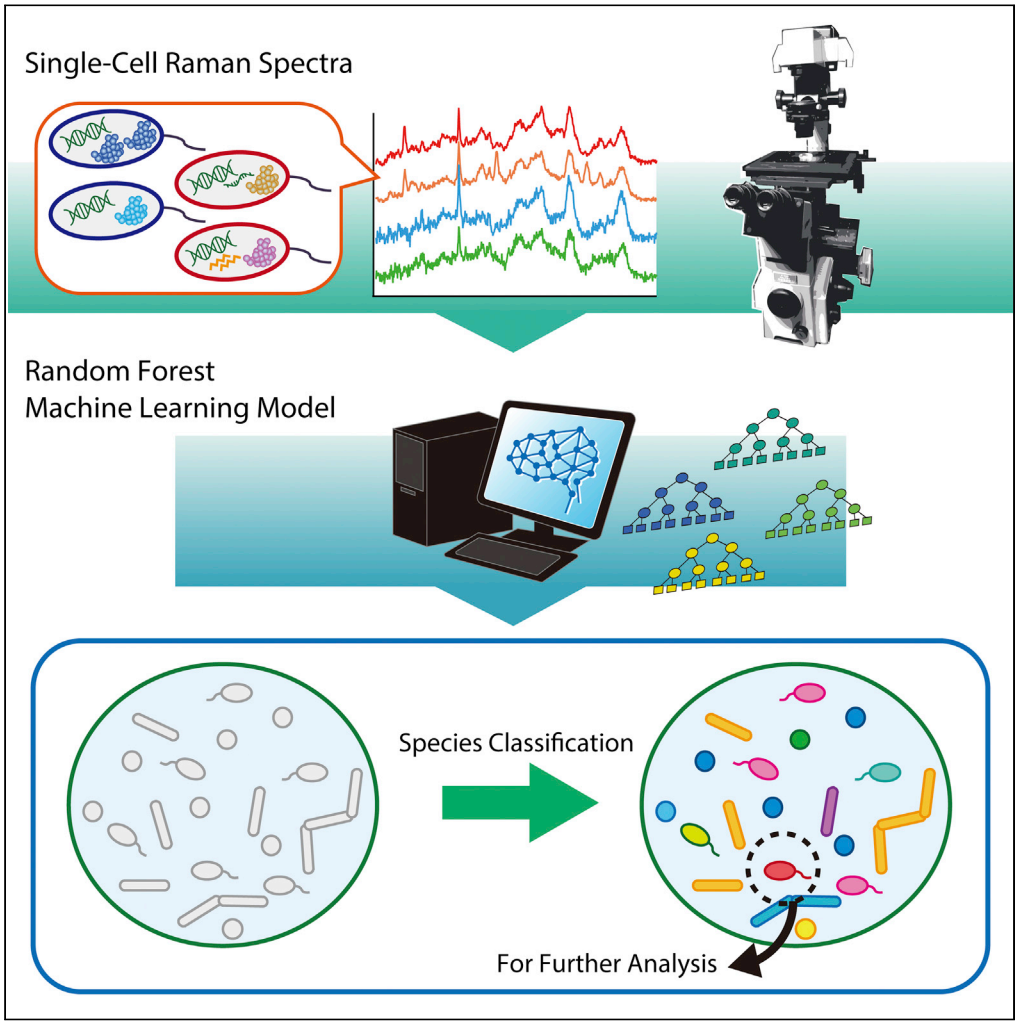
iScience

CellPress
OPEN ACCESS

Article

Machine learning-assisted single-cell Raman fingerprinting for *in situ* and nondestructive classification of prokaryotes



Nanako Kanno, Shingo Kato, Moriya Ohkuma, Motomu Matsui, Wataru Iwasaki, Shinsuke Shigeto

shigeto@kwansei.ac.jp

Highlights

Random forest models classify prokaryotic species with high accuracy of >98%

Both bacteria and archaea are classified using minimally preprocessed Raman data

Feature importance reveals what biomolecules contribute to species classification

Raman marker bands for some archaeal species are discovered

Kanno et al., iScience *24*, 102975
September 24, 2021
https://doi.org/10.1016/j.isci.2021.102975

iScience

CellPress
OPEN ACCESS

Article

Machine learning-assisted single-cell Raman fingerprinting for *in situ* and nondestructive classification of prokaryotes

Nanako Kanno,[1] Shingo Kato,[2] Moriya Ohkuma,[2] Motomu Matsui,[3] Wataru Iwasaki,[3,4] and Shinsuke Shigeto[1,5,*]

SUMMARY

Accessing enormous uncultivated microorganisms (microbial dark matter) in various Earth environments requires accurate, nondestructive classification, and molecular understanding of the microorganisms in *in situ* and at the single-cell level. Here we demonstrate a combined approach of random forest (RF) machine learning and single-cell Raman microspectroscopy for accurate classification of phylogenetically diverse prokaryotes (three bacterial and three archaeal species from different phyla). Our RF classifier achieved a 98.8 ± 1.9% classification accuracy among the six species in pure populations and 98.4% for three species in an artificially mixed population. Feature importance scores against each wavenumber reveal that the presence of carotenoids and structure of membrane lipids play key roles in distinguishing the prokaryotic species. We also find unique Raman markers for an ammonia-oxidizing archaeon. Our approach with moderate data pretreatment and intuitive visualization of feature importance is easy to use for non-spectroscopists, and thus offers microbiologists a new single-cell tool for shedding light on microbial dark matter.

INTRODUCTION

Prokaryotes inhabit a wide variety of environments on Earth from the deep sea to soil to the stratosphere and play essential roles in the entire ecosystem. However, the vast majority (>99%) of them have eluded cultivation in the laboratory, thus constituting "microbial dark matter" (Rinke et al., 2013; Solden et al., 2016; Whitman et al., 1998). Classification and phenotypic characterization of this microbial dark matter will not only advance our understanding of the prokaryotic world, but it is also important for full utilization of their potential in biotechnology (Ling et al., 2015). Over the last decade, culture-independent approaches have been developed in parallel with ongoing efforts to improve the conventional cultivation methodology (Imachi et al., 2011; Ma et al., 2014; Nichols et al., 2010). Among them, single-cell genomics (Stepanauskas, 2012) and metagenomics (Handelsman, 2004) can obtain genome sequences directly from environmental samples and have offered a new view of the phylogenetic tree of microorganisms (Hug et al., 2016). Metatranscriptomics and metabolomics allow researchers to investigate metabolic activities in microbial communities (Kamke et al., 2016; Kim et al., 2015; Lawson et al., 2017). One of the major limitations of these methods is that they are inherently destructive, thus hampering the use of the same microbial cells for subsequent cellular analysis or cultivation efforts. Another limitation is that they are unable to reveal the phenotypic characteristics of individual microbial cells. Therefore, it is highly desired to develop nondestructive and single-cell methods for accurately classifying microorganisms with different phenotypes, physiological states, or activities, whether culturable or not.

