Automatic discrimination of species within the *Enterobacter cloacae* complex using MALDI-TOF Mass Spectrometry and supervised algorithms

Ana Candela[1*], Alejandro Guerrero-López[2*], Miriam Mateos[3], Alicia Gómez-Asenjo[1], Manuel J. Arroyo[4], Marta Hernandez-García[3,5], Rosa del Campo[3,5], Emilia Cercenado[1,6,7], Aline Cuénod[8,9], Gema Méndez[4], Luis Mancera[4], Juan de Dios Caballero[3,5], Laura Martínez-García[3,10], Desirée Gijón[3,5], María Isabel Morosini[3,5], Patricia Ruiz-Garbajosa[3,5], Adrian Egli[8,9], Rafael Cantón[3,5], Patricia Muñoz[1,6,7], David Rodríguez-Temporal[1†] and Belén Rodríguez-Sánchez[1].

[1]Clinical Microbiology and Infectious Diseases Department, Hospital General Universitario Gregorio Marañón, Madrid, Spain, and Institute of Health Research Gregorio Marañón (IiSGM), Madrid, Spain

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

[3]Servicio de Microbiología. Hospital Universitario Ramón y Cajal and Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS), Madrid, Spain.

[4]Clover Bioanalytical Software, Av. del Conocimiento, 41, 18016 Granada, Spain

[5]CIBER en Enfermedades Infecciosas, Madrid, Spain

[6]CIBER de Enfermedades Respiratorias (CIBERES CB06/06/0058), Madrid, Spain.

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

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

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

[10]Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain.

Running Title: *E. cloacae* complex classification with MALDI-TOF MS

[†]**Corresponding author:**

David Rodríguez-Temporal, PhD.

Servicio de Microbiología Clínica y Enfermedades Infecciosas. Hospital General Universitario Gregorio Marañón.Dr. Esquerdo 46. 28007 Madrid, Spain

Phone: +34- 91- 426 7163, Fax: +34- 91- 426 9595

E-mail: david.rodriguez@iisgm.com

*These authors contributed equally to the study

ABSTRACT

The *Enterobacter cloacae* complex (ECC) encompasses heterogeneous clusters of species that have been associated with nosocomial outbreaks. These species may host different acquired antimicrobial resistance and virulence mechanisms and their identification are challenging. This study aims to develop predictive models based on MALDI-TOF MS spectral profiles and machine learning for species-level identification.

A total of 198 ECC and 116 *K. aerogenes* clinical isolates from the University Hospital Ramón y Cajal (Spain) and the University Hospital Basel (Switzerland) were included. The capability of the proposed method to differentiate the most common ECC species (*E. asburiae*, *E. kobei*, *E. hormaechei*, *E. roggenkampii*, *E. ludwigii*, *E. bugandensis*) and *K. aerogenes* was demonstrated by applying unsupervised hierarchical clustering with PCA pre-processing. 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 dataset. Thus, we developed supervised, non-linear predictive models (Support Vector Machine with Radial Basis Function and Random Forest). The external validation of these models with protein spectra from the 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. Similar results were obtained with the MSI database developed recently (https://msi.happy-dev.fr/) except in the case of *E. hormaechei*, which was more accurately identified by Random Forest.

In short, MALDI-TOF MS combined with machine learning demonstrated to be a rapid and accurate method for the differentiation of ECC species.

Keywords: *Enterobacter* species; mass spectrometry; MALDI-TOF MS; Peak analysis; Machine Learning

INTRODUCTION

Enterobacter is a facultative anaerobic Gram-negative genus that can be found as a natural commensal in the gut microbiome of mammals (1). Several species have been associated with nosocomial outbreaks causing urinary tract infection, skin and soft tissue infection, 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: *E. asburiae* (cluster I), *E. kobei* (cluster II), *E. hormaechei* subsp. *hoffmannii* (cluster III), *E. roggenkampii* (cluster IV), *E. ludwigii* (cluster V), *E. hormaechei* subsp. *oharae,* and subsp. *xiangfangensis* (cluster VI), *E. hormaechei* subsp. *hormaechei* (cluster VII), *E. hormaechei* subsp. *steigerwaltii* (cluster VIII), *E. bugandensis* (cluster IX), *E. nimipressuralis* (cluster X), *E. cloacae* subsp. *cloacae* (cluster XI), *E. cloacae* subsp. *dissolvens* (cluster XII), and a heterogeneous group of *E. cloacae* sequences are considered as 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 *K. 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 22 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 multi-locus sequence typing (MLST) and WGS have also been applied (5, 9, 10). Sequence-based diagnostic methods techniques 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 an alternative to sequence-based methods. MALDI-TOF MS has shown to be an excellent methodology for bacterial identification. It can easily identify *E. cloacae*

complex isolates but it showed low discrimination power for the species in this group when using standard analysis and commercial databases with low resolution (11, 12).

This study aimed to develop and validate prediction models for the 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). Besides, *E. hormaechei*, the most encountered ECC species in the 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, two steps were conducted in this study. First, we performed an unsupervised clustering to determine the feasibility of MALDI-TOF MS data for the ECC species identification. Second, we applied a supervised machine learning algorithm with isolates from University Hospital Ramón y Cajal, Madrid -Madrid, Spain - and validated our findings with different ECC isolates from the same hospital and from the University Hospital of Basel, Switzerland.

MATERIALS AND METHODS

Bacterial isolates

Overall, we analyzed 198 clinical isolates belonging to the ECC and nine 116 *K. aerogenes* (formerly *E. aerogenes*). Among them, 164 ECC and 9 *K. aerogenes* were collected in a surveillance study of antimicrobial resistance in the Hospital Universitario Ramón y Cajal -UHRC- (Madrid, Spain) between 2005 and 2018 and identified by partial sequencing of the *hsp60* gene (15). We collected the remaining isolates (34 ECC and 107 *K. aerogenes*) at the University Hospital Basel (UHB; Basel, Switzerland) between 2016 and 2021 and identified the isolates by whole genome sequencing (WGS) using KmerFinder 3.2 (16-18). MALDI-TOF MS spectral profiles of these

isolates were obtained in Basel and submitted to the Hospital General Universitario Gregorio Marañón (Madrid, Spain) for further analysis.

All isolates from UHRC were incubated overnight at 37ºC and metabolically activated after three subcultures on Columbia Blood Agar (bioMérieux, Marcy l'Etoile, France) before their analysis with MALDI-TOF MS at the Hospital General Universitario Gregorio Marañón.

Spectra acquisition using MALDI-TOF MS

We identified the isolates using the MBT Smart MALDI Biotyper (Bruker Daltonics, Bremen, Germany). We spotted all strains from UHRC in duplicate onto the MALDI target plate and overlaid with 1 µl of 70% formic acid. After drying at room temperature, we covered and dried the spots with 1 µl HCCA matrix, according to the manufacturer's indications (Bruker Daltonics). We obtained two spectra in the range of 2,000-20,000 Da on each spot, resulting in 4 spectra per isolate. The isolates from UHB were analysed in one spot per strain and one spectrum from spot was acquired.

