Bacterial Typing and Identification Based on Fourier Transform Infrared Spectroscopy

Huayan Yang

Ningbo University

Fangling Wu

Ningbo University

Fuxin Xu

Ningbo University

Keqi Tang

Ningbo University

Chuanfan Ding

Ningbo University

Haimei Shi (✉ shihaimai@nbu.edu.cn)

Ningbo University

Shaoning Yu (✉ yushaoning@nbu.edu.cn)

Ningbo University

Method Article

Keywords: Bacteria, Typing, Identification, Fourier Transform Infrared Spectroscopy

Posted Date: February 14th, 2020

DOI: <https://doi.org/10.21203/rs.2.23337/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Fourier transform infrared (FT-IR) spectroscopy is a label-free and highly sensitive technique that provides complete information on the chemical composition of biological samples. The bacterial FT-IR signals are extremely specific and highly reproducible fingerprint-like patterns, making FT-IR an efficient tool for bacterial typing and identification. Due to the low cost and high flux, FT-IR has been widely used in hospital hygiene management for infection control, epidemiological studies, and routine bacterial determination of clinical laboratory values. However, the typing and identification accuracy could be affected by many factors, and the bacterial FT-IR data from different laboratories are usually not comparable. A standard protocol is required to improve the accuracy of FT-IR-based typing and identification. Here, we detail the principles and procedures of bacterial typing and identification based on FT-IR spectroscopy, including bacterial culture, sample preparation, instrument operation, spectra collection, spectra preprocessing, and mathematical data analysis. Without bacterial culture, a typical experiment generally takes <2 h.

Introduction

1. Introduction

1.1 overview

Efficient typing and identification of bacteria is important for many applications in microbiology¹⁻³. Convenient bacterial typing and identification techniques are in high demand in hospital hygiene management for infection control, epidemiological studies, the tracking of bacterial cross-transmission, and the monitoring of medical treatment processes, including the use of antibiotics, as well as in the environmental waste treatment and food industries^{1,2,4-7}. Various analytical techniques are currently under investigation for rapid, convenient, and reliable bacterial identification and typing, particularly matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF-MS), Fourier transform infrared (FT-IR) spectroscopy, and Raman spectroscopy^{8,9}. Such spectro-techniques not only reduce costs via automation and high-throughput solutions, but also generate results more rapidly¹⁰⁻¹³. These bacterial spectra are highly specific fingerprint-like signatures that are widely used to differentiate, classify, and identify diverse microbial species and strains¹⁴.

Each bacterial species has a complex cell wall composition that presents a specific fingerprint-like pattern on the FT-IR spectra (**Fig. 1**)^{10,15}. The specific fingerprint-like patterns of bacterial infrared signals make FT-IR a low-cost and high-speed alternative for bacterial typing and identification¹⁶⁻²⁰. The whole bacterial FT-IR spectrum is usually complex and the peaks broad due to contributions from all of the biomolecules present in a bacterial cell²¹. Naumann et al. recommended dividing the bacterial FT-IR spectrum into five sub-ranges for bacterial identification and typing: the membrane amphiphile region (3000-2800 cm⁻¹), the biological infrared-silent region (2800-1800 cm⁻¹), the protein and peptide region (1500-800 cm⁻¹), the protein and fatty acid region (1500-1200 cm⁻¹), and the nucleic acid and polysaccharide region (1200-900

cm⁻¹) ^{16,17}. Some bacterial FT-IR band assignments used in bacterial identification and typing are listed in Table 1.

The bacterial FT-IR signals are disturbed by the unavoidable absorbance of H₂O, which resides in bacteria and the external environment^{17,22}. The effect of water on FT-IR bacterial typing was investigated, with elimination of the protein and lipid region greatly increasing the accuracy of bacterial typing. The polysaccharide and nucleic acid FT-IR region (1200-900 cm⁻¹) are recommended by many investigators for bacterial typing^{22,23}.

The IR-spectroscopic measurements included the traditional absorbance/ transmission (A/T), the attenuated total reflection (ATR) (**Fig. 2**), and diffuse reflectance mode. All measurements can be used in bacterial typing and identification^{10,17,24}. The A/T mode is more common and based on the detection of the transmitted IR radiation, and the bacterial sample is placed in the path of the IR beam and scanned. The diffuse reflectance techniques are based on the reflection of the IR beam after contacting with the sample and receives a distinct designation depending on the reflection process. In ATR mode, the sample is placed on an optically dense crystal of higher relative refractive index. This sampling technique is widely used in the context of bacterial typing due to the associated versatility and no requirement for sample preparation¹⁰.

Various spectral manipulations can be performed to improve the spectral features for spectral interpretation and analysis, as bacterial FT-IR spectra are very similar with subtle biochemical differences²⁵⁻²⁸. The processing of bacterial FT-IR spectra includes spectral preprocessing and multivariate analysis. Baseline correction, spectral normalization, and derivative mathematization of bacterial FT-IR spectra are fundamental steps in spectral preprocessing and can increase the accuracy of bacterial typing and identification^{26,29}. The multivariate statistical approaches can be divided into two types: unsupervised methods (e.g., principal component analysis [PCA], hierarchical cluster analysis [HCA]), and supervised methods (e.g., PCA-linear discriminant analysis, artificial neural networks [ANNs]). The supervised methods require prior knowledge of the bacterial identity^{26,30}. With a set spectra of known bacteria, a model can be trained to identify unknown bacteria.

The aim of this study was to describe the detailed procedures for bacterial identification and typing based on FT-IR spectroscopy, including sample preparation, spectral acquisition, and data analysis.

Table 1| Band assignment in bacterial FT-IR spectra.

Wavenumber (cm ⁻¹)
Assignment
References
~3500

O-H stretching of hydroxyl groups

26,31

~3200

N-H stretching of proteins (amide I)

31

2955~2850

C-H

32,33

1741-1715

C=O stretching of esters, C=O stretching of carbonic acids, nucleic acids

33,34

~1695

Amide I band components

17,33

~1685, ~1675

β -sheets and β -turns

17,33

~1655

Amide I of α -helical structures

17,33

~1637

Amide I of β -pleated sheet structures

17,33

~1548

Amide II

7,33

1515

Tyrosine band

7

1468

C-H deformation of -CH_2

35,36

~1400

C=O Stretching (symmetric) of COO^-

37

1310-1240

Amide III band components of proteins

7

1250-1220

P=O stretching (asymmetric) of PO_2^- phosphodiester

37

1200-900

C-O-C, C-O dominated by ring vibrations of carbohydrates C-O-P, P-O-P

38,31,39

1175-1140

Differences in glycosidic linkage configuration

35

1160-1020

C-O-C glycosidic linkage, C-O-C stretching of carbohydrates, C-O stretching

35

1085

P=O Stretching (symmetric) of PO_2^-

38,40,41

1000-970

Differences in glycosidic linkage configuration

35

720

C-H rocking of $-\text{CH}_2$

33

900-600

Fingerprint region

41

1.2 Development of the protocol

The use of infrared spectroscopy in microbiological analysis was first reported in the 1950s⁴²⁻⁴⁴. However, only after the development of the FT technique and powerful computers did it boost the number of applications of FT-IR in the field of bacteria in the late 1980s and 1990s⁴⁵⁻⁴⁹. Modern multivariate statistical analysis, such as factor analysis and ANNs, have contributed greatly to the utilization of this methodology for bacterial identification and typing at different taxonomic levels²⁶. The revival of IR spectroscopy as a means of characterizing microbial samples has been a subject of different reviews in the past 20 years^{10,27,29,36,50}. Recently, the future perspective of FT-IR for healthcare and clinical applications has been discussed⁵¹. The potential of the methodology as a quick, inexpensive, and high-flux tool for bacterial identification and typing is widely accepted⁵¹.

1.3 Application of the method

Bacterial typing and identification by FT-IR can be applied to the fields of general microbiology^{21,52,53}, rapid identification of life-threatening pathogens^{29,54}, epidemiological investigations and pathogen

screening^{14,55,56}, characterization and screening of microorganisms from the environment^{9,57,58}, and maintenance of strain collections⁵⁹. The audiences include the investigator and operators in microbiology and related fields. Recently, a special instrument (IR biotyper Bruker) was marketed by Bruker Daltonik GmbH (Germany), and more audiences will be interested in this technique.

1.4 Comparison with other methods

Many bacterial identification and typing methods have been developed for specific purposes, but many traditional methods are based on multistep culture-based assays, which is time-consuming and costly. Therefore, multistep culture-based assays are progressively being replaced by molecular biology methods⁶⁰⁻⁶⁴. One of the most well-developed molecular biology methods is polymerase chain reaction (PCR)⁶¹. It has high discriminatory power based on variation in the components, structure, and sequence of bacterial genetic material. Some PCR-based methods, such as pulsed-field gel electrophoresis and whole-genome sequencing, are regarded as the “gold standard”, though some of these methods are semi-automated, with high cost and length of analysis impairing routine implementation⁶⁵. Moreover, typing based on genetic information is not always straightforward, and a gap exists between genotypes and phenotype⁵⁰.

Among newly developed methods, MALDI-TOF-MS and FT-IR have demonstrated promising results for bacterial identification and typing. MALDI-TOF-MS is based on the molecular mass of small peptides and ribosomal proteins in bacteria. The low cost, simplicity, and time-efficiency make MALDI-TOF-MS an attractive option for bacterial identification and typing^{44,66-70}. MALDI-TOF-MS databases have been set-up and are widely used in the identification of clinical bacteria⁷¹⁻⁷³ and fungi⁷⁴⁻⁷⁶, with an accuracy of 90% at the species level. Although promising, the robustness and discriminatory power needs to be improved by a unified process, such as sample preparation and data analysis. Unfortunately, an inherent difference in peak intensity or peak location may be present in the mass spectrum, which is related to independent acquisitions in time and in different devices or laboratories. Moreover, many investigations have reported that the accuracy of MALDI-TOF-MS is limited in some highly similar species, including *Escherichia coli* and *Shigella* species, two important bacteria that cause different clinical diseases⁷⁷⁻⁸⁰.

1.5 Advantages and limitations of the method

The notable advantages of FT-IR methods are¹⁷ that it is uniformly applicable to virtually all microorganisms that can be grown in culture; the results are available within minutes of obtaining adequate samples of the pure culture; IR spectroscopy can classify microorganisms at different levels of taxonomic discrimination without any preselection of strains by other taxonomic criteria; the specificity of the method is generally extremely high, allowing differentiation at the strain and serotype level; it can be used in epidemiological investigations, pathogen screening, hygiene, and therapy control; and it is rapid, cost-effective, non-destructive, in situ detection.

The limitations of FT-IR are that only microorganisms that can be grown in culture and are available as pure cultures can be analyzed; stable results can only be obtained when the microbiological parameters (culture medium, cultivation time, and temperature) are rigorously controlled; the spectral databases from different laboratories are not compatible unless all conditions (instrument, culture media, etc.) are the same; and only a few highly specialized databases have been set-up. Many general applications require a broad and comprehensive database, as in clinical, food, or environmental microbiology, as reference databases need to deal with the diversity in species. Furthermore, multiple studies have shown promising results, but standardization is lacking⁸¹. Random variation between studies can originate from differences in instrumentation, operators, and environmental conditions. For example, before FT-IR studies, instrument settings, sample preparation, and operation mode should be optimized in order to improve the spectral quality and molecular sensitivity¹⁹.