The use of Raman microspectroscopy fulfills these two requirements (i.e., nondestructiveness and single-cell resolution) simultaneously. In this optical technique, rich information on molecular vibrations can be obtained in the form of Raman spectra. The Raman spectrum of an individual microbial cell represents cellular "fingerprints", because it contains contributions from various intracellular biomolecules, such as DNA/RNA, proteins, lipids, and bioactive compounds (Lorenz et al., 2017; Schuster et al., 2000). Unlike fluorescence-based methods (e.g., fluorescence *in situ* hybridization (Amann and Fuchs, 2008; Kubo et al., 2011)), Raman microspectroscopy does not require any probes that could affect the original physiological state of the cell to be introduced into the cell. This label-free character is of great advantage when considering applications to environmental samples consisting of diverse (mostly unknown) microorganisms.

[1]Department of Chemistry, School of Science, Kwansei Gakuin University, 2-1 Gakuen, Sanda, Hyogo 669-1337, Japan

[2]Japan Collection of Microorganisms, RIKEN BioResource Research Center, 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074, Japan

[3]Department of Biological Sciences, Graduate School of Science, the University of Tokyo, 2-11-16 Yayoi, Bunkyo-ku, Tokyo 113-0032, Japan

[4]Department of Integrated Biosciences, Graduate School of Frontier Sciences, the University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-0882, Japan

[5]Lead contact

*Correspondence: shigeto@kwansei.ac.jp

https://doi.org/10.1016/j.isci.2021.102975

However, despite these advantages over the genome sequencing and fluorescence techniques, the implementation of Raman microspectroscopy in facile yet accurate microbial classification (Ho et al., 2019; Lu et al., 2020; Novelli-Rousseau et al., 2018; Uysal Ciloglu et al., 2020) is scant especially for environmental samples.

Here we show that prokaryotic cells (bacteria and archaea from different phyla) in a mixed population can be classified with 98.4% accuracy by using random forest (RF) to learn single-cell Raman spectra measured under less-invasive conditions and subjected to minimum preprocessing. The RF algorithm is known to achieve high classification performance despite its simplicity and can also readily output features that make significant contributions to the species classification (Breiman, 2001). Our approach holds great promise for *in situ* screening of specific bacterial and archaeal cells in environmental samples, as well as for exploring many unknown prokaryotes based on the revealed molecular fingerprints.

RESULTS

Raman spectral data from single cells in aqueous medium

To collect a dataset for machine learning model construction, we measured Raman spectra of optically trapped single cells of six individual prokaryotic species dispersed in phosphate buffer solution (PBS), as shown in Figure 1A (see STAR Methods for details). We performed Raman measurements in triplicate (i.e., three independent batches of the six species). The prokaryotic species used in this study comprises three bacterial and three archaeal species that represent taxonomic, functional, and ecological diversity: *Escherichia coli*, Gram-negative bacterium; *Bacillus subtilis*, Gram-positive bacterium; *Thermus thermophilus*, Gram-negative, hyperthermophilic bacterium; *Thermococcus kodakarensis*, hyperthermophilic, anaerobic archaeon; *Sulfolobus acidocaldarius*, hyperthermophilic, acidophilic archaeon; and *Nitrososphaera viennensis*, ammonia-oxidizing archaeon isolated from soil (Table 1). Archaea are found predominantly in extreme environments such as hot/cold, acidic/basic, highly saline, and high-pressure environments (Baker et al., 2020), but microbiome analysis reveals the abundance of archaea in soil and freshwater and their unique ecological roles (e.g., in the nitrogen cycle (Adair and Schwartz, 2008)). It is thus crucial to include archaeal species in the list of microorganisms to classify, although there are much fewer Raman studies on archaea (Fendrihan et al., 2009; Jehlička et al., 2013; Marshall et al., 2007; Serrano et al., 2015; Spang et al., 2012) than on bacteria.

The experimental Raman spectra of single prokaryotic cells can be affected by many factors. For example, cellular autofluorescence gives rise to a broad background without sharp peaks. The differences in cell size and experimental conditions (e.g., fluctuation of the laser power) may result in varying absolute Raman intensities. To classify prokaryotic species solely based on the shape (i.e., relative intensities) of the Raman spectra reflecting the molecular composition of the cell, these contributions were eliminated through a series of spectral preprocessing (Figure 1B), which include subtraction of the PBS spectrum (composed mainly of water), correction for a wavenumber shift among the data measured on different days (i.e., different conditions of the Raman apparatus), baseline subtraction, and vector normalization (see STAR Methods for further details). The resulting Raman spectra of the six prokaryotic species that combine the so-called fingerprint (~660–1800 cm^{-1}) and CH-stretching (2775–3018 cm^{-1}) regions were gathered to generate a dataset (40 spectra × 6 species) for classification model construction using the RF algorithm (Figure 1C).

Random forest modeling for prokaryotic species classification

The averaged, preprocessed spectra of the prokaryotic species in one of the triplicate datasets (Figure 2A) display high similarity in overall pattern, but differ in the degree of spectral variance and noise level (Figure S1). The variance among the spectra of *N. viennensis* is particularly large because of the small cell size of this archaeon; *N. viennensis* cells are difficult to find even under a phase-contrast microscope. To visualize the similarities among individual Raman spectra, we first used principal component analysis (PCA). The spectra of *N. viennensis* form a well-separated cluster mainly by PC1, but the separation among the other five species is less clear (Figure S2), although they represent taxonomical and functional diversity (bacteria vs. archaea, mesophilic vs. thermophilic, etc). In particular, for *E. coli* and *B. subtilis*, most of the spectra overlap with each other in PC space. This PCA result indicates that unsupervised multivariate analysis of the Raman spectra likely fails to classify a wide array of prokaryotic species with high accuracy.

