bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467040; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1 **Discrimination of species within the *Enterobacter cloacae* complex using MALDI-**
2 **TOF Mass Spectrometry and Fourier-Transform Infrared Spectroscopy coupled**
3 **with Machine Learning tools**

4 **Ana Candela[1*], Miriam Mateos[2*], Alicia Gómez-Asenjo[1], Manuel J. Arroyo[3], Marta**
5 **Hernandez-García[2,4], Rosa del Campo[2,4], Emilia Cercenado[1,5,6], Gema Méndez[3],**
6 **Luis Mancera[3], Juan de Dios Caballero[2,4], Laura Martínez-García[2,7], Desirée**
7 **Gijón[2,4], María Isabel Morosini[2,4], Patricia Ruiz-Garbajosa[2,4], Rafael Cantón[2,4],**
8 **Patricia Muñoz[1,5,6], David Rodríguez-Temporal[1†] and Belén Rodríguez-Sánchez[1].**
9 [1]Clinical Microbiology and Infectious Diseases Department, Hospital General
10 Universitario Gregorio Marañón, Madrid, Spain and Institute of Health Research
11 Gregorio Marañón (IiSGM), Madrid, Spain
12 [2]Servicio de Microbiología. Hospital Universitario Ramón y Cajal and Instituto Ramón y
13 Cajal de Investigación Sanitaria (IRYCIS), Madrid, Spain.
14 [3]Clover Bioanalytical Software, Av. del Conocimiento, 41, 18016 Granada, Spain
15 [4]CIBER en Enfermedades Infecciosas, Madrid, Spain
16 [5]CIBER de Enfermedades Respiratorias (CIBERES CB06/06/0058), Madrid, Spain.
17 [6]Medicine Department, Faculty of Medicine, Universidad Complutense de Madrid,
18 Madrid, Spain
19 [7]Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública
20 (CIBERESP), Madrid, Spain.

21 **Running Title:** *E. cloacae* complex analysis with MALDI-TOF MS and FTIR S

22 **Corresponding Author:**
23 David Rodríguez-Temporal, PhD.
24 Servicio de Microbiología Clínica y Enfermedades Infecciosas. Hospital General
25 Universitario Gregorio Marañón.Dr. Esquerdo 46. 28007 Madrid, Spain
26 Phone: +34- 91- 426 7163, Fax: +34- 91- 426 9595

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467040; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

27 E-mail: david.rodriguez@iisgm.com

28 *These authors contributed equally to the study

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467040; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

29 **ABSTRACT**

30 *Enterobacter cloacae* complex (ECC) encompasses heterogenic genetic clusters of
31 species that have been associated with nosocomial outbreaks. These species may
32 host different acquired antimicrobial susceptibility patterns and their identification is
33 challenging. DNA-based techniques are laborious and require specific equipment.
34 MALDI-TOF MS has showed low accuracy for the discrimination of ECC species. The
35 aim of this study is to develop machine learning predictive models using MALDI-TOF
36 MS and new diagnostic technologies like Fourier-Transform Infrared Spectroscopy
37 (FTIR-S) for species-level identification of these species.

38 A total of 163 ECC clinical isolates were included in the study: 47 for the predictive
39 model development and internal validation and 126 for external validation. All spectra
40 obtained by MALDI-TOF MS and FTIR-S were processed using Clover MS Data
41 Analysis software. Two models were created: Model A for differentiate six ECC species
42 and Model B for *E. hormaechei* subspecies.

43 For MALDI-TOF MS spectra, Model A identified correctly 96.0% of isolates using
44 Random Forest (RF) algorithm, and Model B identified 94.1% using Support Vector
45 Machine (SVM). Regarding FTIR-S, Model A identified 73.0% of isolates by RF and
46 Model B 72.5%. Two new predictive models were created for FTIR-S: Model C for
47 discrimination of *E. hormaechei* from non-*E. hormaechei* (87.3% identification, RF) and
48 Model D for differentiation among non-*E. hormaechei* species (62.7% identification,
49 RF).

50 MALDI-TOF MS combined with machine learning tools could be a rapid and accurate
51 method for species-level identification within the ECC. FTIR-S differentiated general
52 groups of the ECC although discrimination of non-*E. hormaechei* species was poor.

53

54

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467040; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

55 **INTRODUCTION**

56 *Enterobacter* spp. is a facultative anaerobic Gram-negative genus that can be found as
57 a natural commensal in the gut microbiome of mammals (1). Several species have
58 been associated with nosocomial outbreaks causing urinary tract infection, skin and
59 soft tissues infection, pneumonia and bacteraemia (2, 3). *Enterobacter cloacae*
60 complex (ECC) is of particular clinical interest. This group is composed of 13
61 heterogenic genetic clusters according to *hsp60* gene sequencing: *E. asburiae* (cluster
62 I), *E. kobei* (cluster II), *E. hormaechei* subsp. *hoffmannii* (cluster III), *E. roggenkampii*
63 (cluster IV), *E. ludwigii* (cluster V), *E. hormaechei* subsp. *oharae* and subsp.
64 *xiangfangensis* (cluster VI), *E. hormaechei* subsp. *hormaechei* (cluster VII), *E.*
65 *hormaechei* subsp. *steigerwaltii* (cluster VIII), *E. bugandensis* (cluster IX), *E.*
66 *nimipressuralis* (cluster X), *E. cloacae* subsp. *cloacae* (cluster XI), *E. cloacae* subsp.
67 *dissolvens* (cluster XII) and an heterogeneous group of *E. cloacae* sequences
68 considered as cluster XIII. However, the taxonomy of this genus is still under debate (4,
69 5). In fact, *E. aerogenes* has been recently reclassified into the *Klebsiella* genus as *K.*
70 *aerogenes* (6). A more comprehensive study based on whole genome sequencing
71 (WGS) data from ECC isolates yielded a redistribution of the species defined by *hsp60*
72 sequencing (5) into 22 clades (7) and allowed the characterization of new ECC species
73 (8).