1.6 Experimental design

The main experimental steps are bacterial culture, sample preparation, instrument settings, acquisition of spectra, spectral preprocessing, and computational analysis (**Fig. 3**).

1.6.1 Bacterial culture

The strains can be from strain collections or clinical isolates. Only a single colony in culture medium should be used for FT-IR analysis. The parameters, such as age of culture, cultivation medium, incubation time, and the preparation protocol, should be the same. Differences or changes in any of these parameters may lead to false negative results.

Bacterial sample cultivation is started from bacteria of the same age. Pure single colonies are selected for inoculation (a pure culture is essential; if only small colonies are present, an additional streak and incubation is required to increase cell mass). The cell material is spread using an inoculation loop in a way that generates confluent growth. In certain cases, subculturing (on the same medium) may be necessary in order to adapt the organism to the growth medium. The cultivation conditions, including the agar medium used, the temperature, and the incubation time, depend on the organism. These conditions must be as standardized as possible to take advantage of the high selectivity of the IR method. Anaerobic and microaerophilic bacteria (e.g., lactic acid bacteria) are grown in a sealed container in the presence of an oxygen-sorbent (e.g., Anaerocult, Merck; sorbent system should be standardized).

1.6.2 Instrumental parameters

For FT-IR spectrometry, water vapor rotational bands are of low intrinsic half-width and their precise wavenumber position and intensities are dependent on parameters such as partial pressures or temperature. Therefore, the FT-IR spectra can be the results of environmental conditions around the instrument. To collect high-quality FT-IR spectra of bacteria, the FT-IR instrument should be placed in an enclosed space avoiding disturbance of air flow. In addition, vapor contributions should be reduced by purging the instrument with dry air or nitrogen during spectra collection. Furthermore, the particular

instrument will influence the signal-to-noise ratio (SNR); thus, user-specific conditions and some instrumental parameters need to be optimized. Normally, the optimal resolution of the instrument was set to 4 cm⁻¹ or lower and the scan number to 62 scans or more. Background scans of the same sample should have the same parameters. Subsequently, the sample spectra will then undergo a background correction to result in a final spectrum.

1.6.3 Spectral acquisition

Further details regarding spectral acquisition using FT-IR spectroscopy can be found in our previous protocols³³.

1.6.4 Spectra preprocessing

Pre-processing of the spectrum is a significant issue, as the application and effectiveness of preprocessing can greatly affect the results of the computational analysis. The pre-processing of spectra can correct variation related to spectrum acquisition, improve the robustness and accuracy of subsequent multivariate analyses, remove outliers, reduce the dimensionality of the data, and increase data interpretability by correcting issues associated with spectral data acquisition⁸². Pre-processing includes the following processes: spectrum subtracting, baseline correction, derivatization, normalization, and spectrum window selection⁸³. An illustration of the recommended preprocessing workflow based on FT-IR spectra is given in **Fig. 4**.

Normalization: Normalization has been identified as one of the most important pre-processing methods and is commonly applied to minimize the effects of varying optical path lengths on the data, or to compensate for intensity variations in the source. The result of normalization is a spectrum that is simultaneously scaled and offset-corrected⁸⁴. The most popular normalization method is vector normalization. Vector normalization is a widely used method for spectral normalization. Rather than a simple scaling operation, it ensures that all spectra within a dataset have the same vector length of 1. The vector of absorption intensities x can be converted to its corresponding unit vector by dividing each element, x , by the vector length,⁸⁶

[1]

Baseline correction: In transmission or transflection type IR spectroscopy, spectral baselines can be distorted as a result of scattering, absorption by the supporting substrate, changing conditions during data collection, or the variance due to instrumental factors. Such baseline distortions are critical when the absorbance values are systematically evaluated. Although baseline correction methods may rely on distinct principles and algorithms, they have the common objective of minimizing unwanted spectral offsets, broad baseline distortions, positive or negative slopes, and other baseline effects in vibrational spectra. Subtracting the estimation of a background from the un-processed spectrum leads to a more interpretable signal, allowing spectral parameters to be determined more accurately⁸⁵. A popular baseline technique available with Bruker Opus software is rubber-band baseline correction, which stretches the

spectra down so that the minima in the spectral region of interest can be used to fit a convex polygonal line (i.e., the rubber band), which is then subtracted from the original spectrum.

Polynomial baseline correction is intended to be used when distortion of the spectra due to differentiation is not desired⁸⁶. The method is based on a least-squares polynomial curve-fitting function. The degree of the polynomial, k , is modulated in the user input. Equation 2 shows the generalized polynomial formula, where M is the number of terms associated with the polynomial at a particular degree. The measured spectrum with number of data points n becomes the fitted spectrum; a represents the constant, and x represents the indeterminate of the polynomial. The residual sum-of-squares (RSS) between these two is given by Equation (3). These equations show a very standard polynomial least-squares fit. However, the implementation becomes a more useful baseline correction approach by iterating the smoothing 100 times (by default), and having a tolerance for change in RSS between iterations of 0.001 by default.

[2]

[3]

During this process, the small peak from 2,000 to 1,750 cm^{-1} resulting from water vapor can be diminished manually. The peak can be leveled by connecting lines through the selected points and subtracting them from the trace. This method can remove the baseline slope and be offset by an iterative fitting process. The algorithm should be used before resolution enhancement techniques.

Analytical window selection: Bands known to be in the 900–1200 region include the P–O–C and C–O–C stretches of mostly oligo- and polysaccharides^{40,87}. The multiplet of peaks in this cluttered spectral region surely contains many polysaccharide or sugar C–O–C bands. This is important, as the relative amplitudes of these bands is what gives this multiplet of peaks subtle, but very reproducible, band shapes for differentiating species. The polysaccharide and nucleic acid region (900–1200 cm^{-1}) is the only region that is not disturbed by H_2O under certain conditions^{17,22}. The typing results based on this region are much better than with any other region or combination of regions¹⁸.

Derivative analysis: Recently, mathematical tools were applied to minimize confounding effects and to control optical or distorting influences^{27,88–90}. First or second derivative transformations can resolve broad bands overlapping raw spectra, reduce replicate variability, correct baseline shift, and amplify spectral variations^{91–97}. The second derivative spectra are most commonly used for bacterial classification⁹⁸, typing and identification^{91,95,99,100}. Third or higher order derivative spectra tend to distort the graph and are not recommended.

1.6.5 Multivariate data analysis for bacterial typing and identification

Multivariate data analysis involves the application of several statistical methods to determine the similarities and differences between bacterial strains, and to identify the spectral ranges that mostly relate to these similarities and differences¹⁰¹. Different illustrations of multivariate data analysis can be

found in **Fig. 5**. The two-dimensional projection score plots of 7 standard *E. coli* strains analyzed by PCA were shown in **Fig. 5A**. The relationship between these strains analyzed by HCA were shown in **Fig. 5B**. A heat-map overview allow one to visualize the data set in a clear way, and the segregation between bacterial classes can be discriminated in **Fig. 5C**. Each visualization is available for users to choose.

The multivariate data analysis can be divided into supervised and unsupervised methods^{25,27,102}. In unsupervised methods, such as PCA and HCA, are established according to the similarity between spectra⁹. Typing and differentiation of groups of related isolates does not necessarily rely on existing spectral databases or prior knowledge, but the identification of unknown isolates depends on the comparison of spectra with a reference database. Supervised methods are usually used to identify samples, as they use the category membership of samples to define classes²⁸. Partial least squares discriminant analysis (PLSDA) or ANNs are powerful for extracting specific spectral signatures^{27,97,103}. The only commercial database created for bacterial species identification is Bruker Optik, and most studies have relied on in-house databases¹⁰⁴.

Cluster analysis: In cluster analysis, a matrix is calculated that expresses the similarity, or “distance”, between each spectrum and all other spectra of the data. There are several ways to evaluate the distance between or similarity of two spectra. The simplest way is to calculate the EUCLIDEAN distance between vectors. Other methods, such as CORRELATION and COSINE, calculate the Pearson correlation coefficient and the angle between the vectors, respectively. For convenience, the results are also translated into a distance value.

For two spectra, *S* and *R*, in the data hypercube, this distance is defined as the correlation coefficient C_{SR} according to Equation 4.

[4]

Here, the spectra *S* and *R* are represented by one-dimensional vectors of *M* absorbance values, and $\bar{S}$ and $\bar{R}$ are the mean values for each vector. The resulting C_{SR} matrix contains N^2 entries, where *N* is the total number of spectra within the data set. As the matrix is symmetric, only $N(N-1)/2$ spectral distance elements C_{SR} need to be computed.

Subsequently, the two most similar spectra in the hyper-cube are merged into a “cluster”, and a new distance matrix column is calculated for the new cluster and all existing spectra. The process of merging spectra or clusters into new clusters is repeated, and the C_{SR} is recalculated until all spectra have been combined into a few clusters. The mean spectra are extracted for all clusters and used for interpreting the chemical or biochemical differences between clusters. Reasonable noise in the spectral data does not affect the clustering process. In this respect, cluster analysis is much more stable than other methods of multivariate analysis in which an increasing amount of noise accumulates in the less relevant clusters

Hierarchical cluster analysis: HCA and dendrograms are used to show analogies between bacterial spectra. Items closely related are clustered together on a branch with short twigs. The lines parallel to the item axis always connect two clusters (or single items) at a height that indicates at which distance the two clusters were merged into one. Therefore, in dendrograms, the left vertical axis describes the increasing variance or heterogeneity. The magnitude of this heterogeneity is based on the number of spectra in a cluster and the analogies between them.

When starting an agglomerative HCA, the distances calculated for all N items are sorted and the smallest distance between two single items leads to their merge into the first cluster. The second smallest distance is then processed. This is easy as long as only single items are involved. As far as the distance between a cluster and a single item, or the distance between two clusters, the two most useful linkage types are the AVERAGE linkage and WARD's algorithm, which are recommended for a first approach.

Though WARD often results in very nice and biologically meaningful clustering, one of its disadvantages is the distance scale, which varies with the number of items analyzed. In contrast, AVERAGE shows a scale that is independent of the number of items. This allows comparisons of completely different data sets and defining cut-off values for the distance to decide whether isolates are indistinguishable, closely related, or unrelated. Cluster analysis was performed by importing data hypercubes in the Bruker OPUS 3.0 format software package¹⁰⁵.

With regard to isolates and spectra, the most important goal for exploration of the presented data is to find a reasonable cut-off value for the distance to decide which isolates fall into the same cluster and are considered indistinguishable. There are at least two parameters to consider: the cut-off value should be as low as possible to achieve high discriminatory power and is only really discriminatory if spectra that belong together (e.g., the technical replica measurements) are not scattered over different clusters and are coherently together in one cluster. When looking at a data set that contains only isolates that were measured once (only technical replicates present), the correct cut-off value will likely be underestimated. Therefore, we advise the user add at least three independent measurements of some well-known and different isolates.

Principal components analysis: PCA is one of the most common techniques for exploratory analysis. The original data are decomposed into a few principal components (PCs) responsible for most of the variance within the original dataset. The PCs are orthogonal to each other and are generated in a decreasing order of explained variance, so that the first PC represents most of the original data variance, followed by the second PC, and so forth¹⁰⁶. Mathematically, the decomposition takes the form:

$$X = TP^T + E \quad [5]$$

Where X represents the preprocessed data (e.g., preprocessed sample spectra), T is the score, P is the loading, and E is the residual.

