrum (Fig. 2; Table S1). This corresponded to a reduction of 96.0% of the spectra analyzed. Moreover, 160/167 (95.8%) of the strains were represented by reference spectra. In total, 123 (73.7%) of all strains were represented by a single reference spectrum, 31 (18.6%) were represented by two reference spectra, and 3 (1.8%) were represented by three reference spectra. Only seven strains did not yield any reference spectra. Furthermore, obtaining multiple reference spectra for a single strain did not seem to be dependent on the phylogenetic placement of the taxon (Fig. 2). The number of USFs between the spectra of distinct strains was always higher than the

Downloaded from http://msystems.asm.org/ on October 14, 2020 by guest

Downloaded from http://msystems.asm.org/ on October 14, 2020 by guest

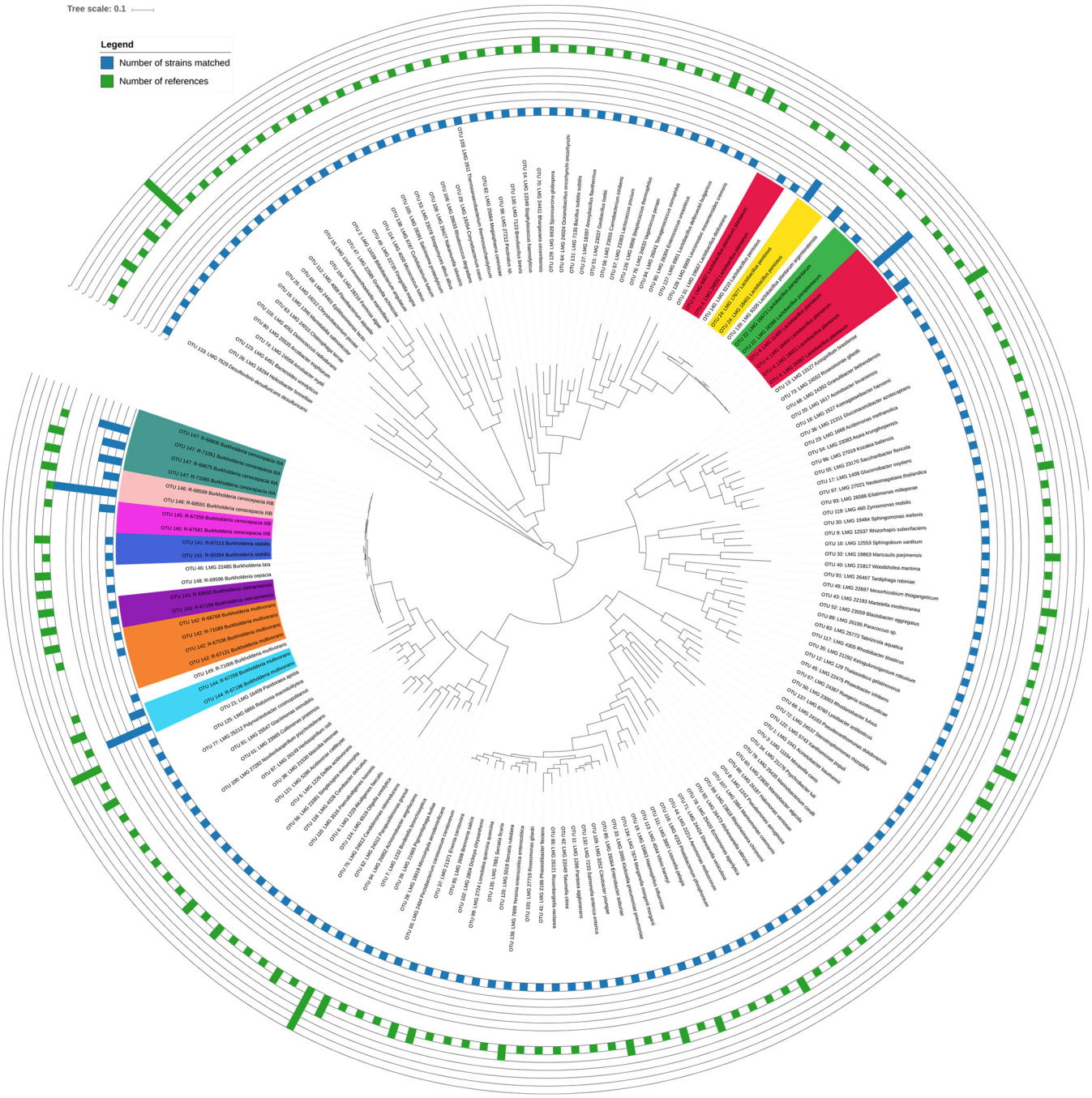


FIG 2 Approximate maximum likelihood phylogenetic tree of strains included in the benchmark data set based on 40 single-copy, conserved marker protein genes. OTUs were defined as groups of strains with an intragroup pairwise genome-wide ANI of >98%. OTU clusters containing more than one strain are highlighted. The number of the references obtained per strain are indicated by green bars, and the number of strains linked to each reference are indicated by blue bars.

number of USFs between the reference spectra originating from the same strain (Fig. S2). To determine the resolution at which SPeDE was able to discriminate between strains, we calculated the genomic distance between distinct strains whose spectra were matched to a single reference (Fig. 3A). Overall, spectra from the seven strains which were not represented in the final set matched the reference spectra from strains with an average pairwise ANI of 98.4% ± 1.1%. Finally, over 93.8% of OIUs represented a single strain (Fig. 3B). Only 13/210 OIUs included multiple strains; of these, 10 had minimal intra-OIU ANI values of >97% and 3 had minimal intra-OIU ANI values of