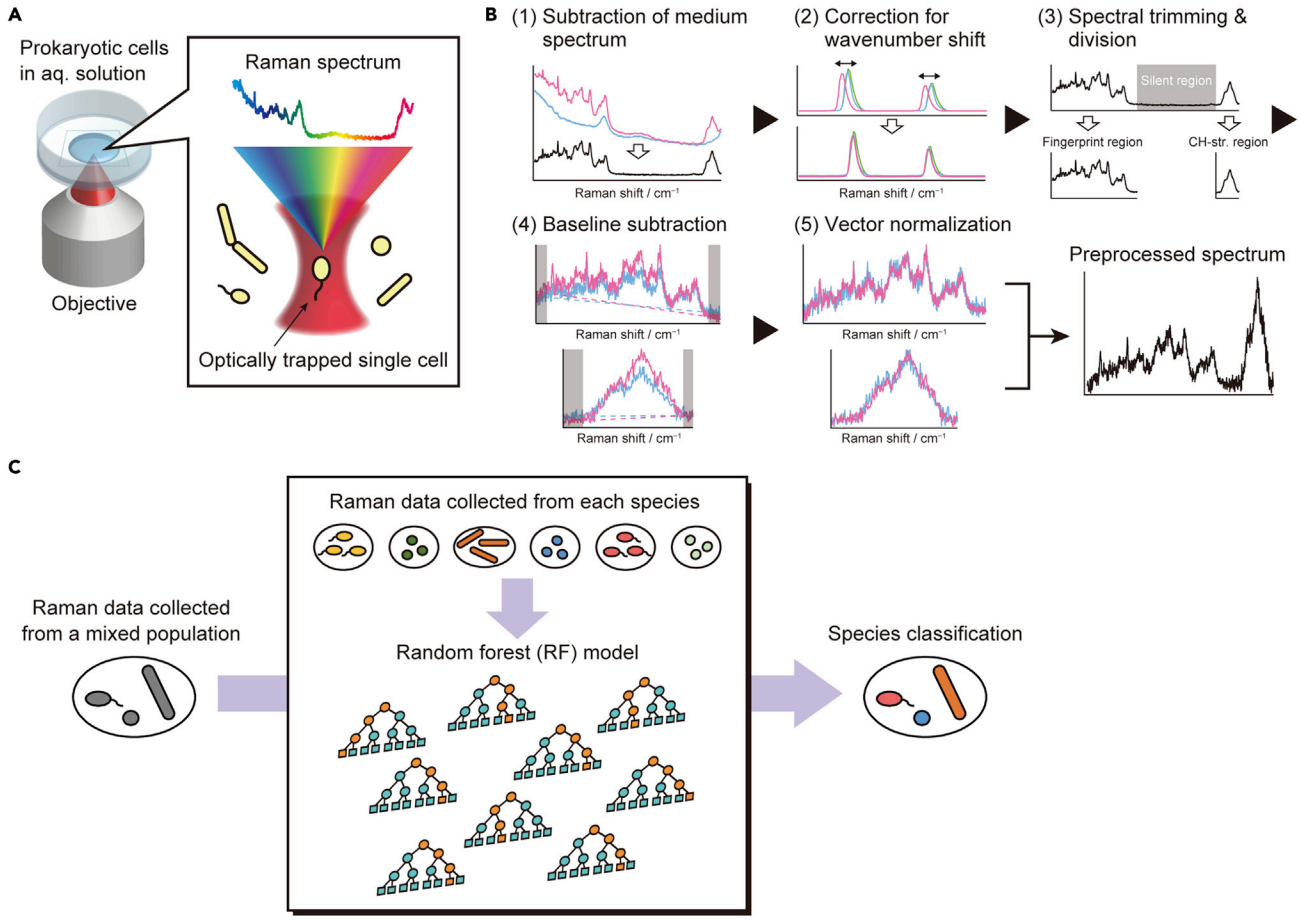


Figure 1. Workflow of the prokaryotic classification method developed in this study, using single-cell Raman microspectroscopy and RF algorithm

(A) Acquisition of the Raman spectra of single prokaryotic cells in aqueous solution (PBS) using an optical tweezer achieved by the same laser beam at 632.8 nm as that used for the Raman excitation (see STAR Methods).

(B) Preprocessing of the measured single-cell Raman spectra: (1) subtraction of the PBS spectrum (blue line) from the cell + PBS spectrum (red line), yielding a difference spectrum (black line); (2) correction for a wavenumber shift that typically occurs among the data taken on different days (red, green, and blue lines); (3) Deletion of the so-called silent region of the Raman spectrum (gray area) and division of the spectrum into the two parts: the fingerprint and CH-stretching regions; (4) subtraction of a linear baseline (dashed lines) that is determined from the edge regions (gray areas); and (5) vector normalization, in which each Raman intensity is divided by the square root of the sum of the squared intensities of the spectrum. The preprocessed spectra in the fingerprint and CH-stretching regions are finally combined.

(C) RF model construction using the preprocessed single-cell Raman data obtained from each prokaryotic species and application to species classification in a mixed population. Training was done exclusively using the Raman spectral data collected from the pure populations of the six species.

Next, we constructed a RF classifier using the above single-cell Raman dataset. The hyper-parameter values of the model were tuned so that high accuracies were achieved in 10-fold cross-validation (see STAR Methods), and the number of decision trees and features were determined to be 1500 and 29, respectively. The out-of-bag error reached 0.79 $\pm$ 0.47% ($\pm$represents standard deviation across 10 train and test splits) when the number of trees was 1500 (Figure S3). Class probabilities averaged over the ten splits are plotted in Figure 2B. The performance breakdown shown as a confusion matrix (Figure 2C) clearly illustrates excellent species classification capability of our RF classifier. It achieves validation accuracy as high as 98.8 $\pm$ 1.9% ($\pm$represents standard deviation across 10 train and test splits). We confirmed on the other two datasets of the triplicate measurements that the six-species classification is reproducible (Figure S4).