For standardization applications, PCA is a fast and reliable tool for determining whether differences exist between the spectra acquired by different systems. The PCA loadings represent the variance in the variable (e.g., wavenumber) direction and are used to detect the variables with the highest importance for the pattern observed in the scores. The PCA scores represent the variance in the sample direction, and they are used to assess similarities/dissimilarities among the samples, detecting clustering patterns.

Feature extraction: Feature extraction can be split into two distinct approaches: feature construction and feature selection²⁶. Briefly, feature construction can be defined as the creation of new features in a data set that can infer otherwise obscured information, such as the previously mentioned linear methods PCA and PLS. This can be especially important for diagnostics, biomarker extraction, and pattern recognition in otherwise homogeneous data sets, and it plays an important role in hyperspectral imaging, as individual pixels can be reduced to single values related to spectral intensity or variance¹⁹. Feature selection approaches extrapolate existing features from the data set, such as specific wavenumbers, that can be used to determine spectral biomarkers and/or feed into diagnostic frameworks¹⁰⁷. Techniques such as genetic algorithm, multivariate curve resolution, and successive projection algorithm are particularly popular as feature-extraction methods, as only informative variables are included in the resultant model¹⁰⁸.

Identification: Identification methods can also be separated into two types: unsupervised and supervised. In unsupervised methods, a similarity comparison algorithm is usually directly used to calculate the spectral distances between unknown bacteria and known bacteria included in a pre-constructed library, or by introducing it in the HCA¹⁰, without artificially correcting parameters and results in the calculation process. By judging the similarity distance, bacteria can be identified at the genus, species, or subspecies level. The Pearson correlation coefficient and Euclidean distance are frequently used in distances measures, and Ward's link algorithm is most commonly employed in HCA to construct clusters¹⁰⁹. The following Pearson correlation coefficient is defined as , in which and represent the intensity values of features in two profiles, and , and represents the number of features in each profile¹¹⁰.

[6]

In supervised methods, initially each sample is artificially grouped into a definite class (genus, species, or strain, etc.) according to prior knowledge, and then the relationship between the data and the considered class will be built using methods such as discriminant analysis (DA) and canonical variate analysis (CVA)¹⁰ with the purpose of forming weighted linear combinations of the data to minimize within group variance and maximize between group variance. The unknown bacteria can then be identified by recognizing the class assigned to it after a projection transformation using the former relationship with its feature data as the input. Compared to the supervised methods, the unsupervised methods have higher efficiency and lower expertise requirements, making them, particularly library-based methods, widely used in various fields for routine bacterial identification¹¹⁰. The precise feature-extraction mentioned above is key for high-quality library construction and accurate identification, and the lower

experimental standardization requirements (e.g., bacterial growth and culture media) of the features result in more stable and reliable identification. A reference strain (e.g., *E. coli* 8739) or standard reference material is often used as quality control to test the accuracy of the model¹¹¹.

ANN analysis: An alternative to creating a classification model that represents the hierarchy is a supervised bioinformatics approach, such as ANN analysis¹¹². Compared to cluster analysis, ANN analysis has the advantage of objective validation of the results. ANNs can be used in identification by analyzing raw data and work well for managing overlapping data because they are a pattern analysis method of advanced multivariate data processing in which large amounts of information are analyzed by training the data in a pattern recognition algorithm to recognize the particular combination of variables in a subset of data¹¹³. The general strategy of ANN analysis includes teaching and optimizing the network models, followed by testing the classifiers with independent (external) validation data sets¹¹⁴. Teaching and internal validation were carried out on the basis of IR spectra with known class assignment using spectra from the database. External validation of the classifier was performed by generating ANN images from infrared spectral maps. Thus, the classifiers were created with database spectra, whereas the model was assessed by comparing the ANN images.⁸¹ The primary advantage of this method is the capability to separately train and validate small and flexible networks that can be combined to build large modular ANN systems^{115,116}.

Reagents

2.1 Reagents

- Ethanol, absolute (AR, Shanghai Titan Scientific Co., Ltd., CAS 64-17-5).
- Water, deionized.
- Sodium chloride (AR Shanghai Titan Scientific Co., Ltd. CAS 7647-14-5).
- Tryptone soy agar (Beijing Land Bridge Technology Co., Ltd.)
- LB broth (Sangon Biotech Co., Ltd., Shanghai).
- Standard strains (China Center of Industrial Culture Collection [CICC] and American Type Culture Collection [ATCC]).
- Isolation strains (Huashan Hospital of Shanghai, Shanghai, China).

2.2 Equipment

- Eppendorf tubes (Shanghai Titan Scientific Co., Ltd.).
- Petri dish (Shanghai Titan Scientific Co., Ltd.).

- Disposable inoculation loops (Shanghai Titan Scientific Co., Ltd.).
- Vortexer XW-80A (Shanghai Huxi Instruments Co., Ltd.).
- Suitable pipettes for volumes from 10 µL to 100 µL (Shanghai Titan Scientific Co., Ltd.).
- Mask (Shanghai Titan Scientific Co., Ltd.).
- Safety glasses (Shanghai Titan Scientific Co., Ltd.).
- Pipette tips (Shanghai Titan Scientific Co., Ltd.).
- ATL-032R digital shaker (Shanghai Chemstar Instruments Co., Ltd. China).
- Centrifuge (Shanghai Boyu Instruments Co., Ltd.).
- Autoclave (Shanghai Boxun Industry & commerce Co., Ltd.).

▲ **CRITICAL** In our laboratory, the FT-IR Laboratory Analyzer was used (ABB, U.S.A.) with GRAMS/32 software (Biotoool, U.S.A.) and Bruker Biotyper (Bremen, Germany) controlled by OPUS and software from Bruker Optik (Bremen, Germany).

Our protocol for the preparation of bacterial samples can also be applied using other FT-IR spectrometers from well-known companies (e.g. Shimadzu [Japan]). Alternative software from commercial or academic sources as described in Box 1 can also be used.

- FT-IR spectroscopy equipment: for a list of available commercial instruments, please refer to **Table 2**.
- CaF₂ liquid cells: demountable transmission cell with spacer can be purchased from Specac (<http://www.specac.com>) or Perkin Elmer (<http://www.perkinelmer.com/>) or from any other company. Cells with two rounded plates can be bought from BioTools under the trade name BioCell (<http://btools.com>).
- Silicon microtiter plates (Bruker Part No. I23258P) and microtiter plate frame.
- Air compressor machine.

2.3 Reagent setup

- TSA medium: 38 g TSA medium powder added to 1 liter of double distilled water. The autoclaved mixture can be stored at room temperature.
- LB broth medium: 25 g LB broth medium powder added to 1 liter of double distilled water. The autoclaved mixture can be stored at 4°C until use.

▲ **CRITICAL** Fresh liquid media or autoclaved liquid media can be stored at 4°C until use (several weeks). For plating, heat TSA solid media until dissolved, cool to 50°C, and plate in Petri dish. Store the agar plates in a refrigerator at 4°C up to a month.

● 0.9% (wt/vol) NaCl solution: 9 g NaCl added to 1 liter of double distilled water. The autoclaved mixture can be stored at 4°C until use.

● 75% (vol/vol) ethanol solution made from 250 ml double distilled water and 750 ml ethanol.

Equipment

Equipment setup

Hardware

● Infrared spectroscopy system. For a list of available commercial instruments, please refer to **Table 2**.

Software

● Spectral acquisition software normally provided by the instrument manufacturer (**Table 2**).

● MATLAB r2009a (<http://www.mathworks.com>)

● Python (<http://www.python.org>)

● Metabo Analyst (<http://www.metaboanalyst.ca>). The latest version of Metabo Analyst (Metabo Analyst 4.0).

Table 2| Instruments, holders, and corresponding data acquisition software.

Manufacturer

Instrument

Software

Website

Bruker

IR Biotyper

OPUS

IR Biotyper Software

Bruker Optics

ALPHA FT-IR spectrometer

OPUS

<http://www.bruker.com/products.html>

ABB

MB3000 FT-IR laboratory analyzer

BOMEM GRAMS/32

<http://new.abb.com/products/measurement-products/analytical>

Thermo Fisher Scientific

Nicolet iS10 FT-IR spectrometer

OMNIC

<http://www.thermoscientific.com/en/products/fourier-transform-infrared-spectroscopy-ftir.html>

Agilent Technologies

Cary 600 series FT-IR spectrometers

Resolutions Pro

<http://crosslab.chem.agilent.com/>

PerkinElmer

Frontier spectrometers Spectrum Two IR spectrometers

Spectrum 10

<http://www.perkinelmer.com/>

Shimadzu

IRTracer-100

LabSolutions IR

<http://www.shimadzu.com/products/index.html>

Procedure

3. Procedure

3.1 Sample preparation ●TIMING ~1 h

1| Bacterial growth

Perform the steps in option A for bacterial cells cultured in solid medium or option B for bacterial cells in liquid medium. ●TIMING 12–24 h

(A) Bacterial cell culture on solid medium

(i) For cell culture of bacteria, such as *E. coli*, on TSA plates, incubate all bacteria for up to 16 h at an optimal temperature (35-37 °C).

▲ **CRITICAL STEP** If concentrations of salts and nutrients in the agar differ, the growth and phenotype of the target organism will also be affected.

! **CAUTION** Dried out agar plates should not be used!

(ii) A small amount of late exponential-phase cells (~10~60 µg dry wt) was carefully removed from confluent colonies on the agar surface using an incubating loop.

▲ **CRITICAL STEP** It is important that the bacteria used for comparative FT-IR spectrometric analysis have entered the same growth phase. Therefore, the user should compare FT-IR spectra of bacterial samples grown under similar conditions to ensure they are at the same growth phase.

! **CAUTION** Samples that present various colony morphologies or colors on the culture plate were rejected to prevent incorporation of contaminated or mixed isolates.

! **CAUTION** Handle all biological samples as a potential source of pathogens. Use appropriate protective attire (lab coats, safety glasses, latex gloves, and other conventional measures applied in the microbiology laboratory) and dispose of all biohazardous materials properly.

! **CAUTION** Changing the agar with the sample will ultimately result in different results.

! CAUTION Do not transfer any agar into the sample, which requires special care when harvesting. Loops should be closed and without sharp ends.

(iii) It is recommended that the bacterial specimen be completely removed medium and suspended in 0.9% NaCl solution. Chemical extraction of bacteria is not required in this protocol.

(optional) Submerge the bacterial cells in the ethanol in the suspension vial, and then remove the bacterial cells from the plastic inoculation loop by flipping. Some bacteria will stay in the loop unsuspended, but this may be compensated for by the mass on and around the loop.

▲ CRITICAL STEP If the bacteria in the loop is too sticky and will not come off, try take a larger amount. If this does not work, cut off the loop with the adherent bacterial cells. A metal cylinder (Bruker Part No. 1851760) can be used to remove the biomass from the loop.

(iv) After closing the suspension vial, vortex to form a homogeneous suspension. Do not leave the vial open during vortexing to avoid release of the solution.

(B) Cell culture in liquid medium ●TIMING 7 h

(i) All strains are cultivated in LB broth at 37 °C for 24 hours with sustained shaking at 170 rpm.

(ii) Harvest bacteria by centrifuging at 1500 g for 10 min at room temperature. Discard the supernatant.

(iii) Gently resuspend the bacterial pellet in 0.9% NaCl solution by pipetting. Centrifuge again at 1500 g for 10 min at room temperature and discard the supernatant. Repeat once until the supernatant is clear.

▲ CRITICAL STEP For strain typing and differentiation, it makes no sense to measure the bacterial growth in different or mixed cultures.

