Journal of Clinical Microbiology

Clinical Microbiology | Full-Length Text

Development and validation of a quick, automated, and reproducible ATR FT-IR spectroscopy machine-learning model for *Klebsiella pneumoniae* typing

Ângela Novais,[1,2] Ana Beatriz Gonçalves,[1,2] Teresa G. Ribeiro,[1,2,3] Ana R. Freitas,[1,2,4] Gema Méndez,[5] Luis Mancera,[5] Antónia Read,[6] Valquíria Alves,[6] Lorena López-Cerero,[7,8] Jesús Rodríguez-Baño,[7,8] Álvaro Pascual,[7,8] Luísa Peixe[1,2,3]

AUTHOR AFFILIATIONS See affiliation list on p. 9.

ABSTRACT The reliability of Fourier-transform infrared (FT-IR) spectroscopy for *Klebsiella pneumoniae* typing and outbreak control has been previously assessed, but issues remain in standardization and reproducibility. We developed and validated a reproducible FT-IR with attenuated total reflectance (ATR) workflow for the identification of *K. pneumoniae* lineages. We used 293 isolates representing multidrug-resistant *K. pneumoniae* lineages causing outbreaks worldwide (2002–2021) to train a random forest classification (RF) model based on capsular (KL)-type discrimination. This model was validated with 280 contemporaneous isolates (2021–2022), using *wzi* sequencing and whole-genome sequencing as references. Repeatability and reproducibility were tested in different culture media and instruments throughout time. Our RF model allowed the classification of 33 capsular (KL)-types and up to 36 clinically relevant *K. pneumoniae* lineages based on the discrimination of specific KL- and O-type combinations. We obtained high rates of accuracy (89%), sensitivity (88%), and specificity (92%), including from cultures obtained directly from the clinical sample, allowing to obtain typing information the same day bacteria are identified. The workflow was reproducible in different instruments throughout time (>98% correct predictions). Direct colony application, spectral acquisition, and automated KL prediction through Clover MS Data analysis software allow a short time-to-result (5 min/isolate). We demonstrated that FT-IR ATR spectroscopy provides meaningful, reproducible, and accurate information at a very early stage (as soon as bacterial identification) to support infection control and public health surveillance. The high robustness together with automated and flexible workflows for data analysis provide opportunities to consolidate real-time applications at a global level.

IMPORTANCE We created and validated an automated and simple workflow for the identification of clinically relevant *Klebsiella pneumoniae* lineages by FT-IR spectroscopy and machine-learning, a method that can be extremely useful to provide quick and reliable typing information to support real-time decisions of outbreak management and infection control. This method and workflow is of interest to support clinical microbiology diagnostics and to aid public health surveillance.

KEYWORDS Fourier-transform infrared spectroscopy, attenuated total reflectance, typing, bacteria, infection control, outbreak, machine-learning, classification model, KL-type, nosocomial, random forest

Fourier-transform infrared (FT-IR) spectroscopy is an analytical technique, where the interaction of the infrared light with the bacterial cell provides a biochemical fingerprint of its composition in main macromolecules. The high resolution, together

Editor Patricia J. Simner, Johns Hopkins University, Batimore, USA

Address correspondence to Ângela Novais, angelasilvanovais@gmail.com.

The authors declare no conflict of interest.

See the funding table on p. 10.

Received 18 September 2023
Accepted 18 December 2023
Published 29 January 2024

with the short time-to-result, and lower cost compared to whole-genome sequencing (WGS) and the simplicity of the procedure make it an attractive tool for bacterial discrimination and typing (1).

Different studies have demonstrated resolution for strain typing in different clinically relevant species, corroborated by WGS analysis (2–7). Moreover, we have provided a comprehensive analysis of the molecular features at the basis of spectral discrimination in different bacterial species (*Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Salmonella enterica*, and *Escherichia coli*), contributing to the establishment of reliable genotypic–phenotypic correlations that support strain typing (2, 8–10). Subsequent studies have demonstrated its usefulness for outbreak management and infection control (11–15).

The application of FT-IR spectroscopy in routine clinical microbiology laboratories for outbreak investigation has been facilitated by the IR Biotyper integrated system launched by Bruker Daltonics (Germany) in 2017. This equipment is based on transmission FT-IR, an acquisition mode where the infrared beam crosses the sample to reach the detector (16). It requires the preparation of a standardized bacterial suspension that needs to be dried uniformly before spectrum acquisition (17, 18). The available software package allows comparative spectral analysis by using a clustering method to infer clonal relatedness; thus, the cutoff definition requires expertise since it is variable according to the data set or the species analyzed (3, 18, 19). Besides, several authors reported difficulties in standardization and reproducibility, with the variation being associated with culture conditions (culture media and incubation time) (17, 19). Available studies have demonstrated reliability in specific and variable experimental conditions, but evaluation of the reproducibility between equipments is still lacking. FT-IR instruments with the attenuated total reflection (ATR) acquisition mode are also widely used, where the infrared light interacts with the sample through an evanescence wave that targets the detector (2, 8–10, 20, 21). In this case, a bacterial colony is directly applied to the ATR crystal, and the spectra are immediately acquired, avoiding additional reagents and time in the preparation of a bacterial suspension. For these reasons, it has been associated with a lower cost, easiness of the procedure, and a higher reproducibility (1).

K. pneumoniae is a critical pathogen identified by the World Health Organization and the European Centre for Disease Prevention and Control, for which the increasing rates of resistance to last-line beta-lactams and other antibiotics are mainly due to nosocomial spread (22). For these reasons, a reliable and quick method for strain typing is especially critical for early and effective infection control. In previous studies, we used FT-IR ATR to demonstrate the accuracy of the technique for *K. pneumoniae* (Kp) capsular (K)-typing by directly testing one bacterial colony (no suspension is required) and using an in-house spectral database and a machine-learning classification model (2, 11, 12, 23). The workflow used requires expertise in FT-IR, multivariate data analysis algorithms, and high-level programming knowledge to work on MATLAB (MathWorks, USA). Over the last years, we have shown that this workflow can reliably support outbreak control (11, 12) and epidemiological surveillance in humans (23) and animals (24, 25).

However, the translation of FT-IR to routine microbiology laboratory workflows for quick bacterial typing depends on (i) high resolution and accuracy for identification of bacterial lineages defined by reference typing methods; (ii) solid demonstration of spectral reproducibility between instruments throughout time; (iii) ability to provide meaningful and immediate information for infection control and surveillance; and (iv) an automated workflow accessible for non-expert users. In this study, we developed a quick, automated, and reproducible FT-IR ATR workflow for the identification of multidrug resistant (MDR) *K. pneumoniae* clinically relevant lineages, which provides meaningful information to support outbreak management and epidemiological surveillance in a simple and user-friendly manner on the same day that it is detected in the laboratory.