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 spectral profiles with the Clover MS Data Analysis software (Clover Biosoft, Granada, Spain). We applied pre-processing pipeline to all protein spectra that consisted of: 1) smoothing -Savitzky-Golay Filter: window length=11; polynomial order=3- and baseline subtraction -Top-Hat filter method with factor=0.02-; 2) creation of an average spectrum per isolate; 3) alignment of the average spectra from different isolates -shift: medium; constant tolerance: 2 Da; linear mass tolerance: 600 ppm-; 4) normalization by Total Ion Current (TIC).

<u>Unsupervised feasibility study</u>

To study the feasibility of MALDI-TOF MS for differentiation of ECC species, we proposed an unsupervised study based on Principal Component Analysis (PCA), and t-distributed Stochastic Neighbour Embedding (t-SNE). For this purpose, an oversampled balanced dataset of each ECC species was used. We included in total, 126 spectra from the 7 ECC species analysed in this study (sourcing from UHRC and UHB), as indicated in **Table 1**.

Supervised model development

Once the feasibility of the study was determined, we proposed the supervised model development. In this case, three different datasets were created: training validation set, internal validation set, and external validation set. The details of these datasets are shown in **Table 1**.

Due to the lack of validation samples of *E. ludwigii* and *E. bugandensis*, these were not included in the development of the supervised model. Therefore, our supervised model was developed to predict the five ECC species: *E. asburiae* (cluster I), *E. kobei* (cluster II), and *E. hormaechei* (clusters III, VI, and VIII considered together), and *E. roggenkampii* (cluster IV). We applied three different supervised models: Partial Least Square-Discriminant Analysis (PLS-DA), Support Vector Machine (SVM) with Linear kernel (SVM-L) and with Radial Basis Function (SVM-R) kernel, and Random Forest (RF). The hyperparameter selection was performed by a 5-fold cross-validation technique.

Finally, we performed two external validations of the predictive models. First, 126 MALDI-TOF MS from UHRC and then 141 MALDI-TOF MS from UHB were blindly classified by the same predictive models using Clover BioSoft v0.6.1. This software uses the scikit-learn 0.23.2 python library to implement all statistical methods used in this study. For reproducibility purposes under FAIR principles, free access to all spectra

and to reproduce the analyses in this work can be found at the following url: https://platform.clovermsdataanalysis.com/public-repository.

MSI Database

Recently, an online database has been developed for the rapid differentiation of ECC species based on their MALDI-TOF MS protein profile (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 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 datasets were also identified using the MSI database as a comparison to the methods proposed in this article. As stated above, MALDI-TOF MS spectra associated to this study have been also made publicly available.

Ethics statement

The Ethics Committee of the Gregorio Marañón Hospital (CEIm) evaluated this project and considered that all the conditions for waiving informed consent were met since the study was conducted with microbiological samples and not with human products. At the University Hospital Basel only anonymized data was used with the purpose of quality control and assay validation. According to the Swiss Human Research Act no specific consent is required in this case. Data was either acquired in routine microbiological diagnostics (excluding cases with a rejected general consent) or used from a previously published dataset (DRIAMS).

RESULTS

Feasibility study

To prove the feasibility of MALDI-TOF MS to differentiate ECC species, an unsupervised hierarchical clustering with PCA and t-SNE pre-processing was performed (**Figure 1**). The protein spectra of the seven ECC 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 containing *E. hormaechei*, a second 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 did *E. asburiae*, *E. ludwigii*, *E. kobei,* and *E. roggenkampii*, although in these four cases some of the spectra were clustered with the wrong species (**Figure 1A, 1B and 1C**).

The implementation of PCA to reduce the dimensionality showed that 14 components were needed to explain 95% of the variance (**Figure 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 solve 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 dataset 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: for PLS-DA 2 components were used, for SVM-L the value of C was 10, and for SVM-R the value of C was 10 and the value of γ was 1000. **Table 2** shows the results obtained for the internal 5-fold cross-validation, which have been further detailed in **Table S1**.

Both *E. hormaechei* and *K. aerogenes* presented the same trend as in the feasibility study and their differentiation was 100% using non-linear 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 analysed ECC species, we also obtained 100% correct classification by the application of non-linear approaches (**Table 2**).

In **Figure 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 dataset 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 dataset sourcing from the same hospital (UHRC). In this case, all *K. aerogenes, E. asburiae,* and *E. kobei* isolates were correctly classified meanwhile one *E. hormaechei* strain was misclassified as *E. kobei* with both algorithms. For *E. roggenkampii*, two isolates were misclassified as *E. hormeachei* and one as *E. kobei* (**Table S2**). The accuracy of the model is shown in **Figures 3A and 3B**.

Since SVM-R and RF algorithms performed equally, both of them were considered for external validation with MALDI-TOF MS spectral profiles obtained at the UHB. In this case, 91.2% of the *E. hormaechei* (n=33), 100% of the *E. roggenkampii* (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*. **Figures 3C and 3C** show the accuracy of both SVM-R and RF for the external validation collection from UHB.

Identification of the ECC isolates using the MSI Database

Finally, the MSI Platform was also used as an identification tool for ECC species to compare the automatic approach developed in this study versus the current state-of-the-art method (19). Among the UHRC isolates, the identification rate for *E. asburiae*, *E. hormaechei,* and *E. kobei* was similar to the rates yielded by the predictive

models developed in this study (**Table 3**). The only difference detected between both methods was that the MSI database correctly classified one more *E. roggenkampii* isolate. As for the protein spectra sourcing from the UHB, 84.8% of the *E. hormaechei* 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, non-linear algorithms (SVM-R and RF) to MALDI-TOF MS spectra allowed the correct species assignment of 100% 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*) sourcing from two different hospitals.

Poor discrimination of *E. cloacae* complex species by MALDI-TOF MS has been previously reported either by using commercial (11, 15) or enriched, in-house databases (20). 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 and correct discrimination of 97.0% *E. hormaechei* isolates (19). This approach can be useful for the discrimination of close-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 non-linear algorithms proposed by our approach.

In this study, we demonstrate the feasibility of 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 as

reported before (20). Secondly, 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 non-linear 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 than the spectral profile used for model training, we performed two validation experiments. First, we carried out a validation with MADI-TOF MS protein spectra sourcing from UHRC. From a total number of samples of 116, both SVM-R and RF only misclassified four isolates, i.e., a 96.5% of accuracy in classifying species within the ECC was yielded. Secondly, we performed an external validation with MALDI-TOF MS sourcing from UHB to simulated a real-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 showed to be overfitted to the UHRC distribution, which was already pointed out by the value of γ value, scoring an 83.7% of accuracy. On the other hand, the current state-of-the-art tool -the MSI database- performed better than SVM-R with 94.9% of accuracy although it was not able to distinguish the *K. aerogenes* (19). However, RF outperformed both approaches with over 96.0% of accuracy in identifying the three species. Hence, 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 analysed collection (15). However, the present study provides the first proof of concept for differentiating ECC species based on machine learning. For a definitive validation, improvement, and implementation of these predictive models, future studies will involve

strains from a more diverse epidemiological and geographical origin and characteristics. Besides, not all analysed ECC species could be represented in the external validation dataset due to the lack of isolates from the species *E. ludwigii* and *E. bugandensis*.