▲ CRITICAL STEP Variations in the incubation time (± 4 h) and growth in different non-selective culture media are not detrimental to reproducibility. However, it may change the typing or differentiation results. Minimize the variables and use fixed conditions: medium, temperature, incubation time, atmosphere, and humidity.

! CAUTION If concentrations of salts and nutrients in the agar differ, the growth and phenotype of the target organism will also be affected.

! CAUTION Prior to use in bacterial culture, all of the experimental materials and reagents should be autoclaved and handled under sterile conditions. Use sterile vessels or Petri dishes.

2| Sample prepared for FT-IR spectroscopy

Transfer the collected bacterial cells to appropriate FT-IR slides.

(A) Transmission mode

(i) The 50 μL of bacteria suspension was smeared on the surface of a CaF_2 slide.

(ii) Sample drying. The bacteria suspension was dried at 40°C for 30 min. The spectra of the samples on these slides can be collected without interference from coating.

! CAUTION Control the exposure temperature and exposure time. When the sample is exposed to a higher temperature or longer time, the bacterial film will swell, crack, and flake off.

(B) ATR mode

(i) Drop the bacterial solution onto the cell and let it form a liquid droplet. At least three spots of the same sample are needed.

! CAUTION Be careful to avoid introducing air bubbles.

(C) Diffuse reflectance mode

(i) Each isolate suspension is pipetted onto the silicon microtiter plate with 20 μL per spot.

(ii) Sample drying: the bacterial suspension was dried at 40°C for 30 min.

(iii) FT-IR analysis was initially performed using the plate without samples to provide a reference reading for each well.

▲ CRITICAL STEP Apply the sample as evenly as possible on the imprinted positions of the microtiter plate, i.e., the sample should cover each imprinted position completely and form a homogeneous layer.

▲ CRITICAL STEP Pipette technical replicates (at least quintuplicates) of each suspension on the target. The software is looking for reproducibility and the discriminatory power of the controls, and it cannot perform this with one spot of each.

! CAUTION To obtain optimal and reproducible results, the sample solution has to cover the imprinted position mark completely. Avoid partial coverage of the sample position.

! CAUTION To avoid scratches on the microtiter plate surface, always apply the sample solution using a pipette with a plastic tip.

? TROUBLESHOOTING

! CAUTION Handle all biological samples as a potential source of pathogens. Use appropriate protective attire (lab coats, safety glasses, latex gloves, and other conventional measures applied in the microbiology laboratory) and dispose of all biohazardous materials properly.

▲ CRITICAL STEP For good quality spectra, a single bacterial colony should be used. Particularly for the identification of bacterial subspecies, it is very important that the user only compares the spectra of

bacterial samples grown under similar conditions.

! CAUTION Be careful when handling the FT-IR slides. Hold the cell at its edges. Do not touch the cell surface, neither the front side nor the rear side. IR-spectroscopy is very sensitive to any contamination and will detect even fingerprints or dust.

! CAUTION Avoid dropping or bending the plate, as it may break. A broken silicon plate has hard and sharp edges like glass. Handle with care!

3| Equipment preparation ●TIMING 2 h (10 bacterial samples)

Different FT-IR instruments can be chosen. The options are transmission mode (option A), ATR mode (option B), or diffuse reflectance mode (option C).

(A) Transmission FT-IR spectroscopy ●TIMING 2 h (10 bacterial samples)

(i) Turn on air-conditioning and a liquid desiccant system in the room in which you are doing the experiment. Control the temperature at 25°C.☒

▲ **CRITICAL STEP** Only operate the FT-IR machine in a well-enclosed room; air movement must be minimized. Maintaining a dry and steady test environment is very important.

? TROUBLESHOOTING

(ii) Turn on the drying equipment for the instrument.☒

▲ **CRITICAL STEP** Dry air or nitrogen should displace water vapor absolutely.

(iii) Turn on the IR source.☒

? TROUBLESHOOTING

(iv) Turn on the FT-IR spectrometer.

▲ **CRITICAL STEP** The IR Biotyper should be permanently switched on.

▲ **CRITICAL STEP** Each time after switching on the instrument, you must wait for ~1 h after power-up for the IR source to stabilize.

? TROUBLESHOOTING

(v) Turn on the computer and monitor, and then start the FT-IR collection program.

(vi) Set the data path to store your data.

(B) ATR FT-IR spectroscopy ●TIMING 1 h (10 bacterial samples)

(i) Open the instrument-operating software.

(ii) Apply instrumental settings.

(iii) Set the data path to store your data.

(iv) Clean the ATR internal reflection element (diamond, zinc selenide, germanium, or silicon) with distilled water and dry it with tissue.

▲ CRITICAL STEP Make sure that the crystal is thoroughly cleaned and dried before background acquisition.

▲ CRITICAL STEP A prepared and dried sample plate can be stored for several days in a dry and dust-free place at room temperature without loss of sample quality.

(v) Place the slide in contact with the ATR internal reflection element.

▲ CRITICAL STEP Ensure that the ATR internal reflection element is completely covered by the sample and that the minimum sample thickness is 3~4 times the depth of penetration to ensure that there is no interference from the substrate.

(C) Diffuse reflectance mode

(i) To start the measurement, click Start Acquisition; the drawer will open and a dialog will request insertion of the microtiter plate.

(ii) Insert the frame with the microtiter plate in the IR Biotyper instrument. After the drawer at the right side of the IR Biotyper instrument opens, put the frame with the inserted microtiter plate in the holding fixture. Make sure that the frame corner labeled A1 is placed in the marked corner of the X/Y-stage (recess with red point). Otherwise, when you evaluate the measurement results afterwards, the sample position and spectrum will not correspond with each other.

(iii) Click Start on the graphical user interface to close the drawer and start the measurement.

! CAUTION During spectrum acquisition, the IR instruments can set a quality control to check the different spectral properties, e.g., absorbance intensity, SNR, and water vapor disturbance. The control sample can be the same standard samples that were inactivation bacteria. Spots with spectra pass this quality test. Failed spectra will not be taken into account for further calculations. (optional) Bruker Infrared Test Standards (Bruker Part No. 1851760) can be bought and used for quality control. two standard samples were usually used (IRTS1 and IRTS2).

! CAUTION The cell should be delivered in a clean, ready-to-use state and packed in a reclosable plastic bag. After use they can be cleaned. For the microtiter plate, if enough free spots are still available, it can be used for another measurement run. You can clean the microtiter plate by covering the surface with

deionized water and carefully rubbing the plate with a sponge or cleaning tissue. To remove any fatty residues, repeat the cleaning with 60% (v/v) isopropanol. To dry the microtiter plate, use clean pressurized air (oil-free) or a clean and dust-free cleaning cloth.

? TROUBLESHOOTING

(iv) Turn on the drying equipment for the instrument.

▲ **CRITICAL STEP** Dry air or nitrogen should displace water vapor absolutely.

(v) Turn on the IR source.

? TROUBLESHOOTING

(vi) Turn on the FT-IR spectrometer.

▲ **CRITICAL STEP** The IR Biotyper should be permanently switched on.

▲ **CRITICAL STEP** Each time after switching on the instrument, you must wait for ~1 h after power-up for the IR source to stabilize.

? TROUBLESHOOTING

(vii) Turn on the computer and monitor, and then start the FT-IR collection program.

(viii) Set the data path to store your data.

! **CAUTION** The cell should be delivered in a clean, ready-to-use state and packed in a reclosable plastic bag. After use they can be cleaned. For the microtiter plate, if enough free spots are still available, it can be used for another measurement run. You can clean the microtiter plate by covering the surface with deionized water and carefully rubbing the plate with a sponge or cleaning tissue. To remove any fatty residues, repeat the cleaning with 60% (v/v) isopropanol. To dry the microtiter plate, use clean pressurized air (oil-free) or a clean and dust-free cleaning cloth.

4| Collecting FT-IR spectra ●TIMING 1–3 h

Set the spectrum collection parameters based on the chosen FT-IR instrument. The options are transmission mode (option A), ATR mode (option B), Biotyper (option C), or A/R (option D).

(A) Transmission mode

Spectrometer resolution: 4 cm⁻¹; spectral range: 4,000 ~ 500 cm⁻¹; scan number: 64.

(B) ATR mode

FT-IR-ATR spectra acquired from 4,000–500 cm⁻¹ with a resolution of 4 cm⁻¹ and 32 scans.

(C) Diffuse reflectance mode

Spectra collected over the wavelength range of 4000 to 600 cm^{-1} at a rate of 20 scans per second with a resolution of 4 cm^{-1} . To improve the SNR, 128 spectra were co-added and averaged.

5| Collect a background spectrum without IR cell.

This spectrum is used to remove spectral signals that originate from air, moisture (water vapor), and coating materials on the reflecting mirrors along the IR radiation path from the spectra of the protein and buffer in order to subtract the background noise.

▲ **CRITICAL STEP** If the water vapor is not purged completely by dry air, there will be peaks in the regions of 1,500–1,200 cm^{-1} and 4,000–3,500 cm^{-1} . If these peaks appear, you should check the tubing and ensure no air leakage is occurring. You should also check the liquid desiccant system and drying equipment, and then purge for a longer period of time.☒

▲ **CRITICAL STEP** Air bubbles and/or empty areas will result in poor absorption spectra caused by fringing from channel spectra, dispersion of the IR beam by bubbles, and/or incorrect absorbance from lack of sample.

! **CAUTION** A background spectrum should be taken if atmospheric changes occur (e.g., if a door is suddenly opened).

?TROUBLESHOOTING

6| Collect a subtractive spectrum.

▲ **CRITICAL STEP** It is important to acquire the spectrum for the buffer before acquiring the spectrum for the sample.

7| Collect a spectrum for one type of bacteria.

▲ **CRITICAL STEP** The sample cell must be handled with care. Make sure the bacterial film does not flake off.

8| Save the raw data on the hard drive or other media.☒

▲ **CRITICAL STEP** Always save the data immediately to prevent loss if there is a power failure.

9| Clean the cell when finished.

Wash the sample cell with ethyl alcohol (~5 mL volume), and then with an excess of water (~10 mL volume). Finally, gently wipe the windows with lens cleaning paper to keep them clean.

! CAUTION The overriding principle of experimental disposal is that all waste should be decontaminated, autoclaved, or incinerated within the laboratory to guarantee "zero leaking" of infectious biohazards.

3.2 FT-IR spectrum pre-processing ●TIMING 1~2h

10| Calculate IR absorbance spectra.

Upon completion of all measurements, use the function in the software to compute the absorbance spectra.

?TROUBLESHOOTING

11| Subtract reference spectra.

Subtract the reference spectra from the bacterial spectra to remove the background signals (signals from water vapor, CO₂ gas, NaCl, etc.).

▲ CRITICAL STEP To obtain high-quality bacterial IR spectra, the background spectra must be carefully and adequately subtracted from the obtained sample spectrum. Any uncompensated absorption bands in the original spectra will be enhanced by the derivative analysis due to degradation of the SNR.

12| Perform baseline correction.

This is a pre-processing step to account and correct for noise and sloping baseline effects. Rubber band baseline correction is a convex polygonal line in which edges are "troughs" within the spectrum. Manual point baseline correction picks the wavenumber locations of the polygonal line and can subtract the overlapped area from the absorption spectra.

13| Savitsky-Golay differentiation

- (i) From the applications menu, select derivative function (first differentiation or second differentiation).
- (ii) Use a baseline-corrected seven-point Savitsky–Golay derivative function.
- (iii) Save the second-derivative spectrum.