74 Discrimination of the ECC at the species level is usually performed by DNA-based
75 methods. The most commonly targeted gene is *hsp60*, although multilocus sequence
76 analysis (MLSA) and WGS have also been applied (5, 9, 10). DNA-based diagnostic
77 methods techniques are laborious and require specific equipment. Therefore, new
78 emerging techniques such as MALDI-TOF MS (Matrix Assisted Laser
79 Desorption/Ionization Time-of-flight Mass Spectrometry) have been proposed as an
80 alternative to molecular methods. MALDI-TOF MS has shown to be an excellent
81 methodology for bacterial identification. It can easily identify *E. cloacae* complex

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467040; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

82 isolates but it showed low discrimination power for the species in this group when using
83 standard analysis (11, 12).

84 Fourier-Transform Infrared Spectroscopy (FTIR-S) has also emerged recently as a
85 rapid method for bacterial typing (13), although there are still few studies for the
86 discrimination of close-related species. These new technologies, coupled to a machine
87 learning approach has a great potential for fast and cost-effective bacterial typing and
88 discrimination of closely related species.

89 The aim of this study is to develop and validate prediction models for the
90 differentiation of the species within the ECC using MALDI-TOF MS and FTIR-S spectra
91 combined with machine learning tools. This task is important because of their diverse
92 implications in human pathologies and their involvement in nosocomial outbreaks (4).
93 Besides, *E. hormaechei*, the most commonly encountered ECC species in the clinical
94 settings, has been correlated with enhanced acquisition of antimicrobial resistance
95 mechanisms and the expression of virulence factors (14-16).

96

97 **MATERIALS AND METHODS**

98 **Bacterial isolates**

99 Overall, 163 clinical isolates belonging to the ECC and nine *K. aerogenes* (formerly *E.*
100 *aerogenes*) were analysed. They belonged to a surveillance study of antimicrobial
101 resistances were collected at the Hospital Universitario Ramón y Cajal (Madrid, Spain)
102 between 2005 and 2018 and identified by partial sequencing of the *hsp60* gene (17).
103 All the analysed isolates host different carbapenemases.

104 All isolates were incubated overnight at 37ºC and metabolically activated after
105 three subcultures on Columbia Blood Agar (bioMérieux, Marcy l'Etoile, France).

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467040; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

106 Isolates were analysed with MALDI-TOF MS and FTIR-S at the Hospital General
107 Universitario Gregorio Marañón (Madrid, Spain).

108 **Bacterial identification and spectra acquisition using MALDI-TOF MS**

109 Identification was performed using the MBT Smart MALDI Biotyper (Bruker Daltonics,
110 Bremen, Germany). All strains were spotted in duplicate onto the MALDI target plate
111 and overlaid with 1 µl of 70% formic acid. After drying at room temperature, spots were
112 covered with 1 µl HCCA matrix, prepared according to the manufacturer's indications
113 (Bruker Daltonics) and allowed to dry. Two spectra in the range of 2,000-20,000 Da
114 were obtained on each spot, resulting in 4 spectra per isolate.

115 **Spectra acquisition using Fourier-Transform Infrared Spectroscopy**

116 All strains were also analysed by FTIR-S using an IR Biotyper (Bruker Daltonics)
117 according to the manufacturer's instructions. The procedure consisted on the
118 resuspension of a 1 µl loopful of bacteria in 50 µl of 70% ethanol. After a brief vortexing
119 step, 50 µl of sterile water were added and the mixture was vortexed again. Then, 15 µl
120 of this preparation were transferred onto the 96-well silicon plate and allowed to dry at
121 40ºC for 20 minutes approximately. Each isolate was analysed in triplicates -technical
122 replicates- and two standards (Bruker Infrared Test Standard 1 and 2, Bruker
123 Daltonics) were included in each run and only spectra with valid target quality controls
124 were further analysed.

125 **Data processing of MALDI-TOF MS protein spectra and development of**
126 **predictive models**

127 After spectra acquisition, 47 isolates representing all *Enterobacter* species were used
128 as a test set for the evaluation of the predictive model for species differentiation (Table
129 1). MALDI-TOF MS protein spectra were processed with Clover MS Data Analysis
130 software (Clover Biosoft, Granada, Spain). A pre-processing pipeline was applied to

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467040; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

131 the MALDI-TOF MS spectra that consisted of: 1) smoothing (Savitzky-Golay Filter:
132 window length=11; polynomial order=3) and baseline substraction (Top-Hat filter
133 method with factor=0.02); 2) creation of an average spectrum per isolate; 3) alignment
134 of average spectra from different isolates (shift: medium; constant tolerance: 2 Da;
135 linear mass tolerance: 600 ppm); 4) normalization by Total Ion Current (TIC).