The present study provides promising results for differentiating ECC species based on machine learning and MALDI-TOF MS protein spectra. It also highlights the facts that MALDI-TOF MS data should be linked to WGS data in order to allow future work and providing 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. The use of spectra analysis tools is becoming user-friendly and easy to apply and its 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 by the community freely.

CONFLICT OF INTERESTS

The authors have no conflict of interests.

ACKNOWLEDGMENTS

This work was supported by the projects PI15/01073 and PI18/00997 from the Health Research Fund (FIS. Instituto de Salud Carlos III. Plan Nacional de I+D+I 2013-2016) of the Carlos III Health Institute (ISCIII, Madrid, Spain) partially financed by the European Regional Development Fund (FEDER) ‘A way of making Europe’. MM was funded by the Community of Madrid (Programa de Garantía Juvenil, PEJD-2017-PRE/BMD-5106) and DRT with a postdoc contract from the Intramural Funding

Program of the IiSGM. BRS (CPII19/00002) is a recipient of a Miguel Servet contract supported by the FIS. AGL was funded by a predoc contract from the Intramural Funding Program of the IiSGM.

AUTHOR CONTRIBUTIONS

Ana Candela: experimental part, formal analysis, data collection, validation, visualization, writing – original draft preparation and review/editing. Alejandro Guerrero-López: formal analysis, data collection, validation, visualization, writing – original draft preparation and review/editing. Miriam Mateos and Alicia Gómez-Asenjo: experimental part, formal analysis, and data collection. Manuel J. Arroyo, Gema Méndez, and Luis Mancera: data analysis, validation, writing – original draft preparation and review/editing. Marta Hernández-García, Rosa del Campo: experimental part, formal analysis, and writing, submission of isolates, original draft preparation, and review/editing. Aline Cuénod and Adrian Egli: submission of isolates, original draft preparation, and review/editing. Juan de Dios Caballero, Laura Martínez-García, Desirée Gijón, María Isabel Morosini, Patricia Ruiz-Garbajosa, Rafael Cantón and Patricia Muñoz: validation, writing and review/editing. Emilia Cercenado: conceptualization, formal analysis, validation, writing, and review/editing. David Rodríguez-Temporal: conceptualization, formal analysis, validation, original draft preparation and review/editing; Belén Rodríguez-Sánchez: conceptualization, project administration, formal analysis, supervision, validation, visualization, original draft preparation and review/ editing.

FIGURE LEGENDS

Figure 1. A. Dendrogram built with 126 MALDI-TOF MS spectra. **B**. PCA of feasibility study spectra. **C.** t-SNE of study spectra. **D**. PCA Eigenvalues showing the variance of each component.

Figure 2. MALDI-TOF MS Euclidean distance between species by Random Forest classifier.

Figure 3. ROC Curves and AUC values for both SVM-R (A and C) and RF models (B and D) were applied to the external validations with MALDI-TOF MS protein spectra from UHRC (A and B) and UHB (C and D).

REFERENCES

1. Mezzatesta ML, Gona F, Stefani S. 2012. Enterobacter cloacae complex: clinical impact and emerging antibiotic resistance. Future Microbiol 7:887-902.
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.
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 31:2951-5.
4. Davin-Regli A, Lavigne JP, Pages JM. 2019. Enterobacter spp.: Update on Taxonomy, Clinical Aspects, and Emerging Antimicrobial Resistance. Clin Microbiol Rev 32.
5. Hoffmann H, Roggenkamp A. 2003. Population genetics of the nomenspecies Enterobacter cloacae. Appl Environ Microbiol 69:5306-18.
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.
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.
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.
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.

379 10. Hoffmann H, Stindl S, Ludwig W, Stumpf A, Mehlen A, Heesemann J, Monget D,
380 Schleifer KH, Roggenkamp A. 2005. Reassignment of enterobacter dissolvens to
381 Enterobacter cloacae as E. cloacae subspecies dissolvens comb. nov. and emended
382 description of Enterobacter asburiae and Enterobacter kobei. Syst Appl Microbiol
383 28:196-205.
384 11. Pavlovic M, Konrad R, Iwobi AN, Sing A, Busch U, Huber I. 2012. A dual approach
385 employing MALDI-TOF MS and real-time PCR for fast species identification within the
386 Enterobacter cloacae complex. FEMS microbiology letters 328:46-53.
387 12. De Florio L, Riva E, Giona A, Dedej E, Fogolari M, Cella E, Spoto S, Lai A, Zehender G,
388 Ciccozzi M, Angeletti S. 2018. MALDI-TOF MS Identification and Clustering Applied to
389 Enterobacter Species in Nosocomial Setting. Frontiers in microbiology 9:1885.
390 13. Barnes AI, Paraje MG, del CBP, Albesa I. 2001. Molecular properties and metabolic
391 effect on blood cells produced by a new toxin of Enterobacter cloacae. Cell Biol Toxicol
392 17:409-18.
393 14. Paauw A, Caspers MPM, Leverstein-van Hall MA, Schuren FHJ, Montijn RC, Verhoef J,
394 Fluit AC. 2009. Identification of resistance and virulence factors in an epidemic
395 Enterobacter hormaechei outbreak strain. Microbiology (Reading) 155:1478-1488.
396 15. Mateos M, Hernandez-Garcia M, Del Campo R, Martinez-Garcia L, Gijon D, Morosini
397 MI, Ruiz-Garbajosa P, Canton R. 2020. Emergence and Persistence over Time of
398 Carbapenemase-Producing Enterobacter Isolates in a Spanish University Hospital in
399 Madrid, Spain (2005-2018). Microbial drug resistance.
400 16. Hasman H, Saputra D, Sicheritz-Ponten T, Lund O, Svendsen CA, Frimodt-Moller N,
401 Aarestrup FM. 2014. Rapid whole-genome sequencing for detection and
402 characterization of microorganisms directly from clinical samples. J Clin Microbiol
403 52:139-46.
404 17. Larsen MV, Cosentino S, Lukjancenko O, Saputra D, Rasmussen S, Hasman H, Sicheritz-
405 Ponten T, Aarestrup FM, Ussery DW, Lund O. 2014. Benchmarking of methods for
406 genomic taxonomy. J Clin Microbiol 52:1529-39.
407 18. Clausen P, Aarestrup FM, Lund O. 2018. Rapid and precise alignment of raw reads
408 against redundant databases with KMA. BMC Bioinformatics 19:307.
409 19. Godmer A, Benzerara Y, Normand AC, Veziris N, Gallah S, Eckert C, Morand P, Piarroux
410 R, Aubry A. 2021. Revisiting Species Identification within the Enterobacter cloacae
411 Complex by Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass
412 Spectrometry. Microbiol Spectr 9:e0066121.
413 20. Wang YQ, Xiao D, Li J, Zhang HF, Fu BQ, Wang XL, Ai XM, Xiong YW, Zhang JZ, Ye CY.
414 2018. Rapid Identification and Subtyping of Enterobacter cloacae Clinical Isolates Using
415 Peptide Mass Fingerprinting. Biomed Environ Sci 31:48-56.