MATERIALS AND METHODS

Bacterial isolates

We selected a set of 293 well-characterized isolates representative of the main clinically relevant *K. pneumoniae* lineages frequently associated with multidrug resistance and hospital-acquired infections, some of them responsible for outbreaks in the Iberian Peninsula and elsewhere. They included (i) 176 contemporaneous isolates identified in four hospitals from North of Portugal (*n* = 135; 2016–2021) and one reference hospital from South of Spain (*n* = 41; 2017–2021); (ii) 108 isolates from our previous study where we were able to discriminate 19 KL-types (2002–2015) (2); and (iii) nine isolates previously characterized from poultry (2019–2020) to enrich poorly represented classes (24) (Table S1). This bacterial collection was selected to capture a high coverage and diversity of genetic backgrounds and KL-types, hence avoiding a bias in the FT-IR-based classification model. The clonal relationship between isolates had been previously established by pulsed field gel electrophoresis, multi-locus sequence typing (MLST), and/or core genome MLST (cgMLST) using Ridom + SeqSphere software (2, 12, 23, 24, 26) (unpublished data). These isolates are defined as the *training set* and were used to train a new machine-learning classification model with increased coverage and representativeness for *K. pneumoniae* clones and KL-types.

In addition, a set of another 280 *K. pneumoniae* isolates collected systematically between 2021 and 2022 in two hospitals from North of Portugal were used to validate the machine-learning classification model and constitute the *validation set*. They were divided into positive controls (*n* = 204, isolates belonging to KL-types included in the model) and negative controls (*n* = 76, isolates belonging to KL-types not included in the model). These isolates were used as an external collection of isolates to validate the model and the workflow developed and to test the reproducibility of the method in different culture media (Table S2). Sequencing of the *wzi* gene was used as the reference method to infer KL-types according to the Pasteur database (https://bigsdb.pasteur.fr/klebsiella/). WGS was performed in a subset of the validation set (*n* = 38/280; 14%) to confirm discrepancies between *wzi* sequencing and FT-IR spectroscopy, to identify potentially new clone/KL correlations, or to establish cgMLST types.

FT-IR spectra acquisition and automated analysis

Isolates were grown in standardized culture conditions regarding the medium (Mueller-Hinton 2 from Biomérieux, France) and incubation (37°C for 18 h). Afterward, a colony was directly spread in the ATR accessory of the FT-IR instrument (Spectrum 2 from PerkinElmer, USA) and air-dried. Three technical replicate spectra per strain were acquired in less than 5 minutes, in the region from 4,000 cm^{-1} to 600 cm^{-1}, with 4 cm^{-1} resolution and 16 scan co-additions. This procedure was repeated for at least one biological replicate per isolate (an independent culture obtained in the same conditions on a different day) (Fig. 1).

Spectra were analyzed using an automated and user-friendly Clover MS Data analysis software developed for spectral data analysis by Clover Bioanalytical Software (https://www.clovermsdataanalysis.com; Granada, Spain). It uses the Python package *scikit-learn* to perform machine-learning algorithms. The existing workflow was adapted to optimize preprocessing, sample replicate analysis, and spectral analysis algorithms for FT-IR-based typing. FT-IR spectra were imported and preprocessed using the standard normal variate and Savitzky-Golay filter (window length: 9; polynomial order: 2; derivative order: 2), as previously established (2, 12). A region between 1,200 cm^{-1} and 900 cm^{-1} was selected to create a peak matrix that is used to create the classification algorithm.

A new machine-learning classification model based on random forest (RF) analysis was built using a total of 293 *K. pneumoniae* strains belonging to 33 KL-types (Table 1). The RF algorithm was created with the RandomForestClassifier and uses bagging and feature randomness to create an uncorrelated forest of decision trees, the prediction of which is more accurate than for any of the individual trees. The following parameters

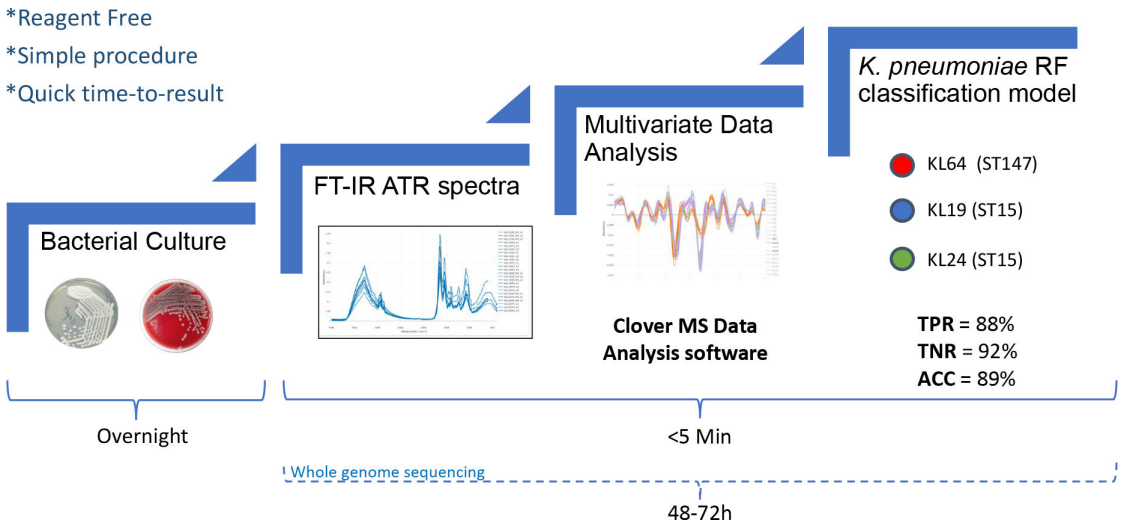

FIG 1 Workflow for typing using FT-IR ATR and a machine-learning model. Main steps of our workflow for *Klebsiella pneumoniae* typing and comparison of the time-to-result for FT-IR ATR typing with that of the gold-standard method (whole-genome sequencing using a short-read approach), both starting from the isolated bacterial culture. TPR = true positive rate; TNR = true negative rate; ACC = accuracy.

were used: number of estimators or number of trees (n = 200); maximum features (n = 12), indicating the number of features included in each tree, where normally a large number leads to better but slower training performance but can lead to overfitting; minimum split size (n = 2), specifying the minimum number of samples required to split an internal node; and minimum samples per leaf (n = 1), which is the minimum number of samples required to be considered a leaf node. The internal validation of the models was performed using the leave-one-out k-fold cross-validation algorithm.