136 The processed spectra were included in two peak matrices used to develop two
137 predictive models: 1) Model A, aiming for the differentiation of the 6 main species -*E.*
138 *asburiae* (cluster I), *E. kobei* (cluster II), *E. hormaechei* (clusters III, VI and VIII
139 considered together), *E. roggenkampii* (cluster IV), *E. ludwigii* (cluster V) and *E.*
140 *cloacae* subsp. *dissolvens* (cluster XII)- and 2) Model B, for the differentiation of the
141 three *E. hormaechei* clusters (Figure 1). These two peak matrices were used as input
142 data for three machine learning algorithms: Partial Least Square-Discriminant Analysis
143 (PLS-DA), Support Vector Machine (SVM) and Random Forest (RF). Internal k-fold
144 cross validation using k=10 was performed.

145 **Data processing of FTIR-S spectra and development of predictive models**

146 Following the same approach applied for the analysis of the *Enterobacter* isolates with
147 MALDI-TOF MS, the same group of 47 isolates was used as test set for the creation of
148 a predictive model based on FTIR-S spectra. For this purpose, each isolate was
149 analysed in triplicates - technical replicates - in two consecutive days - biological
150 replicates -. Therefore, the total number of spectra per isolate was six. The spectra
151 were processed using Clover MS Data Analysis with the following pipeline: 1) Standard
152 Normal Variation; 2) Smoothing (Savitzky-Golay Filter: window length=9; polynomial
153 order=2; second derivative); 3) Maximum absorbance normalization.

154 As for MALDI-TOF MS spectra analysis, discrimination of the six major ECC
155 species was attempted, as well as differentiation of the three *E. hormaechei* clusters.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467040; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

156 For this purpose, the algorithms SVM and RF were also applied using internal k-fold
157 cross validation of k=10.

158 **Classification of the *E. cloacae* complex isolates using the developed models for**
159 **MALDI-TOF MS spectra**

160 MALDI-TOF MS spectra from the external validation set (n=126) were processed using
161 the pipeline described above and blindly classified by the predictive models using
162 Clover MS Data Analysis software.

163 **Classification of the *E. cloacae* complex isolates using the developed model for**
164 **FTIR-S spectra**

165 FTIR-S spectra from the external validation set (n=126) were processed using the
166 pipeline described above and blindly classified by the predictive models using Clover
167 MS Data Analysis software.

168 **Ethics statement**

169 The Ethics Committee of the Gregorio Marañón Hospital (CEIm) evaluated this project
170 and considered that all the conditions for waiving informed consent were met, since the
171 study was conducted with microbiological samples and not with human products.

172

173 **RESULTS**

174 **MALDI-TOF MS predictive models**

175 Different models were developed by the implementation of the available algorithms
176 (PLS-DA, SVM and RF) to the test set. These models were validated internally by
177 cross-validation (k-fold=10) and externally by the automatic classification of the 126
178 protein spectra included in the validation set (Table 1). For Model A evaluation,
179 Random Forest (RF) yielded the highest rates of correct classification for the most

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467040; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

180 common *Enterobacter* species (Model A): 95.7% and 96.0% for internal and external
181 validation, respectively (Table 2). The RF plot for internal validation is showed in Figure
182 2. Only 5/126 isolates were misidentified by Model A using RF (Table S1): *E.*
183 *hormaechei* identified as *E. roggenkampii* (n=1), *E. roggenkampii* identified as *E.*
184 *hormaechei* (n=2), *E. roggenkampii* identified as *E. kobei* (n=1) and *E. kobei* identified
185 as *E. roggenkampii* (n=1). PLS-DA and SVM algorithms also provided high rates of
186 correct classification in the internal validation although PLS-DA showed lower
187 discriminative power in the external validation (70.6%) -Table 2-.

188 Regarding Model B, designed to differentiate *E. hormaechei* clusters III, VI and
189 VIII, the implementation of PLS-DA yielded 77.8% correct classification in the internal
190 validation (Table 2). Thus, this model was not considered suitable for external
191 validation. However, SVM provided a correct classification of 94.4% and 94.1% of the
192 isolates in the internal and external validation, respectively (Tables S2-S3). Three *E.*
193 *hormaechei* isolates out of 51 were misidentified: one strain from cluster VIII was
194 identified as cluster VI, and two strains from cluster VI were identified as cluster VIII. All
195 cluster III strains were correctly classified (Table S3). Besides, RF algorithm reached
196 94.4% of correct classification (Figure 3) and allowed species-level identification of
197 80.4% of the isolates from the validation set (Table S4).

198 **FTIR-S predictive models**

199 The implementation of FTIR-S allowed the development of classification models based
200 on SVM and RF. Internal cross-validation (k=10) and external validation of Model A
201 using RF reached 98.5% (Table S5) and 73.0% correct classification, respectively. The
202 rate of accuracy yielded by SVM was lower in both internal and external validation
203 (Table 3). For Model B, correct discrimination of the *E. hormaechei* clusters was
204 achieved in 100% and 72.5% of the cases in the internal and external validation,
205 respectively. In that model, also SVM obtained lower identification rates.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467040; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

206 Due to the low percentages of correct classification obtained in the external
207 validation of Model A and Model B, two more predictive models were evaluated for
208 FTIR-S spectra: 1) Model C: for the differentiation of *E. hormaechei* vs non-*E.*
209 *hormaechei* isolates; 2) Model D: for discrimination among non-*E. hormaechei* species.
210 Applying RF algorithm, Model C and D reached 99.1% and 99.0% correct classification
211 of the isolates in the internal validation and 87.3% and 62.7% in the external validation,
212 respectively (Table 4).