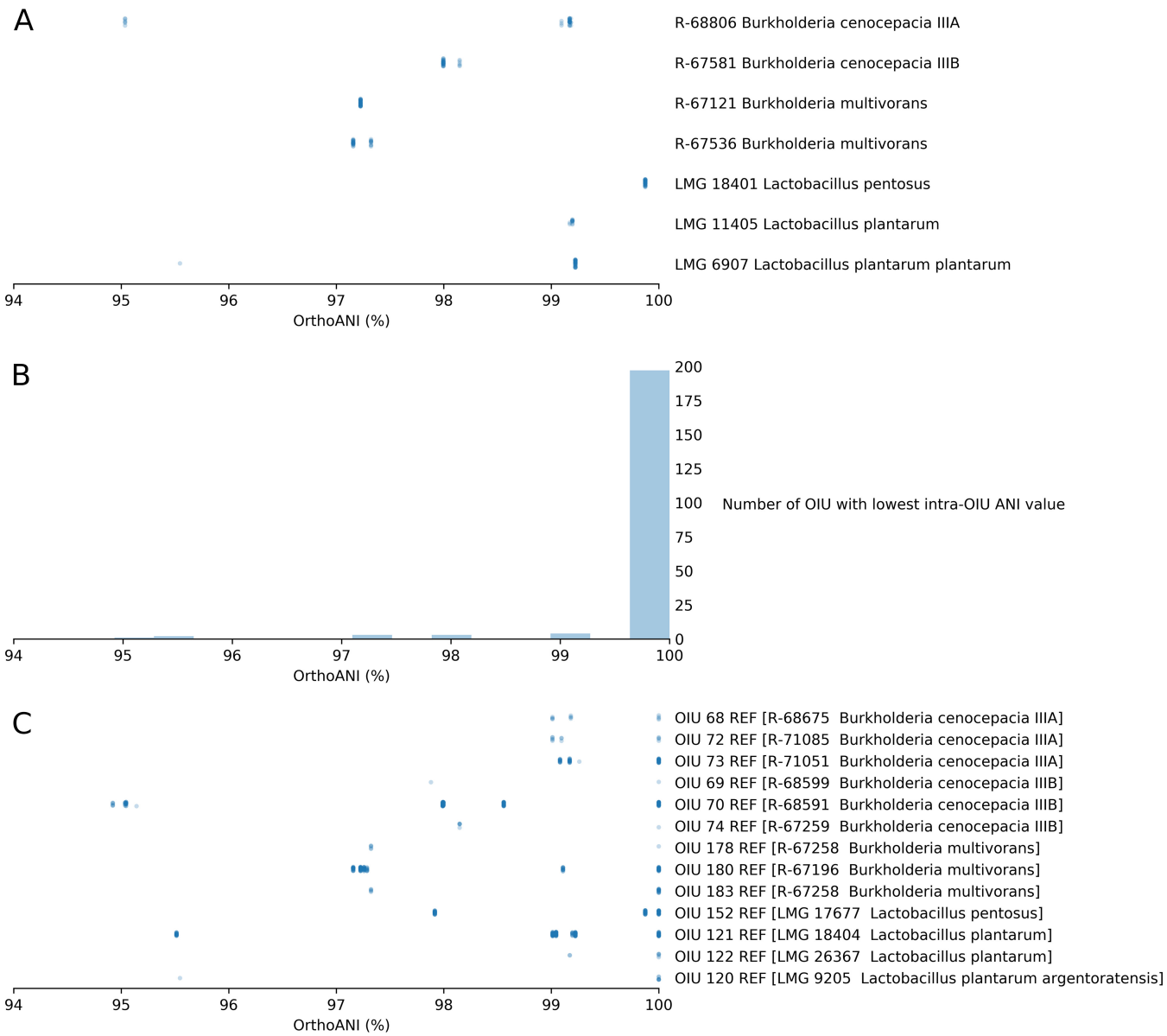


FIG 3 Accuracy of MALDI-TOF mass spectra matching by SPeDE. (A) Genomic similarity of the 7 strains of the benchmark data set for which no reference spectrum was retained to strains within the same OIU. Genomic similarities expressed by ANI values are shown. (B) Distribution of the lowest ANI values within 210 OIUs. Bins have a width of 0.36%. (C) Genomic similarity within OIUs composed of more than one strain. Each data point corresponds to the ANI value between a pair of strains contained within the OIU.

>94.9% (Fig. 3C). We were thus able to reduce a complex data set of over 5,000 experimentally acquired spectra to a final set of 210 OIUs without a significant loss of diversity.

Robustness of SPeDE dereplication to biological sample variation on a set of closely related strains. MALDI-TOF MS measurements can be sensitive to various experimental factors, such as instrument calibration, but also to the physiological state of the cultures (28). To test the sensitivity of MALDI-TOF MS spectrum dereplication to experimental artifacts introduced during culturing in a high-throughput work flow, we generated MALDI-TOF MS spectra from at least 5 independent cultures for each of 25 strains of *Lactobacillus brevis* in our culture collection. Dereplication using randomly selected spectra from only one culture per strain (50 spectra in total) yielded on average 22.25 reference spectra, while adding spectra from 5 cultures (250 spectra) resulted in 31.29 reference spectra (Fig. S3). These data indicate that SPeDE is only moderately sensitive to biological or technical variation.

Downloaded from http://msystems.asm.org/ on October 14, 2020 by guest

TABLE 1 SPeDE accuracy compared to PPMC/UPGMA and cosine/UPGMA clustering[a]

Method	No. of OIUs	No. of strains retained	Reduction rate (%)	% of redundant references/cluster	OTU coverage	Lowest ANI within an OIU	No. of OIUs with multiple OTUs	Total analysis time (h)[b]
SPeDE	210	160	96	23.8	149	94.92	13	2
94.5% PPMC/UPGMA[c]	337	143	93.4	51	147	77.33	15	8
92% cosine/UPGMA[d]	152	152	97	0	145	91.89	5	2–3

[a]The benchmark data set consisted of 5,228 spectra, 167 strains, and 149 OTUs. PPMC/UPGMA, Pearson product moment correlation/unweighted pair group method with arithmetic mean; OIU, operational isolation unit; OTU, operational taxonomic unit (ANI > 98%).
[b]Including data import, data processing, and interpretation of results.
[c]Method as described in Ghyselinck et al. (21).
[d]Method as described in Strejcek et al. (22).

Performance of SPeDE dereplication to sample variation. To determine if dereplication analysis by SPeDE is robust under conditions representative of those of an isolation campaign which aims at dereplicating spectra of conspecific isolates, we generated MALDI-TOF mass spectra for 78 strains of the BCCM/LMG Bacteria Collection, representing 8 genera and 47 species, from samples taken before and after lyophilization and subculturing. Each measurement was done in duplicate (2 spots per extract), for a final set consisting of 312 spectra (Table S2). This set contained spectra from 4 species of *Enterococcus*, 4 species of *Klebsiella*, 13 species of *Lactobacillus*, 2 species of *Lactococcus*, 6 species of *Leuconostoc*, 6 species of *Streptococcus*, 3 species of *Weissella*, and 9 species of *Xanthomonas*. To account for intraspecies variability, this set also included spectra for several strains of most species (Table S5).