▲CRITICAL STEP The derivative spectra show the details much better, allowing the differences between different strains to be visualized. In particular, second-derivative spectra (for which ease of visual inspection is often inverted around the frequency axis) are well suited for identifying component bands in a complex spectral region.

▲CRITICAL STEP The sharp and narrow absorption peaks from residual water vapor would be greatly enhanced by the second-derivative analysis and distort the protein amide I bands. Discard points below a threshold, and then fit the remaining points with a straight line.

14| Smoothing and data normalization. This step can be done using min-max normalization or vector normalization.

(Optional) Scale the variables: this could be done by standardization (normalization of variables to zero mean and unit s.d.) or by normalization to a 0–1 range.

▲ CRITICAL STEP Prior to conducting any kind of data analysis, it is important to assess the overall data quality and determine if there are any obvious outliers. Smooth and normalize the amide I band area. The application of smoothing is to remove the possible white noise, which results from using the second derivative function. As the spectra were collected on different instruments, even at the same protein concentration, the SNRs were quite different, especially at lower concentrations, and the application of smoothing and normalization were necessary.

! CAUTION As excessive smoothing can build up side lobes and periodic noise, which may be confused with true spectral features, the amount of smoothing should be kept to a minimum.

15| After second-derivative calculation.

Truncate the second-derivative spectrum to 900 and 1200 cm^{-1} sections by selecting the Utilities function from the applications menu and then selecting the ZAP function.

16| Baseline.

A baseline can be easily obtained by connecting the two most positive points within the selected region.

17| Save the final results.

Save the results in a comma-separated value or CSV file (.csv) or a text file (.txt).

?TROUBLESHOOTING

3.3 Bacterial typing, classification, and identification ●TIMING 2–5 h

18| Create a new folder in your file system.

This will contain the preprocessed files.

19| Assemble the data set.

(1) Execute the following steps in Excel:

(i) Open the text file created in the previous step. This will appear as a table in Excel. The last column of this table is filled with zeros, which means all spectra are initially assigned the same class.

(ii) Next, enter the class labels in this column (take care not to miss any boxes). Use a different number for each class, starting from 0 or 1. Use numbers only.

(iii) Save the spreadsheet as a 'Text (tab-delimited)' file. Choose a data analysis procedure appropriate to your analysis goal. Here, we cover clustering and identification. Options A and B (unsupervised classification) are suitable for bacterial typing at the strain and substrain level. Options C and D (supervised classification) are suitable for bacterial identification.

▲ **CRITICAL STEP** Each list file was saved in a .csv or a text file (.txt) format containing several columns, with the first column corresponding to peak positions (cm^{-1}) and the remaining columns corresponding to peak intensities. The first line of each peak list file is reserved for column labels. The replicates should share the same class label. At least five replicates are required in each class.

! **CAUTION** The data (except class labels) must not contain non-numeric values. Missing values or features should be given zeros.

(2) (Optional) Software (IR-data-assemble) has been programmed by our lab and is provided freely.

Execute IR-data-assemble:

(i) Go to 'Macro' and select 'Run macro'.

(ii) Find the file 'IR-data-assemble' and then click 'Open'.

(iii) In the 'Path' field, enter the complete path for your raw files.

(iv) In the 'Path destination' field, enter the complete path for the 'dat' folder created. Click 'OK'. This will convert all raw files into preprocessed 'Pirouette .dat' files.

(v) click 'Go!', then specify a location and file name to store your merged data set, and click 'Save'.

▲ **CRITICAL STEP** At least five replicates are required for each group.

20| Multivariate analysis

Analysis of the raw spectra produced by the procedure shown above can be performed with bioinformatics tools (**Box 1**). We use hierarchical clustering as a simple and powerful tool (**Box 1**).

We describe two options for achieving this: using a home-made tool (A) or MetaboAnalyst (B).

(A) Home-made tool: The "home-made tool" is used to establish a library that can be used to identify individual sample bacteria at different levels. We provide this home-made tool as supporting information. First, several individual spectra are averaged to generate the reference spectrum of the strain to establish the library. The Pearson correlation coefficient was used in the comparison of spectral profiles¹¹⁷, and the score derived from the distance is always used to indicate the similarity between the unknown and known

bacteria. The matching scores of these spectra can be divided into ranges: highly probable species identification, probable species identification, reliable genus identification and unreliable identification according to the experimental results. Finally, a confusion matrix can be used to evaluate the typing performance of the library ¹¹⁸.

(B) MetaboAnalyst: MetaboAnalyst uses a step-wise data processing pipeline that guides users through all of the major data processing steps, beginning with data-type selection, formatting, “cleansing,” and normalization. It then guides users into using (or exploring) a number of data analysis options ranging from basic univariate (i.e., *t* tests, analysis of variance [ANOVA]) and multivariate (i.e., PCA) methods for advanced machine learning approaches, such as random forest and support vector machine (SVM) classification.

(i) You can enter <https://www.metaboanalyst.ca/> in the address bar and the browser navigates to a corresponding page. Click on the Statistical Analysis button.

(ii) Upload the files, choose Data Type: spectral bins; Format: samples in columns (unpaired); Data file: .CSV.

(iii) Data integrity checking is done to ensure that the data meets the basic requirements for meaningful downstream analysis.

▲ **CRITICAL STEP** MetaboAnalyst (<http://www.metaboanalyst.ca>) is a comprehensive Web application for metabolomic data analysis and interpretation. It provides a variety of data processing and normalization procedures and supports a number of data analysis and data visualization tasks using a range of univariate and multivariate methods, such as PCA, heatmap clustering, and machine learning methods. The software was carefully designed to enable researchers with little statistical or computational background to perform complex data analysis procedures commonly used in metabolomic studies.

(iv) Low-quality data filtering. Select the default “Interquantile range (IQR)” and click the Process button.

(v) Select “Normalization by sum” for sample normalization. After normalization, click the Proceed button.

▲ **CRITICAL STEP** Differences in thickness or concentration can sometimes be the most prominent source of spectral variation between samples, often masking the biochemical differences. To minimize these effects, the spectra in the dataset are scaled to match a specific criterion. Normalization to a particular peak can be applied when such a peak is consistently present in all spectra in the dataset.

(vi) Click the Submit button to proceed to the Data Analysis Exploration page. Five major analysis categories are presented: univariate analysis, multivariate analysis, significant feature identification, cluster analysis, and classification & feature selection. Listed under each of these five categories are many other subcategories containing individual methods. The most particular approaches are useful for

clustering, such as PCA and dendrogram, which are listed under the Multivariate Analysis category, and heatmaps, which is listed under the Cluster Analysis category. The user can choose the preferred mathematical method.

(vii) Click Download to generate a .pdf report (Analysis_Report.pdf).

You have the choice between two views: distance matrix and dendrogram. The distance matrix shows all distances of the spectra to each other sorted according to the chosen HCA, whereas the dendrogram gives a more condensed, tree-like overview of the spectral relations. In general, we recommend using the dendrogram view first.

Box 1 COMPUTATIONAL TOOLS FOR BACTERIAL FT-IR PATTERN ANALYSIS

For bacterial data analysis, we use the IR software (ABB). Using this software, we pre-process spectra by applying default parameters, whereby we apply FT-IR spectra compressing, the Savitsky–Golay smoothing and second-derivative method, and rubber band baseline correction method. For data normalization, we use ‘maximum norm’.

For species and subspecies typing, we hierarchically cluster the FT-IR spectra of bacterial species and subspecies. Each FT-IR spectrum in a dataset is compared to the other spectra. On the basis of the distance values, we produce a dendrogram using the appropriate function of the statistical toolbox of Biotyper. Taking a list of FT-IR signals and their intensities into consideration, dendrograms are produced by similar scoring of a set of FT-IR spectra. A correlation function is applied to calculate distance values. In general, species with distance levels <0.15 are reliably classified. However, this value can be changed by the user.

Clearly, the FT-IR spectra analysis for bacterial samples is not restricted to the described software package that is applied in our laboratory. Instead of hierarchical clustering, model construction can also be applied. The user may also test alternative classification procedures and software from other companies. Statistical tools, such as Matlab, are described by Al-Holy et al.⁹⁸. Moreover, specific software for bacterial identification and analysis has been developed by a number of academic research groups.

Box 2 COMPUTATIONAL TOOLS FOR BACTERIAL IDENTIFICATION

Computational tools for bacterial identification include five main steps: spectral preprocessing, feature-extraction, library construction, distance measuring, and score and prediction.

Spectral preprocessing: The raw spectral profile data always present complex features. In addition to the true signal, two main disturbances can be revealed in the spectra: baseline drift, and noise that comes from electric and/or chemical interference. Spectra obtained from different conditions, such as different instrument parameters or different amounts of sample, can have different spectral intensities, which should be deleted for an equal weighted analysis. The valuable feature is usually expressed as a peak value or derivative data regarding the raw data. Consequently, the raw profile data need to be

preprocessed to reduce redundant and noise information and obtain valuable signals. Baseline correction, smoothing, normalization, first or second derivative, and peak picking are commonly used in this step before mathematization¹⁰⁹.

Feature-extraction: As mentioned above, feature construction methods, such as PCA and PLS, or feature selection methods, such as genetic algorithm, multivariate curve resolution, and successive projection algorithm, can be employed in this step to extract the most informative variables that will be used later in the resultant model. In this case, the amount of data is greatly reduced and the model execution more efficient.

Library construction: Library construction can also be described as the creation of reference spectra for known bacteria of interest. In general, a number of spectra data are collected, with the aim of accurately describing the known bacteria, and the reference spectrum is created using the features extracted from this spectra by an average or other linear combination algorithm¹¹⁹. The precise feature-extraction, meaning qualifies features and just the right amount of features selected in this step, plays an important role in increasing the discriminatory power of library-based approaches, and the lower experimental standardization requirements (e.g., bacterial growth and culture media) of the features allow more stable and reliable identification¹⁰. According to the importance of these features in distinguishing one strain from another or to categorize one strain within a group, different weights of these features can be assigned and recorded in the database to achieve more accurate identification.

Distance measure: The similarity between the unknown and known bacteria can be calculated through a distance measure algorithm, most commonly the Pearson correlation coefficient^{17,109,110}. In this algorithm, values with the same feature make a pair, and all values presenting different features of the unknown and reference spectra are compared in these pairs. A higher score means a shorter distance and will be obtained when the values in each pair are closer. Bacterial identification often requires analysis of the entire spectrum rather than one or a few biomarker features, and the different weights assigned to the spectrum features allows a more accurate comparison. In addition, strain identification is often facilitated with software employing advanced statistical analyses¹¹⁰.

Score and prediction: A score derived from the distance is always used to indicate the similarity between the unknown and known bacteria. In general, a higher score indicates greater similarity, and the highest matching scores and related spectra in the library will be selected as the final results among all of the comparisons between the unknown spectrum and the spectra in the library. Next, a prediction is made according to the score, with a higher score resulting in a more precise result. The highest matching scores of these comparisons with the BiotyperTM database were divided into ranges: highly probable species identification (>2.3), probable species identification (2.0–2.3), reliable genus identification (1.7–2.0), and unreliable identification (<1.7)¹²⁰.

Troubleshooting

4. Troubleshooting

Troubleshooting advice can be found in **Table 6**.