With a view to specific detection of archaeal cells in environmental samples, we also trained a binary RF classifier to distinguish between bacteria (*E. coli*, *B. subtilis*, and *T. thermophilus*) and archaea (*T. kodakarensis*, *S. acidocaldarius*, and *N. viennensis*) using the same dataset as above. The results

Table 1. Prokaryotic strains used in this study

Species	Domain	Phylum	Characteristics
Escherichia coli	Bacteria	*Proteobacteria*	Gram-negative, facultative anaerobic[a], mesophilic
Bacillus subtilis	Bacteria	*Firmicutes*	Gram-positive, facultative anaerobic[a], mesophilic
Thermus thermophilus	Bacteria	*Deinococcus-Thermus*	Gram-negative, Aerobic, hyperthermophilic
Thermococcus kodakarensis	Archaea	*Euryarchaeota*	Anaerobic, hyperthermophilic
Sulfolobus acidocaldarius	Archaea	*Crenarchaeota*	Aerobic, acidophilic, hyperthermophilic
Nitrososphaera viennensis	Archaea	*Thaumarchaeota*	Aerobic, mesophilic, ammonia-oxidizer

[a]*E. coli* and *B. subtilis* were grown under aerobic conditions.

(Figure S5) are as good as those of the six-class classification (Figure 2C), yielding 100% validation accuracy and 91.9% accuracy in the mixed population (as shown in the following).

Classification of prokaryotic species in a mixed population

As a proof-of-concept of our approach toward *in situ* classification of prokaryotes based on single-cell Raman spectra, we applied a RF classifier constructed using all the 240 spectra as training data to a mixture of the three species *B. subtilis*, *T. thermophilus*, and *N. viennensis*. The former two are heterotrophic bacteria, whereas the latter is an autotrophic archaeon (ammonia oxidizer). We chose these species because they are easily distinguishable on the basis of their cell morphology: *B. subtilis* has a rod shape, *T. thermophilus* a long rod shape, and *N. viennensis* a small irregular spherical shape (Figure 3A).

We acquired Raman spectra of ~20 cells from each species in the mixed population and used the six-class model to classify a total of 62 spectra. As can be seen from the performance breakdown (Figure 3B), our RF model classifies the three species with 98.4% accuracy. Misclassification occurs only in *N. viennensis*, which was falsely predicted to be *S. acidocaldarius*. Because the present classification is essentially based on Raman spectral patterns rather than images, it can be extended to classification among prokaryotic species with similar cell morphologies.

As seen from Figure 3A, *N. viennensis*, whose cell size is typically below 1 μm (Tourna et al., 2011), shows much noisier spectra than do the other species. One may therefore suspect that the difference in the noise level of single-cell Raman spectra could be responsible at least for the successful classification of *N. viennensis* (Figures 2C and 3B). To assess the effect of the difference in the noise level, we artificially added Gaussian noise to the preprocessed Raman spectra of the five species, except *N. viennensis*, such that the mean of spectral variance across the entire spectral range becomes equal (Figure S6A), and repeated RF modeling and classification using the new dataset with noise added. Although the classification accuracy was somewhat lower overall, we found no marked performance degradation in the identification of *N. viennensis* (Figures S6B and S6C). In addition, the loading spectrum of PC1, which makes a clear distinction between *N. viennensis* and the others in the score plot, appears to exhibit Raman-like features and not a random noise pattern (Figure S7). Taken together, we conclude that the prokaryotic species were classified by Raman fingerprints rather than by the magnitude of noise.

Important Raman spectral features of high discrimination power

In the previous sections, we have only looked at the performance of machine learning classification. What features play important roles in the classification? Answering this question will lead to the discovery of potential Raman markers for identifying specific prokaryotic cells and characterizing their functions. Although it is possible to somehow obtain quantities corresponding to important spectral features that contribute to classification using other algorithms such as support vector machine and convolutional neural network (CNN), these quantities can be extracted in a more straightforward manner using the RF algorithm. The top 50 most important features (wavenumbers) extracted from the six-class classification result are shown in Figure 2A and Table S1. The features mentioned below are excerpted and summarized in Table 2. Most of them are distributed across the fingerprint region and can be associated with proteins, DNA/RNA, lipids,

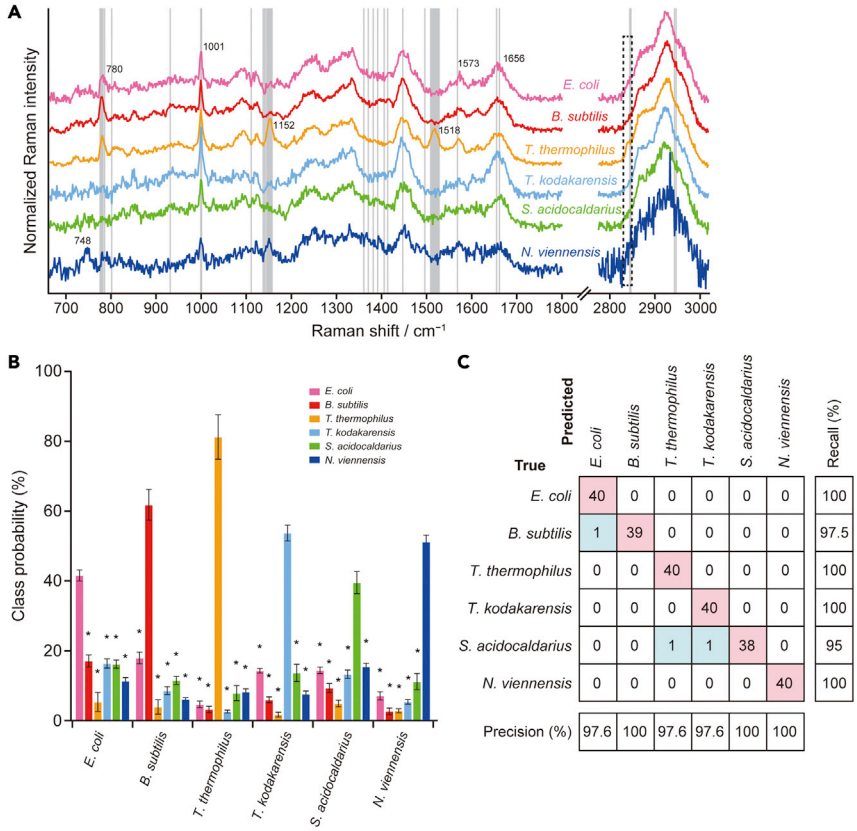


Figure 2. Construction of an RF classifier using the single-cell Raman dataset and its performance on species classification

(A) Average of the preprocessed Raman spectra for each of the six prokaryotic species (*E. coli*, *B. subtilis*, *T. thermophilus*, *T. kodakarensis*, *S. acidocaldarius*, and *N. viennensis*). Top 50 most important features are shown as vertical lines. The dashed rectangular box indicates the region for which the CH_2-stretching band intensities were calculated (see Figure 4B).

