**Zurich Open Repository and Archive**
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2023

# Automatic Discrimination of Species within the Enterobacter cloacae Complex Using Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry and Supervised Algorithms

Candela, Ana ; Guerrero-López, Alejandro ; Mateos, Miriam ; Gómez-Asenjo, Alicia ; Arroyo, Manuel J ; Hernandez-García, Marta ; Del Campo, Rosa ; Cercenado, Emilia ; Cuénod, Aline ; Méndez, Gema ; Mancera, Luis ; de Dios Caballero, Juan ; Martínez-García, Laura ; Gijón, Desirée ; Morosini, María Isabel ; Ruiz-Garbajosa, Patricia ; Egli, Adrian ; Cantón, Rafael ; Muñoz, Patricia ; Rodríguez-Temporal, David ; Rodríguez-Sánchez, Belén

DOI: https://doi.org/10.1128/jcm.01049-22

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-259008
Journal Article
Published Version

Originally published at:
Candela, Ana; Guerrero-López, Alejandro; Mateos, Miriam; Gómez-Asenjo, Alicia; Arroyo, Manuel J; Hernandez-García, Marta; Del Campo, Rosa; Cercenado, Emilia; Cuénod, Aline; Méndez, Gema; Mancera, Luis; de Dios Caballero, Juan; Martínez-García, Laura; Gijón, Desirée; Morosini, María Isabel; Ruiz-Garbajosa, Patricia; Egli, Adrian; Cantón, Rafael; Muñoz, Patricia; Rodríguez-Temporal, David; Rodríguez-Sánchez, Belén (2023). Automatic Discrimination of Species within the Enterobacter cloacae Complex Using Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry and Supervised Algorithms. Journal of Clinical Microbiology, 61(4):e0104922.
DOI: https://doi.org/10.1128/jcm.01049-22

AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Clinical Microbiology®

# Automatic Discrimination of Species within the *Enterobacter cloacae* Complex Using Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry and Supervised Algorithms

Ana Candela,[a,b] Alejandro Guerrero-López,[c] Miriam Mateos,[d,l] Alicia Gómez-Asenjo,[a,b] Manuel J. Arroyo,[e] Marta Hernandez-García,[d,f,l] Rosa del Campo,[d,f,l] Emilia Cercenado,[a,b,g,h] Aline Cuénod,[i,j] Gema Méndez,[e] Luis Mancera,[e] Juan de Dios Caballero,[d,f,l] Laura Martínez-García,[d,k,l] Desirée Gijón,[d,f,l] María Isabel Morosini,[d,f,l] Patricia Ruiz-Garbajosa,[d,f,l] Adrian Egli,[i,j] Rafael Cantón,[d,f,l] Patricia Muñoz,[a,b,g,h] David Rodríguez-Temporal,[a,b] Belén Rodríguez-Sánchez[a,b]

[a] Clinical Microbiology and Infectious Diseases Department, Hospital General Universitario Gregorio Marañón, Madrid, Spain

[b] Institute of Health Research Gregorio Marañón, Madrid, Spain

[c] Department of Signal Theory and Communication, University Carlos III of Madrid, Madrid, Spain

[d] Servicio de Microbiología, Hospital Universitario Ramón y Cajal, Madrid, Spain

[e] Clover Bioanalytical Software, Granada, Spain

[f] CIBER en Enfermedades Infecciosas, Madrid, Spain

[g] CIBER de Enfermedades Respiratorias, CIBERES CB06/06/0058, Madrid, Spain

[h] Medicine Department, Faculty of Medicine, Universidad Complutense de Madrid, Madrid, Spain

[i] Applied Microbiology Research, Department of Biomedicine, University of Basel, Basel, Switzerland

[j] Division of Clinical Bacteriology and Mycology, University Hospital Basel, Basel, Switzerland

[k] Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública, Madrid, Spain

[l] Instituto Ramón y Cajal de Investigación Sanitaria, Madrid, Spain

Ana Candela and Alejandro Guerrero-López contributed equally to this study. Author order was determined alphabetically.

**ABSTRACT** The *Enterobacter cloacae* complex (ECC) encompasses heterogeneous clusters of species that have been associated with nosocomial outbreaks. These species may have different acquired antimicrobial resistance and virulence mechanisms, and their identification is challenging. This study aims to develop predictive models based on matrix-assisted laser desorption ionization–time of flight mass spectrometry (MALDI-TOF MS) profiles and machine learning for species-level identification. A total of 219 ECC and 118 *Klebsiella aerogenes* clinical isolates from three hospitals were included. The capability of the proposed method to differentiate the most common ECC species (*Enterobacter asburiae*, *Enterobacter kobei*, *Enterobacter hormaechei*, *Enterobacter roggenkampii*, *Enterobacter ludwigii*, and *Enterobacter bugandensis*) and *K. aerogenes* was demonstrated by applying unsupervised hierarchical clustering with principal-component analysis (PCA) preprocessing. We observed a distinctive clustering of *E. hormaechei* and *K. aerogenes* and a clear trend for the rest of the ECC species to be differentiated over the development data set. Thus, we developed supervised, nonlinear predictive models (support vector machine with radial basis function and random forest). The external validation of these models with protein spectra from two participating hospitals yielded 100% correct species-level assignment for *E. asburiae*, *E. kobei*, and *E. roggenkampii* and between 91.2% and 98.0% for the remaining ECC species; with data analyzed in the three participating centers, the accuracy was close to 100%. Similar results were obtained with the Mass Spectrometric Identification (MSI) database developed recently (https://msi.happy-dev.fr) except in the case of *E. hormaechei*, which was more accurately identified with the random forest algorithm. In short, MALDI-TOF MS combined with machine learning was demonstrated to be a rapid and accurate method for the differentiation of ECC species.

**Editor** Patricia J. Simner, Johns Hopkins University

**Copyright** © 2023 American Society for Microbiology. All Rights Reserved.

Address correspondence to David Rodríguez-Temporal, david.rodriguez@iisgm.com.

The authors declare no conflict of interest.

**Received** 20 July 2022
**Returned for modification** 14 August 2022
**Accepted** 28 February 2023
**Published** 4 April 2023

Downloaded from https://journals.asm.org/journal/jcm on 05 April 2024 by 89.206.81.44.