Table 3| Troubleshooting

Step	Problem	Solution
2	Insufficient sample purity	Improve the complex isolation/purification procedure
5	There is a very low-IR absorbance signal or no signal	Check the quantity of bacteria. Ensure there are no bubbles or empty cell areas when collecting the sample spectrum.
4	Bacterial specimens not dry enough	Although proteins and peptides are the majority of bacteria, the FT-IR absorbance of this region is disturbed by H ₂ O. Reduction and stable control of H ₂ O absorbance are key procedures for protecting the polysaccharide and nucleic acid region from being disturbed by H ₂ O.
4	Unwanted water vapor band in the spectra	Check if the desiccant cartridge in the microplate compartment or the spectrometer compartment is exhausted and needs to be replaced or regenerated.
6		

Different reference and experimental spectra of the same bacterial subspecies

Grow bacteria in the same media. Verify that bacteria are collected during the same growth (log or stationary) phase. Eliminate noticeable media-dependent peaks using appropriate data analysis software.

6

Low-quality FT-IR spectra

Low number of bacterial cells or bacterial samples are still in the lag phase or entered the death phase. Culture cells again, resuspend bacteria in less volume for chemical treatment.

6

Great variability between the four independently acquired spectra

Contamination of different clones cultured in liquid culture or growing on the solid media.

The presence of culture medium adhering to the bacterial colony.

One or the same strain of bacteria cultured in different conditions.

20

What is the cut-off for dendrograms?

The user has to decide on the limit for related vs. unrelated. The software can help by taking into account the variability of the replicate measurements. This is the same problem in ALL typing methods. How many bases are allowed to be different in an MLST?

20

Samples are processed with low reproducibility.

Check if the sample is too thick or too thin in spots. Bubbles in the sample will also increase the variance within the technical replicates.

20

Two identical strains appear to be different.

One strain could have been freshly transferred from blood to lysogeny broth (LB) agar, whereas the other was passaged on LB agar throughout.

Check the bacteria culture conditions (medium; time; temperature).

Time Taken

5. Timing

Sample preparation (cell culture on solid medium or in liquid medium, and CaF_2 slide preparation): 1~2 h. Steps 1–5 can be done in advance.

Acquisition of spectrum (parameter setting): 1~3 h. Steps 6–11 (machine setup): 3–4 h (the amount of time depends on air humidity and drying equipment in the laboratory).

FT-IR spectrum pre-processing 1~2 h

Bacterial typing and classification 2–5 h

Anticipated Results

6. Anticipated results

6.1 Typing of *E. coli*

Hierarchical clustering based on spectral information contained in the spectral range of $1,200\text{--}900\text{ cm}^{-1}$, the polysaccharide region, was used for typing. The results in 2D score plot was shown in **Fig. 6** and the dendrogram in **Fig. 7**. A clear discrimination of different species could be observed in **Fig. 6**, and 12 distinct clusters were produced in **Fig. 7** for the discrimination of 12 strains of *E. coli*.

6.2 Identification of *E. coli* and *Shigella*

A FT-IR spectra library can be employed to identify individual bacteria at different levels. Usually, several individual bacterial spectra are averaged to generate the reference spectrum of the strain to establish the library, and the Pearson correlation coefficient was used in spectral comparison. The score derived from the bacteria spectral distance is always used to indicate the similarity between the unknown and known bacteria. After comparison, the reference strain with the highest correlation score for the comparison was regarded to be the result of the identification. The library-based identification of 207 FT-IR spectra of *E. coli* and *Shigella* strains was shown in **Fig. 8**. The identification accuracy was 95.2% for distinguishing between *E. coli* and *Shigella* at the genus level. For species and strain level, the identification accuracy are 91.8% and 81.2% for all selected strains, respectively.

References

References

1. Sabat, A.J. et al. Overview of molecular typing methods for outbreak detection and epidemiological surveillance. *Eurosurveillance* **18**, 17-30 (2013).

2. Jin, N., Zhang, D. & Martin, F.L. Fingerprinting microbiomes towards screening for microbial antibiotic resistance. *Integr. Biol. (Camb)* **9**, 406-417 (2017).
3. Sintchenko, V., Iredell, J.R. & Gilbert, G.L. Pathogen profiling for disease management and surveillance. *Nat. Rev. Microbiol.* **5**, 464-470 (2007).
4. Garon, D., Kaddoumi, A.E., Carayon, A. & Amiel, C. FT-IR Spectroscopy for Rapid Differentiation of *Aspergillus flavus*, *Aspergillus fumigatus*, *Aspergillus parasiticus* and Characterization of Aflatoxigenic Isolates Collected from Agricultural Environments. *Mycopathologia* **170**, 131-42 (2010).
5. Alvarez-Ordóñez, A. & Prieto, M. *Fourier Transform Infrared Spectroscopy in Food Microbiology*, (2012).
6. Shapaval, V. et al. FTIR spectroscopic characterization of differently cultivated food related yeasts. *Analyst* **138**, 4129-4138 (2013).
7. Alvarez-Ordóñez, A., Mouwen, D.J.M., López, M. & Prieto, M. Fourier transform infrared spectroscopy as a tool to characterize molecular composition and stress response in foodborne pathogenic bacteria. *Journal of Microbiological Methods* **84**, 369-378 (2011).
8. Ji, S.Y., Dong, H.J., Hassan, M. & Ilev, I.K. Signature Infrared Bacteria Spectra Analyzed by an Advanced Integrative Computational Approach Developed for Identifying Bacteria Similarity. *IEEE J. Sel. Top. Quant.* 7200908 (2018).
9. Muhamadali, H. et al. Rapid, accurate, and comparative differentiation of clinically and industrially relevant microorganisms via multiple vibrational spectroscopic fingerprinting. *Analyst* **141**, 5127-5136 (2016).
10. Quintelas, C., Ferreira, E.C., Lopes, J.A. & Sousa, C. An Overview of the Evolution of Infrared Spectroscopy Applied to Bacterial Typing. *Biotechnol. J.* **13**, 1700449 (2018).
11. Forrester, J.B., Valentine, N.B., Su, Y.F. & Johnson, T.J. Chemometric analysis of multiple species of *Bacillus* bacterial endospores using infrared spectroscopy: Discrimination to the strain level. *Analytica Chimica. Acta.* **651**, 24-30 (2009).
12. Kirschner, C., Maquelin, K., Pina, P., Thi, N.A.N. & Naumann, D. Classification and Identification of Enterococci: a Comparative Phenotypic, Genotypic, and Vibrational Spectroscopic Study. *J.Clin.Microbiol.* **39**, 1763 (2001).
13. Butler, H.J. et al. Using Raman spectroscopy to characterize biological materials. *Nat. Protoc.* **11**, 664-687 (2016).
14. Almasoud, N. et al. Rapid discrimination of *Enterococcus faecium* strains using multiple physicochemical analytical techniques. *Anal. Meth.* **8**, (2016).

15. Pasternak, C.A. *Microbial cell walls and membranes*, (1980).
16. Amiel, C. et al. Identification and Classification of bacteria by fourier transform infrared spectroscopy (FTIR), *Spectroscopy of Biological Molecules: Modern Trends*. Springer, (1997).
17. Naumann, D. Infrared Spectroscopy in Microbiology. *Encyclopedia of Analytical Chemistry*, John Wiley & Sons, Ltd. (2006).
18. Burgula, Y. et al. Review of mid-infrared fourier transform-infrared spectroscopy application for bacterial detection. *J. Rapid. Meth. Aut. Mic.* **15**, 146-175 (2010).
19. Baker, M.J. et al. Using Fourier transform IR spectroscopy to analyze biological materials. *Nat. Protoc.* **9**, 1771-1791 (2014).
20. De, B.S., Speeckaert, M.M. & Delanghe, J.R. Applications of mid-infrared spectroscopy in the clinical laboratory setting. *Crit. Rev. Clin. Lab. Sci.* **55**, 1-20 (2018).
21. Freitas, A.R., Rodrigues, C. & Peixe, L. Fourier transform infrared spectroscopy: unlocking fundamentals and prospects for bacterial strain typing. *Eur. J. Clin. Microbiol.* **38**, 427-448 (2019).
22. Shi, H. et al. The strategy for correcting interference from water in Fourier transform infrared spectrum based bacterial typing. *Talanta* **208**, 120347 (2020).
23. Zarnowiec, P., Mizera, A., Chrapek, M., Urbaniak, M. & Kaca, W. Chemometric analysis of attenuated total reflectance infrared spectra of *Proteus mirabilis* strains with defined structures of LPS. *Innate. Immun.* **22**, 325-335 (2016).
24. Griffiths, P. & Haseth, J. *Fourier Transform Infrared Spectrometry*, Second Edition. (2006).
25. Davis, R. & Mauer, L.J. Fourier Transform Infrared (FT-IR) Spectroscopy: A Rapid Tool for Detection and Analysis of Foodborne Pathogenic Bacteria. *Eur. J. Clin. Microbiol.* **2**, 1582-1594 (2010).
26. Trevisan, J., Angelov, P.P., Carmichael, P.L., Scott, A.D. & Martin, F.L. Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives. *Analyst* **137**, 3202-3215 (2012).
27. Wenning, M. & Scherer, S. Identification of microorganisms by FTIR spectroscopy: perspectives and limitations of the method. *Appl. Microbiol.Biot.* **97**, 7111-7120 (2013).
28. Kaca, W.A., Czerwonka, G., Lechowicz, L. & Zarnowiec, P. Fourier Transform Infrared Spectroscopy (FTIR) as a Tool for the Identification and Differentiation of Pathogenic Bacteria. *Curr Med.Chem.* **22**, (2015).
29. Preisner, O., Lopes, J.o.A., Guiomar, R., Machado, J. & Menezes, J.C. Fourier transform infrared (FT-IR) spectroscopy in bacteriology: towards a reference method for bacteria discrimination. *Anal. Bioanal.*

Chem. **387**, 1739 (2007).

30. Goodacre, R., Timmins, E.M., Rooney, P.J., Rowland, J.J. & Kell, D.B. Rapid identification of *Streptococcus* and *Enterococcus* species using diffuse reflectance-absorbance Fourier transform infrared spectroscopy and artificial neural networks. *Fems Microbiol. Lett.* **140**, 233-239 (2010).
31. Maquelin, K. et al. Vibrational Spectroscopic Studies of Microorganisms. *John Wiley & Sons* (2002).
32. Carlos, C., Maretto, D.A., Poppi, R.J., Sato, M.I.Z. & Ottoboni, L.M.M. Fourier transform infrared microspectroscopy as a bacterial source tracking tool to discriminate fecal *E. coli* strains. *Microchem. J.* **99**, 15-19 (2015).
33. Yang, H., Yang, S., Kong, J., Dong, A. & Yang, S. Obtaining information about protein secondary structures in aqueous solution using Fourier transform IR spectroscopy. *Nat. Protoc.* **10**, 382-396 (2015).
34. Kong, J. & Yu, S. Fourier Transform Infrared Spectroscopic Analysis of Protein Secondary Structures. *Acta. Bioch. Bioph. Sin* **39**, 549 (2007).
35. Movasaghi, Z., Rehman, S. & Rehman, I.U. Fourier transform infrared (FTIR) spectroscopy of biological tissues. *Appl. Spectrosc. Rev.* **43**, 134-179 (2008).
36. Maquelin, K. et al. Identification of medically relevant microorganisms by vibrational spectroscopy. *J. Microbiol Meth.* **51**, 255-271 (2002).
37. Yu, C. & Irudayaraj, J. Spectroscopic characterization of microorganisms by Fourier transform infrared microspectroscopy. *Biopolymers* **77**, 368-377 (2010).
38. Grunert, T. et al. Analysis of *Staphylococcus aureus* wall teichoic acid glycoepitopes by Fourier Transform Infrared Spectroscopy provides novel insights into the staphylococcal glycode. *Sci. Rep.* **8**, 1889 (2018).
39. Bosch, A. et al. Characterization of *Bordetella pertussis* growing as biofilm by chemical analysis and FT-IR spectroscopy. *Appl. Microbiol. Biot.* **71**, 736 (2006).
40. Socrates, G. Infrared and Raman Characteristic Group Frequencies: Tables and Charts, 3rd Edition. *Proteomics* (2004).
41. Movasaghi, Z., Rehman, S. & ur Rehman, D.I. Fourier Transform Infrared (FTIR) Spectroscopy of Biological Tissues. *Appl. Spectrosc. Rev.* **43**, 134-179 (2008).
42. Randall, H.M., Smith, D.W., Colm, A.C. & Nungester, W.J. Correlation of biologic properties of strains of *Mycobacterium* with infra-red spectrums. I. Reproducibility of extracts of *M. tuberculosis* as determined by infra-red spectroscopy. *American Review of Tuberculosis* **63**, 372 (1951).