416

417

418

419

420

421

Table 1. Number of ECC isolates used for the unsupervised feasibility study and the supervised model development.

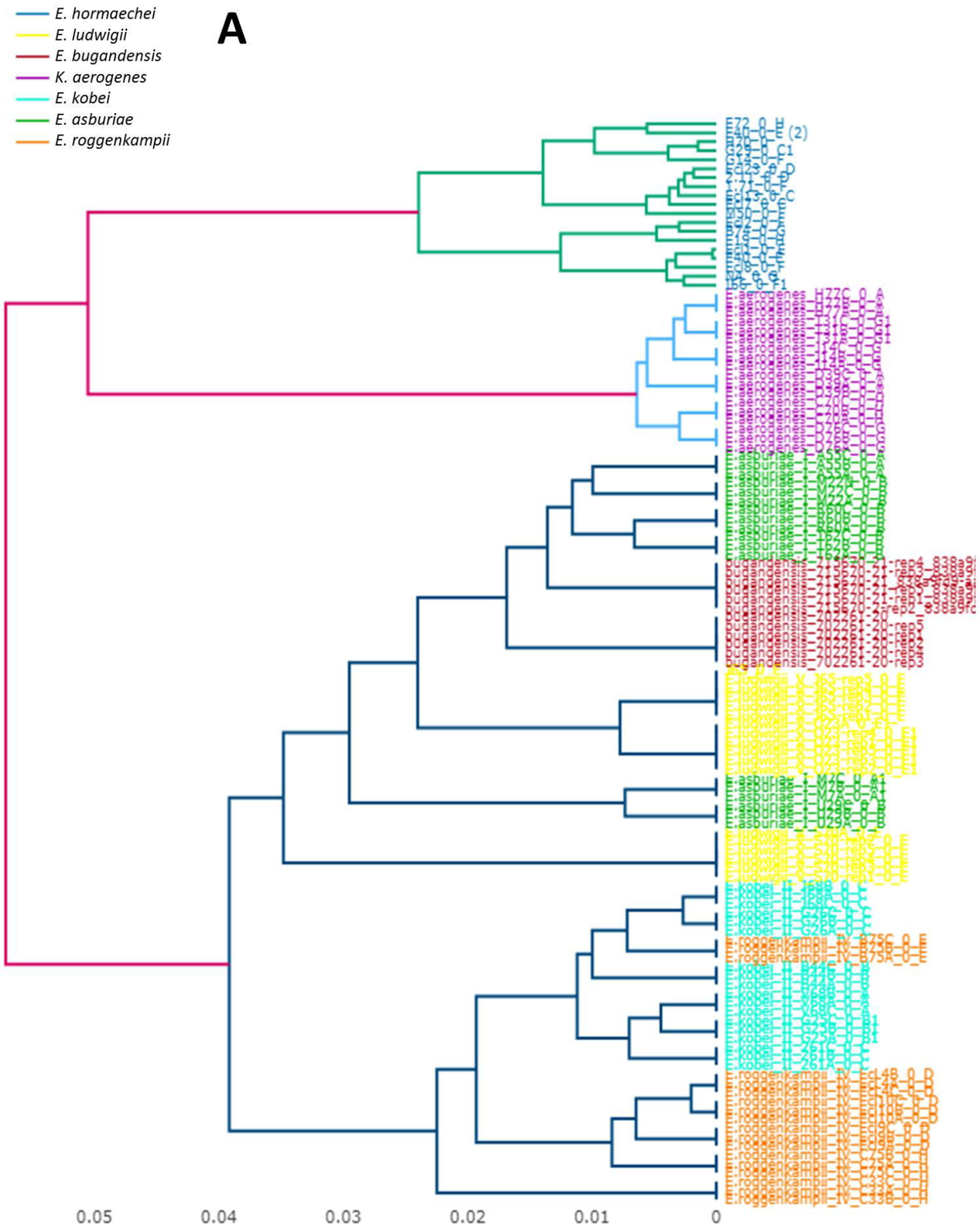
ECC species	Unsupervised study	Supervised study		
	Number of balanced samples	Development dataset (UHRC)	Validation dataset (UHRC)	External validation dataset (UHB)
K. aerogenes	18	18	3	107
E. asburiae	18	18	1	0
E. bugandensis	18	0	0	0
E. hormaechei	18	18	51	33
E. kobei	18	18	9	0
E. ludwigii	18	0	0	0
E. roggenkampii	18	18	62	1
Total	**126**	**90**	**126**	**141**

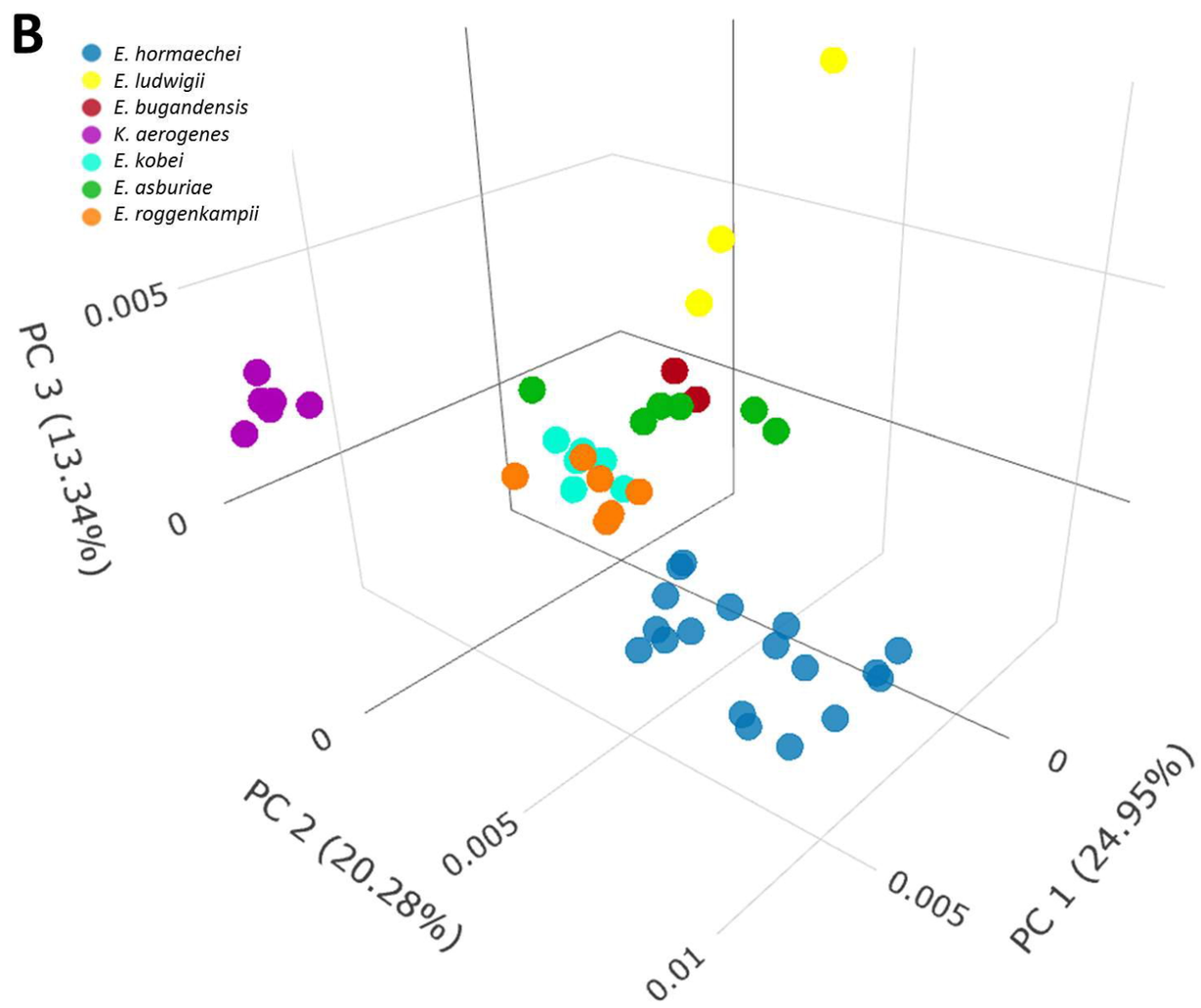
Table 2. Accuracy results for internal 5-fold cross-validation over development dataset (90 spectral profiles). PLS-DA: Partial Least Squares-Discriminant Analysis; SVM-L: Support Vector Machine-Linear kernel; SVM-R: Support Vector Machine-Radial Basis Function kernel; RF: Random Forest.