10.1128/jcm.01049-22

**KEYWORDS** *Enterobacter* species, mass spectrometry, MALDI-TOF MS, peak analysis, machine learning

*Enterobacter* is a genus of facultative anaerobic Gram-negative organisms that can be found as natural commensals in the gut microbiome of mammals (1). Several species have been associated with nosocomial outbreaks, causing urinary tract infections, skin and soft tissue infections, pneumonia, and bacteremia (2, 3). *Enterobacter cloacae* complex (ECC) is of particular clinical interest. This group is composed of 13 heterogenic genetic clusters according to *hsp60* gene sequencing, i.e., *Enterobacter asburiae* (cluster I), *Enterobacter kobei* (cluster II), *Enterobacter hormaechei* subsp. *hoffmannii* (cluster III), *Enterobacter roggenkampii* (cluster IV), *Enterobacter ludwigii* (cluster V), *E. hormaechei* subsp. *oharae* and *E. hormaechei* subsp. *xiangfangensis* (cluster VI), *E. hormaechei* subsp. *hormaechei* (cluster VII), *E. hormaechei* subsp. *steigerwaltii* (cluster VIII), *Enterobacter bugandensis* (cluster IX), *Enterobacter nimipressuralis* (cluster X), *Enterobacter cloacae* subsp. *cloacae* (cluster XI), *E. cloacae* subsp. *dissolvens* (cluster XII), and a heterogeneous group of *E. cloacae* sequences are considered cluster XIII. However, the taxonomy of this genus is still under debate (4, 5). In fact, *Enterobacter aerogenes* has been recently reclassified into the *Klebsiella* genus as *Klebsiella aerogenes* (6). A more comprehensive study based on whole-genome sequencing (WGS) data from ECC isolates yielded a redistribution of the species defined by *hsp60* sequencing (5) into different clades (7) and allowed the characterization of new ECC species (8).

Discrimination of the ECC at the species level is usually performed by sequence-based methods. The most commonly targeted gene is *hsp60*, although multilocus sequence typing (MLST) and WGS have also been applied (5, 9, 10). Sequence-based diagnostic methods are laborious and require specific equipment. Therefore, new emerging techniques such as matrix-assisted laser desorption ionization–time of flight mass spectrometry (MALDI-TOF MS) have been proposed as alternatives to sequence-based methods. MALDI-TOF MS has been shown to be an excellent methodology for bacterial identification. It can easily identify ECC isolates, but it showed low discrimination power for the species in this group when using standard analyses and commercial databases with low resolution (11, 12).

This study aimed to develop and validate prediction models for automatic species differentiation within the ECC using MALDI-TOF MS and supervised learning algorithms. This task is important because of the diverse implications of ECC species in human pathologies and their involvement in nosocomial outbreaks (4). In addition, *E. hormaechei*, the ECC species most commonly encountered in clinical settings, has been correlated with the enhanced acquisition of antimicrobial resistance mechanisms and the expression of virulence factors (13, 14). To achieve this goal, three steps were conducted in this study. First, we performed an unsupervised clustering to determine the feasibility of using MALDI-TOF MS data for ECC species identification. Second, we applied a supervised machine learning algorithm with isolates from University Hospital Ramón y Cajal (UHRC) (Madrid, Spain) and validated our findings with different ECC isolates from the same hospital and from the University Hospital Basel (UHB) (Basel, Switzerland). Finally, the developed model was validated in the participating centers by the analysis of a batch of 23 ECC isolates.

## MATERIALS AND METHODS

**Bacterial isolates.** Overall, we analyzed 219 clinical isolates belonging to the ECC and 118 *K. aerogenes* (formerly *Enterobacter aerogenes*) isolates. Among them, 164 ECC isolates and 9 *K. aerogenes* isolates were collected in a surveillance study of antimicrobial resistance in the UHRC (Madrid, Spain) between 2005 and 2018 and were identified by partial sequencing of the *hsp60* gene (15). A second set of 141 isolates (34 ECC isolates and 107 *K. aerogenes* isolates) were collected at the UHB (Basel, Switzerland) between 2016 and 2021 and were identified by WGS using KmerFinder v3.2 (16–18). MALDI-TOF MS profiles of these isolates were obtained at the UHB and submitted to the Hospital General Universitario Gregorio Marañón (HGM) (Madrid, Spain) for further analysis. Finally, a batch of 23 isolates (21 ECC isolates and 2 *K. aerogenes* isolates) were collected at the HGM in 2022 for validation purposes.

All isolates from UHRC and HGM were incubated overnight at 37°C and metabolically activated after three subcultures on Columbia blood agar (bioMérieux, Marcy l'Etoile, France) before their analysis with MALDI-TOF MS at the HGM.

Downloaded from https://journals.asm.org/journal/jcm on 05 April 2024 by 89.206.81.44.

**TABLE 1** Number of ECC isolates used for the unsupervised feasibility study and the supervised model development

| | | No. of isolates in supervised study | | | |
|---|---|---|---|---|---|
| Species | No. of balanced samples in unsupervised study | Development dataset (UHRC) | Validation dataset (UHRC) | External validation dataset (UHB) | Validation dataset (HGM) |
| *K. aerogenes* | 18 | 18 | 3 | 107 | 2 |
| *E. asburiae* | 18 | 18 | 1 | 0 | 3 |
| *E. bugandensis* | 18 | 0 | 0 | 0 | 0 |
| *E. hormaechei* | 18 | 18 | 51 | 33 | 15 |
| *E. kobei* | 18 | 18 | 9 | 0 | 2 |
| *E. ludwigii* | 18 | 0 | 0 | 0 | 0 |
| *E. roggenkampii* | 18 | 18 | 62 | 1 | 1 |
| Total | 126 | 90 | 126 | 141 | 23 |

**Spectrum acquisition using MALDI-TOF MS.** We identified the isolates using the MALDI Biotyper$^{smart}$ (Bruker Daltonics, Bremen, Germany). We spotted all strains from UHRC in duplicate onto the MALDI target plate and overlaid 1 $\mu$L of 70% formic acid. After drying at room temperature, we covered and dried the spots with 1 $\mu$L $\alpha$-cyano-4-hydroxycinnamic acid (HCCA) matrix, according to the manufacturer's instructions (Bruker Daltonics). Each spot was read twice in the range of 2,000 to 20,000 Da, resulting in 4 composite spectra per isolate. The isolates from UHB were analyzed in the daily routine; therefore, 1 spot per strain was analyzed, and 1 composite spectrum from the spot was obtained.