External validation of the RF classification model

The new RF classification model was validated using the *validation set* (Table S2) and *wzi* sequencing as the reference method to infer the KL-type. Spectra from these isolates acquired in the same conditions as described above were queried in the RF model created, and the first two predicted categories (KL-type 1 and KL-type 2) with their corresponding probability scores (P1 and P2) were considered for analysis and interpretation. We calculated the accuracy, the sensitivity, and the specificity of the method, taking into consideration the true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) results. The congruence of FT-IR spectroscopy with *wzi* sequencing for KL-typing was measured using Simpon's index of diversity, the adjusted Rand, and the adjusted Wallace coefficients according to http://www.comparingpartitions.info/ (27).

Evaluation of the reproducibility and repeatability

Reproducibility among culture media was evaluated by testing the same workflow (spectra acquisition, preprocessing, and validation) in a subset of the *validation set* (n = 101/280) after growth in Columbia agar with 5% sheep blood (Biomérieux, France) at 37°C for 18 h and comparison of the prediction results using the same criteria for the interpretation and validation of prediction, as described above. We chose this culture medium since it is commonly used in clinical microbiology laboratories for bacterial isolation from several clinical samples.

We also tested the repeatability and reproducibility of the FT-IR ATR workflow in the same or different FT-IR ATR instrument models from the same manufacturer using a subset of 90 isolates from the *training set* that represent 21 KL-types (Table S3). Repeatability was assessed by testing different biological replicates in different time-points

TABLE 1 Nature and diversity of classes used to build the RF model[a]

Class	KL-type	*wzi*	ST/cgMLST
1	KL2	2	ST14
	KL2	72	ST25
2	KL9	9	ST22
3	KL13	243	ST11
4	KL14	14	ST54
5	KL15	50	ST37
	KL15	50	ST16
6	KL16	16	ST14
7	KL17	137	ST101
8	KL19	19	ST15
9	KL21	262	ST323
	KL21	262	ST6449
10	KL22.37	37	ST449
	KL22.37	22	ST109
11	KL23	83	ST39
	KL23	82	ST280
12	KL23-*like*[b]	732	ST15
13	KL24	24	ST15
	KL24	101	ST45
14	KL25	141	ST11
15	KL27	27	ST11
	KL27	27	ST528
	KL27	27	ST6393
	KL27	187	ST392
16	KL30	273	ST2328
17	KL38	82	ST17
18	KL45	45	ST111
	KL45	45	ST1/5115
19	KL48	151	ST15
20	KL60	201	ST253
	KL60	201	ST392
21	KL62	94	ST348
22	KL64	64	ST147
23	KL81	81	ST252
24	KL102	173	ST307
25	KL105	75	ST11
26	KL106	29	ST258
27	KL107	154	ST258
28	KL110	89	ST15
	KL110	89	ST716
	KL110	199	ST35
29	KL112	93	ST15
30	KL125	177	ST378
	KL125	177	ST219
31	KL127	202	ST11
32	KL151	143	ST405
33	KL163	150	ST336

[a]KL = capsular locus; ST = sequence type; cgMLST = core genome multi-locus sequence typing. [b] KL-type similar to KL23 not yet characterized biochemically.

(Equipment 1_TM; Equipment 1_TM2) by the same operator on the same instrument (PerkinElmer Spectrum 2), under the same experimental conditions (culturing and FT-IR spectra acquisition). For the same subset, reproducibility between instruments was

tested by comparing the prediction results for spectra acquired in a different instrument model (PerkinElmer Frontier—Equipment 2) using the same experimental conditions.

RESULTS

Improvement of database coverage and robustness

The updated RF classification model includes >2,000 spectra from 293 *K. pneumoniae* isolates belonging to 33 different KL-types. This represents a ~70% increase compared to our previous database (from 19 to 33 KL-types) due to the inclusion of 14 new KL-types (KL9, KL21, KL13, KL15, KL22.37, KL23-like, KL25, KL30, KL38, KL45, KL81/KL120, KL102, and KL125/KL114) in the current model (Table 1) (2). Besides, we improved the robustness of most previously existing classes by increasing the number of isolates per class on average by 130% (13%–300%), especially those that were poorly represented before (KL2, KL27, KL63, or KL107). The updated RF classification model, including the 33 KL-types, allowed 90% correct predictions in an internal cross-validation step (Fig. S1).

In most classes (n = 21/33; 64%), each KL-type was linked to a unique ST (e.g., KL19-ST15, KL107-ST302, KL64-ST147, KL105-ST11, and KL62-ST348) and, occasionally, unique cgMLST types representing lineages circulating in wide geographic areas (Table S1). In some cases (KL23, KL24, KL27, and KL38), isolates were associated with diverse ST, O-, or cgMLST-types (Table 2). We thus hypothesized that FT-IR could discriminate O-antigen variation. To evaluate this possibility, we created partial-least squares discriminant analysis (PLS-DA) models to improve discrimination within KL23/KL38, KL24, and KL27. These models yielded 96%–100% correct predictions in an internal cross-validation test (Fig. 2; Fig. S2). Furthermore, among isolates predicted *in silico* as KL23 by KAPTIVE (https://kaptive.holtlab.net), FT-IR distinguished a variant profile (designated KL23-*like*) suggesting capsular biochemical variation (Table 2). Thus, by increasing the discriminatory power within KL-types, we have the potential to distinguish by FT-IR up to 36 *K. pneumoniae* lineages that are frequently associated with multidrug resistance patterns, high transmissibility, colonization, and/or persistence (Table S1).

Validation of the RF classification model

The positive controls from the validation set (n = 204 isolates) represented 22 out of the 33 KL-types. Most (90%; n = 183/204) of these isolates were identified correctly, and a large proportion of these (95%; n = 175/184) yielded a probability score (P1) >25% and a P1-P2 difference >10% (Table S2). Henceforth, we set these parameters to distinguish TP from FP results. In accordance, the Simpon's index of diversity for FT-IR was 0.894 (CI = 0.872–0.916), the adjusted Rand was 0.911, and the adjusted Wallace was 0.947 (0.926–0.968). False-negative results represented 12% of the sample, and a fraction of these (32%; n = 8/25) corresponded to isolates that were correctly identified but did not meet the set criteria. The remaining belonged to variable (n = 9) KL-types, such as KL23-*like* (31%; n = 4/13) or KL30 (27%; n = 3/11), being more frequently misidentified. As explained above, differentiation between KL23-*like* isolates has already been improved in a specific model (96% correct predictions, Fig. S2C). It is of note that 56% of false-negative results were correctly identified when re-evaluated after re-isolation in Columbia agar plates with 5% sheep blood.