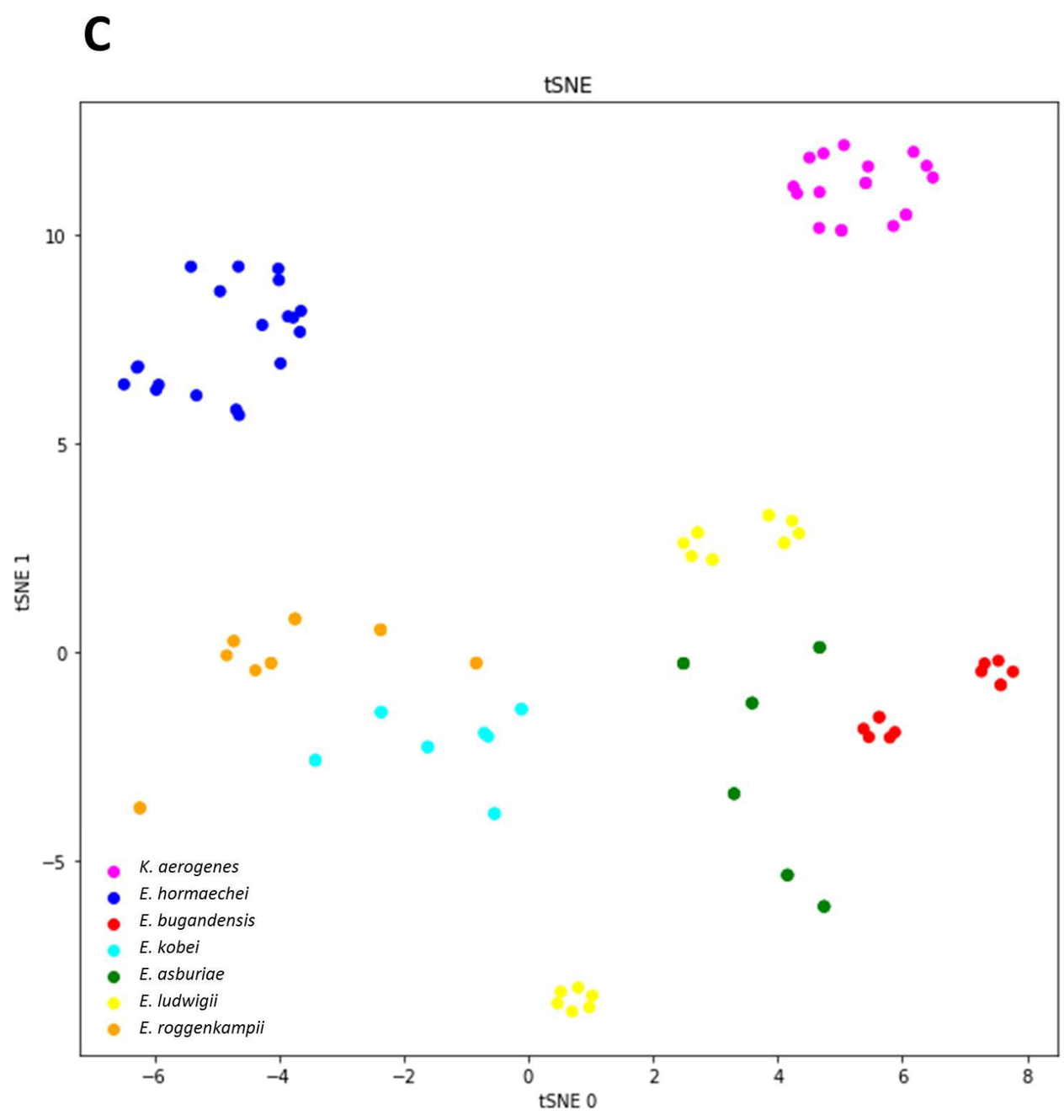
Algorithm	*E. asburiae*		*E. hormaechei*		*E. kobei*		*E. roggenkampii*		*K. aerogenes*	
PLS-DA	9/18	50%	18/18	100%	5/18	27.8%	4/18	22.2%	18/18	100%
SVM-L	6/18	33.3%	18/18	100%	3/18	16.7%	6/18	33.3%	14/18	77.8%
SVM-R	**18/18**	**100%**	**18/18**	**100%**	**18/18**	**100%**	**18/18**	**100%**	**18/18**	**100%**
RF	**18/18**	**100%**	**18/18**	**100%**	**18/18**	**100%**	**18/18**	**100%**	**18/18**	**100%**