43. Levine, S., Stevenson, H.J.R., Chambers, L.A. & Kenner, B.A. Infrared spectrophotometry of enteric bacteria. *J. Bacteriol.* **65**, 10-15 (1953).
44. Bordner, R.H. et al. Bacterial identification by infrared spectrophotometry. *J. Bacteriol.* **72**, 593 (1956).
45. Naumann, D., Helm, D. & Labischinski, H. Microbiological characterizations by FT-IR spectroscopy. *Nature* **351**, 81-2 (1991).
46. Helm, D., Labischinski, H. & Naumann, D. Elaboration of a procedure for identification of bacteria using Fourier-Transform IR spectral libraries: A stepwise correlation approach. *Journal of Microbiological Methods* **14**, 127-142 (1991).
47. Classification and identification of bacteria by fourier transform ir spectroscopy. *J Microbiol. Meth.* **137**, 69-80 (1991).
48. Nelson, W.H. *Modern Techniques for Rapid Microbiological Analysis*, Wiley-VCH (1991).
49. Naumann, D., Keller, S., Helm, D., Schultz, C. & Schrader, B. FT-IR spectroscopy and FT-Raman spectroscopy are powerful analytical tools for the non-invasive characterization of intact microbial cells. *J. Mol. Struct.* **347**, 399-405 (1995).
50. Novais, Â., Freitas, A.R., Rodrigues, C. & Peixe, L. Fourier transform infrared spectroscopy: unlocking fundamentals and prospects for bacterial strain typing. *Eur. J. Clin. Microbiol.* **38**, 427-448 (2019).
51. Finlayson, D., Rinaldi, C. & Baker, M. Is Infrared Spectroscopy Ready for the Clinic? *Analy.Chem.* **91**(2019).
52. Stuart, B. *FTIR of Biomolecules*, Wiley-VCH Verlag GmbH & Co. KGaA. (2006).
53. Bury, D. et al. Spectral classification for diagnosis involving numerous pathologies in a complex clinical setting: A neuro-oncology example. *Spectrochim. Acta A.* **206**, 89-96 (2019)
54. Byrne, H.J., Knief, P., Keating, M.E. & Bonnier, F. Spectral pre and post processing for infrared and Raman spectroscopy of biological tissues and cells. *Chem. Soc. Rev.* **45**, 1865-1878 (2016).
55. Bosch, A. et al. Fourier transform infrared spectroscopy for rapid identification of nonfermenting gram-negative bacteria isolated from sputum samples from cystic fibrosis patients. *J. Clin. Microbiol.* **46**, 2535-2546 (2008).
56. Salman, A. et al. Detection of antibiotic resistant Escherichia Coli bacteria using infrared microscopy and advanced multivariate analysis. *Analyst* **142**, 2136-2144 (2017).

57. Gram, L. et al. Food spoilage—interactions between food spoilage bacteria. *Int. J. Food. Microbiol.* **78**, 79-97 (2002).
58. Ahmad, S. et al. Identification of fungal phytopathogens using Fourier transform infrared-attenuated total reflection spectroscopy and advanced statistical methods. *J. Biomed. Opt.* **17**, 017002 (2012).
59. Mantsch, H. & Chapman, D. *Infrared Spectroscopy of Biomolecules*. Wiley–Liss, (1996).
60. Willey, B.M. et al. Practical approach to the identification of clinically relevant Enterococcus species. *Diagn. Micro. Infect. Dis.* **34**, 165-171 (1999).
61. Joensen, K.G., Tetzschner, A.M.M., Iguchi, A., Aarestrup, F.M. & Scheutz, F. Rapid and Easy In Silico Serotyping of Escherichia coli Isolates by Use of Whole-Genome Sequencing Data. *J. Clin. Microbiol.* **53**, 2410-2426 (2015).
62. Jauneikaite, E. et al. Current methods for capsular typing of Streptococcus pneumoniae. *J. Microbiol. Meth.* **113**, 41-49 (2015).
63. Fratamico, P.M. et al. Advances in Molecular Serotyping and Subtyping of Escherichia coli. *Front. Microbiol.* **7**(2016).
64. Yoshida, C.E. et al. The Salmonella In Silico Typing Resource (SISTR): An Open Web-Accessible Tool for Rapidly Typing and Subtyping Draft Salmonella Genome Assemblies. *Plos One* **11**(2016).
65. Schwartz, D.C. & Cantor, C.R. Separation of Yeast Chromosome-Sized Dnas by Pulsed Field Gradient Gel-Electrophoresis. *Cell* **37**, 67-75 (1984).
66. Freiwald, A. & Sauer, S. Phylogenetic classification and identification of bacteria by mass spectrometry. *Nat. Protoc.* **4**, 732-742 (2009).
67. Dixon, P. et al. A systematic review of matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry compared to routine microbiological methods for the time taken to identify microbial organisms from positive blood cultures. *Eur. J. Clin. Microbiol. Infect. Dis.* **34**, 863-876 (2015).
68. Deak, E. et al. Comparison of the Vitek MS and Bruker Microflex LT MALDI-TOF MS platforms for routine identification of commonly isolated bacteria and yeast in the clinical microbiology laboratory. *Diagn. Micro. Infect. Dis.* **81**, 27-33 (2015).
69. Sauget, M., Valot, B., Bertrand, X. & Hocquet, D. Can MALDI-TOF Mass Spectrometry Reasonably Type Bacteria? *Trends Microbiol.* **25**, 447-455 (2017).
70. Fernández-Álvarez, C., Torres-Corral, Y. & Santos, Y.J.J.o.p. Use of ribosomal proteins as biomarkers for identification of Flavobacterium psychrophilum by MALDI-TOF mass spectrometry. *J.*

Proteomics. **170**, 59-69 (2018).

71. Teramoto, K. et al. Phylogenetic classification of *Pseudomonas putida* strains by MALDI-MS using ribosomal subunit proteins as biomarkers. *Anal. Chem.* **79**, 8712-8719 (2007).
72. Sato, H. et al. Characterization of the *Lactobacillus casei* group based on the profiling of ribosomal proteins coded in S10-spc-alpha operons as observed by MALDI-TOF MS. *Syst. Appl. Microbiol.* **35**, 447-454 (2012).
73. Fernández-No, I., D'Yaz-Bao, M., Cepeda, A., Barros-Velázquez, J. & Calo-Mata, P.J.F.m. Characterisation and profiling of *Bacillus subtilis*, *Bacillus cereus* and *Bacillus licheniformis* by MALDI-TOF mass fingerprinting. *Food. Microbiol.* **33**, 235-242 (2013).
74. Quiles-Melero, I., Garcia-Rodriguez, J., Gómez-López, A., Mingorance, J.J.E.j.o.c.m. & diseases, i. Evaluation of matrix-assisted laser desorption/ionisation time-of-flight (MALDI-TOF) mass spectrometry for identification of *Candida parapsilosis*, *C. orthopsilosis* and *C. metapsilosis*. *Eur. J. Clin Microbiol. Infect. Dis.* **31**, 67-71 (2012).
75. Biswas, S. & Rolain, J.-M.J.J.o.m.m. Use of MALDI-TOF mass spectrometry for identification of bacteria that are difficult to culture. *J. Microbiol. Meth.* **92**, 14-24 (2013).
76. Stübiger, G. et al. Characterization of yeasts and filamentous fungi using MALDI lipid phenotyping. *J. Microbiol. Meth.* **130**, 27-37 (2016).
77. He, Y., Li, H., Lu, X., Stratton, C.W. & Tang, Y.-W.J.J.o.c.m. Mass spectrometry biotyper system identifies enteric bacterial pathogens directly from colonies grown on selective stool culture media. *J. Clin. Microbiol.* **48**, 3888-3892 (2010).
78. Martiny, D. et al. Comparison of the Microflex LT and Vitek MS systems for routine identification of bacteria by matrix-assisted laser desorption ionization–time of flight mass spectrometry. *J. Clin. Microbiol.* **50**, 1313-1325 (2012).
79. Khot, P.D. & Fisher, M.A.J.J.o.c.m. Novel approach for differentiating *Shigella* species and *Escherichia coli* by matrix-assisted laser desorption ionization–time of flight mass spectrometry. *J. Clin. Microbiol.* **51**, 3711-3716 (2013).
80. Paauw, A. et al. Rapid and reliable discrimination between *Shigella* species and *Escherichia coli* using MALDI-TOF mass spectrometry. *Int. J. Med. Microbiol.* **305**, 446-452 (2015).
81. Romanolo, K.F., Gorski, L., Wang, S. & Lauzon, C.R. Rapid Identification and Classification of *Listeria* spp. and Serotype Assignment of *Listeria monocytogenes* Using Fourier Transform-Infrared Spectroscopy and Artificial Neural Network Analysis. *PLoS One* **10**, e0143425 (2015)..