Dereplication analysis by SPeDE using a PPMC threshold value of 50% yielded 97 reference spectra representing 72 strains. In total, 48 strains were represented by a single reference spectrum and 24 were represented by 2 reference spectra. Six strains did not yield reference spectra, but in all cases, the spectra were matched to reference spectra of the same species. These included 1 strain of *Streptococcus salivarius*, 2 strains of *Lactobacillus rhamnosus*, 2 strains of *Weissella confusa*, and 1 strain of *Enterococcus faecalis*. Of the 24 strains which yielded two reference spectra, 13 yielded spectra which clustered by condition (i.e., before or after lyophilization). For the remaining 11 cases, 3 spectra from either condition were matched to a single reference and the remaining spectrum was retained as a singleton reference. Hierarchical clustering of the reference spectra based on the number of USFs grouped all the spectra from a given strain together (Fig. S4), indicating that small spectral variations are responsible for the moderate decay in the dereplication ratio. Importantly, no spectra were matched to references outside species boundaries, highlighting the capacity of SPeDE to reliably discriminate spectra at the species level or lower.

Performance of SPeDE dereplication compared to other methods. We compared the outcome of SPeDE dereplication for the benchmark data set to that of previously described methods based on (i) the global Pearson product moment correlation and the unweighted pair group method with arithmetic mean (UPGMA) clustering proposed by Ghyselinck and colleagues (21) and (ii) the cosine similarity/UPGMA clustering described by Strejcek and colleagues (22). The results are given in Table S3 and Table S4, respectively. The method of Ghyselinck et al. (21) returned the most OIUs, retrieved the least diversity from the benchmark data set, and had the overall longest processing time (Table 1). The method of Strejcek and colleagues (22), based on a proposed cosine similarity cutoff of 92%, was the best performing in terms of overall dereplication ratio, with a reduction of 97% of the spectra for the benchmark data set. However, the analysis also returned 15 OIUs which included multiple genetically distinct strains, with 1 OIU merging strains sharing less than 92% genome-wide ANI (Table 1). SPeDE was the second-best performer in terms of the dereplication ratio but was the most sensitive, retrieving the most diversity from the sample (149/149 OTUs) and returning the fewest OIUs containing multiple strains. SPeDE was also the fastest method in terms of both the run time and the manual data processing required before data analysis (Table 1).

Downloaded from http://msystems.asm.org/ on October 14, 2020 by guest

DISCUSSION

We developed SPeDE, a fast, accurate, and low-memory program and algorithm for processing large spectral data sets. This is accomplished by identifying unique features of MALDI-TOF MS spectra instead of relying on global similarity measures. The algorithm groups spectra into OIUs and outputs a reference spectrum for each OIU. We optimized and validated the algorithm on a set of 5,228 spectra representing 167 strains belonging to 132 genera across six phyla and obtained 210 OIUs. SPeDE can accurately assign the MALDI-TOF mass spectra of bacterial strains sharing a minimum of 98% genome-wide average nucleotide identity. Dereplication using SPeDE yields a number of OIUs comparable to the number generated by previously published methods and is more accurate (21, 22). More relevant for an application in culturomics, SPeDE was also able to recover the most diversity from a sample.

One of the major limitations of MALDI-TOF MS analysis is the variability of measurements. We first tested SPeDE on a set of spectra obtained from multiple cultures of a large set of *Lactobacillus brevis* strains. SPeDE dereplication resulted in a small increase in the number of reference spectra with an increasing number of cultures included in the data set (Fig. S3). These results show that small artifacts introduced during culturing, sample preparation, and/or data acquisition have an impact on the efficiency of SPeDE. However, this impact is small: increasing the number of reference spectra from 22.5 on average when considering only spectra representing a single culture (50 spectra), to 31.29 for spectra representing 5 independent cultures (250 spectra).

To be useful in the frame of large-scale culturomics experiments, MALDI-TOF MS dereplication must also be robust to the sample variation introduced by the storage of cultures. We thus tested SPeDE on the spectra of bacterial cultures before and after lyophilization and subculturing. Experimental variability affects the dereplication ratio, introducing an excess of reference spectra but without sacrificing precision. In an analysis using global similarity to dereplicate mass spectra, Strejcek and colleagues also attributed the decreased overall performance of cluster-based dereplication to biological variation, a confounding factor inherent to MALDI-TOF MS measurements (22). Possible reasons for the variability in the mass spectra of strains grown on the same medium and analyzed using the same sample preparation method are not well described, but the total incubation time of cultures appears to be a critical factor affecting reproducibility (28, 29). For most culturomics experiments, isolates would be obtained from cultures incubated under identical conditions and processed using a linear work flow, thus minimizing sample variation. Rapid analysis and dereplication are crucial when performing large-scale culturomics experiments when long-term storage of cultures is limited or not feasible. We argue that the high-throughput capacity, speed, and low cost of a dereplication pipeline built around MALDI-TOF MS and SPeDE vastly outweigh the benefits of 16S rRNA gene sequence-based dereplication techniques.

In conclusion, SPeDE is a fully automated, scalable algorithm which can be run on a single workstation. SPeDE has been designed to deal with large quantities of data like those generated in culturomics experiments and can process more than 5,000 MALDI-TOF mass spectra in less than an hour on a Linux or Windows laptop computer. SPeDE can be run from a graphical user interface or from the command line and can be fully integrated in a culturomics pipeline designed to automatically retrieve the cultivable diversity from complex samples.

MATERIALS AND METHODS

MALDI-TOF MS data sets. The strains used in the benchmark data set were derived from the BCCM/LMG Bacteria Collection and the research collection of LM-UGent. In total 167 different strains were used. The bacterial set consisted of 6 major bacterial phyla (*Actinobacteria*, *Bacteroidetes*, *Deinococcus*, *Firmicutes*, *Proteobacteria*, and *Epsilonbacteraeota*) representing 16 classes and 132 genera. The strains used in the lyophilization data set to estimate the effect of biological variability were derived from the BCCM/LMG Bacteria Collection. In total, 79 strains covering four major bacterial phyla were used. The 25 strains used in the *Lactobacillus brevis* study were obtained from the BCCM/LMG Bacteria Collection

Downloaded from http://msystems.asm.org/ on October 14, 2020 by guest

and the research collection of LM-UGent and were isolated from 22 different isolation sources. An overview of the cultures used in all sets is given in Table S5 in the supplemental material.