**Data processing of MALDI-TOF MS protein spectra and development of predictive models.** For both feasibility and supervised studies, we processed all MALDI-TOF MS profiles with the Clover MS data analysis software (Clover BioSoft, Granada, Spain). We applied a preprocessing pipeline, which consisted of (i) smoothing (Savitzky-Golay filter: window length=11, polynomial order=3, and baseline subtraction; top-hat filter method: factor=0.02); (ii) creation of an average spectrum per isolate; (iii) alignment of the average spectra from different isolates (shift: medium; constant tolerance: 2 Da; linear mass tolerance: 600 ppm); and (iv) normalization by total ion current (TIC), to all protein spectra.

**(i) Unsupervised feasibility study.** To study the feasibility of using MALDI-TOF MS for differentiation of ECC species, we proposed an unsupervised study based on principal-component analysis (PCA) and *t*-distributed stochastic neighbor embedding (t-SNE). For this purpose, an oversampled balanced data set for each ECC species was used. We included a total of 126 spectra from the 7 ECC species analyzed in this study (sourced from UHRC and UHB), as indicated in Table 1.

**(ii) Supervised model development.** Once the feasibility of the study was determined, we proposed the supervised model development. In this case, three different data sets were created, i.e., a training validation set, an internal validation data set, and an external validation data set. The details of these data sets are shown in Table 1.

Due to the lack of validation samples for *E. ludwigii* and *E. bugandensis*, these species were not included in the development of the supervised model. Therefore, our supervised model was developed to predict four ECC species, namely, *E. asburiae* (cluster I), *E. kobei* (cluster II), *E. hormaechei* (clusters III, VI, and VIII considered together), and *E. roggenkampii* (cluster IV). We applied four different supervised models, i.e., partial least-squares discriminant analysis (PLS-DA), support vector machine (SVM) with linear (SVM-L) kernel and SVM with radial basis function (SVM-R) kernel, and random forest (RF). The hyperparameter selection was performed with a 5-fold cross-validation technique.

Finally, we performed two external validations of the predictive models. First, 126 isolates from UHRC and 141 isolates from UHB were blindly classified by the same predictive models using Clover BioSoft software v0.6.1. Later, a batch of 23 isolates from HGM were sent to UHRC and UHB for external validation and study of the reproducibility of the developed models. The software applied uses the scikit-learn v0.23.2 Python library to implement all statistical methods used in this study. For reproducibility purposes under findability, accessibility, interoperability, and reusability (FAIR) principles, free access to all spectra to reproduce the analyses developed in this study is provided.

**(iii) Evaluation of the MSI database.** Recently, an online database was developed for the rapid differentiation of ECC species based on their MALDI-TOF MS protein profiles (19). This database has free access (https://msi.happy-dev.fr) and has been built using protein spectra from 42 ECC isolates characterized by sequencing of the *hsp60* gene. This identification method is considered the state-of-the-art method for the identification of ECC isolates at the species level. Therefore, both external validation data sets were also identified using the Mass Spectrometric Identification (MSI) database as a comparison to the methods proposed in this article. As stated above, MALDI-TOF MS spectra associated with this study have been made publicly available.

**Ethics statement.** The Ethics Committee of the HGM evaluated this project and considered that all of the conditions for waiving informed consent were met since the study was conducted with microbiological samples and not with human products. At the UHB, only anonymized data were used for the purpose of quality control and assay validation; according to the Swiss Human Research Act, no specific consent is required in such cases. Data either were acquired in routine microbiological diagnostics (excluding cases with a rejected general consent) or were used from a previously published data set (Database of Resistance Information on Antimicrobials and MALDI-TOF Mass Spectra [DRIAMS]).

Downloaded from https://journals.asm.org/journal/jcm on 05 April 2024 by 89.206.81.44.