(B) Averaged class probabilities in 10-fold cross-validation. Error bars represent ±SD (n = 10). The asterisks represent statistically significant differences between the probability of being predicted to be in the true class and those in other classes, with Welch's t test ($P < 0.05$).

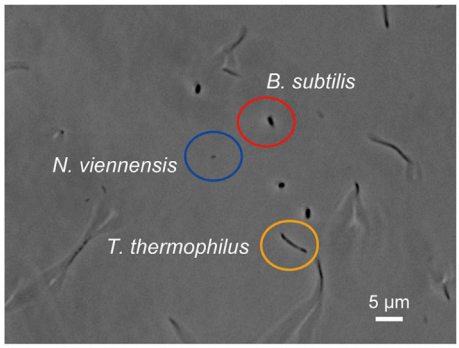
(C) Confusion matrix, **C**, for six strain classes. Each entry of the confusion matrix, C_{ij}, represents the total number of spectra known to be in class i and predicted by the RF classifier to be in class j in 10-fold cross-validation. Correct classification results are shown in red boxes on the diagonal, and misclassification results in blue boxes. Also shown are the precision and recall rates in percentage.

and carotenoids. The other two independent datasets yielded overall similar distributions of important features (Figure S4).

The wavenumber of the highest importance (1000.7 cm^{-1}) coincides with the ring-breathing band of the phenylalanine residues, a characteristic Raman band of proteins. The averaged Raman spectra (Figure 2A) reveal a variation in the intensity of this band among the six species, with that for *T. kodakarensis* being the highest and that for *N. viennensis* being the lowest. The intensity of the phenylalanine band is considered to reflect the total protein abundance in a microbial cell (Noothalapati Venkata and Shigeto, 2012). Wavenumbers 1661.5 and 1655.4 cm^{-1} are also attributable to proteins. The corresponding Raman band is known as the amide I, which arises mainly from the C=O stretching vibration in the peptide bond. Again, the intensity of this band is highest in *T. kodakarensis* and lowest in *N. viennensis*. The present result suggests that protein abundance is a key factor in microbial discrimination.

The next important features center in regions 1137.5–1157.3 and 1508.8–1527.7 cm^{-1}. The Raman bands in these regions are well-known signatures of carotenoids (Horiue et al., 2020; Withnall et al., 2003;

A



B

True \ Predicted	*E. coli*	*B. subtilis*	*T. thermophilus*	*T. kodakarensis*	*S. acidocaldarius*	*N. viennensis*	Recall (%)
B. subtilis	0	21	0	0	0	0	100
T. thermophilus	0	0	20	0	0	0	100
N. viennensis	0	0	0	0	1	20	95.2
Precision (%)	-	100	100	-	-	100	

Figure 3. Application of an RF classifier constructed using all the single-cell Raman spectra as training data to the classification of three prokaryotic species (*B. subtilis*, *T. thermophilus*, and *N. viennensis*) in a mixed population

(A) Phase-contrast micrograph of the mixed prokaryotic population, where a rod-shaped *B. subtilis* cell (red circled), a long *T. thermophilus* cell (yellow circled), and a tiny *N. viennensis* cell (blue circled) can be identified. Scale bar = 5 μm.

(B) Performance breakdown. Correct classification results are shown in red boxes, and misclassification results in blue boxes. The values in the table represent the number of spectra (cells). The precision and recall rates are shown in percentage.

Zheng et al., 2013), whose intensities are enhanced owing to the (pre)resonance Raman effect. In typical carotenoids, the band at ~1152 cm^{-1} is assigned to the C–C stretching mode, and that at ~1518 cm^{-1} the in-phase C=C stretching mode in the conjugated chain. Both bands are observed in the Raman spectra of *T. thermophilus*. This observation is consistent with the report that *T. thermophilus* produces yellow carotenoids (Oshima and Imahori, 1974). It is very likely that this species was accurately differentiated from the others by whether or not the carotenoid pigment is present.

With regard to the carotenoid Raman bands, there is a puzzling observation worth noting. In *N. viennensis*, a Raman band of comparable intensity to *T. thermophilus* is clearly seen at 1152 cm^{-1}, exactly the same wavenumber as the C–C stretching band of carotenoids in *T. thermophilus*. Nevertheless, no appreciable band is observed at 1518 cm^{-1} (Figure 4A), indicating that the 1152 cm^{-1} band is not due to carotenoids. This argument agrees with the fact that the gene sets associated with carotenoid biosynthesis are not found in the published genome information. In addition to the 1152 cm^{-1} band, *N. viennensis* cells show prominent Raman bands at ~748 cm^{-1}. Neither nitrite nor nitrate (products of ammonia oxidation) can account for these bands. The 1152 and 748 cm^{-1} bands, though unassigned at present, could be used as specific markers for *N. viennensis* and possibly for ammonia-oxidizing microorganisms.

DNA/RNA Raman bands also take a significant part in the classification. Wavenumbers 775.7–786.0 cm^{-1} and 1568.4 cm^{-1} correspond to Raman bands arising from the pyrimidine ring (cytosine, thymine, or uracil) and the purine ring (adenine and guanine), respectively. At ~780 cm^{-1}, there is an additional contribution of the DNA/RNA backbone (O–P–O stretching) mode. Our results (Figure 2A) indicate that the bacterial species show higher intensities of the DNA/RNA bands compared to the archaeal species.

Four of the top 50 important features are located in the higher wavenumber (>2800 cm^{-1}) region. This region encompasses CH_2/CH_3 stretching bands of major macro-biomolecules such as proteins and lipids. The profile of the broad CH-stretching band peaking at ~2930 cm^{-1} (Figure 2A) resembles that of the proteins derived previously with a multivariate curve resolution technique (Hsu et al., 2015; Yasuda et al., 2019), suggesting that the CH-stretching band of the six species is dominated by proteins with minor contributions of lipids. Interestingly, however, it is around the shoulder at ~2850 cm^{-1} coinciding with the CH_2

Table 2. Important features (wavenumbers) that make significant contributions to RF classification of the prokaryotic species

Wavenumber (cm^{-1})	Assignment[a]	Molecular components
775.7–786.0	Pyrimidine ring (C, T, and U), O-P-O backbone	DNA/RNA
999.0, 1000.7	Phenylalanine ring breathing	Proteins
1137.5–1157.3	C–C stretch	Carotenoids
1508.8–1527.7	C=C stretch	Carotenoids
1568.4	Purine ring (A and G)	DNA/RNA
1655.4, 1661.5	Amide I	Proteins
2844.3, 2846.8	CH_2 symmetric stretch	Lipids