MALDI-TOF MS sample preparation and data acquisition. (i) Preparation of MALDI-TOF-MS samples. The strains included in the benchmark data set were subcultured 3 times, prior to harvesting of the cell material grown under the conditions stated by the BCCM/LMG Bacteria Collection catalogue (http://bccm.belspo.be/catalogues/lmg-catalogue-search). The strains included in the lyophilization study were subcultured 3 times before and after preservation according to the protocol described by Peiren et al. (30). For the preparation of cell extracts, a 1-μl loopful of bacterial cells was suspended in 300 μl of Milli-Q water and vortexed to a homogeneous suspension. Next, 900 μl of absolute ethanol (EtOH) was added, the components were mixed by inversion, and the mixture was centrifuged for 3 min at 4°C (14,000 rpm). Samples were stored at −20°C. At the time of analysis, samples were centrifuged as described above, supernatants were discarded, and centrifugation was repeated to remove the residual EtOH, followed by air drying for 5 min at room temperature. The pellet was suspended in 40 μl of 70% formic acid in water and mixed by vortexing. Finally, 40 μl of acetonitrile was added and the mixture was vortexed. The extract was centrifuged for 2 min at 4°C (14,000 rpm) to remove the cell debris, and the supernatant was transferred to a new tube.

The strains included in the *Lactobacillus brevis* study were passaged twice on MRS agar medium (Oxoid, UK) and incubated at 28°C for 3 days. For each strain, six colonies were transferred to a different well of a 96-well deep-well plate containing 1 ml of MRS broth (Oxoid, UK) and subcultured after 3 days of incubation at 28°C using a Viaflo 96/384 pipetting robot (Integra, UK). For the preparation of samples, the cultures were centrifuged for 10 min at 4°C (14,000 rpm). The cell pellets were suspended in 500 μl Milli-Q water, and this process was repeated twice to remove residual medium components from the cell suspension. After the second washing step, the cell pellets were suspended in 100 μl Milli-Q water. Sample preparation methods using solvent extraction tend to yield more consistent MALDI-TOF MS spectra, but whole-cell suspensions or smears yield spectra of equivalent quality for many Gram-positive bacteria, including lactic acid bacteria (31, 32). Whole-cell suspensions were used here instead of protein extracts because they significantly reduce the sample preparation time, are more likely to be adopted in a high-throughput isolation work flow (22), and tend to introduce more variability than protein extracts.

(ii) MALDI-TOF MS data acquisition. Bacterial cell extracts (1 μl) of the lyophilization study samples were spotted on a target plate (Bruker Daltonik, Bremen, Germany) in duplicate, and samples of the benchmark study were spotted seven to eight times and dried in air at room temperature. The sample spot was overlaid with 1 μl of matrix solution (10 mg/ml α-cyano-4-hydroxycinnamic acid in acetonitrile-water-trifluoroacetic acid [TFA] [50:47.5:2.5]). Each target plate comprised one spot of pure matrix solution, used as a negative control, and one spot of Bacterial Test Standard (Bruker Daltonik, Bremen, Germany), used for calibration. The target plate was measured automatically on a Bruker Microflex LT/SH (lyophilization study) or Bruker Microflex LT/SH s-Smart (benchmark study) platform (Bruker Daltonik, Bremen, Germany). The target plates of the benchmark study were measured 4 times, thus obtaining a total of 28 to 32 replicate spectra for each strain. The spectra were obtained in linear, positive ion mode using FlexControl (v3.4) software according to the manufacturer's recommended settings (Bruker Daltonik, Bremen, Germany). Each final spectrum resulted from the sum of the spectra generated at random positions to a maximum of 240 shots per spectrum.

For *Lactobacillus brevis* cultures, bacterial cell suspensions were spotted (1 μl) on a target plate (Applied Biosystems, USA) in 4 replicates and dried in air at room temperature. The sample spot was overlaid with 1 μl of matrix solution (5 mg/ml α-cyano-4-hydroxycinnamic acid in acetonitrile-water-TFA [50:47.5:2.5]). Each target plate comprised one spot of pure matrix solution, used as a negative control, and one spot of Bacterial Test Standard (Bruker Daltonik, Bremen, Germany), used for calibration. The target plate was measured automatically on a 4800 Plus MALDI TOF/TOF analyzer (Applied Biosystems, USA). The spectra were obtained in the linear, positive ion mode using and covered a mass range of 2 to 20 kDa. Each final spectrum resulted from the sum of the spectra generated at random positions to a maximum of 2,000 shots per spectrum. The laser intensity was set to between 4,200 and 5,700 procedure defined units (pdu). The mass spectra were retrieved as t2dfiles from the 4800 Plus MALDI TOF/TOF analyzer via the 4000 series Explorer software. Data Explorer (v4.0) software (Applied Biosystems, USA) was used to convert the t2dfiles into text files.

(iii) Bruker Biotyper identification. The spectra were compared to those in the Bruker MBT 7712 MSP library using MBT Compass Explorer software according to the manufacturer's settings (Bruker Daltonics, Bremen, Germany) to verify the authenticity of the strains. The scores obtained were reported to be of high-confidence identification (score, >2.0), low-confidence identification (score, 1.70 to 1.99), and no identification possible (score, <1.70).

(iv) Mass spectrum preprocessing. Mass spectra were converted to text format using the FlexAnalysis batch processing tool (Bruker Daltonics, Germany). Peak lists were generated using the continuous wavelet transformation (CWT) peak detection algorithm described by Du and colleagues with the following parameter settings: a signal-to-noise ratio of 3 and a relative amplitude threshold of 0.0001 (33). The R script used to generate the peak list is available at https://github.com/LM-UGent/SPeDE/tree/master/data_preprocessing/peak_calling. To control for small variations in *m/z* axis values between runs, raw spectra were normalized to a fixed *m/z* axis, referred to as "regridding." The weighted average of the raw spectrum data was calculated for each *m/z* value of the fixed grid. The script used to regrid the spectra is available at https://github.com/LM-UGent/SPeDE/tree/master/data_preprocessing/regridding.