213

214 **DISCUSSION**

215 In this study, the implementation of classifying algorithms to MALDI-TOF spectra
216 allowed the correct species assignment of 96.0% isolates belonging to the main ECC
217 species and 94.1% of the *E. hormaechei* clusters III, VI and VIII. On the other hand, the
218 discrimination power of FTIR S for the same categories was 73.0% and 72.5% and in
219 87.3% of the cases this technology discriminated correctly between *E. hormaechei* and
220 non-*E. hormaechei* isolates.

221 Poor discrimination of *E. cloacae* complex species by MALDI-TOF MS has been
222 previously reported either by using commercial (11, 17) and enriched, in-house
223 databases (18). However, a recent study from a research group with extensive
224 experience in MALDI-TOF MS and the creation of expanded libraries reported 92.0%
225 correct species-level identification by implementing a specific in-house library enriched
226 with well-characterized ECC strains and correct discrimination of 97.0% *E. hormaechei*
227 isolates (19). This approach can be useful for discrimination of close-related species
228 but the construction of a database is cumbersome and requires highly trained
229 personnel.

230 A more rapid approach for the discrimination of ECC species is the
231 implementation of classification models. A previous study already demonstrated that

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467040; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

232 the application of hierarchical clustering to protein spectra provided a clear
233 discrimination of *E. hormaechei* isolates (clusters III, VI and VIII) from the rest of the
234 species from the complex in the study developed by Wang et al. (18).

235 In this study, different supervised algorithms were implemented for the
236 discrimination of six main ECC species and for differentiation of the three *E.*
237 *hormaechei* clusters (III, VI and VIII). For MALDI-TOF MS spectra, RF algorithm
238 showed highest discrimination capacity both in the cross-validation (95.7%) and in the
239 external validation of the model (96.0%) and SVM demonstrated a high suitability for
240 the differentiation of the *E. hormaechei* clusters (94.4% correct classification in the
241 cross-validation and 94.1% in the external validation). Although a lower specificity in
242 the classification of *E. hormaechei* subspecies was found, the two developed models
243 have shown high discrimination between species of the ECC. Therefore, MALDI-TOF
244 MS combined with machine learning tools could be a rapid and accurate method for
245 species-level identification within the ECC. Only five (3.9%) isolates from the main
246 species and three (5.8%) from the *E. hormaechei* clusters were misclassified, which
247 shows the high specificity of this methodology.

248 The implementation of FTIR-S and RF modelling allowed the correct
249 classification of 98.5% -cross-validation- and 73.0% -external validation- of the spectra
250 from the main species and 100% and 72.5% of the *E. hormaechei* clusters. The low
251 rate of correct classifications encouraged us to developed different models for the
252 discrimination of *E. hormaechei* isolates from the rest of the ECC species. In this case,
253 99.1% and 87.3% correct classifications were achieved in the cross-validation and the
254 external validation, respectively. Discrimination of non-*E. hormaechei* species was low
255 using RF (62.7%) and even lower with SVM. Our hypothesis is that FTIR-S accuracy is
256 so high that it may discriminate the analysed isolates based on their sequence type or
257 their genetic lineage since this methodology is designed mainly for bacterial typing.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467040; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

258 Unfortunately, the strains of this study were not characterized at that level. This
259 approach has been previously observed in the study by Vogt et al. (20).

260 One of the limitations of this study was the fact that all isolates host
261 carbapenemases because they source from a previously analysed collection (17).
262 Although this feature does not hamper the results from the present work, further
263 studies will be performed including both resistant and susceptible isolates.

264 Although further algorithms and modelling is needed in order to implement
265 FTIR-S for the discrimination of *E. cloacae* complex species, MALDI-TOF MS has
266 demonstrated to be a rapid and cost-effective method for this task. The use of spectra
267 analysis tools is becoming user-friendly and easy to apply and its use may provide
268 species-level identification in a fast and inexpensive way.

269

270 **ACKNOWLEDGMENTS**

271 This work was supported by the projects PI15/01073 and PI18/00997 from the Health
272 Research Fund (FIS. Instituto de Salud Carlos III. Plan Nacional de I+D+I 2013-2016)
273 of the Carlos III Health Institute (ISCIII, Madrid, Spain) partially financed by the
274 European Regional Development Fund (FEDER) ‘A way of making Europe’.

275 MM was funded by the Community of Madrid (Programa de Garantía Juvenil, PEJD-
276 2017-PRE/BMD-5106) and DRT with a postdoc contract from the Intramural Funding
277 Program of the IiSGM. BRS (CPII19/00002) is a recipient of a Miguel Servet contract
278 supported by the FIS.