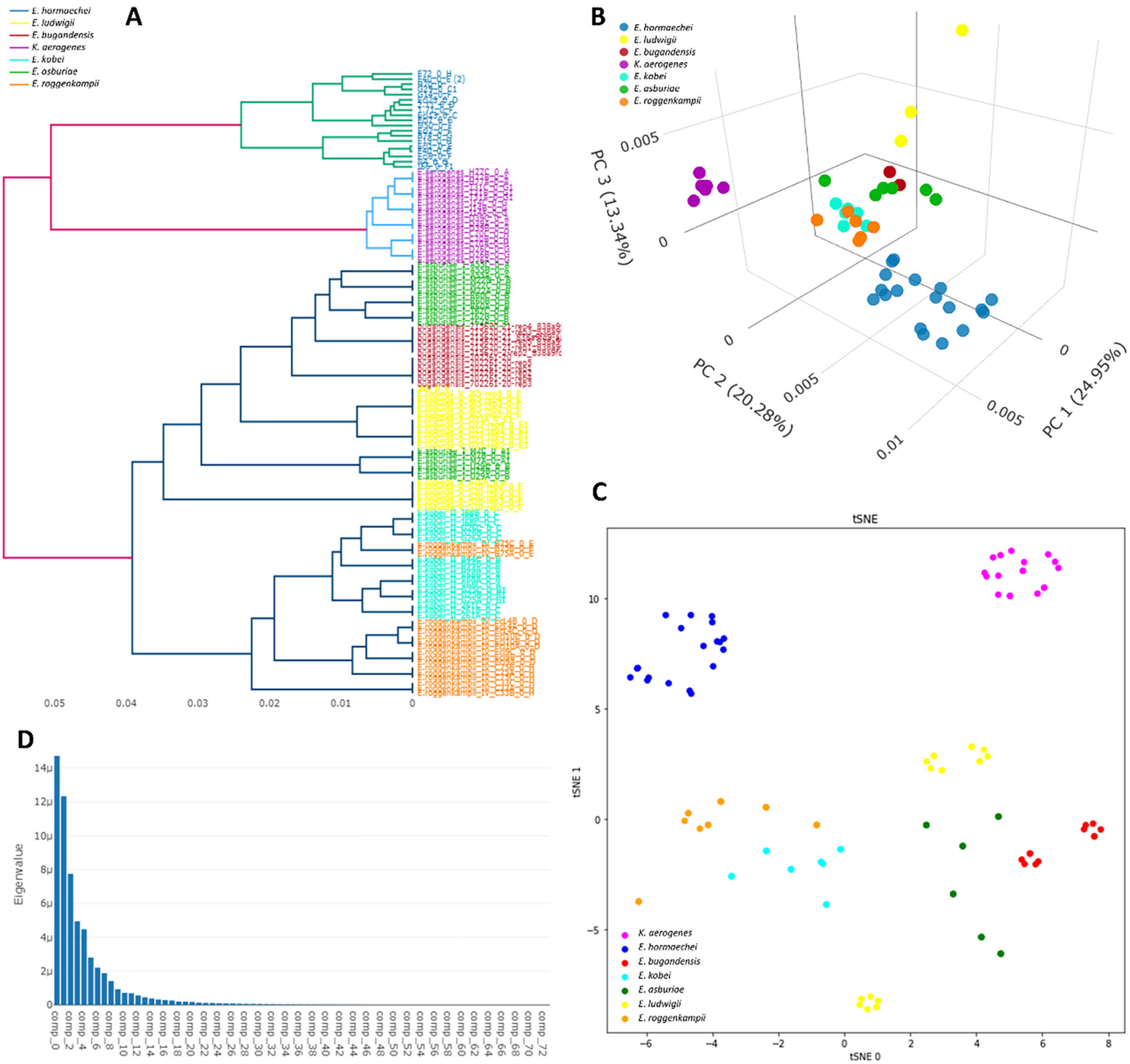

**FIG 1** Comparative analysis of 126 MALDI-TOF MS spectra from *K. aerogenes* and different ECC species using unsupervised methods. (A) Dendrogram built with 126 spectra using Euclidean distance and Ward metric. Spectra from *K. aerogenes* (purple) and *E. hormaechei* (dark blue) are completely separated from the other species. (B) PCA of feasibility study spectra. *K. aerogenes* and *E. hormaechei* are completely separated. Visualization in three dimensions was needed for separation of *E. ludwigii* (yellow). (C) t-SNE analysis of study spectra. (D) PCA eigenvalues showing the variance of each component A total of 14 components were needed to reach 95% of the total data variance.

**Data availability.** For reproducibility purposes under FAIR principles, all spectra to reproduce the analyses developed in this study are available in the Clover Repository (https://platform.clovermsdataanalysis.com/repository/collection/EHGM001).

## RESULTS

**Feasibility study.** To prove the feasibility of using MALDI-TOF MS to differentiate ECC species, an unsupervised hierarchical clustering with PCA and t-SNE preprocessing was performed (Fig. 1). The protein spectra of the seven species (*E. asburiae*, *E. kobei*, *E. hormaechei*, *E. roggenkampii*, *E. ludwigii*, *E. bugandensis*, and *K. aerogenes*), which are equally represented in the model, were compared. The dendrogram built with these data showed three main clusters, one cluster containing *E. hormaechei*, a second

Downloaded from https://journals.asm.org/journal/jcm on 05 April 2024 by 89.206.81.44.

**TABLE 2** Accuracy results for internal 5-fold cross-validation over development data set (90 spectral profiles)

| | No. identified/total no. (%) | | | | |
|---|---|---|---|---|---|
| **Algorithm** | ***E. asburiae*** | ***E. hormaechei*** | ***E. kobei*** | ***E. roggenkampii*** | ***K. aerogenes*** |
| PLS-DA | 9/18 (50) | 18/18 (100) | 5/18 (27.8) | 4/18 (22.2) | 18/18 (100) |
| SVM-L | 6/18 (33.3) | 18/18 (100) | 3/18 (16.7) | 6/18 (33.3) | 14/18 (77.8) |
| SVM-R | 18/18 (100) | 18/18 (100) | 18/18 (100) | 18/18 (100) | 18/18 (100) |
| RF | 18/18 (100) | 18/18 (100) | 18/18 (100) | 18/18 (100) | 18/18 (100) |

cluster with *K. aerogenes*, and a third cluster with the rest of the species. Inside the latter cluster, *E. bugandensis* strains were clustered together and so were *E. asburiae*, *E. ludwigii*, *E. kobei*, and *E. roggenkampii*, although in those four cases some of the spectra were clustered with the wrong species (Fig. 1A to C).

The implementation of PCA to reduce the dimensionality showed that 14 components were needed to explain 95% of the variance (Fig. 1D). This fact and the relatively accurate classification of ECC species using an unsupervised algorithm demonstrated the potentiality of MALDI-TOF MS to differentiate ECC species.

**Supervised models based on MALDI-TOF MS.** To address the limitations of unsupervised learning, we added the label knowledge to the training phase by using supervised algorithms such as PLS-DA, SVM, and RF. We trained these models using the development data set shown in Table 1 and selected their hyperparameters by a 5-fold cross-validation technique. This cross-validation process led to the next hyperparameter selection; 2 components were used for PLS-DA, the value of *C* was 10 for SVM-L, and the value of *C* was 10 and the value of $\gamma$ was 1,000 for SVM-R. Table 2 shows the results obtained for the internal 5-fold cross-validation, which have been further detailed in Table S1 in the supplemental material.

*E. hormaechei* and *K. aerogenes* presented the same trend as in the feasibility study, and their differentiation was 100% using nonlinear approaches (SVM-R and RF) Only the implementation of a linear approach (SVM-L) yielded lower results for *K. aerogenes* (Table 2). For the rest of the analyzed ECC species, we also obtained 100% correct classification by the application of nonlinear approaches (Table 2).

In Fig. 2, the distance between samples calculated by the RF classifier is shown. We detected a unique cluster for each species. Due to the results presented in Table 2, only SVM-R and RF were considered for further analysis. Table 3 shows the results of SVM-R and RF for the validation data set collected at UHRC and UHB.

Both algorithms, SVM-R and RF, yielded the same results in the external validation performed on the MALDI-TOF MS spectra from the validation data set sourced from the same hospital (UHRC). In that case, all *K. aerogenes*, *E. asburiae*, and *E. kobei* isolates were correctly classified, whereas 1 *E. hormaechei* strain was misclassified as *E. kobei* with both algorithms. For *E. roggenkampii*, 2 isolates were misclassified as *E. hormaechei* and 1 as *E. kobei* (see Table S2). The accuracy of the model is shown in Fig. 3A and B.