Genome sequencing, assembly, and analysis. (i) DNA extraction. Genomic DNA for genome sequencing was extracted using the procedure of Gevers et al. (34), Wilson (35), or Pitcher et al. (36) or using a Maxwell 16 tissue DNA purification kit (catalog number AS1030; Promega, Madison, WI, USA) and

Downloaded from http://msystems.asm.org/ on October 14, 2020 by guest

a Maxwell 16 instrument (catalog number AS2000; Promega). Gram-positive bacterial cultures were incubated with 5 mg of lysozyme (Serva) and 40 μl mutanolysin (5,000 U/ml; Sigma) dissolved in 110 μl of TE buffer (10 mM Tris Cl, 1 mM EDTA). DNA integrity and purity were evaluated on a 1.0% (wt/vol) agarose gel and by spectrophotometric measurements at 234, 260, and 280 nm. A Quantus fluorimeter and a QuantiFluor One double-stranded DNA system (Promega) were used to measure the DNA concentration.

(ii) Genome sequencing, assembly, and ANI calculations. Paired-end 2 $\times$ 150-bp libraries were prepared at the Wellcome Trust Human Genome Center (Oxford, UK) using a NEBNext DNA library kit for Illumina (New England Biolabs, Ipswich, MA, USA) and sequenced on an Illumina HiSeq 4000 instrument. Sequencing reads were prepared for assembly by adapter trimming and read filtering using the Trimmomatic tool (37). Reads with phred scores below 30 were removed, and nonpaired reads were discarded. Reads were assembled using the SPAdes (v3.10.1) program (38) and kmer lengths of 21, 33, 55, 77, and 99. Short contigs (<500 bp) or contigs with an average genome coverage of <50% were removed. To rule out possible contamination or mislabeling of samples, 16S rRNA gene sequences were extracted from the assembled genomes using the Barrnap (v0.6) program (https://github.com/tseemann/barrnap) and compared to available sequences for the strain.

Pairwise average nucleotide identity (ANI) values were calculated in two steps. First, genome distance values (corresponding approximately to 1 − ANI) were calculated between all possible genome pairs using Mash (v2.0) software (39). For genome pairs with Mash distance values of <0.1, genome distances were refined by calculating ANI values using OrthoANI (v0.90) software (40). Final genome distance values are given as percent ANI, with scores below 90% corresponding to [1 − (mash distance)] $\times$ 100. A matrix of ANI values is available in Data Set S1 in the supplemental material. OTUs were determined using the genome distance between pairs of strains in the benchmarking set as 1 − ANI and performing hierarchical clustering in R using the hclust function and the ward.D2 method. The dendrogram was cut using the R function cutree with a height parameter of 2, yielding groups of strains where all intragroup pairwise ANI values were >98%. As a reference, the commonly accepted threshold for species delineation is ANI values of 95 to 96% (26).

(iii) Phylogenetic analyses. Forty single-copy, conserved marker protein sequences were extracted from assembled genomes using FetchMG software (41). FetchMG automatically extracts the sequences of 40 universal gene markers which were found in a single copy in bacterial genomes and which have been used to reconstruct robust phylogenies across the tree of life (42). Core protein sequences were aligned with the Muscle program (43). Each alignment was trimmed to remove poorly aligned regions using the Trimal program (44) and concatenated into a superalignment. This superalignment was used to create an approximate maximum likelihood phylogenetic tree using FastTree (v2) software with the JTT+CAT model (45). The resulting phylogenetic tree was annotated and visualized using the iTol web service (46).

Description of the SPeDE algorithm. (i) Overview. The objective of culturomics experiments is to isolate a maximal number of distinct or new taxa. To this purpose, SPeDE is based on unicity measures of MALDI-TOF mass spectra instead of global similarity measures. The algorithm relies on peak matching coupled to spectrum similarity in an area around the peaks to determine unique spectral features (USFs). Determination of the number of unique features, or unicity, allows for a higher resolution than standard matching algorithms to differentiate between profiles of different strains or taxa. This approach also circumvents the need for extensive manual preprocessing, minimizing the risk for technical errors. The most informative spectrum from all redundant profiles (i.e., the one containing the highest number of USFs overall) is selected as the reference profile to which all other profiles are matched. Subtle peak differences (e.g., *m/z* shifts) can therefore be easily detected and can improve the discrimination of otherwise similar profiles. The overall work flow of the SPeDE algorithm is presented in Fig. 1.

(ii) Data input. The SPeDE algorithm processes two data files per sample: (i) a file containing the one-dimensional raw spectrum of intensities observed to a fixed *m/z* axis and (ii) a file containing the list of peaks detected in the raw spectrum, containing for each detected peak the *m/z* value and the S/N ratio. Optionally, an existing set of reference profiles may also be added for incremental dereplication scenarios. A file containing metadata information which will be automatically parsed in the output (i.e., strain information and/or MALDI-TOF MS-based identification data) can be provided.

(iii) Data processing. *(a) Step 1: quality control.* The quality of the samples is assessed based on the peak signal strength. Only peaks with an S/N of >30 are taken into consideration for this step. Samples are considered to be of good quality if the spectrum contains five or more peaks with an S/N of >30 (green samples). Samples containing one to four peaks with an S/N of >30 are considered to be of low quality (orange). Samples containing no peaks with an S/N of >30 are considered to be of poor quality (red). Only good-quality samples are processed in the following steps. Low-quality samples are ignored for initial processing but are matched to reference spectra in the final step (step 5).

(b) Step 2: USFs. Each pair of good-quality samples is compared to determine the number of unique spectral features (USFs) for each spectrum. First, a peak-based comparison is carried out and peaks are matched if they fall within a predefined peak accuracy window, calculated as the *m/z* difference (in parts per million) between peaks. Second, peaks that match and peaks that are considered to be unique for one of the spectra are validated by calculating the Pearson product moment correlation (PPMC) in a local area around each peak. This local PPMC around a pair of matched peaks is efficient in finding peak shifts yet still robust to variations in peak intensities. Three scenarios are possible after these two steps: (i) peaks that are unmatched but that have a local PPMC above a predefined threshold are considered matched, (ii) peaks that are matched but that have a local PPMC below a threshold are considered unmatched and marked as a USF, and (iii) peaks that are unmatched and that have a local PPMC below

Downloaded from http://msystems.asm.org/ on October 14, 2020 by guest

the threshold are marked as a USF. The number of USFs between each pair of samples is stored in a USF matrix. Note that, in contrast to similarity matrices, USF matrices are not symmetric.