279

280 **AUTHOR CONTRIBUTIONS**

281 Ana Candela: experimental part, formal analysis, data collection, validation,
282 visualization, writing – original draft preparation and review/editing. Miriam Mateos and

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467040; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

283 Alicia Gómez-Asenjo: experimental part, formal analysis and data collection. Manuel J.
284 Arroyo, Gema Méndez, and Luis Mancera: data analysis, validation, writing – original
285 draft preparation and review/editing. Marta Hernández-García, Rosa del Campo:
286 experimental part, formal analysis and writing, submission of isolates, original draft
287 preparation and review/editing. Juan de Dios Caballero, Laura Martínez-García,
288 Desirée Gijón, María Isabel Morosini, Patricia Ruiz-Garbajosa, Rafael Cantón and
289 Patricia Muñoz: validation, writing and review/editing. Emilia Cercenado:
290 conceptualization, formal analysis, validation, writing and review/editing. David
291 Rodríguez-Termporal: conceptualization, formal analysis, validation, original draft
292 preparation and review/editing; Belén Rodríguez-Sánchez: conceptualization, project
293 administration, formal analysis, supervision, validation, visualization, original draft
294 preparation and review/ editing.

295 **FIGURE LEGENDS**

296 **Figure 1.** *E. cloacae* complex identification algorithm according to models developed.

297 **Figure 2.** MALDI-TOF MS spectra in Random Forest distance plot for Model A.

298 **Figure 3.** MALDI-TOF MS spectra in Random Forest distance plot for Model B.

299

300

301

302

303

304

305

306

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467040; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

307 **REFERENCES**

308 1. Mezzatesta ML, Gona F, Stefani S. 2012. Enterobacter cloacae complex: clinical impact
309 and emerging antibiotic resistance. Future Microbiol 7:887-902.
310 2. Akbari M, Bakhshi B, Najar Peerayeh S. 2016. Particular Distribution of Enterobacter
311 cloacae Strains Isolated from Urinary Tract Infection within Clonal Complexes. Iran
312 Biomed J 20:49-55.
313 3. Kremer A, Hoffmann H. 2012. Prevalences of the Enterobacter cloacae complex and its
314 phylogenetic derivatives in the nosocomial environment. Eur J Clin Microbiol Infect Dis
315 31:2951-5.
316 4. Davin-Regli A, Lavigne JP, Pages JM. 2019. Enterobacter spp.: Update on Taxonomy,
317 Clinical Aspects, and Emerging Antimicrobial Resistance. Clin Microbiol Rev 32.
318 5. Hoffmann H, Roggenkamp A. 2003. Population genetics of the nomenspecies
319 Enterobacter cloacae. Appl Environ Microbiol 69:5306-18.
320 6. Tindall BJ, Sutton G, Garrity GM. 2017. Enterobacter aerogenes Hormaeche and
321 Edwards 1960 (Approved Lists 1980) and Klebsiella mobilis Bascomb et al. 1971
322 (Approved Lists 1980) share the same nomenclatural type (ATCC 13048) on the
323 Approved Lists and are homotypic synonyms, with consequences for the name
324 Klebsiella mobilis Bascomb et al. 1971 (Approved Lists 1980). Int J Syst Evol Microbiol
325 67:502-504.
326 7. Sutton GG, Brinkac LM, Clarke TH, Fouts DE. 2018. Enterobacter hormaechei subsp.
327 hoffmannii subsp. nov., Enterobacter hormaechei subsp. xiangfangensis comb. nov.,
328 Enterobacter roggenkampii sp. nov., and Enterobacter muelleri is a later heterotypic
329 synonym of Enterobacter asburiae based on computational analysis of sequenced
330 Enterobacter genomes. F1000Res 7:521.
331 8. Wu W, Feng Y, Zong Z. 2019. Characterization of a strain representing a new
332 Enterobacter species, Enterobacter chengduensis sp. nov. Antonie Van Leeuwenhoek
333 112:491-500.
334 9. Singh NK, Bezdan D, Checinska Sielaff A, Wheeler K, Mason CE, Venkateswaran K.
335 2018. Multi-drug resistant Enterobacter bugandensis species isolated from the
336 International Space Station and comparative genomic analyses with human pathogenic
337 strains. BMC Microbiol 18:175.
338 10. Hoffmann H, Stindl S, Ludwig W, Stumpf A, Mehlen A, Heesemann J, Monget D,
339 Schleifer KH, Roggenkamp A. 2005. Reassignment of enterobacter dissolvens to
340 Enterobacter cloacae as E. cloacae subspecies dissolvens comb. nov. and emended
341 description of Enterobacter asburiae and Enterobacter kobei. Syst Appl Microbiol
342 28:196-205.
343 11. Pavlovic M, Konrad R, Iwobi AN, Sing A, Busch U, Huber I. 2012. A dual approach
344 employing MALDI-TOF MS and real-time PCR for fast species identification within the
345 Enterobacter cloacae complex. FEMS Microbiol Lett 328:46-53.
346 12. De Florio L, Riva E, Giona A, Dedej E, Fogolari M, Cella E, Spoto S, Lai A, Zehender G,
347 Ciccozzi M, Angeletti S. 2018. MALDI-TOF MS Identification and Clustering Applied to
348 Enterobacter Species in Nosocomial Setting. Front Microbiol 9:1885.
349 13. Quintelas C, Ferreira EC, Lopes JA, Sousa C. 2018. An Overview of the Evolution of
350 Infrared Spectroscopy Applied to Bacterial Typing. Biotechnol J 13.
351 14. Garinet S, Fihman V, Jacquier H, Corvec S, Le Monnier A, Guillard T, Cattoir V, Zahar JR,
352 Woerther PL, Carbonnelle E, Wargnier A, Kerneis S, Morand PC. 2018. Elective
353 distribution of resistance to beta-lactams among Enterobacter cloacae genetic
354 clusters. J Infect 77:178-182.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467040; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