82. Lasch, P. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. *Chemometr. Intell. Lab.* **117**, 100-114 (2012).
83. Siqueira, L.F.S. & Lima, K.M.G. A decade (2004 – 2014) of FTIR prostate cancer spectroscopy studies: An overview of recent advancements. *Trac. Trends. Anal. Chem.* **82**, 208-221 (2016).
84. Martens, H., Nielsen, J.P. & Engelsen, S.B. Light Scattering and Light Absorbance Separated by Extended Multiplicative Signal Correction. Application to Near-Infrared Transmission Analysis of Powder Mixtures. *Anal. Chem.* **75**, 394-404 (2003).
85. Mazet, V., Carteret, C., Brie, D., Idier, J.m. & Humbert, B. Background removal from spectra by designing and minimising a non-quadratic cost function. *Chemometr. Intell. Lab.* **76**, 121-133 (2005).
86. Lieber, C.A. & Mahadevanjansen, A. Automated method for subtraction of fluorescence from biological Raman spectra. *Applied Spectroscopy* **57**, 1363-1367 (2003).
87. Baldauf, N.A., Rodriguezromo, L.A., Yousef, A.E. & Rodriguezsaona, L.E. Differentiation of Selected Salmonella enterica Serovars by Fourier Transform Mid-Infrared Spectroscopy. *Appli. Spectrosc.* **60**, 592-598 (2006).
88. Hamel, L. & Brown, C.W. Bayesian Probability Approach to Feature Significance for Infrared Spectra of Bacteria. *Appli. Spectrosc.* **66**, 48-59 (2012).
89. Sharaha, U. et al. Using Infrared Spectroscopy and Multivariate Analysis to Detect Antibiotics' Resistant Escherichia coli Bacteria. *Anal. Chem.* **89**, 8782-8790 (2017).
90. Comparato Filho, O.O., Morais, F.V., Bhattacharjee, T., Castilho, M.L. & Raniero, L. Rapid identification of Paracoccidioides lutzii and P. Brasiliensis using Fourier Transform Infrared spectroscopy. *J. Mol. Struc.* **1177**, 152-159 (2019).
91. Lin, M., Al-Holy, M., Al-Qadiri, H., Kang, D.H. & Rasco, B.A. Discrimination of Intact and Injured Listeria monocytogenes by Fourier Transform Infrared Spectroscopy and Principal Component Analysis. *J. Agr. Food. Chem.* **52**, 5769 (2004).
92. Rebuffo-Scheer, C.A., Schmitt, J. & Scherer, S. Differentiation of Listeria monocytogenes Serovars by Using Artificial Neural Network Analysis of Fourier-Transformed Infrared Spectra. *Appl. Environ. Microbiol.* **73**, 1036-1040 (2007).
93. Rebuffo-Scheer, C.A., Dietrich, J., Wenning, M. & Scherer, S. Identification of five Listeria species based on infrared spectra (FTIR) using macrosamples is superior to a microsample approach. *Anal. Bioanal. Chem.* **390**, 1629-1635 (2008).
94. Al-Qadiri, H.M., Al-Alami, N.I., Lin, M., Al-Holy, M. & Rasco, B.A. Studying of the bacterial growth phases using fourier transform infrared spectroscopy and multivariate analysis. *J. Rapid Meth. Aut. Mic.*

16, 73-89 (2010).

95. Al-Qadiri, H.M., Lin, M., Al-Holy, M.A., Cavinato, A.G. & Rasco, B.A. Studying of the bacterial growth phases using fourier transform infrared spectroscopy and multivariate analysis. *J. Rapid Meth. Aut. Mic.* **16**, 73-89 (2010).
96. Smith, B.C. Fundamentals of fourier transform infrared spectroscopy, second edition. *Crc Press* (2011).
97. Grunert, T. et al. Rapid and Reliable Identification of Staphylococcus aureus Capsular Serotypes by Means of Artificial Neural Network-Assisted Fourier Transform Infrared Spectroscopy. *J.Clin. Microbiol.* **51**, 2261-2266 (2013).
98. Al-Holy, M.A., Lin, M., Al-Qadiri, H., Cavinato, A.G. & Rasco, B.A. Classification of foodborne pathogens by fourier transform infrared spectroscopy and pattern recognition techniques. *J. Rapid. Meth. Aut. Microbiol.* **14**, 189-200 (2010).
99. Lorin-Latxague, C. & Melin, A.-M. Radical induced damage of Micrococcus luteus bacteria monitored using FT-IR spectroscopy. *Spectroscopy* **19**(2005).
100. Al-Qadiri, H.M., Al-Alami, N.I., Al-Holy, M.A., Rasco, B.A.J.J.o.a. & chemistry, f. Using Fourier transform infrared (FT-IR) absorbance spectroscopy and multivariate analysis to study the effect of chlorine-induced bacterial injury in water. *J. Agric. Food Chem.* **56**, 8992-8997 (2008).
101. Singh, I., Juneja, P., Kaur, B. & Kumar, P. Pharmaceutical Applications of Chemometric Techniques. *Isrn. Analy. Chem.* **2013**, 13 (2013).
102. Sekhar, S., Ohri, M. & Chakraborti, A. Biofilms: An evolving and universal evasive strategy of bacterial pathogens. "Current Research, Technology and Education Topics in Applied Microbiology and Microbial Biotechnology". 855-859 (2011).
103. Campos, J. et al. Discrimination of non-typhoid Salmonella serogroups and serotypes by Fourier Transform Infrared Spectroscopy: A comprehensive analysis. *Int. J. Food. Microbiol.* **285**, 34-41 (2018).
104. Chong, J., Wishart, D.S. & Xia, J. Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis. *Current. Protocols. In Bioinformatics.* **68**, e86 (2019).
105. Somorjai, R.L. et al. A Data-Driven, Flexible Machine Learning Strategy for the Classification of Biomedical Data. *Current. Protocols. In Bioinformatics.* (2004).
106. Wold, S. Principal component analysis. *Chemometrics & Intell. Lab. Syst.* **2**, 37-52 (1987).
107. Trevisan, J. et al. Measuring similarity and improving stability in biomarker identification methods applied to Fourier-transform infrared (FTIR) spectroscopy. *J. Biophotonics.* **7**, 254-265.

108. Aline, D.S.M., de Melo, M.C.N., Cidral, T.A. & de Lima, K.M.G.J.J.o.M.M. Feature selection strategies for identification of *Staphylococcus aureus* recovered in blood cultures using FT-IR spectroscopy successive projections algorithm for variable selection: A case study. *J. Microbiol. Meth.* **98**, 26-30.
109. Mariey, L., Signolle, J.P., Amiel, C. & Travert, J. Discrimination, classification, identification of microorganisms using FTIR spectroscopy and chemometrics. *Vib. spectrosc.* **26**, 151-159 (2001).
110. Sandrin, T.R., Goldstein, J.E. & Schumaker, S. MALDI TOF MS profiling of bacteria at the strain level: A review. *Mass. Spectrom. Rev.* **32**, 188-217 (2013).
111. Pailhoriès, H., Daure, S., Eveillard, M., Joly-Guillou, M.-L. & Kempf, M. Using Vitek MALDI-TOF mass spectrometry to identify species belonging to the *Acinetobacter calcoaceticus*–*Acinetobacter baumannii* complex: a relevant alternative to molecular biology? *Diagnostic. Mic. Infec. Dis.* **83**, 99-104 (2015).
112. Annette, N. A novel procedure for strain classification of fungal mycelium by cluster and artificial neural network analysis of Fourier transform infrared (FTIR) spectra. *Analyst* **134**, 1215-1223 (2009).
113. Burgula, Y., Khali, D., Kim, S., Krishnan, S.S. & Mauer, L.J. Review of mid-infrared Fourier transform-infrared spectroscopy applications for bacterial detection. *Rapid. Meth. Aut. Mic.* **15**, 146-175 (2010).
114. Lasch, P. et al. FT-IR Hyperspectral Imaging and Artificial Neural Network Analysis for Identification of Pathogenic Bacteria. *Anal.Chem.* **90**, 8896-8904 (2018).
115. Auda, G. & Kamel, M.S. Modular Neural Network Classifiers: A Comparative Study, *J. Robot. Syst.* 117-129 (1998).
116. Rigelsford, J. Pattern recognition: concepts, methods, and applications. *Assembly. autom.* **22** 4. (2001).
117. Mantini, D. et al. LIMPIC: a computational method for the separation of protein MALDI-TOF-MS signals from noise. *BMC. Bioinformatics.* **8**, 101-0.
118. Zhou, J., Yang, Y., Zhang, M. & Xing, H.J.S.C. Constructing ECOC based on confusion matrix for multiclass learning problems. *Sci. china. Inf. Sci.* **59**, 1-14.
119. Normand, A. C. et al. Assessment of various parameters to improve MALDI-TOF MS reference spectra libraries constructed for the routine identification of filamentous fungi. *BMC. Microbiol.* **13**, 76 (2013).
120. Erler, R. et al. VibrioBase: A MALDI-TOF MS database for fast identification of *Vibrio* spp. that are potentially pathogenic in humans. *Syst. Appl. Microbiol.* **38**, 16-25 (2015).

Acknowledgements

Acknowledgements

This work was supported by the National Key Research and Development Plan of China (No. 2018YFF0212501), National Natural Science Foundation of China (No. 21927805 and 31470786), and Science and Technology Commission of Shanghai (No. 19441903800).

Figures

Figure 1

Fig. 1 Typical bacterial FT-IR spectrum and the overlap of H₂O and CO₂. The black (lipids region, 3700-2800 cm⁻¹), blue (membrane amphiphile region, 3000-2800 cm⁻¹), purple (the protein and peptide region, 1500-1200 cm⁻¹), brown (the protein and fatty acid region, 1500-1200 cm⁻¹), and yellow (the nucleic acid and polysaccharide region, 1200-900 cm⁻¹) contours represent different regions of the bacterial FT-IR spectrum. The grey contours represent the FT-IR spectrum of H₂O.

Figure 2

Fig. 2 The main IR radiation mode. (A) Transmittance and typical bacterial FT-IR spectra, (B) attenuated total reflectance (ATR), (C) reflectance

Figure 3

Fig. 3 Workflow of bacterial typing and classification based on FT-IR spectral acquisition and computational analysis.

Figure 4

Fig. 4 Combination of pre-processing steps. The raw spectral data were processed using the following sequence of pre-processing steps: (i) subtract the background spectra and obtain an absorbance spectra sample; (ii) baseline adjustment; (iii) derivative; (iv) normalization; (v) spectral window selection.

Figure 5

Fig. 5 Results presentation for FT-IR based bacterial typing and identification. (A) The two-dimensional projection score plots of 7 standard *E. coli* strains analyzed by PCA. (B) HCA of the second derivative FT-IR spectra, illustrating the relationship between these strains. (C) A heat-map overview of these strains with hierarchical clustering.

Figure 6

Fig. 6 Hierarchical cluster analysis of the second derivative FT-IR spectra, illustrating the relationship between different species.

Figure 7

Fig. 7 Typing scheme for selected *E. coli* strains. In a hierarchical cluster analysis (HCA), the cut-off value defines up to which distance the spectra are considered to be in the same cluster. The green spot indicates that the strain is clustered correctly.

	A	B	C	D	E	F	G	H	I	G	K	L	M	N	O	P	Q	R	S	T	U	V	W
<i>E.coli</i> 10372(A)	9																						
<i>E.coli</i> 10389(B)		4												5									
<i>E.coli</i> 10411(C)			9																				
<i>E.coli</i> 10413(D)				9																			
<i>E.coli</i> 10661(E)					7														2				
<i>E.coli</i> 10663(F)					5	4																	
<i>E.coli</i> 10668(G)							9																
<i>E.coli</i> 21530(H)								7	2														
<i>E.coli</i> 21531(I)									9														
<i>E.coli</i> 25922(G)										6		3											
<i>E.coli</i> 35218(K)	1										8												
<i>E.coli</i> 44102(L)					1					2		6											
<i>E.coli</i> 8099(M)													9										
<i>E.coli</i> 8739(N)					2									6								1	
<i>S.boydii</i> 51522(O)															6				3				
<i>S.boydii</i> A346(P)																9							
<i>S.dysenteriae</i> 51105(Q)																	9						
<i>S.flexneri</i> 1825(R)																		9					
<i>S.flexneri</i> 51572(S)															4				5				
<i>S.flexneri</i> A1129(T)																				9			
<i>S.somnei</i> 51592(U)					3																	6	
<i>S.somnei</i> A1130(V)															1							7	1
<i>S.somnei</i> A592(W)					3																		6

Figure 8

Fig. 8 Confusion matrix of library-based classifiers at the strain level. Each row of the matrix represents the nine sample data for one strain, and each column represents one strain assigned by the classifier. The *E. coli* and *Shigella* strains are segmented by dotted lines. Different species are marked by different background colors, and misclassified numbers are written in red font.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [BacterialTypingProtocol.docx](#)