(c) Step 3: reference spectra. Once all pairs of good-quality samples are compared, the USF matrix is sorted on the basis of the sum of USFs per spectrum. Spectra containing the highest number of USFs have the lowest index number. Reference spectra are then selected by iteration over the sorted USF matrix by applying the following criterion: a spectrum is a reference spectrum if and only if it has at least one USF compared to all previously evaluated spectra. This approach results in the spectrum with the highest number of USFs in all matched spectra to be chosen as the reference spectrum of an operational isolation unit (OIU). OIUs are defined as clusters of spectra which cannot be distinguished from one another and which likely represent a single operational taxonomic unit.

(d) Step 4: OIUs. Spectra not marked as a reference are further matched by iterating over the index of the USF matrix: a spectrum is matched to the reference spectrum with the lowest index to which the spectrum has no USF. Spectra not marked as references are matched to existing reference spectra, and all spectra matched to a given reference are considered an OIU.

(e) Step 5: match low-quality spectra to the obtained reference spectrum. Low-quality spectra (i.e., spectra with <5 peaks with an S/N of >30) are matched to the set of references by Dice coefficient comparison. Peaks are matched if they fall within the peak accuracy window (700 ppm), and spectra are matched to all references resulting in a Dice coefficient of $>70\%$.

(iv) Output. The default output format is a CSV table matching all samples to the references. Optionally, a USF matrix can be exported. Code to generate a dendrogram based on sample distance is available in a Jupyter notebook at https://github.com/LM-UGent/SPeDE/tree/master/output_dendrogram. Input files are a USF matrix and a table containing the percentage of samples matched to the references. A bar plot shows the abundance of each OIU in the set of samples. The dendrogram can be exported in PDF and Newick formats.

(v) Implementation and availability. SPeDE is implemented in Python (v3) software. A graphical user interface was developed for Microsoft Windows, but the software can be run as a command line tool under Windows, Linux, or MacOS. The data analysis performed for this study was done on a Windows computer with an Intel Core i5-4210 central processing unit and 8 GB of random-access memory. An installer for installation of all required modules for Windows computers is provided at https://github.com/LM-UGent/SPeDE. The SPeDE source code is freely distributed under the MIT license and is available at https://github.com/LM-UGent/SPeDE.

Validation on a data set of 5,228 spectra representing 167 strains and parameter optimization. To set the optimal threshold used for the peak accuracy window and the local PPMC used to determine a USF, the benchmark data set was analyzed using local PPMC threshold values (the −l flag in the command-line version of SPeDE) ranging from 1 to 100 in increments of 2. To determine the optimal value of the peak accuracy window (the −d flag), the values tested ranged from 500 to 1,000 in increments of 25.

We counted true positives as sample spectra matched to a reference spectrum originating from a strain within the same OTU. OTUs were defined as groups of strains sharing at least 98% genome-wide ANI (see above for details on ANI calculations). False positives were recovered as sample spectra matched to a reference spectrum outside of the expected OTU. Precision was calculated as number of samples with true-positive results/(number of samples with true-positive results + number of samples with false-positive results). The dereplication ratio was determined as the number of OTUs divided by the number of observed OIUs.

Analysis of the *Lactobacillus brevis* data set. Spectra were analyzed with SPeDE with default parameters (local PPMC threshold = 50 and PPMC window = 700), including for each run 2 technical replicates per culture, randomly selected without replacement with the random.choice function of the Numpy package (47), and one or more randomly selected cultures per strain. The analysis was repeated 100 times for each condition, and results were plotted using the R package ggplot (48).

Comparison to other dereplication methods. To compare the performance of the SPeDE algorithm with that of conventional clustering approaches, the spectra were imported in BioNumerics (v7.6.2) software (Applied Maths, Belgium). The similarity between spectra was expressed using PPMC. UPGMA was used as a hierarchical clustering algorithm to obtain OIUs. The dendrogram was further processed by grouping branches at 94.75% similarity, as proposed by Ghyselinck and colleagues (21). Subsequently, spectra were imported in R via the MALDIquant Foreign package and analyzed according to the method described by Strejcek and colleagues (22).

Data availability. All mass spectrometry data and genome assemblies used in this study are available at https://doi.org/10.5281/zenodo.3066838.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mSystems.00437-19.

FIG S1, PDF file, 0.2 MB.
FIG S2, PDF file, 0.3 MB.
FIG S3, PDF file, 0.1 MB.
FIG S4, PDF file, 0.2 MB.
TABLE S1, PDF file, 0.6 MB.
TABLE S2, PDF file, 0.3 MB.

Downloaded from http://msystems.asm.org/ on October 14, 2020 by guest

TABLE S3, PDF file, 2.9 MB.
TABLE S4, PDF file, 2.8 MB.
TABLE S5, PDF file, 0.2 MB.
DATA SET S1, XLS file, 0.3 MB.

ACKNOWLEDGMENTS

This work was funded by the Geconcerteerde Onderzoeksacties (GOA) of Ghent University (grant BOF15/GOA/006). The Oxford Genomics Centre at the Wellcome Centre for Human Genetics is funded by Wellcome Trust grant reference 203141/Z/16/Z. The BCCM/LMG Bacteria Collection is supported by the Federal Public Planning Service—Science Policy, Belgium.

We thank the Oxford Genomics Centre at the Wellcome Centre for Human Genetics for the generation and initial processing of the sequencing data. We thank the 2017-2018 computational biology class (UGent, B.S. and M.S. in computer science) and M.S. students for scrutinizing the performance and memory footprint of SPeDE software. We thank the technical staff from the BCCM/LMG Bacteria Collection, especially Leentje Christiaens, Jolien Claeys, Marjan De Wachter, Sofie Leyman, Cindy Snauwaert, Liesje Willekens, and Chris Willems, for the cooperative work in preparing DNA for whole-genome sequencing, performing MALDI-TOF MS analysis, and providing the strains used in this study. We also thank three anonymous referees for their expertise and constructive comments.