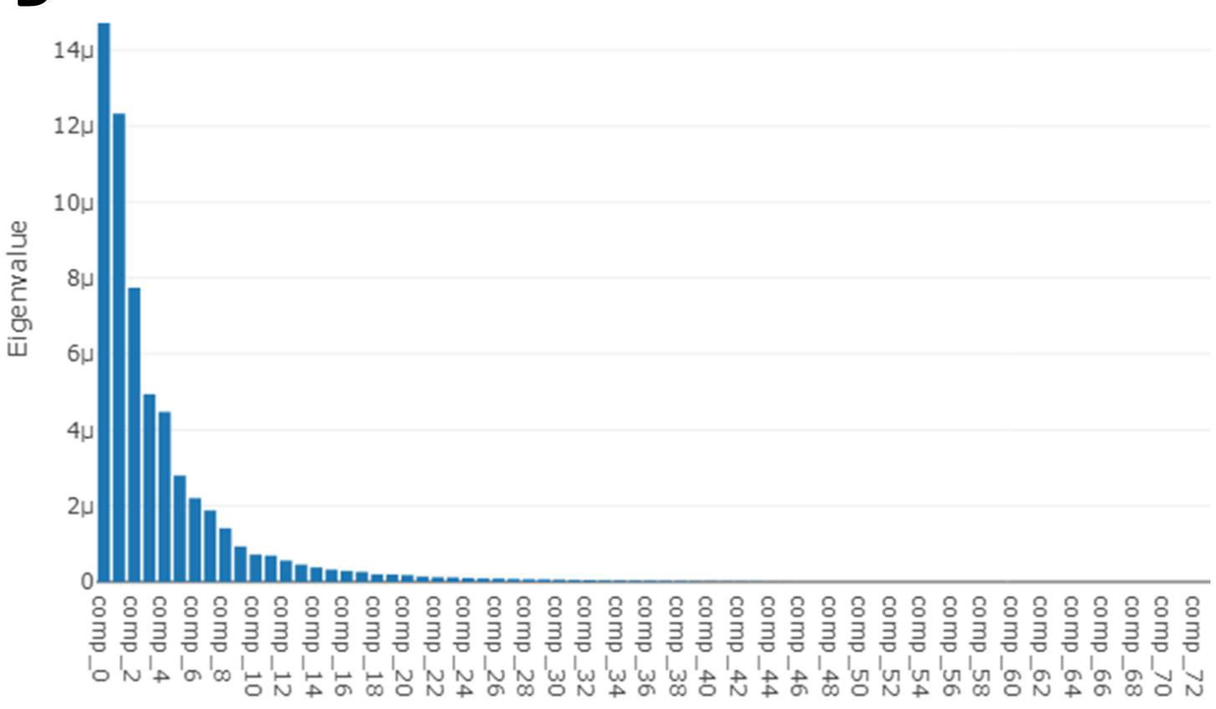
Table 3. Accuracy results for the validation dataset from UHRC and UHB, and the identification accuracy obtained by the MSI database. The MSI platform, specific for the identification of ECC species (19), was also applied for the classification of ECC species.

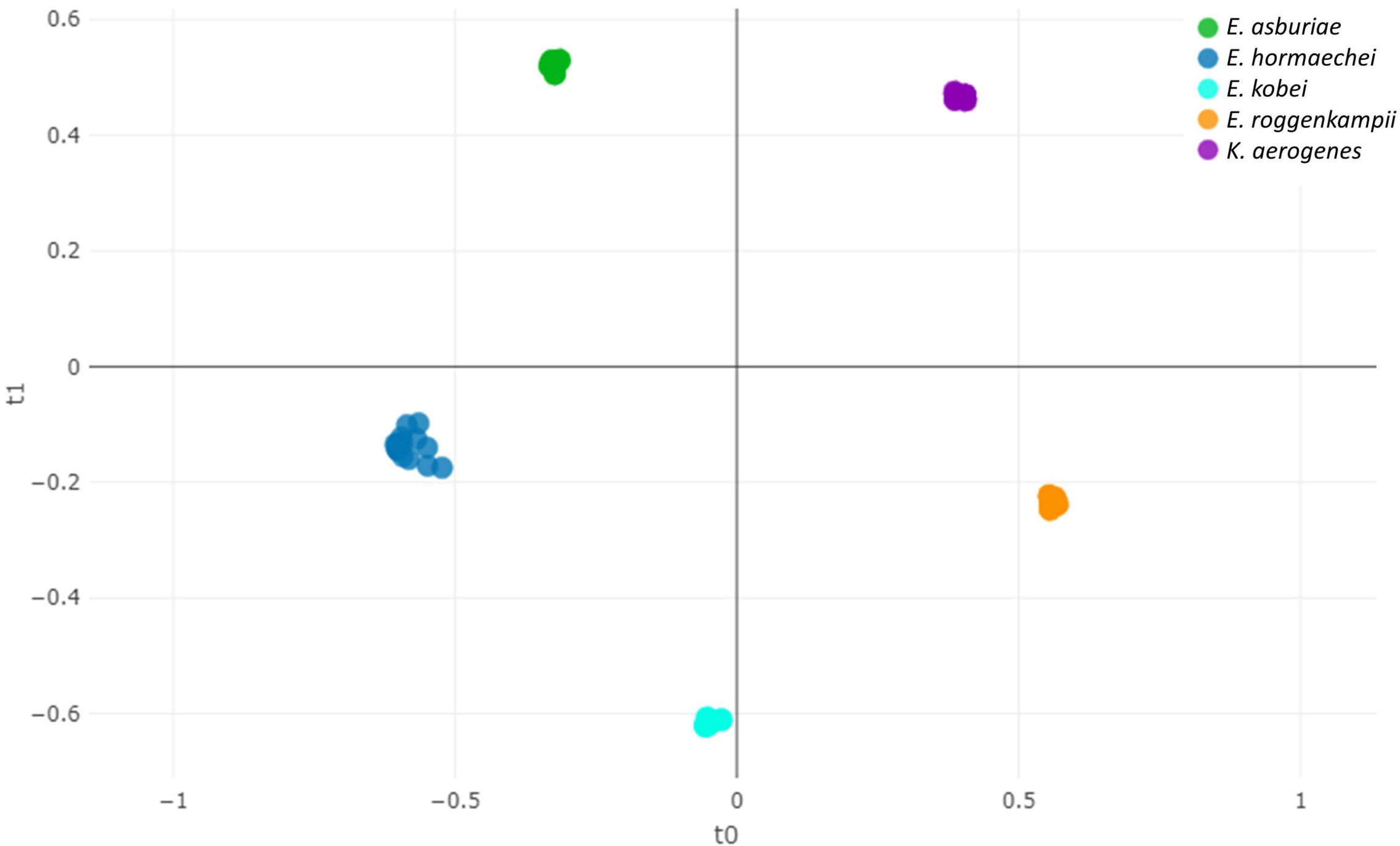
Algorithm	*E. asburiae*		*E. hormaechei*		*E. kobei*		*E. roggenkampii*		*K. aerogenes*	
UHRC										
SVM-R	1/1	100%	50/51	98.0%	9/9	100%	59/62	95.2%	3/3	100%
RF	1/1	100%	50/51	98.0%	9/9	100%	59/62	95.2%	3/3	100%
MSI	1/1	100%	50/51	98.0%	9/9	100%	60/62	96.8%	-	-
UHB										
SVM-R	-	-	15/33	44.1%	-	-	1/1	100%	102/107	95.3%
RF	-	-	31/33	91.2%	-	-	1/1	100%	105/107	98.1%
MSI	-	-	28/33	84.8%	-	-	1/1	100%	-	-