Since the SVM-R and RF algorithms performed equally, both of them were considered for external validation with MALDI-TOF MS profiles obtained at the UHB. In this case, 91.2% of the *E. hormaechei* isolates (*n* = 33), 100% of the *E. roggenkampii* isolates (*n* = 1), and 98.1% of the *K. aerogenes* isolates were correctly classified by RF, as shown in Table 3. The application of SVM-R yielded lower results for *E. hormaechei* and *K. aerogenes*. Figure 3C and D show the accuracy of both SVM-R and RF for the external validation collection from UHB.

Finally, for the 23 isolates from HGM used for the reproducibility study, 100% correct identification was obtained in all centers using the RF algorithm (Table 4). In the case of the SVM-R algorithm, 100% correct classification was obtained for strains analyzed in UHRC and UHB, whereas 1 strain of *E. kobei* was misclassified as *E. roggenkampii* in HGM.

**Identification of the ECC isolates using the MSI database.** The MSI platform was also used as an identification tool for ECC species, to compare the automatic approach developed in this study with the current state-of-the-art method (19). Among the UHRC isolates, the identification rates for *E. asburiae*, *E. hormaechei*, and *E. kobei* were

Downloaded from https://journals.asm.org/journal/jcm on 05 April 2024 by 89.206.81.44.

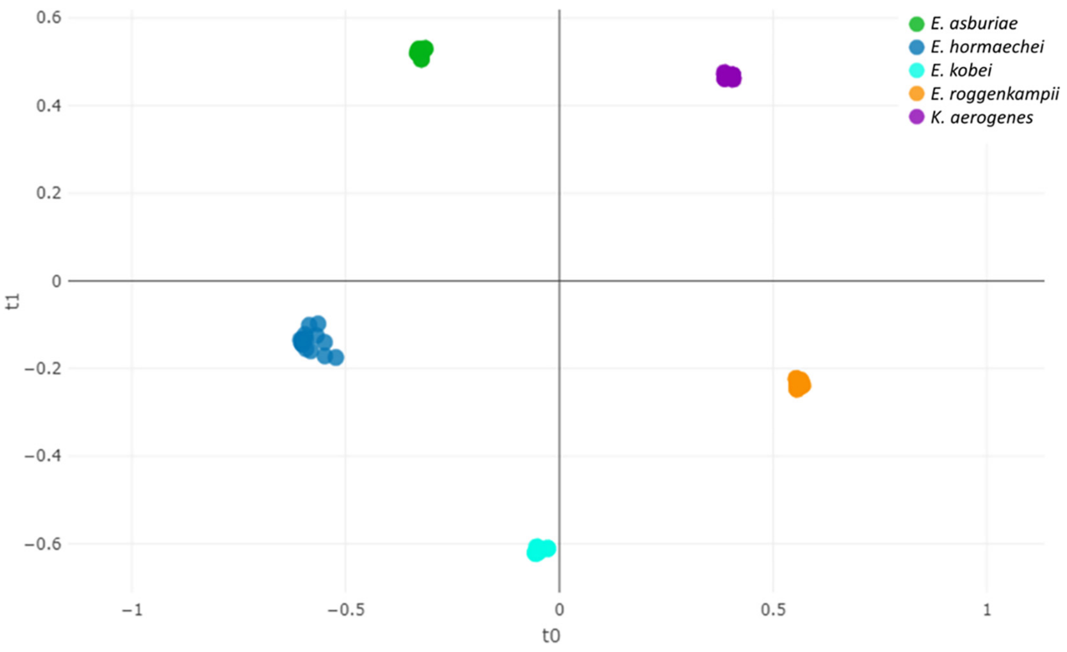

**FIG 2** RF distance plot using Euclidean distance for MALDI-TOF MS spectra from the species included in the classification model.

similar to the rates yielded by the predictive models developed in this study (Table 3). The only difference detected between the two methods was that the MSI database correctly classified 1 more *E. roggenkampii* isolate. As for the protein spectra sourced from the UHB, 84.8% of the *E. hormaechei* isolates and 100% of the *E. roggenkampii* isolates ($n$ = 1) were correctly identified with the MSI platform. In this case, the MSI platform provided a lower rate of correct identifications for *E. hormaechei* than the RF algorithm (Table 3).

## DISCUSSION

In this study, the implementation of supervised, nonlinear algorithms (SVM-R and RF) with MALDI-TOF MS spectra allowed the correct species assignment for 100% of isolates belonging to two ECC species (*E. asburiae* and *E. kobei*) and between 91.2% and 98.1% for *E. hormaechei*, *E. roggenkampii*, and *K. aerogenes* (formerly *E. aerogenes*) sourced from three different hospitals.

Poor discrimination of ECC species by MALDI-TOF MS using either commercial (11, 15) or enriched, in-house (20) databases was reported previously. However, a recent study from a research group with broad experience in MALDI-TOF MS and the creation of expanded libraries reported 92.0% correct species-level identification by implementing a specific in-house library enriched with well-characterized ECC strains, with correct discrimination of

**TABLE 3** Accuracy results for the validation data set from UHRC and UHB and the identification accuracy obtained with the MSI database[a]

| | No. identified/total no. (%) | | | | |
|---|---|---|---|---|---|
| **Center and algorithm** | ***E. asburiae*** | ***E. hormaechei*** | ***E. kobei*** | ***E. roggenkampii*** | ***K. aerogenes*** |
| UHRC | | | | | |
| SVM-R | 1/1 (100) | 50/51 (98.0) | 9/9 (100) | 59/62 (95.2) | 3/3 (100) |
| RF | 1/1 (100) | 50/51 (98.0) | 9/9 (100) | 59/62 (95.2) | 3/3 (100) |
| MSI | 1/1 (100) | 50/51 (98.0) | 9/9 (100) | 60/62 (96.8) | NA |
| UHB | | | | | |
| SVM-R | NA | 15/33 (44.1) | NA | 1/1 (100) | 102/107 (95.3) |
| RF | NA | 31/33 (91.2) | NA | 1/1 (100) | 105/107 (98.1) |
| MSI | NA | 28/33 (84.8) | NA | 1/1 (100) | NA |

[a] NA, not applicable.