355 15. Barnes AI, Paraje MG, del CBP, Albesa I. 2001. Molecular properties and metabolic
356 effect on blood cells produced by a new toxin of Enterobacter cloacae. Cell Biol Toxicol
357 17:409-18.
358 16. Paauw A, Caspers MPM, Leverstein-van Hall MA, Schuren FHJ, Montijn RC, Verhoef J,
359 Fluit AC. 2009. Identification of resistance and virulence factors in an epidemic
360 Enterobacter hormaechei outbreak strain. Microbiology (Reading) 155:1478-1488.
361 17. Mateos M, Hernandez-Garcia M, Del Campo R, Martinez-Garcia L, Gijon D, Morosini
362 MI, Ruiz-Garbajosa P, Canton R. 2020. Emergence and Persistence over Time of
363 Carbapenemase-Producing Enterobacter Isolates in a Spanish University Hospital in
364 Madrid, Spain (2005-2018). Microb Drug Resist.
365 18. Wang YQ, Xiao D, Li J, Zhang HF, Fu BQ, Wang XL, Ai XM, Xiong YW, Zhang JZ, Ye CY.
366 2018. Rapid Identification and Subtyping of Enterobacter cloacae Clinical Isolates Using
367 Peptide Mass Fingerprinting. Biomed Environ Sci 31:48-56.
368 19. Godmer A, Benzerara Y, Normand AC, Veziris N, Gallah S, Eckert C, Morand P, Piarroux
369 R, Aubry A. 2021. Revisiting Species Identification within the Enterobacter cloacae
370 Complex by Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass
371 Spectrometry. Microbiol Spectr 9:e0066121.
372 20. Vogt S, Loffler K, Dinkelacker AG, Bader B, Autenrieth IB, Peter S, Liese J. 2019. Fourier-
373 Transform Infrared (FTIR) Spectroscopy for Typing of Clinical Enterobacter cloacae
374 Complex Isolates. Front Microbiol 10:2582.

375
376
377
378
379
380
381
382
383
384
385
386
387

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467040; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

388 **Table 1.** Number of isolates of each species used for model creation and validation for
389 MALDI-TOF MS and FTIR S.

Species included in the study	Number of strains used for model development	Number of strains used for external validation
K. aerogenes	6	3
E. asburiae	6	1
E. dissolvens	1	0
E. hormaechei III	6	13
E. hormaechei VI	6	22
E. hormaechei VIII	6	16
E. kobei	6	9
E. ludwigii	3	0
E. roggenkampii	7	62
Total	**47**	**126**

390

391

392 **Table 2.** Internal and external validation for Model A and Model B in all algorithms used
393 for MALDI-TOF MS spectra.

	Model A		Model B	
Algorithm	**Internal validation**	**External validation**	**Internal validation**	**External validation**
PLS-DA	91.5%	70.6%	77.8%	NR
SVM	91.5%	92.1%	94.4%	94.1%
RF	95.7%	96.0%	94.4%	80.4%

PLS-DA, Partial Least Square Discriminant Analysis; SVM, Support Vector Machine; RF, Random Forest; NR, Not Realized

394

395

396

397

398

399

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467040; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

400

401 **Table 3.** Internal and external validation for Model A and B for all algorithms used for
402 FTIR-S spectra.

	Model A		Model B	
Algorithm	**Internal validation**	**External validation**	**Internal validation**	**External validation**
SVM	95.1%	69.8%	99.2%	72.5%
RF	98.5%	73.0%	100%	72.5%

403

404

405

406

407

408

409

410 **Table 4.** Internal and external validation for new FTIR-S models.

	Model C		Model D	
Algorithm	**Internal validation**	**External validation**	**Internal validation**	**External validation**
SVM	96.7%	88.1%	95.1%	56.0%
RF	99.1%	87.3%	99.0%	62.7%