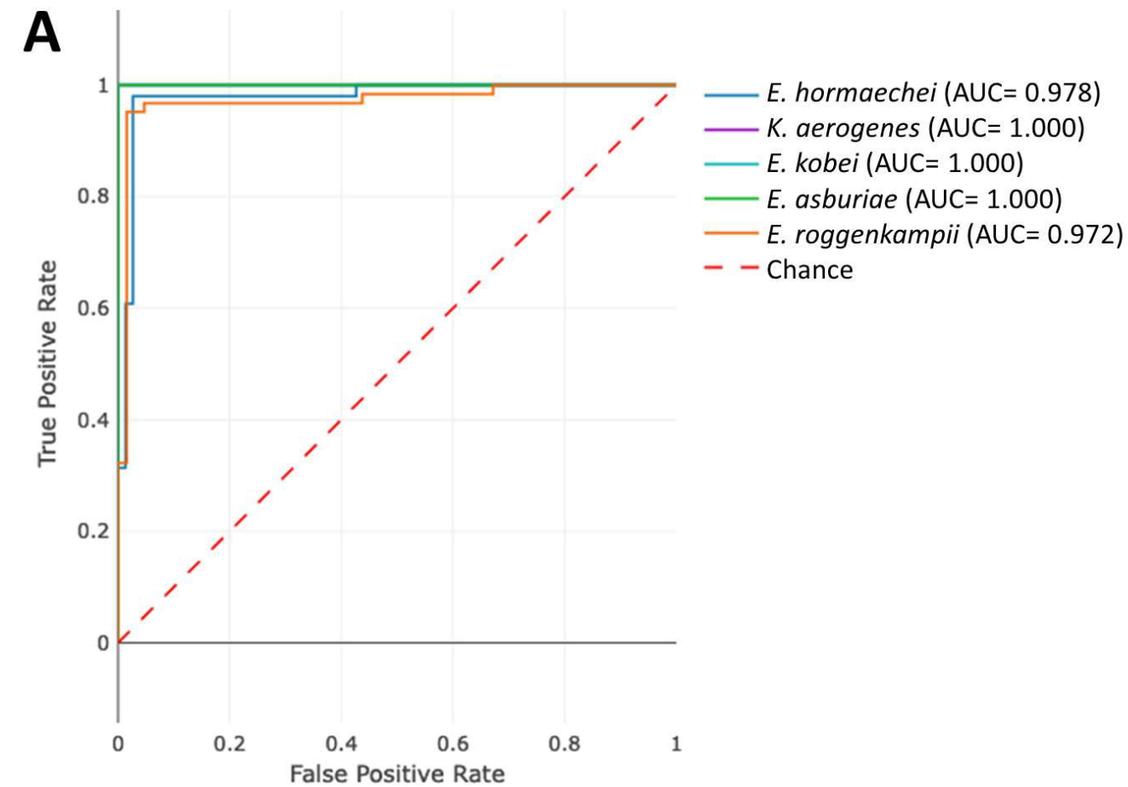

A
E. hormaechei
E. ludwigii
E. bugandensis
K. aerogenes
E. kobei
E. asburiae
E. roggenkampii
0.05
0.04
0.03
0.02
0.01
0

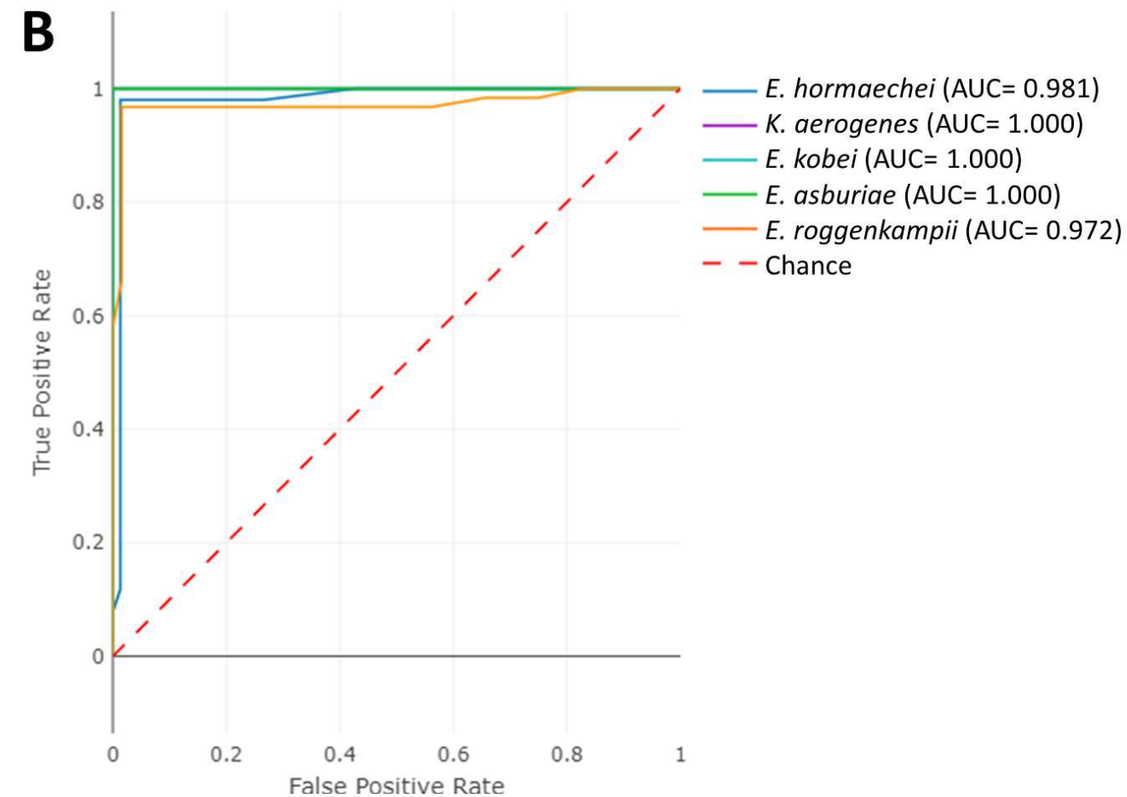

B
E. hormaechei
E. ludwigii
E. bugandensis
K. aerogenes
E. kobei
E. asburiae
E. roggenkampii
PC 3 (13.34%)
PC 2 (20.28%)
PC 1 (24.95%)
0.005
0
0.01