Downloaded from https://journals.asm.org/journal/jcm on 05 April 2024 by 89.206.81.44.

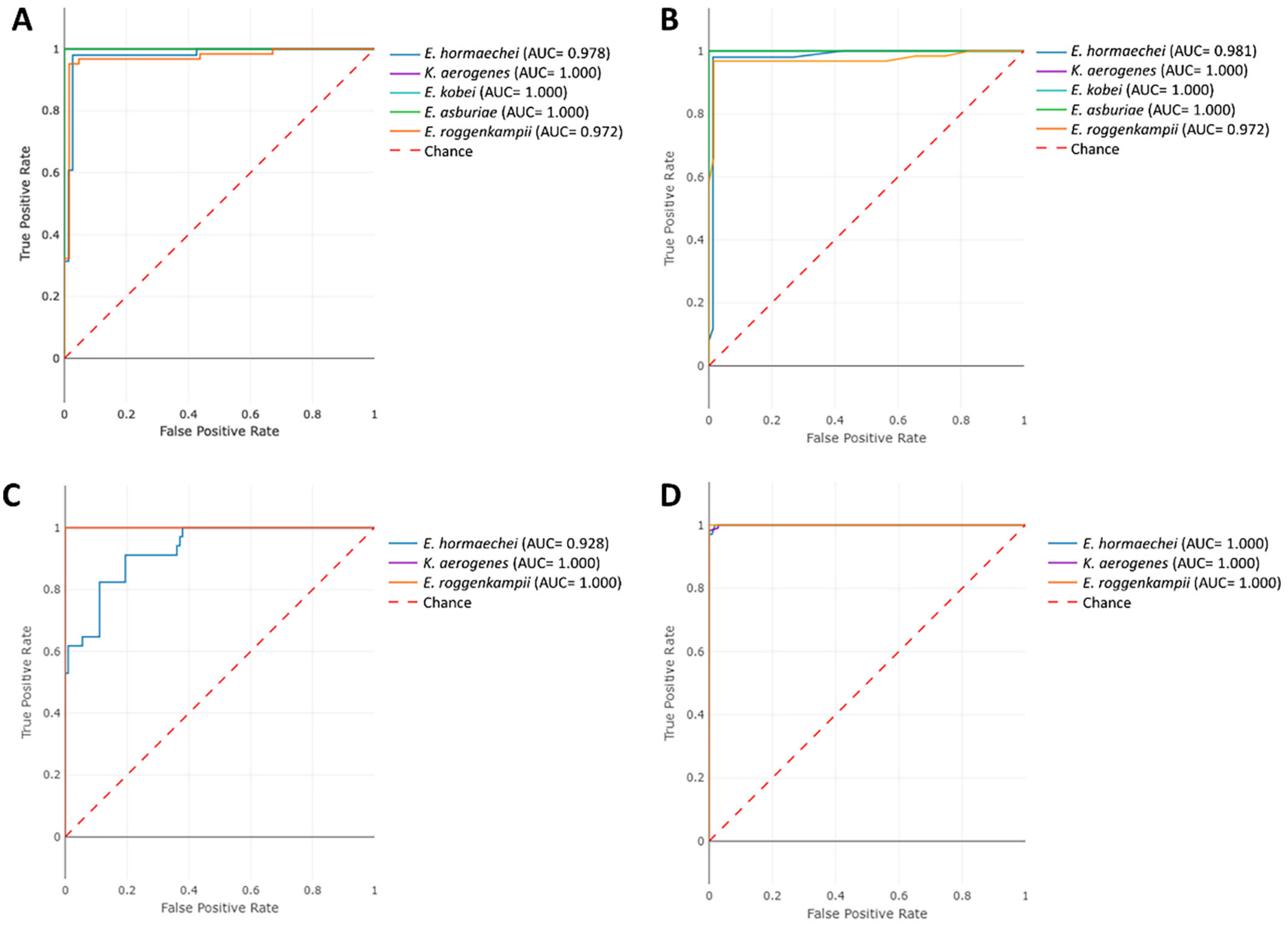

**FIG 3** Receiver operating characteristic curve (ROC) and area under the curve (AUC) values obtained with different supervised algorithms for the two participating centers. (A) SVM-R model validated with UHRC strains. (B) RF model validated with UHRC strains. (C) SVM-R model validated with UHB strains. (D) RF model validated with UHB strains.

97.0% of *E. hormaechei* isolates (19). This approach can be useful for the discrimination of closely related species, but the construction of a database is cumbersome and requires highly trained personnel. The implementation of the MSI platform allowed 94.9% correct species-level identification of 155 ECC protein spectra in this study. This rate was slightly lower than the obtained with the nonlinear algorithms proposed by our approach.

In this study, we demonstrate the feasibility of using MALDI-TOF MS to identify species within the ECC. First, hierarchical clustering showed that it is possible to differentiate between species using the information contained in MALDI-TOF MS spectra, as reported previously (20). Second, a supervised study using machine learning algorithms yielded the correct classification of all ECC species. Therefore, different supervised classification algorithms were implemented to correctly provide species assignment of ECC species. The internal validation experiment demonstrated that nonlinear approaches, such as SVM-R or RF, were needed to correctly identify all species. Both models perfectly classified all samples in internal cross-validation.

To further demonstrate that the model can perform in different scenarios with data different from the spectral profiles used for model training, we performed three validation experiments. First, we carried out a validation with MALDI-TOF MS protein spectra sourced from UHRC. Of a total of 126 samples, both SVM-R and RF assigned only 4 isolates to species different than those assigned by molecular techniques, i.e., 96.5% accuracy in classifying species within the ECC was obtained. Second, we performed an external validation with MALDI-TOF MS spectra sourced from UHB to simulate a real-

Downloaded from https://journals.asm.org/journal/jcm on 05 April 2024 by 89.206.81.44.