TABLE 2 Typing data of isolates from closely related KL-types discriminated by specific models[a]

Submodels	KL-type	O-type	ST	*wzi*	N° isolates
1	KL24	O1v1	ST15	24	10
	KL24	O2a	ST45	101	13
2	KL27	O2	ST11	27	11
	KL27	O4	ST392	187	12
3	KL23	O1	ST39	83	8
	KL23-*like*	O2afg	ST15	732	9
	KL38	O2	ST17	82	6

[a]KL-type = capsular type; O-type = O antigen type; ST = sequence type.

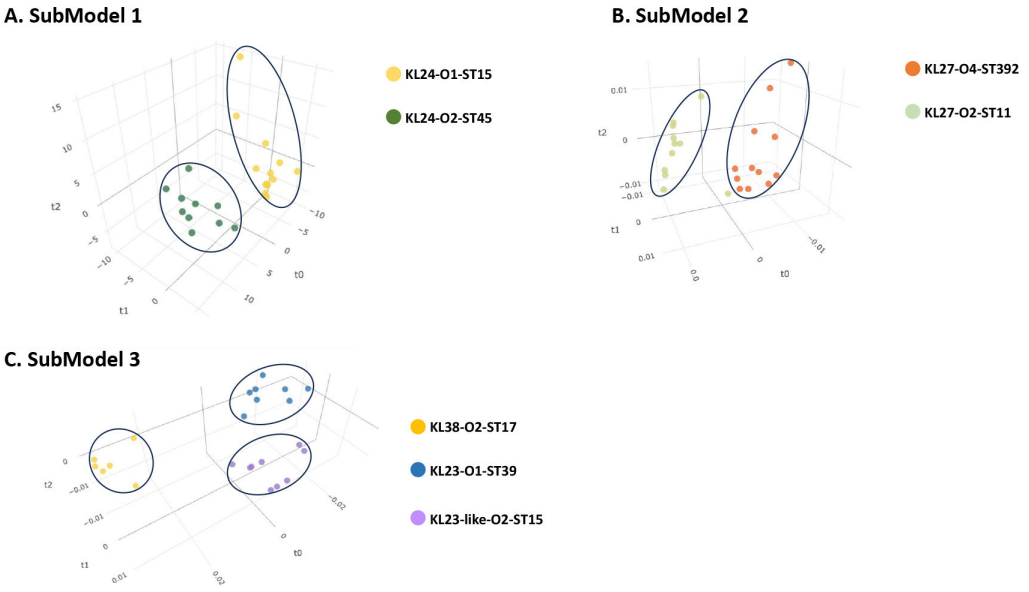

FIG 2 Specific models for the discrimination of closely related KL-types. (A) Discrimination between lineages carrying KL24. (B) Discrimination between lineages carrying KL27. (C) Discrimination of lineages carrying KL23 or KL38. All models were obtained using the partial-least square discriminant analysis method (A with four latent variables; B and C with three latent variables) and the region between 900 cm^{-1} and 1,200 cm^{-1} of the spectra. They were validated with the leave-one-out internal cross-validation method (96–100% correct predictions).

The negative controls from the validation set consisted of 76 isolates belonging to 33 different KL-types/*wzi* alleles absent from the RF model (*n* = 1–15 isolates each; average 4). Using the set criteria, 92% (*n* = 70/76) of these were correctly excluded. False positives (*n* = 6) were recorded for five different KL-types. All of them yielded high scores and represent KL-types not yet characterized biochemically (e.g., two isolates with the *cps* genotyped as KL139 were classified as KL27) (Table S2).

Considering the whole validation set, we obtained an accuracy rate of 89%, a sensitivity of 88%, and a specificity of 92% with the established workflow. Colony application, spectral acquisition, and automated KL prediction through Clover MS Data Analysis software yielded a time-to-result was of 5 min/isolate.

Repeatability and reproducibility of the workflow

We obtained 98% of correct predictions from biological replicates over time (Equipment 1) and 100% from spectra obtained in a different FT-IR instrument (Equipment 2) (Table S3).

Using Columbia Agar plates with 5% sheep blood and the same prediction scores (> or =25% for P1 and P1–P2 > 10%), the accuracy (86%) and sensitivity (87%) were similar to those obtained with Mueller-Hinton, but the specificity was lower (83%) (Table S4). In accordance, Simpon's index of diversity for FT-IR was 0.903 (CI = 0.872–0.934), the adjusted Rand was 0.831, and the adjusted Wallace was 0.881 (0.901–0.947). Sporadic false-negative (*n* = 10) and false-positive (*n* = 4) results were observed for strains belonging to 10 different KL-types. Of note, a high proportion (60%) of false negatives resulted correctly predicted when tested in the Mueller-Hinton media.

DISCUSSION

In this study, we developed a quick, automated, and reproducible FT-IR ATR workflow for typing up to 36 clinically relevant *K. pneumoniae* lineages frequently associated with multidrug resistance. Though KL2 is included in the model, hypervirulent genetic backgrounds and other typical KL-types (e.g., KL1) were not represented since they

are infrequent in nosocomial infections in Europe (28). The method is based on the recognition of biochemical patterns associated with the KL-type and, in some cases, of variable KL- and O-type combinations of specific lineages within the same (ST15-KL24 or ST15-KL112) or different STs (ST15-KL24-O1 from ST45-KL24-O2 or ST11-KL27-O2 from ST392-KL27-O4). These data correlate well with those of lineages defined by whole-genome sequencing, though not always at the core genome MLST level (this study) (4, 29). Since intrahospital transmission is dominated by a few highly transmissible clonal lineages carrying the same capsular locus (2, 22, 30–33), assigning isolates to the same KL-type is highly suggestive of genetic relatedness and enough to support effective and real-time infection control (11, 12).

In fact, the high accuracy and sensitivity (~88%) obtained assure that few closely related isolates eventually involved in an outbreak will be missed, while most (if not all) unrelated isolates are discarded. In the context of an outbreak or cluster investigation, early elimination of isolates that are different from each other, as soon as they are detected by an antibiogram or other phenotypic methods, is a very useful tool for infection control teams. For these reasons, we propose the use of FT-IR as a screening tool for clustering and identification of closely related isolates upfront WGS (i) to support early infection control measures based on typing information obtained at the same time as bacterial identification and (ii) to reduce the number of isolates to be sequenced by WGS for a deeper epidemiological analysis. This would decrease the workload, time, and cost associated with typing (Fig. 1), not only for *K. pneumoniae* but also for other species of public health interest such as *S. enterica*, *A. baumannii,* or *E. coli* for which proof-of-concept studies are available (8–10, 18). Furthermore, the method can also be useful for public health surveillance in humans (23), animals (24, 25, 34), or water environments (35).