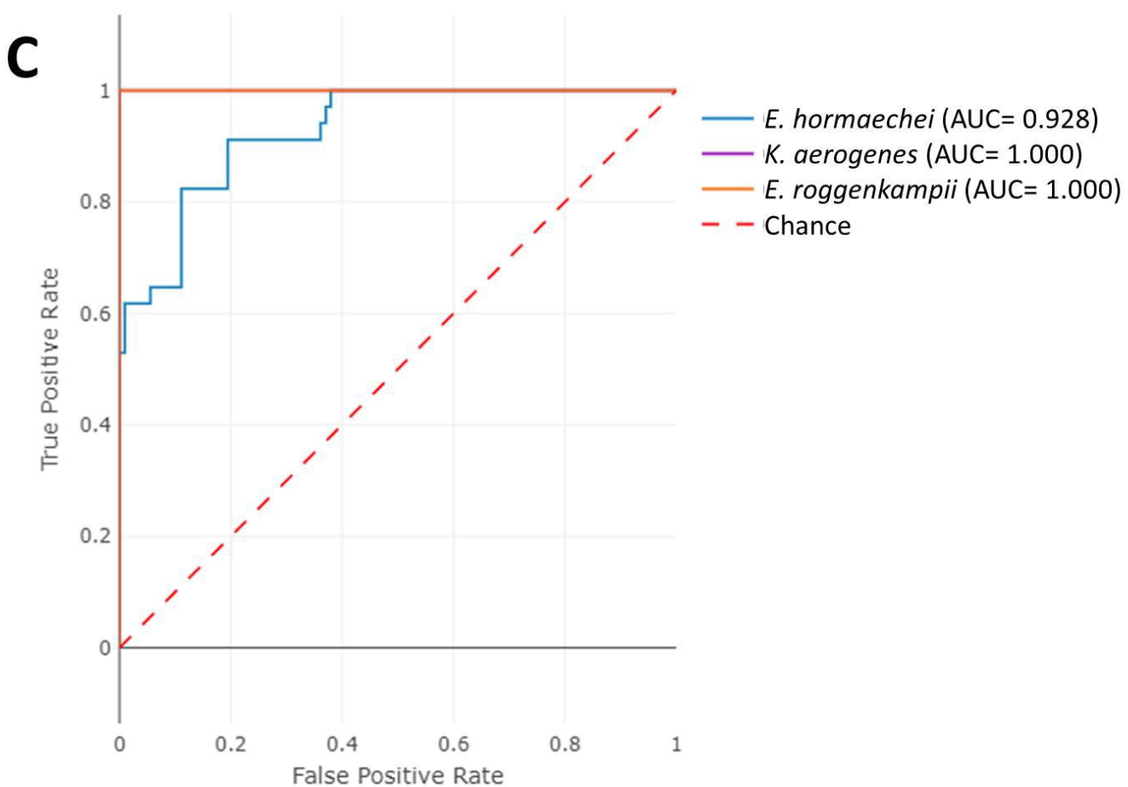

C
tSNE
tSNE 1
tSNE 0
K. aerogenes
E. hormaechei
E. bugandensis
E. kobei
E. asburiae
E. ludwigii
E. roggenkampii

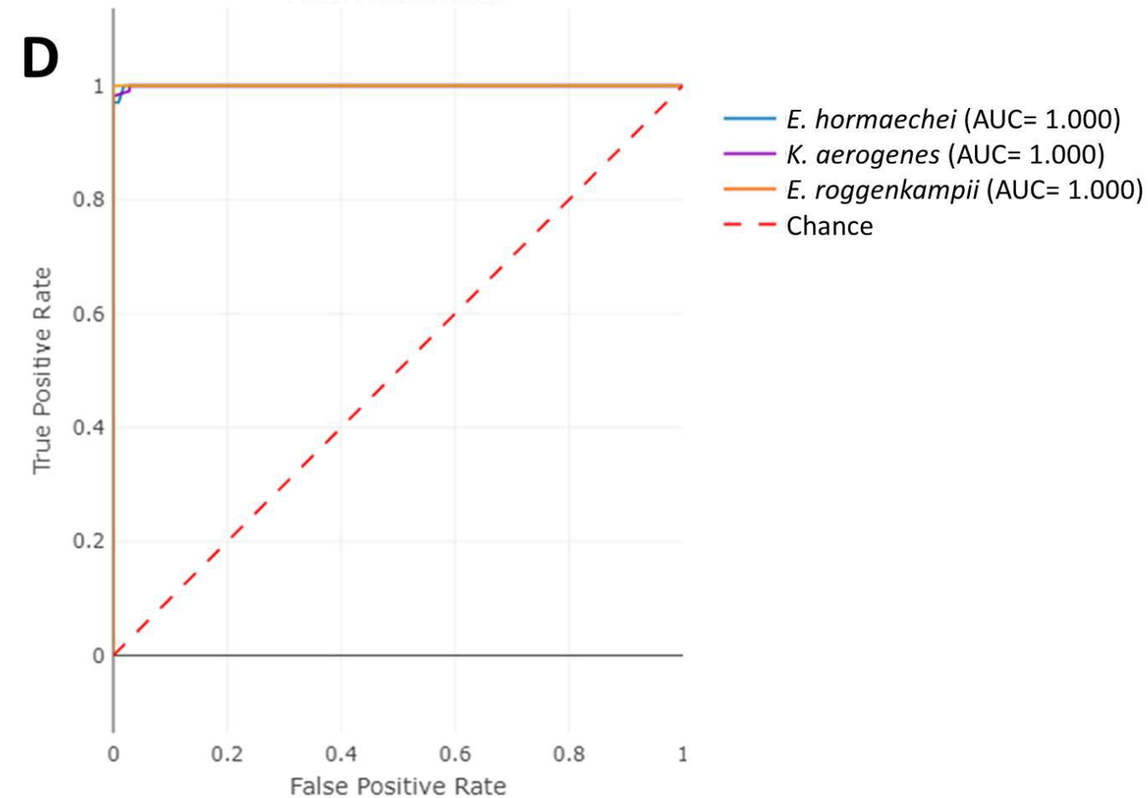

D
Eigenvalue
14µ
12µ
10µ
8µ
6µ
4µ
2µ
0
comp_0
comp_2
comp_4
comp_6
comp_8
comp_10
comp_12
comp_14
comp_16
comp_18
comp_20
comp_22
comp_24
comp_26
comp_28
comp_30
comp_32
comp_34
comp_36
comp_38
comp_40
comp_42
comp_44
comp_46
comp_48
comp_50
comp_52
comp_54
comp_56
comp_58
comp_60
comp_62
comp_64
comp_66
comp_68
comp_70
comp_72


E. asburiae
E. hormaechei
E. kobei
E. roggenkampii
K. aerogenes
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
t1
−1
−0.5
0
0.5
1
t0


A
True Positive Rate
False Positive Rate
0
0.2
0.4
0.6
0.8
1
E. hormaechei (AUC= 0.978)
K. aerogenes (AUC= 1.000)
E. kobei (AUC= 1.000)
E. asburiae (AUC= 1.000)
E. roggenkampii (AUC= 0.972)
Chance


B
True Positive Rate
False Positive Rate
0
0.2
0.4
0.6
0.8
1
E. hormaechei (AUC= 0.981)
K. aerogenes (AUC= 1.000)
E. kobei (AUC= 1.000)
E. asburiae (AUC= 1.000)
E. roggenkampii (AUC= 0.972)
Chance


C
True Positive Rate
False Positive Rate
0
0.2
0.4
0.6
0.8
1
E. hormaechei (AUC= 0.928)
K. aerogenes (AUC= 1.000)
E. roggenkampii (AUC= 1.000)
Chance


D
True Positive Rate
False Positive Rate
0
0.2
0.4
0.6
0.8
1
E. hormaechei (AUC= 1.000)
K. aerogenes (AUC= 1.000)
E. roggenkampii (AUC= 1.000)
Chance