**TABLE 4** Reproducibility study of 23 isolates from HGM identified in parallel in the three participating centers, UHRC, UHB, and HGM

| | **No. identified/total no. (%)** | | |
|---|---|---|---|
| **Algorithm and species** | **UHRC** | **UHB** | **HGM** |
| SVM-R | | | |
| *E. asburiae* | 3/3 (100) | 3/3 (100) | 3/3 (100) |
| *E. hormaechei* | 15/15 (100) | 15/15 (100) | 15/15 (100) |
| *E. kobei* | 2/2 (100) | 2/2 (100) | 1/2 (50) |
| *E. roggenkampii* | 1/1 (100) | 1/1 (100) | 1/1 (100) |
| *K. aerogenes* | 2/2 (100) | 2/2 (100) | 2/2 (100) |
| SVM-R total | 23/23 (100) | 23/23 (100) | 22/23 (95.6) |
| | | | |
| RF | | | |
| *E. asburiae* | 3/3 (100) | 3/3 (100) | 3/3 (100) |
| *E. hormaechei* | 15/15 (100) | 15/15 (100) | 15/15 (100) |
| *E. kobei* | 2/2 (100) | 2/2 (100) | 2/2 (100) |
| *E. roggenkampii* | 1/1 (100) | 1/1 (100) | 1/1 (100) |
| *K. aerogenes* | 2/2 (100) | 2/2 (100) | 2/2 (100) |
| RF total | 23/23 (100) | 23/23 (100) | 23/23 (100) |

world scenario. These MALDI-TOF MS protein spectra originated in a different epidemiological scenario and were processed by operators from the UHB. In this case, SVM-R was shown to be overfitted to the UHRC distribution, which was already pointed out by the $\gamma$ value, with 83.7% accuracy. Finally, we validated the models by analyzing the same batch of 23 isolates in the three centers. In this case, 100% correct species-level classification was achieved using the RF algorithm. These results show that the models have not been overfitted to the spectra sourced from only one center. This phenomenon occurs when a machine learning model has learned the training data too well and does not recognize new data with small variations, yielding a lower accuracy than expected. The current state-of-the-art tool, i.e., the MSI database, performed better than SVM-R, with 94.9% accuracy, although it was not able to distinguish *K. aerogenes* (19). However, RF outperformed both approaches, with >96.0% accuracy in identifying the three species. Therefore, it is demonstrated that supervised machine learning algorithms are feasible, and indeed applicable, in microbiology laboratories.

One limitation of this study was the fact that all UHRC isolates were carbapenemase-producing isolates, because this was the source of the previously analyzed collection (15). However, the present study provides the first proof of concept for differentiating ECC species based on machine learning. For definitive validation, improvement, and implementation of these predictive models, future studies will involve strains with more diverse epidemiological and geographical origins and characteristics. In addition, not all analyzed ECC species could be represented in the external validation data set due to the lack of isolates from the species *E. ludwigii* and *E. bugandensis*.

This study shows promising results for differentiating ECC species based on machine learning and MALDI-TOF MS protein spectra. It also highlights the fact that MALDI-TOF MS data should be linked to WGS data in order to allow future work and to provide a reference standard. The MALDI-TOF MS and machine learning approach has been demonstrated to be a rapid and cost-effective method, suitable for correct species-level assignment of closely related species, as in the case of ECC species. The use of spectrum analysis tools is becoming user-friendly and easy to apply, and their use may provide species-level identification in a fast and inexpensive way. Once the model is validated with a comprehensive number of ECC species, an open Web application will be deployed to be used freely by the community.

## SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

**SUPPLEMENTAL FILE 1**, PDF file, 0.4 MB.

Downloaded from https://journals.asm.org/journal/jcm on 05 April 2024 by 89.206.81.44.

## ACKNOWLEDGMENTS

This work was supported by projects PI15/01073 and PI18/00997 from the Health Research Fund (FIS) (Instituto de Salud Carlos III, Plan Nacional de I+D+I 2013 to 2016) of the Carlos III Health Institute (ISCIII) (Madrid, Spain), partially financed by the European Regional Development Fund (FEDER) program A Way of Making Europe. M. Mateos was funded by the Community of Madrid (Programa de Garantía Juvenil, grant PEJD-2017-PRE/BMD-5106) and D. Rodríguez-Temporal by a postdoctoral contract from the Intramural Funding Program of the Institute of Health Research Gregorio Marañón (IiSGM). B. Rodríguez-Sánchez is a recipient of a Miguel Servet contract supported by the FIS (grant CPII19/00002). A. Guerrero-López was funded by a predoctoral contract from the Intramural Funding Program of the IiSGM.

A. Candela: experimental part, formal analysis, data collection, validation, visualization, and writing (original draft preparation and review/editing); A. Guerrero-López: formal analysis, data collection, validation, visualization, and writing (original draft preparation and review/editing); M. Mateos and A. Gómez-Asenjo: experimental part, formal analysis, and data collection; M. J. Arroyo, G. Méndez, and L. Mancera: data analysis, validation, and writing (original draft preparation and review/editing); M. Hernández-García and R. del Campo: experimental part, formal analysis, submission of isolates, and writing (original draft preparation and review/editing); E. Cercenado: conceptualization, formal analysis, validation, and writing (review/editing); A. Cuénod and A. Egli: submission of isolates and writing (original draft preparation and review/editing); J. D. Caballero, L. Martínez-García, D. Gijón, M. I. Morosini, P. Ruiz-Garbajosa, R. Cantón, and P. Muñoz: validation and writing (review/editing); D. Rodríguez-Temporal: conceptualization, formal analysis, validation, and writing (original draft preparation and review/editing); B. Rodríguez-Sánchez: conceptualization, project administration, formal analysis, supervision, validation, visualization, and writing (original draft preparation and review/editing).

We have no conflict of interests.