Pattern recognition techniques are increasingly being explored in other microbiology diagnostic areas such as MALDI-TOF MS-based species differentiation or antibiotic resistance prediction (36–39). Similar strategies have also been used for FT-IR-based serotyping in *Streptococcus pneumoniae* (40), *S. enterica* (41), or *Staphylococcus aureus* (6) and differentiation of *Enterococcus* sp. (21). and yeasts (20). These applications are based on machine-learning classification models (using O or K antigens as classes) that are trained using a well-known spectral data set that is validated by challenging with new input data. We used RF considering the low risk of overfitting with the training set and the easiness to determine feature importance (42). We are aware that speed of analysis might be compromised in larger data sets, but improved RF algorithms might represent a solution (43). Hence, machine-learning will be crucial for future developments of the method, which include (i) the expansion of current databases for other lineages, including hypervirulent *K. pneumoniae*, as well as for other clinically relevant bacteria, (ii) validation in larger samples and in real-time contexts, and (iii) exploring the adequacy and limits of spectral databases that represent a historical record of an institution or a given geographic region. Therefore, a classifier, such as the one created here, must be periodically retrained and adapted to accommodate the *K. pneumoniae* lineages prevalent in the local area, the specific needs and strategies of a given setting or institution, and remain responsive to changes in the bacterial population over time. Hence, larger-scale validation studies are currently underway to optimize spectral databases and/or models, which will be openly shared with the scientific community to foster continued improvement and innovation. Once a machine-learning algorithm is trained and accessible through a user-friendly platform, users can employ the established workflow to obtain typing information without the need for expertise in spectral data analysis, similar to the experience with MALDI-TOF MS.

Notably, we showed that the spectral information required for typing is stable across time, instruments, and culture media. Robustness of ATR FT-IR has been previously demonstrated for yeast identification (44) and is associated with direct colony analysis and the use of a classification model, which prevents the inconsistencies associated with sample preparation and cluster cutoff definition described for IR Biotyper (Bruker

Daltonics, Germany) (17–19). False-negative results belonged to scattered isolates from different KL-types, most of which were correctly predicted in a different culture medium. Thus, to maximize both speed and sensitivity, we recommend testing directly in Columbia Agar with 5% sheep blood and re-test poorly predicted isolates in Mueller-Hinton, after overnight culturing. Moreover, when misidentifications occurred with highly related KL-types, subsequent models improved discrimination and accuracy (e.g., KL23-like), a strategy that has been used previously (45). On the other hand, most false-positive results were obtained for non-characterized KL-types, suggesting a high relatedness to known capsules. Hence, FT-IR spectral information can also be used to confirm or disregard *in silico* predictions based on capsule genotype (*cps*) (46) or eventually to depict evolutionary events involving the capsule that can occur *in vivo* (47–49).

The workflow developed is comparable to that of MALDI-TOF MS, using directly the bacterial colony and obtaining the result in <5 min, including from the Columbia Agar culture isolated directly from the clinical sample. Not only the simplicity of the protocol and automated data analysis make this technique suitable for non-expert users, but also the extraordinary short time-to-response represents a great advantage when compared with that of in-house implemented whole-genome sequencing (usually 48–72 h) (Fig. 1). Furthermore, the possibility to obtain typing information the same day the bacteria are identified constitutes a hallmark of infection control. The Clover MS Data analysis software is simple and flexible and does not require knowledge on spectral data analysis, allowing non-expert users to type through a user-friendly workflow (36, 37). Developments from this study (e.g., spectral processing workflow, algorithm development, and data visualization) were already incorporated into the software, which is available to potential users by subscription. Different entry-level FT-IR ATR instruments from different manufacturers (e.g., PerkinElmer, Thermo-Fisher, and Shimadzu) can be used. The cost of these instruments is lower than that of other specialized equipments (IR Biotyper, MALDI-TOF MS, Illumina, and MinION), and the costs of the reagents are negligible, turning the method especially attractive for low-resource settings (44).

In conclusion, we demonstrated that FT-IR ATR spectroscopy is an accurate, quick, and reproducible tool providing meaningful and accurate information at a very early stage (at the same time as bacterial identification) to support infection control and public health surveillance. Furthermore, the high robustness of the established workflow together with the availability of spectral databases and/or ML models through flexible and user-friendly platforms (Clover MS Data analysis or others) will facilitate adoption of the method and provide opportunities to enhance and consolidate real-time applications at a global level.

ACKNOWLEDGMENTS

This work is financed by national funds from FCT, Fundação para a Ciência e a Tecnologia, I.P., in the scope of the project UIDP/04378/2020 and UIDB/04378/2020 of the Research Unit on Applied Molecular Biosciences, UCIBIO and the project LA/P/0140/2020 of the Associate Laboratory Institute for Health and Bioeconomy, i4HB. Part of the work was supported by funds from the University of Porto's proof-of-concept funding program, BIP Proof. Â. Novais and A.B. Gonçalves are supported by national funds from FCT (Fundação para a Ciência e a Tecnologia, I.P.) through the Scientific Employment Stimulus Program (https://doi.org/10.54499/2021.02252.CEECIND/CP1662/CT0009) and a PhD fellowship (2020.09440.BD), respectively. T. G. Ribeiro is supported by UCIBIO, UIDP/QUI/04378/2020, with the financial support of the FCT/ MCTES through national funds.

AUTHOR AFFILIATIONS

[1]UCIBIO, Applied Molecular Biosciences Unit, Department of Biological Sciences, Faculty of Pharmacy, University of Porto, Porto, Portugal

[2]Associate Laboratory i4HB - Institute for Health and Bioeconomy, Faculty of Pharmacy, University of Porto, Porto, Portugal
[3]CCP, Culture Collection of Porto, Faculty of Pharmacy, University of Porto, Porto, Portugal
[4]1H-TOXRUN, One Health Toxicology Research Unit, University Institute of Health Sciences, CESPU, CRL, Gandra, Portugal
[5]CLOVER Bioanalytical Software, Granada, Spain
[6]Clinical Microbiology Laboratory, Local Healthcare Unit, Matosinhos, Portugal
[7]Unidad Clínica de Enfermedades Infecciosas y Microbiología, Hospital Universitario Vírgen Macarena, Instituto de Biomedicina de Sevilla (IBIS; CSIC/Hospital Virgen Macarena/Universidad de Sevilla), Sevilla, Spain
[8]Departamentos de Microbiología y Medicina, Universidad de Sevilla, Sevilla, Spain

AUTHOR ORCIDs

Ângela Novais http://orcid.org/0000-0003-1171-0326
Ana Beatriz Gonçalves http://orcid.org/0000-0003-0013-7851
Teresa G. Ribeiro http://orcid.org/0000-0003-0433-9485
Lorena López-Cerero http://orcid.org/0000-0001-8950-4384
Jesús Rodríguez-Baño http://orcid.org/0000-0001-6732-9001
Luísa Peixe http://orcid.org/0000-0001-5810-8215