We declare no conflict of interest.

REFERENCES

1. Urbaniak C, Sielaff AC, Frey KG, Allen JE, Singh N, Jaing C, Wheeler K, Venkateswaran K. 2018. Detection of antimicrobial resistance genes associated with the International Space Station environmental surfaces. Sci Rep 8:814. https://doi.org/10.1038/s41598-017-18506-4.
2. Shreiner AB, Kao JY, Young VB. 2015. The gut microbiome in health and in disease. Curr Opin Gastroenterol 31:69–75. https://doi.org/10.1097/MOG.0000000000000139.
3. Busby PE, Soman C, Wagner MR, Friesen ML, Kremer J, Bennett A, Morsy M, Eisen JA, Leach JE, Dangl JL. 2017. Research priorities for harnessing plant microbiomes in sustainable agriculture. PLoS Biol 15:e2001793. https://doi.org/10.1371/journal.pbio.2001793.
4. Falony G, Joossens M, Vieira-Silva S, Wang J, Darzi Y, Faust K, Kurilshikov A, Bonder MJ, Valles-Colomer M, Vandeputte D, Tito RY, Chaffron S, Rymenans L, Verspecht C, De Sutter L, Lima-Mendez G, D'hoe K, Jonckheere K, Homola D, Garcia R, Tigchelaar EF, Eeckhaudt L, Fu J, Henckaerts L, Zhernakova A, Wijmenga C, Raes J. 2016. Population-level analysis of gut microbiome variation. Science 352:560–564. https://doi.org/10.1126/science.aad3503.
5. Lorenz P, Eck J. 2005. Metagenomics and industrial applications. Nat Rev Microbiol 3:510–516. https://doi.org/10.1038/nrmicro1161.
6. Gilbert JA, Dupont CL. 2011. Microbial metagenomics: beyond the genome. Annu Rev Mar Sci 3:347–371. https://doi.org/10.1146/annurev-marine-120709-142811.
7. Staley JT, Konopka A. 1985. Microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol 39:321–346. https://doi.org/10.1146/annurev.mi.39.100185.001541.
8. Nunoura T, Chikaraishi Y, Izaki R, Suwa T, Sato T, Harada T, Mori K, Kato Y, Miyazaki M, Shimamura S, Yanagawa K, Shuto A, Ohkouchi N, Fujita N, Takaki Y, Atomi H, Takai K. 2018. A primordial and reversible TCA cycle in a facultatively chemolithoautotrophic thermophile. Science 359:559–563. https://doi.org/10.1126/science.aao3407.
9. Lynch MDJ, Neufeld JD. 2015. Ecology and exploration of the rare biosphere. Nat Rev Microbiol 13:217–229. https://doi.org/10.1038/nrmicro3400.
10. Pham VHT, Kim J. 2012. Cultivation of unculturable soil bacteria. Trends Biotechnol 30:475–484. https://doi.org/10.1016/j.tibtech.2012.05.007.
11. Zhu B, Ibrahim M, Cui Z, Xie G, Jin G, Kube M, Li B, Zhou X. 2016. Multi-omics analysis of niche specificity provides new insights into ecological adaptation in bacteria. ISME J 10:2017–2075. https://doi.org/10.1038/ismej.2015.251.
12. Lagier J-C, Armougom F, Million M, Hugon P, Pagnier I, Robert C, Bittar F, Fournous G, Gimenez G, Maraninchi M, Trape J-F, Koonin EV, La Scola B, Raoult D. 2012. Microbial culturomics: paradigm shift in the human gut microbiome study. Clin Microbiol Infect 18:1185–1193. https://doi.org/10.1111/1469-0691.12023.
13. Lagier J-C, Khelaifia S, Alou MT, Ndongo S, Dione N, Hugon P, Caputo A, Cadoret F, Traore SI, Seck EH, Dubourg G, Durand G, Mourembou G, Guilhot E, Togo A, Bellali S, Bachar D, Cassir N, Bittar F, Delerce J, Mailhe M, Ricaboni D, Bilen M, Dangui Nieko NPM, Dia Badiane NM, Valles C, Mouelhi D, Diop K, Million M, Musso D, Abrahão J, Azhar EI, Bibi F, Yasir M, Diallo A, Sokhna C, Djossou F, Vitton V, Robert C, Rolain JM, La Scola B, Fournier P-E, Levasseur A, Raoult D. 2016. Culture of previously uncultured members of the human gut microbiota by culturomics. Nat Microbiol 1:16203. https://doi.org/10.1038/nmicrobiol.2016.203.
14. Spitaels F, Wieme AD, Vandamme P 2016. MALDI-TOF MS as a novel tool for dereplication and characterization of microbiota in bacterial diversity studies, p 235–256. *In* Applications of mass spectrometry in microbiology. Springer International Publishing, Cham, Switzerland.
15. Dieckmann R, Graeber I, Kaesler I, Szewzyk U, von Döhren H. 2005. Rapid screening and dereplication of bacterial isolates from marine sponges of the Sula Ridge by intact-cell-MALDI-TOF mass spectrometry (ICM-MS). Appl Microbiol Biotechnol 67:539–548. https://doi.org/10.1007/s00253-004-1812-2.
16. Bai Y, Müller DB, Srinivas G, Garrido-Oter R, Potthoff E, Rott M, Dombrowski N, Münch PC, Spaepen S, Remus-Emsermann M, Hüttel B, McHardy AC, Vorholt JA, Schulze-Lefert P 2015. Functional overlap of the Arabidopsis leaf and root microbiota. Nature 528:364–369. https://doi.org/10.1038/nature16192.
17. Clark AE, Kaleta EJ, Arora A, Wolk DM. 2013. Matrix-assisted laser desorption ionization-time of flight mass spectrometry: a fundamental shift in the routine practice of clinical microbiology. Clin Microbiol Rev 26:547–603. https://doi.org/10.1128/CMR.00072-12.
18. Sandrin TR, Goldstein JE, Schumaker S. 2013. MALDI TOF MS profiling of bacteria at the strain level: a review. Mass Spectrom Rev 32:188–217. https://doi.org/10.1002/mas.21359.
19. Santos IC, Hildenbrand ZL, Schug KA. 2016. Applications of MALDI-TOF MS in environmental microbiology. Analyst 141:2827–2837. https://doi.org/10.1039/c6an00131a.
20. Welker M, Moore E. 2011. Applications of whole-cell matrix-assisted laser-desorption/ionization time-of-flight mass spectrometry in systematic microbiology. Syst Appl Microbiol 34:2–11. https://doi.org/10.1016/j.syapm.2010.11.013.