## REFERENCES

1. Mezzatesta ML, Gona F, Stefani S. 2012. *Enterobacter cloacae* complex: clinical impact and emerging antibiotic resistance. Future Microbiol 7: 887–902. https://doi.org/10.2217/fmb.12.61.
2. Akbari M, Bakhshi B, Najar Peerayeh S. 2016. Particular distribution of *Enterobacter cloacae* strains isolated from urinary tract infection within clonal complexes. Iran Biomed J 20:49–55. https://doi.org/10.7508/ibj.2016.01.007.
3. Kremer A, Hoffmann H. 2012. Prevalences of the *Enterobacter cloacae* complex and its phylogenetic derivatives in the nosocomial environment. Eur J Clin Microbiol Infect Dis 31:2951–2955. https://doi.org/10.1007/s10096-012-1646-2.
4. Davin-Regli A, Lavigne JP, Pages JM. 2019. *Enterobacter* spp.: update on taxonomy, clinical aspects, and emerging antimicrobial resistance. Clin Microbiol Rev 32:e00002-19. https://doi.org/10.1128/CMR.00002-19.
5. Hoffmann H, Roggenkamp A. 2003. Population genetics of the nomenspecies *Enterobacter cloacae*. Appl Environ Microbiol 69:5306–5318. https://doi.org/10.1128/AEM.69.9.5306-5318.2003.
6. Tindall BJ, Sutton G, Garrity GM. 2017. *Enterobacter aerogenes* Hormaeche and Edwards 1960 (Approved Lists 1980) and *Klebsiella mobilis* Bascomb et al. 1971 (Approved Lists 1980) share the same nomenclatural type (ATCC 13048) on the Approved Lists and are homotypic synonyms, with consequences for the name *Klebsiella mobilis* Bascomb et al. 1971 (Approved Lists 1980). Int J Syst Evol Microbiol 67:502–504. https://doi.org/10.1099/ijsem.0.001572.
7. Sutton GG, Brinkac LM, Clarke TH, Fouts DE. 2018. *Enterobacter hormaechei* subsp. *hoffmannii* subsp. nov., *Enterobacter hormaechei* subsp. *xiangfangensis* comb. nov., *Enterobacter roggenkampii* sp. nov., and *Enterobacter muelleri* is a later heterotypic synonym of *Enterobacter asburiae* based on computational analysis of sequenced *Enterobacter* genomes. F1000Res 7:521. https://doi.org/10.12688/f1000research.14566.2.
8. Wu W, Feng Y, Zong Z. 2019. Characterization of a strain representing a new *Enterobacter* species, *Enterobacter chengduensis* sp. nov. Antonie Van Leeuwenhoek 112:491–500. https://doi.org/10.1007/s10482-018-1180-z.
9. Singh NK, Bezdan D, Checinska Sielaff A, Wheeler K, Mason CE, Venkateswaran K. 2018. Multi-drug resistant *Enterobacter bugandensis* species isolated from the International Space Station and comparative genomic analyses with human pathogenic strains. BMC Microbiol 18:175. https://doi.org/10.1186/s12866-018-1325-2.
10. Hoffmann H, Stindl S, Ludwig W, Stumpf A, Mehlen A, Heesemann J, Monget D, Schleifer KH, Roggenkamp A. 2005. Reassignment of *Enterobacter dissolvens* to *Enterobacter cloacae* as *E. cloacae* subspecies *dissolvens* comb. nov. and emended description of *Enterobacter asburiae* and *Enterobacter kobei*. Syst Appl Microbiol 28:196–205. https://doi.org/10.1016/j.syapm.2004.12.010.
11. Pavlovic M, Konrad R, Iwobi AN, Sing A, Busch U, Huber I. 2012. A dual approach employing MALDI-TOF MS and real-time PCR for fast species identification within the *Enterobacter cloacae* complex. FEMS Microbiol Lett 328:46–53. https://doi.org/10.1111/j.1574-6968.2011.02479.x.
12. De Florio L, Riva E, Giona A, Dedej E, Fogolari M, Cella E, Spoto S, Lai A, Zehender G, Ciccozzi M, Angeletti S. 2018. MALDI-TOF MS identification and clustering applied to *Enterobacter* species in nosocomial setting. Front Microbiol 9:1885. https://doi.org/10.3389/fmicb.2018.01885.
13. Barnes AI, Paraje MG, Battán PDC, Albesa I. 2001. Molecular properties and metabolic effect on blood cells produced by a new toxin of *Enterobacter cloacae*. Cell Biol Toxicol 17:409–418. https://doi.org/10.1023/A:1013704801570.
14. Paauw A, Caspers MPM, Leverstein-van Hall MA, Schuren FHJ, Montijn RC, Verhoef J, Fluit AC. 2009. Identification of resistance and virulence factors in an epidemic *Enterobacter hormaechei* outbreak strain. Microbiology (Reading) 155:1478–1488. https://doi.org/10.1099/mic.0.024828-0.
15. Mateos M, Hernandez-Garcia M, Del Campo R, Martinez-Garcia L, Gijon D, Morosini MI, Ruiz-Garbajosa P, Canton R. 2020. Emergence and persistence over time of carbapenemase-producing *Enterobacter* isolates in a Spanish university hospital in Madrid, Spain (2005–2018). Microb Drug Resist 27:895–903. https://doi.org/10.1089/mdr.2020.0265.
16. Hasman H, Saputra D, Sicheritz-Ponten T, Lund O, Svendsen CA, Frimodt-Moller N, Aarestrup FM. 2014. Rapid whole-genome sequencing for detection and characterization of microorganisms directly

Downloaded from https://journals.asm.org/journal/jcm on 05 April 2024 by 89.206.81.44.

from clinical samples. J Clin Microbiol 52:139–146. https://doi.org/10.1128/JCM.02452-13.

17. Larsen MV, Cosentino S, Lukjancenko O, Saputra D, Rasmussen S, Hasman H, Sicheritz-Ponten T, Aarestrup FM, Ussery DW, Lund O. 2014. Benchmarking of methods for genomic taxonomy. J Clin Microbiol 52:1529–1539. https://doi.org/10.1128/JCM.02981-13.
18. Clausen P, Aarestrup FM, Lund O. 2018. Rapid and precise alignment of raw reads against redundant databases with KMA. BMC Bioinformatics 19:307. https://doi.org/10.1186/s12859-018-2336-6.
19. Godmer A, Benzerara Y, Normand AC, Veziris N, Gallah S, Eckert C, Morand P, Piarroux R, Aubry A. 2021. Revisiting species identification within the *Enterobacter cloacae* complex by matrix-assisted laser desorption ionization-time of flight mass spectrometry. Microbiol Spectr 9:e00661-21. https://doi.org/10.1128/Spectrum.00661-21.
20. Wang YQ, Xiao D, Li J, Zhang HF, Fu BQ, Wang XL, Ai XM, Xiong YW, Zhang JZ, Ye CY. 2018. Rapid identification and subtyping of *Enterobacter cloacae* clinical isolates using peptide mass fingerprinting. Biomed Environ Sci 31:48–56. https://doi.org/10.3967/bes2018.005.

Downloaded from https://journals.asm.org/journal/jcm on 05 April 2024 by 89.206.81.44.