FUNDING

Funder	Grant(s)	Author(s)
MEC | Fundação para a Ciência e a Tecnologia (FCT)	UIDP/04378/2020, UIDB/04378/2020, LA/P/0140/2020	Ângela Novais Ana Beatriz Gonçalves Teresa G. Ribeiro Ana R. Freitas Luísa Peixe
MEC | Fundação para a Ciência e a Tecnologia (FCT)	https://doi.org/10.54499/2021.02252.CEECIND/CP1662/CT0009	Ângela Novais
MEC | Fundação para a Ciência e a Tecnologia (FCT)	2020.09440.BD	Ana Beatriz Gonçalves

AUTHOR CONTRIBUTIONS

Ângela Novais, Conceptualization, Formal analysis, Funding acquisition, Methodology, Software, Supervision, Validation, Writing – original draft, Writing – review and editing | Ana Beatriz Gonçalves, Formal analysis, Methodology, Validation, Writing – review and editing | Teresa G. Ribeiro, Funding acquisition, Investigation, Methodology, Writing – review and editing | Ana R. Freitas, Funding acquisition, Investigation, Methodology, Writing – review and editing | Gema Méndez, Software, Writing – review and editing | Luis Mancera, Software, Writing – review and editing | Antónia Read, Resources, Writing – review and editing | Valquíria Alves, Resources, Writing – review and editing | Lorena López-Cerero, Investigation, Methodology, Resources, Writing – review and editing | Jesús Rodríguez-Baño, Resources, Writing – review and editing | Álvaro Pascual, Resources, Writing – review and editing | Luísa Peixe, Conceptualization, Funding acquisition, Project administration, Validation, Writing – review and editing

DATA AVAILABILITY

All spectra used in this study can be accessed through the Clover Garden Repository, and the instructions and source code used to process the data and build the machine-learning RF model are deposited in GITHUB and ZENODO.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Supplementary figures (JCM01211-23-s0001.docx). Fig. S1 and S2.
Supplementary tables (JCM01211-23-s0002.xlsx). Tables S1 to S4.

REFERENCES

1. Novais Â, Freitas AR, Rodrigues C, Peixe L. 2019. Fourier transform infrared spectroscopy: unlocking fundamentals and prospects for bacterial strain typing. Eur J Clin Microbiol Infect Dis 38:427–448. https://doi.org/10.1007/s10096-018-3431-3
2. Rodrigues C, Sousa C, Lopes JA, Novais Â, Peixe L, Dorrestein PC. 2020. A front line on *Klebsiella pneumoniae* capsular polysaccharide knowledge: Fourier transform infrared spectroscopy as an accurate and fast typing tool. mSystems 5:e00386-19. https://doi.org/10.1128/mSystems.00386-19
3. Hu Y, Zhu K, Jin D, Shen W, Liu C, Zhou H, Zhang R. 2023. Evaluation of IR Biotyper for carbapenem-resistant *Pseudomonas aeruginosa* typing and its application potential for the investigation of nosocomial infection. Front Microbiol 14:1068872. https://doi.org/10.3389/fmicb.2023.1068872
4. Teng ASJ, Habermehl PE, van Houdt R, de Jong MD, van Mansfeld R, Matamoros SPF, Spijkerman IJB, van Meer MPA, Visser CE. 2022. Comparison of fast Fourier transform infrared spectroscopy biotyping with whole genome sequencing-based genotyping in common nosocomial pathogens. Anal Bioanal Chem 414:7179–7189. https://doi.org/10.1007/s00216-022-04270-6
5. Jun SY, Kim YA, Lee S-J, Jung W-W, Kim H-S, Kim S-S, Kim H, Yong D, Lee K. 2023. Performance comparison between Fourier-transform infrared spectroscopy-based IR Biotyper and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry for strain diversity. Ann Lab Med 43:174–179. https://doi.org/10.3343/alm.2023.43.2.174
6. Johler S, Stephan R, Althaus D, Ehling-Schulz M, Grunert T. 2016. High-resolution subtyping of *Staphylococcus aureus* strains by means of Fourier-transform infrared spectroscopy. Syst Appl Microbiol 39:189–194. https://doi.org/10.1016/j.syapm.2016.03.003
7. Vogt S, Löffler K, Dinkelacker AG, Bader B, Autenrieth IB, Peter S, Liese J. 2019. Fourier-transform infrared (FTIR) spectroscopy for typing of clinical *Enterobacter cloacae* complex isolates. Front Microbiol 10:2582. https://doi.org/10.3389/fmicb.2019.02582
8. Silva L, Grosso F, Rodrigues C, Ksiezarek M, Ramos H, Peixe L. 2021. The success of particular *Acinetobacter baumannii* clones: accumulating resistance and virulence inside a sugary shield. J Antimicrob Chemother 76:305–311. https://doi.org/10.1093/jac/dkaa453
9. Campos J, Sousa C, Mourão J, Lopes J, Antunes P, Peixe L. 2018. Discrimination of non-typhoid *Salmonella* serogroups and serotypes by Fourier transform infrared spectroscopy: a comprehensive analysis. Int J Food Microbiol 285:34–41. https://doi.org/10.1016/j.ijfoodmicro.2018.07.005
10. Sousa C, Novais Â, Magalhães A, Lopes J, Peixe L. 2013. Diverse high-risk B2 and D *Escherichia coli* clones depicted by Fourier transform infrared spectroscopy. Sci Rep 3:3278. https://doi.org/10.1038/srep03278
11. Silva L, Rodrigues C, Lira A, Leão M, Mota M, Lopes P, Novais Â, Peixe L. 2020. Fourier transform infrared (FT-IR) spectroscopy typing: a real-time analysis of an outbreak by carbapenem-resistant *Klebsiella pneumoniae*. Eur J Clin Microbiol Infect Dis 39:2471–2475. https://doi.org/10.1007/s10096-020-03956-y
12. Novais Â, Ferraz RV, Viana M, da Costa PM, Peixe L. 2022. NDM-1 introduction in Portugal through a ST11 KL105 *Klebsiella pneumoniae* widespread in Europe. Antibiotics (Basel) 11:92. https://doi.org/10.3390/antibiotics11010092
13. Lombardo D, Cordovana M, Deidda F, Pane M, Ambretti S. 2021. Application of Fourier transform infrared spectroscopy for real-time typing of *Acinetobacter baumannii* outbreak in intensive care unit. Future Microbiol 16:1239–1250. https://doi.org/10.2217/fmb-2020-0276
14. Wang-Wang JH, Bordoy AE, Martró E, Quesada MD, Pérez-Vázquez M, Guerrero-Murillo M, Tiburcio A, Navarro M, Castellà L, Sopena N, Casas I, Saludes V, Giménez M, Cardona P-J. 2022. Evaluation of Fourier transform infrared spectroscopy as a first-line typing tool for the identification of extended-spectrum β-lactamase-producing *Klebsiella pneumoniae* outbreaks in the hospital setting. Front Microbiol 13:897161. https://doi.org/10.3389/fmicb.2022.897161
15. Dinkelacker AG, Vogt S, Oberhettinger P, Mauder N, Rau J, Kostrzewa M, Rossen JWA, Autenrieth IB, Peter S, Liese J. 2018. Typing and species identification of clinical *Klebsiella* isolates by Fourier transform infrared spectroscopy and matrix-assisted laser desorption ionization–time of flight mass spectrometry. J Clin Microbiol 56:e00843-18. https://doi.org/10.1128/JCM.00843-18
16. Lasch P, Naumann D. Infrared spectroscopy in microbiology, p 1–32. In Encyclopedia of analytical chemistry. John Wiley & Sons, Ltd, Chichester.
17. Martak D, Valot B, Sauget M, Cholley P, Thouverez M, Bertrand X, Hocquet D. 2019. Fourier-transform infrared spectroscopy can quickly type Gram-negative bacilli responsible for hospital outbreaks. Front Microbiol 10:1440. https://doi.org/10.3389/fmicb.2019.01440
18. Hu Y, Zhou H, Lu J, Sun Q, Liu C, Zeng Y, Zhang R. 2021. Evaluation of the IR Biotyper for *Klebsiella pneumoniae* typing and its potentials in hospital hygiene management. Microb Biotechnol 14:1343–1352. https://doi.org/10.1111/1751-7915.13709
19. Rakovitsky N, Frenk S, Kon H, Schwartz D, Temkin E, Solter E, Paikin S, Cohen R, Schwaber MJ, Carmeli Y, Lellouche J. 2020. Fourier transform infrared spectroscopy is a new option for outbreak investigation: a retrospective analysis of an extended-spectrum-beta-lactamase-producing *Klebsiella pneumoniae* outbreak in a neonatal intensive care unit. J Clin Microbiol 58:e00098-20. https://doi.org/10.1128/JCM.00098-20
20. Lam LMT, Dufresne PJ, Longtin J, Sedman J, Ismail AA. 2019. Reagent-free identification of clinical yeasts by use of attenuated total reflectance Fourier transform infrared spectroscopy. J Clin Microbiol 57:e01739-18. https://doi.org/10.1128/JCM.01739-18
21. Nitrosetein T, Wongwattanakul M, Chonanant C, Leelayuwat C, Charoensri N, Jearanaikoon P, Lulitanond A, Wood BR, Tippayawat P, Heraud P. 2021. Attenuated total reflection Fourier transform infrared spectroscopy combined with chemometric modelling for the classification of clinically relevant Enterococci. J Appl Microbiol 130:982–993. https://doi.org/10.1111/jam.14820
22. David S, Reuter S, Harris SR, Glasner C, Feltwell T, Argimon S, Abudahab K, Goater R, Giani T, Errico G, et al. 2019. Epidemic of carbapenem-resistant *Klebsiella pneumoniae* in Europe is driven by nosocomial spread. Nat Microbiol 4:1919–1929. https://doi.org/10.1038/s41564-019-0492-8