[a](Carey, 1982; Huang et al., 2005; Krafft et al., 2003; Puppels et al., 1990).

symmetric stretching band of lipids, that is important in the present classification. The intensity at ~2850 cm^{-1} varies considerably among the six species (Figure 4B). Gram-negative bacteria *E. coli* and *T. thermophilus* show an obvious shoulder, whereas Gram-positive bacterium *B. subtilis* and archaeon *T. kodakarensis* do not. *S. acidocaldarius* is somewhat intermediate. Statistical tests also confirmed that the 2850 cm^{-1} intensity is significantly ($P < 0.05$) lower in *B. subtilis* and *T. kodakarensis* than the others. The higher intensity at ~2850 cm^{-1} for Gram-negative bacteria is probably because they have an outer membrane in addition to a cell membrane. Archaea possess an S-layer consisting of proteins as the cell wall, but do not possess an outer membrane (Albers and Meyer, 2011). Furthermore, archaeal membrane lipids are based on isoprene chains (Jain et al., 2014), which contain fewer CH_2 moieties than does the bacterial counterpart. These differences in absolute amounts and compositions of membrane lipids between bacteria and archaea account well for the observed variation in the Raman intensity at ~2850 cm^{-1}.

The exception is *N. viennensis*, which shows ~2850 cm^{-1} intensity, as high as *T. thermophilus*. We reason that, being by far the smallest in cell size among the six species studied, *N. viennensis* must have a larger membrane-lipid abundance (represented by CH_2) relative to protein abundance (represented by CH_3). Igisu and co-workers reported on the basis of micro-FTIR spectroscopic analysis that the CH_2/CH_3 IR absorbance ratio can be used to distinguish between bacteria and archaea (Igisu et al., 2009, 2012). Our results are in line with the previous study in that the relative intensity of the CH_2 stretching mode could differ depending on microorganisms, but is also in contrast because our RF classification results suggest that clear distinction between bacteria and archaea may not be always possible using the CH_2 stretching band alone. It is most likely species-dependent.

DISCUSSION

We have applied a combination of single-cell Raman microspectroscopy and RF machine learning strategy for classification of six prokaryotic species chosen from a variety of phyla and for identification in a mixed population envisaging environmental samples. Recently, CNN, a deep learning technique, was used to classify Raman spectra of 30 clinically relevant bacteria including methicillin-resistant *Staphylococcus aureus* (Ho et al., 2019) and of 14 microorganisms (two bacteria, five archaea, and seven fungi) (Lu et al., 2020). The significance of our work compared to these studies is threefold.

First, we aimed at developing an accurate classification model applicable to a complex environmental system composed of multiple prokaryotic species. As shown in Table 1, the prokaryotic species used in this work cover microbial-ecologically relevant species that belong to *Proteobacteria*, *Firmicutes*, *Deinococcus-Thermus*, *Crenarchaeota*, *Euryarchaeota*, and *Thaumarchaeota*, although focus in the previous studies was placed primarily on pathogenic or human-related microorganisms only from *Proteobacteria* and *Firmicutes* (Ho et al., 2019; Lu et al., 2020). The classification accuracy exceeding 98% achieved by our approach for the mixed prokaryotic population gives hope for *in situ* identification of specific prokaryotic groups (e.g., archaea) at the single-cell level without the need for time-consuming, destructive analysis.

RF is among ensemble learning methods, in which an ensemble of weak learners trained on many samples created by bootstrap aggregation (or bagging) (Breiman, 1996) is combined to achieve high classification accuracy. Another type of ensemble learning method, boosting, converts weak learners to strong ones by

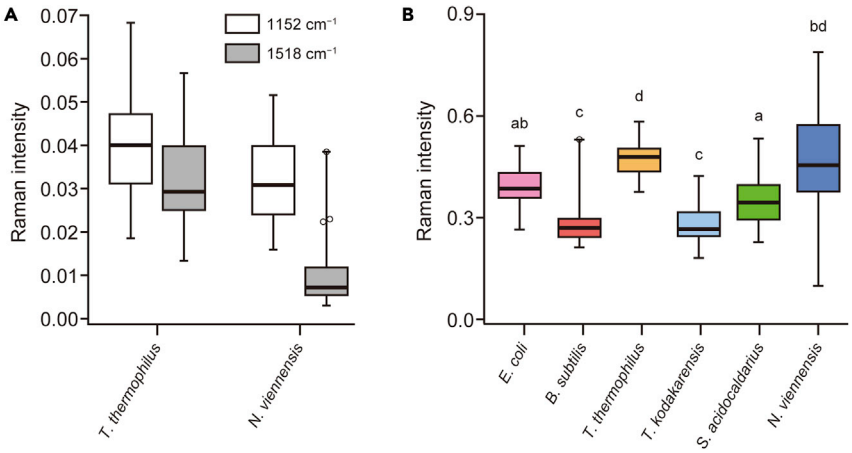


Figure 4. Quantitative analysis of the important spectral features that contribute to the RF classification of the six prokaryotic species

(A) Comparison of the Raman intensities at ~1152 cm^{-1} (white boxes) and ~1518 cm^{-1} (gray boxes) between *T. thermophilus* and *N. viennensis*. The intensity plotted here was evaluated from each single-cell spectrum (smoothed using 15-point Savitzky–Golay polynomial filter of degree 2) by subtracting the minimum value in 1134.2–1173.8 cm^{-1} from the maximum value in 1145.7–1154.0 cm^{-1} or by subtracting the minimum value in 1497.8.7–1537.1 cm^{-1} from the maximum value in 1516.7–1519.8 cm^{-1}.

(B) Comparison of the CH-stretching Raman intensities at ~2850 cm^{-1} among the six prokaryotic species studied. The intensity plotted here was evaluated from each single-cell spectrum (smoothed using 15-point Savitzky–Golay polynomial filter of degree 2) by summing up the intensities in 2830.5–2849.3 cm^{-1} (the region indicated by the dashed rectangular box in Figure 2A). Different letters indicate statistically significant differences ($P < 0.05$) among the six species by a nonparametric Kruskal–Wallis test followed by a post hoc Dunn–Holland–Wolfe multiple comparison tests.

adjusting the weights of the data that are misclassified by previous learners. Previous studies comparing RF and boosting methods such as AdaBoost and XGBoost reported that careful selection of the learning method and tuning of their hyper-parameters improved predictive performances (Huang et al., 2021). To achieve excellent classification accuracy even in more challenging microbial classifications, we could leverage these various types of learning methods.