411

412

413

414

415

416

417

418

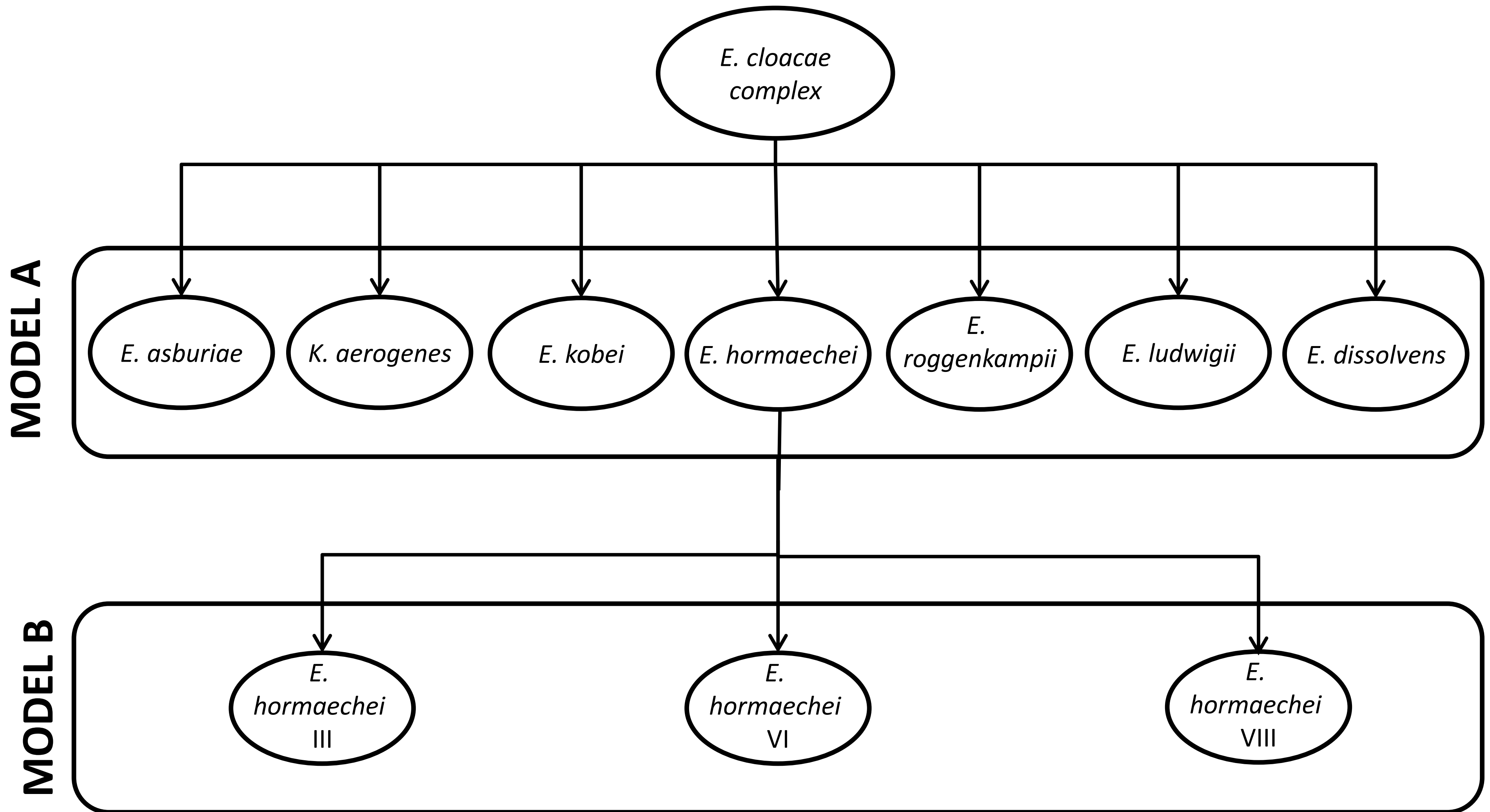

Figure 1. *E. cloacae* complex identification algorithm according to models developed.