23. Guerra AM, Lira A, Lameirão A, Selaru A, Abreu G, Lopes P, Mota M, Novais Â, Peixe L. 2020. Multiplicity of carbapenemase-producers three years after a KPC-3-producing *Klebsiella pneumoniae* ST147-K64 hospital outbreak. Antibiotics (Basel) 9:806. https://doi.org/10.3390/antibiotics9110806
24. Mourão J, Ribeiro-Almeida M, Novais C, Magalhães M, Rebelo A, Ribeiro S, Peixe L, Novais Â, Antunes P. 2023. From farm to fork: persistence of clinically relevant multidrug-resistant and copper-tolerant *Klebsiella pneumoniae* long after colistin withdrawal in poultry production. Microbiol Spectr 11:e0138623. https://doi.org/10.1128/spectrum.01386-23
25. Ribeiro-Almeida M, Mourão J, Novais Â, Pereira S, Freitas-Silva J, Ribeiro S, Martins da Costa P, Peixe L, Antunes P. 2022. High diversity of pathogenic *Escherichia coli* clones carrying *mcr - 1* among gulls underlines the need for strategies at the environment–livestock–human interface. Environ Microbiol 24:4702–4713. https://doi.org/10.1111/1462-2920.16111
26. Jünemann S, Sedlazeck FJ, Prior K, Albersmeier A, John U, Kalinowski J, Mellmann A, Goesmann A, von Haeseler A, Stoye J, Harmsen D. 2013. Updating benchtop sequencing performance comparison. Nat Biotechnol 31:294–296. https://doi.org/10.1038/nbt.2522
27. Carriço JA, Silva-Costa C, Melo-Cristino J, Pinto FR, de Lencastre H, Almeida JS, Ramirez M. 2006. Illustration of a common framework for relating multiple typing methods by application to macrolide-resistant *Streptococcus pyogenes*. J Clin Microbiol 44:2524–2532. https://doi.org/10.1128/JCM.02536-05
28. Russo TA, Marr CM. 2019. Hypervirulent *Klebsiella pneumoniae*. Clin Microbiol Rev 32:e00001-19. https://doi.org/10.1128/CMR.00001-19
29. Wendel AF, Peter D, Mattner F, Weiss M, Hoppenz M, Wolf S, Bader B, Peter S, Liese J. 2022. Surveillance of *Enterobacter cloacae* complex colonization and comparative analysis of different typing methods on a neonatal intensive care unit in Germany. Antimicrob Resist Infect Control 11:54. https://doi.org/10.1186/s13756-022-01094-y
30. Rodrigues C, Desai S, Passet V, Gajjar D, Brisse S. 2022. Genomic evolution of the globally disseminated multidrug-resistant *Klebsiella pneumoniae* clonal group 147. Microb Genom 8:000737. https://doi.org/10.1099/mgen.0.000737
31. Rodrigues C, Lanza VF, Peixe L, Coque TM, Novais Â. 2023. Phylogenomics of globally spread clonal groups 14 and 15 of *Klebsiella pneumoniae*. Microbiol Spectr 11:e0339522. https://doi.org/10.1128/spectrum.03395-22
32. Follador R, Heinz E, Wyres KL, Ellington MJ, Kowarik M, Holt KE, Thomson NR. 2016. The diversity of *Klebsiella pneumoniae* surface polysaccharides. Microb Genom 2:e000073. https://doi.org/10.1099/mgen.0.000073
33. Tsai C-C, Lin J-C, Chen P-C, Liu E-M, Tsai Y-K, Yu C-P, Li J-J, Wang C-H, Fung C-P, Lin F-M, Chang F-Y, Siu LK. 2023. A 20-year study of capsular polysaccharide seroepidemiology, susceptibility profiles, and virulence determinants of *Klebsiella pneumoniae* from bacteremia patients in Taiwan. Microbiol Spectr 11:e0035923. https://doi.org/10.1128/spectrum.00359-23
34. Eisenberg T, Rau J, Westerhüs U, Knauf-Witzens T, Fawzy A, Schlez K, Zschöck M, Prenger-Berninghoff E, Heydel C, Sting R, Glaeser SP, Pulami D, van der Linden M, Ewers C. 2017. *Streptococcus agalactiae* in elephants – a comparative study with isolates from human and zoo animal and livestock origin. Vet Microbiol 204:141–150. https://doi.org/10.1016/j.vetmic.2017.04.018
35. Tata A, Marzoli F, Cordovana M, Tiengo A, Zacometti C, Massaro A, Barco L, Belluco S, Piro R. 2023. A multi-center validation study on the discrimination of *Legionella pneumophila* sg.1, *Legionella pneumophila* sg. 2-15 and *Legionella* non-pneumophila isolates from water by FT-IR spectroscopy. Front Microbiol 14:1150942. https://doi.org/10.3389/fmicb.2023.1150942
36. Gato E, Arroyo MJ, Méndez G, Candela A, Rodiño-Janeiro BK, Fernández J, Rodríguez-Sánchez B, Mancera L, Arca-Suárez J, Beceiro A, Bou G, Oviaño M. 2023. Direct detection of carbapenemase-producing *Klebsiella pneumoniae* by MALDI-TOF analysis of full spectra applying machine learning. J Clin Microbiol 61:e0175122. https://doi.org/10.1128/jcm.01751-22