Second, our Raman measurement is as less invasive as possible to prokaryotic cells, and our preprocessing of the measured Raman spectra is also minimally demanding for non-spectroscopists. Lu and co-workers used ~16 $mW/\mu m^2$ laser power and 60–90 s exposure time in their 785 nm-excited single-cell Raman measurements for CNN-based classification of microorganisms (Lu et al., 2020). In contrast, the laser power and exposure time we used were only ~3.8 $mW/\mu m^2$ at 632.8 nm and 30 s, respectively. Low invasiveness is crucial for the feasibility of downstream analysis.

We carefully examined what preprocessing was integral to highly accurate species classification and found that noise reduction using methods like smoothing (Barton et al., 2018), derivative spectra (Xie et al., 2005), and singular value decomposition (SVD) (Huang et al., 2011; van Manen et al., 2004; Yasuda et al., 2019) was not necessary (see Figure 1B and STAR Methods). Smoothing and derivative spectra may result in underestimation of sharp Raman bands such as the phenylalanine band at 1001 cm^{-1}. SVD has proven to be effective particularly for Raman imaging data taken with a short exposure time (Huang et al., 2011; van Manen et al., 2004; Yasuda et al., 2019), but it requires some knowledge and experience about Raman spectroscopy to determine how many SV components should be retained and could possibly discard minor but important features as noise. The RF classifications using the dataset with SVD denoising (Figure S8) turned out to perform only slightly better (99.6 $\pm$ 1.3% accuracy) for the six-class model validation and somewhat worse (95.2% accuracy) for the mixed population than those using the dataset without any noise reduction. Furthermore, step 2 in our spectral preprocessing (Figure 1B) allows one to use spectra measured on different days, mitigating the experimental burden that as much data as possible must be taken at once. Polynomial fitting is often used in baseline subtraction (step 4), but we found that the Raman data processed with fourth-order polynomial fitting (Figure S9) yielded high classification accuracy (98.3 $\pm$ 2.8%)

comparable to that obtained with linear fitting (98.8 ± 1.9%). We thus adopted a linear baseline, which is less arbitrary in terms of the spectral regions to be included in the fitting than polynomial fitting. Using a longer excitation wavelength (e.g., 785 nm (Lu et al., 2020)) may be effective in suppressing autofluorescence, although theoretically the Raman scattering probability is lower.

Third, we interpreted the outcome of RF classification on a molecular basis. In their CNN classification, Lu and co-workers devised a method called occlusion-based Raman spectra feature extraction to investigate the wavenumbers that contribute to the classification (Lu et al., 2020), but the extracted features were not fully analyzed in the molecular, phylogenetic, or physiological context. Our RF-based approach directly provides insight into the property and structure of prokaryotic cells that render them distinguishable.

Spectral variance can result not only from cell-to-cell variation in autofluorescence background and measurement conditions that were removed during our preprocessing (see Figure 1B and STAR Methods), but also from that in physiological state such as growth stage and growth conditions (Lorenz et al., 2017; Xie et al., 2005). Microorganisms are known to accumulate different kinds of storage material inside the cell and they may considerably affect Raman spectral patterns (Ciobotă et al., 2010; Miyaoka et al., 2014; Noothalapati Venkata et al., 2011). Investigating these effects on spectral variance will help us further improve the accuracy of species classification. In conclusion, we believe that our method, which takes advantage of minimally invasive and easily operated Raman microspectroscopy and a machine learning technique, can be a useful addition to the toolbox of microbiologists exploring microbial dark matter.

Limitations of the study

There will be more and more demand for screening specific prokaryotic cells from the environment as omics analyses unravel diverse, novel activities of environmental prokaryotes (Gutleben et al., 2018; Kaster and Sobol, 2020). Discrimination based on phylogenetic characteristics, physiologically active substances, and cellular structure and detection of unique spectra that do not fit in reference spectral database are expected, when coupled with omics analyses, to facilitate picking unknown microorganisms from various environments. However, we have not yet demonstrated the applicability of our method to a sample containing prokaryotic species whose reference Raman data are unavailable. In such a case, unsupervised learning such as PCA with x-means clustering could be used to estimate the number of clusters (prokaryotic species, including unknown ones) that exist in the sample and to know whether or not our reference Raman spectra are classified into those clusters.

We also need to extend the study to more prokaryotic species so that it can be applied to real environmental samples, and to investigate the effects of laser irradiation and the physiological states of cells (e.g., exponential vs. stationary phases) in detail, both of which are currently underway in our laboratory.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Confocal Raman microspectroscopy
  - Data preprocessing
  - Random forest training and test
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102975.

ACKNOWLEDGMENTS

This work was supported by the Japan Society for the Promotion of Science KAKENHI Grant Numbers JP19H05679 (Post-Koch Ecology), JP19H05681, and JP19H05689, and in part by the Japan Science and Technology Agency Exploratory Research for Advanced Technology (ERATO) Grant Number JPMJER1502.

AUTHOR CONTRIBUTIONS

N.K. and S.S. conceived and designed the research. S.S. supervised the project. N.K. performed the experiments and analyzed the data. M.M. and W.I. helped data analysis. S.K. and M.O. provided the samples. N.K. and S.S. wrote the manuscript with input from S.K. and M.M. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 16, 2021
Revised: July 27, 2021
Accepted: August 9, 2021
Published: September 24, 2021

REFERENCES

Adair, K.L., and Schwartz, E. (2008). Evidence that ammonia-oxidizing archaea are more abundant than ammonia-oxidizing bacteria in semiarid soils of northern Arizona, USA. Microb. Ecol. *56*, 420–426.

Albers, S.-V., and Meyer, B.H. (2011). The archaeal cell envelope. Nat. Rev. Microbiol. *9*, 414–426.

Amann, R., and Fuchs, B.M. (2008). Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques. Nat. Rev. Microbiol. *6*, 339–348.

Baker, B.J., De Anda, V., Seitz, K.W., Dombrowski, N., Santoro, A.E., and Lloyd, K.G. (2020). Diversity, ecology and evolution of Archaea. Nat. Microbiol. *5*, 887–900.