Downloaded from http://msystems.asm.org/ on October 14, 2020 by guest

21. Ghyselinck J, Van Hoorde K, Hoste B, Heylen K, De Vos P 2011. Evaluation of MALDI-TOF MS as a tool for high-throughput dereplication. J Microbiol Methods 86:327–336. https://doi.org/10.1016/j.mimet.2011.06.004.
22. Strejcek M, Smrhova T, Junkova P, Uhlik O. 2018. Whole-cell MALDI-TOF MS versus 16S rRNA gene analysis for identification and dereplication of recurrent bacterial isolates. Front Microbiol 9:1294. https://doi.org/10.3389/fmicb.2018.01294.
23. Furlan JPR, Pitondo-Silva A, Braz VS, Gallo IFL, Stehling EG. 2019. Evaluation of different molecular and phenotypic methods for identification of environmental Burkholderia cepacia complex. World J Microbiol Biotechnol 35:39. https://doi.org/10.1007/s11274-019-2614-0.
24. Rudrik JT, Soehnlen MK, Perry MJ, Sullivan MM, Reiter-Kintz W, Lee PA, Pettit D, Tran A, Swaney E. 2017. Safety and accuracy of matrix-assisted laser desorption ionization–time of flight mass spectrometry for identification of highly pathogenic organisms. J Clin Microbiol 55:3513–3529. https://doi.org/10.1128/JCM.01023-17.
25. Dec M, Puchalski A, Urban-Chmiel R, Wernicki A. 2016. 16S-ARDRA and MALDI-TOF mass spectrometry as tools for identification of Lactobacillus bacteria isolated from poultry. BMC Microbiol 16:105. https://doi.org/10.1186/s12866-016-0732-5.
26. Richter M, Rosselló-Móra R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A 106:19126–19131. https://doi.org/10.1073/pnas.0906412106.
27. Kim M, Oh H-S, Park S-C, Chun J. 2014. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst Evol Microbiol 64:346–351. https://doi.org/10.1099/ijs.0.059774-0.
28. Arnold RJ, Karty JA, Ellington AD, Reilly JP. 1999. Monitoring the growth of a bacteria culture by MALDI-MS of whole cells. Anal Chem 71:1990–1996. https://doi.org/10.1021/ac981196c.
29. Goldstein JE, Zhang L, Borror CM, Rago JV, Sandrin TR. 2013. Culture conditions and sample preparation methods affect spectrum quality and reproducibility during profiling of Staphylococcus aureus with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Lett Appl Microbiol 57:144–150. https://doi.org/10.1111/lam.12092.
30. Peiren J, Buyse J, De Vos P, Lang E, Clermont D, Hamon S, Bégaud E, Bizet C, Pascual J, Ruvira MA, Macián MC, Arahal DR. 2015. Improving survival and storage stability of bacteria recalcitrant to freeze-drying: a coordinated study by European culture collections. Appl Microbiol Biotechnol 99:3559–3571. https://doi.org/10.1007/s00253-015-6476-6.
31. van Veen SQ, Claas ECJ, Kuijper EJ. 2010. High-throughput identification of bacteria and yeast by matrix-assisted laser desorption ionization-time of flight mass spectrometry in conventional medical microbiology laboratories. J Clin Microbiol 48:900–907. https://doi.org/10.1128/JCM.02071-09.
32. Alatoom AA, Cunningham SA, Ihde SM, Mandrekar J, Patel R. 2011. Comparison of direct colony method versus extraction method for identification of Gram-positive cocci by use of Bruker Biotyper matrix-assisted laser desorption ionization-time of flight mass spectrometry. J Clin Microbiol 49:2868–2873. https://doi.org/10.1128/JCM.00506-11.
33. Du P, Kibbe WA, Lin SM. 2006. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. Bioinformatics 22:2059–2065. https://doi.org/10.1093/bioinformatics/btl355.
34. Gevers D, Huys G, Swings J. 2001. Applicability of rep-PCR fingerprinting for identification of Lactobacillus species. FEMS Microbiol Lett 205:31–36. https://doi.org/10.1111/j.1574-6968.2001.tb10921.x.
35. Wilson K. 2001. Preparation of genomic DNA from bacteria. Curr Protoc Mol Biol 56:2.4.1–2.4.5. https://doi.org/10.1002/0471142727.mb0204s56.
36. Pitcher DG, Saunders NA, Owen RJ. 1989. Rapid extraction of bacterial genomic DNA with guanidium thiocyanate. Lett Appl Microbiol 8:151–156. https://doi.org/10.1111/j.1472-765X.1989.tb00262.x.
37. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
38. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.
39. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 17:132. https://doi.org/10.1186/s13059-016-0997-x.
40. Lee I, Ouk Kim Y, Park S-C, Chun J. 2016. OrthoANI: an improved algorithm and software for calculating average nucleotide identity. Int J Syst Evol Microbiol 66:1100–1103. https://doi.org/10.1099/ijsem.0.000760.
41. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, Coelho LP, Arumugam M, Tap J, Nielsen HB, Rasmussen S, Brunak S, Pedersen O, Guarner F, de Vos WM, Wang J, Li J, Doré J, Ehrlich SD, Stamatakis A, Bork P. 2013. Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods 10:1196–1199. https://doi.org/10.1038/nmeth.2693.
42. Ciccarelli FD. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283–1287. https://doi.org/10.1126/science.1123061.
43. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340.
44. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973. https://doi.org/10.1093/bioinformatics/btp348.
45. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490. https://doi.org/10.1371/journal.pone.0009490.
46. Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 44:W242–W245. https://doi.org/10.1093/nar/gkw290.
47. Oliphant TE. 2006. A guide to Numpy. Tregol Publishing.
48. Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York, NY.

Downloaded from http://msystems.asm.org/ on October 14, 2020 by guest