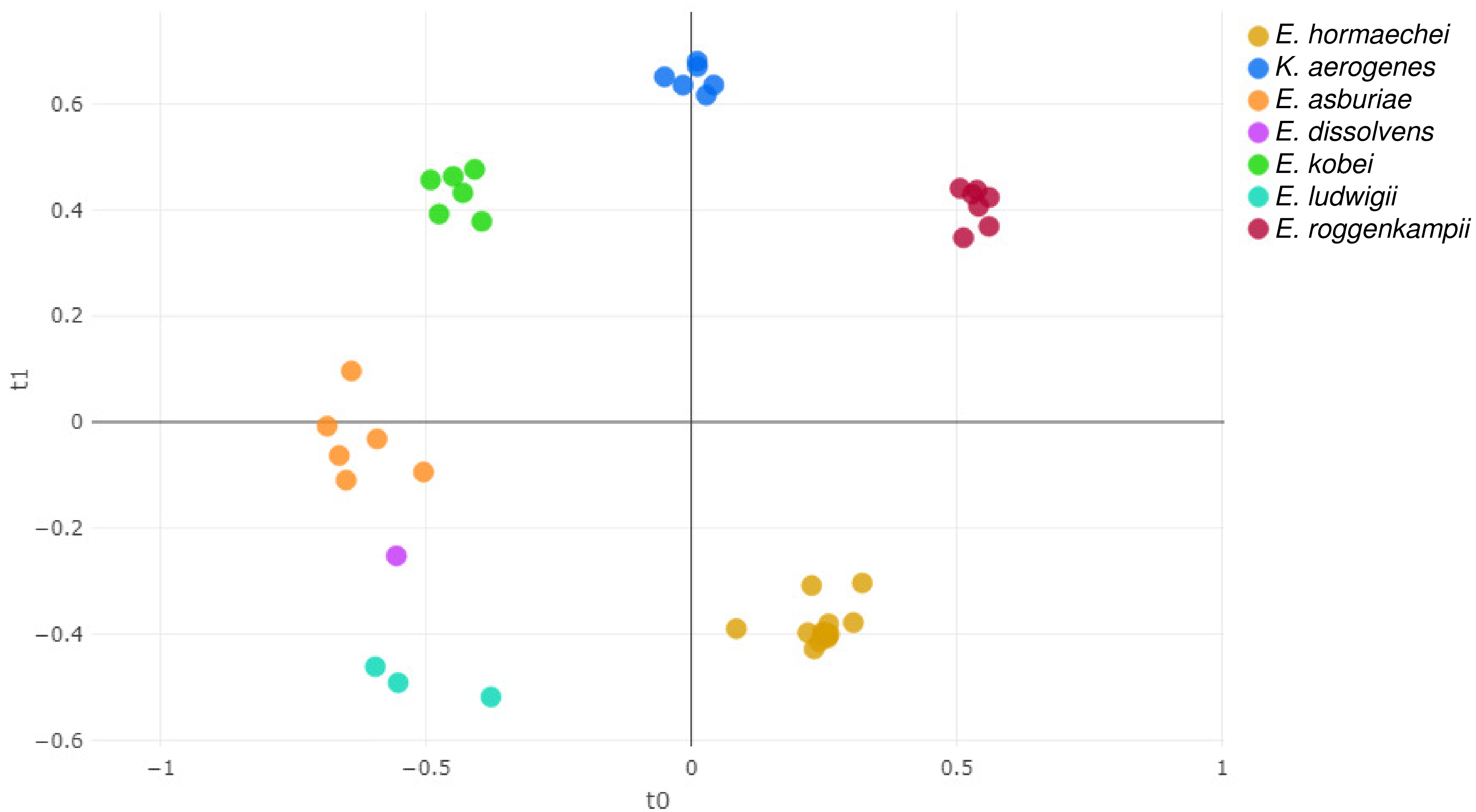

Figure 2. MALDI-TOF MS spectra in Random Forest distance plot for Model A.

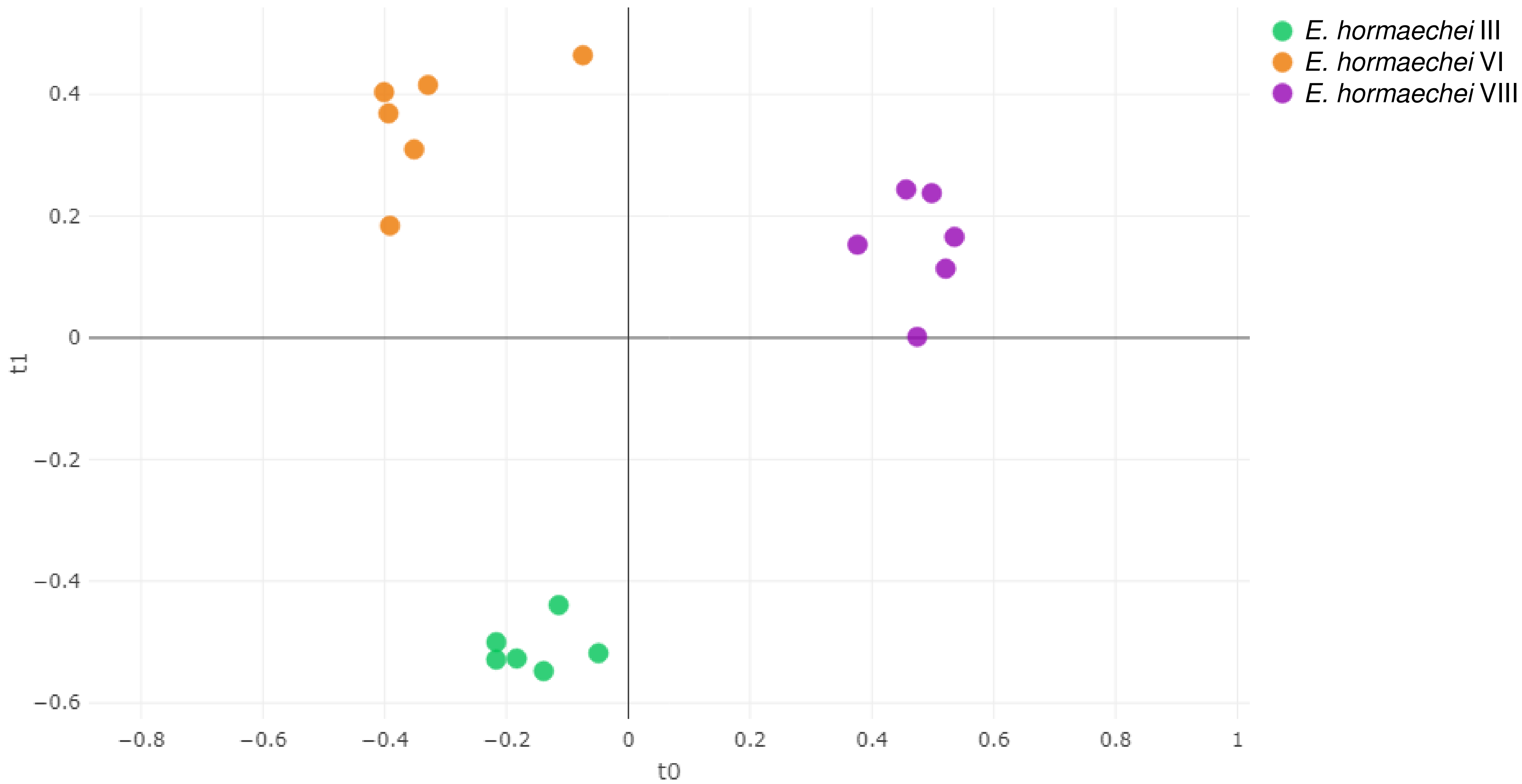

Figure 3. MALDI-TOF MS spectra in Random Forest distance plot for Model B.