37. Candela A, Guerrero-López A, Mateos M, Gómez-Asenjo A, Arroyo MJ, Hernandez-García M, Del Campo R, Cercenado E, Cuénod A, Méndez G, Mancera L, Caballero J de D, Martínez-García L, Gijón D, Morosini MI, Ruiz-Garbajosa P, Egli A, Cantón R, Muñoz P, Rodríguez-Temporal D, Rodríguez-Sánchez B. 2023. Automatic discrimination of species within the *Enterobacter cloacae* complex using matrix-assisted laser desorption ionization–time of flight mass spectrometry and supervised algorithms. J Clin Microbiol 61:e0104922. https://doi.org/10.1128/jcm.01049-22
38. Weis C, Cuénod A, Rieck B, Dubuis O, Graf S, Lang C, Oberle M, Brackmann M, Søgaard KK, Osthoff M, Borgwardt K, Egli A. 2022. Direct antimicrobial resistance prediction from clinical MALDI-TOF mass spectra using machine learning. Nat Med 28:164–174. https://doi.org/10.1038/s41591-021-01619-9
39. Ballard SA, Sherry NL, Howden BP. 2023. Public health implementation of pathogen genomics: the role for accreditation and application of ISO standards. Microb Genom 9:mgen001097. https://doi.org/10.1099/mgen.0.001097
40. Passaris I, Mauder N, Kostrzewa M, Burckhardt I, Zimmermann S, van Sorge NM, Slotved H-C, Desmet S, Ceyssens P-J. 2022. Validation of Fourier transform infrared spectroscopy for serotyping of *Streptococcus pneumoniae*. J Clin Microbiol 60:e0032522. https://doi.org/10.1128/jcm.00325-22
41. Cordovana M, Mauder N, Join-Lambert O, Gravey F, LeHello S, Auzou M, Pitti M, Zoppi S, Buhl M, Steinmann J, et al. 2022. Machine learning-based typing of *Salmonella enterica* O-serogroups by the Fourier-transform infrared (FTIR) spectroscopy-based IR Biotyper system. J Microbiol Methods 201:106564. https://doi.org/10.1016/j.mimet.2022.106564
42. Dou B, Zhu Z, Merkurjev E, Ke L, Chen L, Jiang J, Zhu Y, Liu J, Zhang B, Wei G-W. 2023. Machine learning methods for small data challenges in molecular science. Chem Rev 123:8736–8780. https://doi.org/10.1021/acs.chemrev.3c00189
43. Yates D, Islam MZ. 2021. FastForest: increasing random forest processing speed while maintaining accuracy. Inf Sci 557:130–152. https://doi.org/10.1016/j.ins.2020.12.067
44. Lam LMT, Ismail AA, Lévesque S, Dufresne SF, Cheng MP, Vallières É, Luong M-L, Sedman J, Dufresne PJ. 2022. Multicenter evaluation of attenuated total reflectance Fourier transform infrared (ATR-FTIR) spectroscopy-based method for rapid identification of clinically relevant yeasts. J Clin Microbiol 60:e0139821. https://doi.org/10.1128/JCM.01398-21
45. Sousa C, Silva L, Grosso F, Lopes J, Peixe L. 2014. Development of a FTIR-ATR based model for typing clinically relevant *Acinetobacter baumannii* clones belonging to ST98, ST103, ST208 and ST218. J Photochem Photobiol B 133:108–114. https://doi.org/10.1016/j.jphotobiol.2014.02.015
46. Lam MMC, Wick RR, Watts SC, Cerdeira LT, Wyres KL, Holt KE. 2021. A genomic surveillance framework and genotyping tool for *Klebsiella pneumoniae* and its related species complex. Nat Commun 12:4188. https://doi.org/10.1038/s41467-021-24448-3
47. Aihara M, Nishida R, Akimoto M, Gotoh Y, Kiyosuke M, Uchiumi T, Nishioka M, Matsushima Y, Hayashi T, Kang D. 2021. Within-host evolution of a *Klebsiella pneumoniae* clone: selected mutations associated with the alteration of outer membrane protein expression conferred multidrug resistance. J Antimicrob Chemother 76:362–369. https://doi.org/10.1093/jac/dkaa439
48. Ernst CM, Braxton JR, Rodriguez-Osorio CA, Zagieboylo AP, Li L, Pironti A, Manson AL, Nair AV, Benson M, Cummins K, Clatworthy AE, Earl AM, Cosimi LA, Hung DT. 2020. Adaptive evolution of virulence and persistence in carbapenem-resistant *Klebsiella pneumoniae*. Nat Med 26:705–711. https://doi.org/10.1038/s41591-020-0825-4
49. Lee H, Baek JY, Kim SY, Jo H, Kang K, Ko J-H, Cho SY, Chung DR, Peck KR, Song J-H, Ko KS. 2018. Comparison of virulence between matt and mucoid colonies of *Klebsiella pneumoniae* coproducing NDM-1 and OXA-232 isolated from a single patient. J Microbiol 56:665–672. https://doi.org/10.1007/s12275-018-8130-3