Barton, S.J., Ward, T.E., and Hennelly, Bryan M. (2018). Algorithm for optimal denoising of Raman spectra. Anal. Methods *10*, 3759–3769.

Breiman, L. (1996). Bagging predictors. Mach. Learn. *24*, 123–140.

Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32.

Carey, P.R. (1982). Biochemical Applications of Raman and Resonance Raman Spectroscopies (Academic Press).

Ciobotă, V., Burkhardt, E.-M., Schumacher, W., Rösch, P., Küsel, K., and Popp, J. (2010). The influence of intracellular storage material on bacterial identification by means of Raman spectroscopy. Anal. Bioanal. Chem. *397*, 2929–2937.

Fendrihan, S., Musso, M., and Stan-Lotter, H. (2009). Raman spectroscopy as a potential method for the detection of extremely halophilic archaea embedded in halite in terrestrial and possibly extraterrestrial samples. J. Raman Spectrosc. *40*, 1996–2003.

Gutleben, J., Chaib De Mares, M., van Elsas, J.D., Smidt, H., Overmann, J., and Sipkema, D. (2018). The multi-omics promise in context: from sequence to microbial isolate. Crit. Rev. Microbiol. *44*, 212–229.

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. Microbiol. Mol. Biol. Rev. *68*, 669–685.

Ho, C.S., Jean, N., Hogan, C.A., Blackmon, L., Jeffrey, S.S., Holodniy, M., Banaei, N., Saleh, A.A.E., Ermon, S., and Dionne, J. (2019). Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. Nat. Commun. *10*, 4927.

Horiue, H., Sasaki, M., Yoshikawa, Y., Toyofuku, M., and Shigeto, S. (2020). Raman spectroscopic signatures of carotenoids and polyenes enable label-free visualization of microbial distributions within pink biofilms. Sci. Rep. *10*, 7704.

Hsu, J.-F., Hsieh, P.-Y., Hsu, H.-Y., and Shigeto, S. (2015). When cells divide: label-free multimodal spectral imaging for exploratory molecular investigation of living cells during cytokinesis. Sci. Rep. *5*, 17541.

Huang, C.-K., Ando, M., Hamaguchi, H., and Shigeto, S. (2012). Disentangling dynamic changes of multiple cellular components during the yeast cell cycle by in vivo multivariate Raman imaging. Anal. Chem. *84*, 5661–5668.

Huang, C.K., Hamaguchi, H., and Shigeto, S. (2011). *In vivo* multimode Raman imaging reveals concerted molecular composition and distribution changes during yeast cell cycle. Chem. Commun. *47*, 9423–9425.

Huang, X., Li, C., Tan, K., Wen, Y., Guo, F., Li, M., Huang, Y., Sun, C.Q., Gozin, M., and Zhang, L. (2021). Applying machine learning to balance performance and stability of high energy density materials. iScience *24*, 102240.

Huang, Y.-S., Karashima, T., Yamamoto, M., and Hamaguchi, H. (2005). Molecular-level investigation of the structure, transformation, and bioactivity of single living fission yeast cells by time- and space-resolved Raman spectroscopy. Biochemistry *44*, 10009–10019.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., et al. (2016). A new view of the tree of life. Nat. Microbiol. *1*, 16048.

Igisu, M., Takai, K., Ueno, Y., Nishizawa, M., Nunoura, T., Hirai, M., Kaneko, M., Naraoka, H., Shimojima, M., Hori, K., et al. (2012). Domain-level identification and quantification of relative prokaryotic cell abundance in microbial communities by micro-FTIR spectroscopy. Environ. Microbiol. Rep. *4*, 42–49.

Igisu, M., Ueno, Y., Shimojima, M., Nakashima, S., Awramik, S.M., Ohta, H., and Maruyama, S. (2009). Micro-FTIR spectroscopic signatures of Bacterial lipids in Proterozoic microfossils. Precamb. Res. *173*, 19–26.

Imachi, H., Aoi, K., Tasumi, E., Saito, Y., Yamanaka, Y., Saito, Y., Yamaguchi, T., Tomaru, H., Takeuchi, R., Morono, Y., et al. (2011). Cultivation of methanogenic community from subseafloor sediments using a continuous-flow bioreactor. ISME J. *5*, 1913–1925.

Jain, S., Caforio, A., and Driessen, A.J. (2014). Biosynthesis of archaeal membrane ether lipids. Front. Microbiol. *5*, 641.

Jehlička, J., Edwards, H.G.M., and Oren, A. (2013). Bacterioruberin and salinixanthin carotenoids of extremely halophilic archaea and bacteria: a Raman spectroscopic study. Spectrochim. Acta A *106*, 99–103.

Kamke, J., Kittelmann, S., Soni, P., Li, Y., Tavendale, M., Ganesh, S., Janssen, P.H., Shi, W., Froula, J., Rubin, E.M., et al. (2016). Rumen metagenome and metatranscriptome analyses of

low methane yield sheep reveals a *Sharpea*-enriched microbiome characterised by lactic acid formation and utilisation. Microbiome *4*, 56.

Kaster, A.K., and Sobol, M.S. (2020). Microbial single-cell omics: the crux of the matter. Appl. Microbiol. Biotechnol. *104*, 8209–8220.

Kim, Y.-M., Nowack, S., Olsen, M.T., Becraft, E.D., Wood, J.M., Thiel, V., Klapper, I., Kühl, M., Fredrickson, J.K., Bryant, D.A., et al. (2015). Diel metabolomics analysis of a hot spring chlorophototrophic microbial mat leads to new hypotheses of community member metabolisms. Front. Microbiol. *6*, 209.

Krafft, C., Knetschke, T., Siegner, A., Funk, R.H.W., and Salzer, R. (2003). Mapping of single cells by near infrared Raman microspectroscopy. Vib. Spectrosc. *32*, 75–83.

Kubo, K., Knittel, K., Amann, R., Fukui, M., and Matsuura, K. (2011). Sulfur-metabolizing bacterial populations in microbial mats of the Nakabusa hot spring. Japan. Syst. Appl. Microbiol. *34*, 293–302.

Lawson, C.E., Wu, S., Bhattacharjee, A.S., Hamilton, J.J., McMahon, K.D., Goel, R., and Noguera, D.R. (2017). Metabolic network analysis reveals microbial community interactions in anammox granules. Nat. Commun. *8*, 15416.

Ling, L.L., Schneider, T., Peoples, A.J., Spoering, A.L., Engels, I., Conlon, B.P., Mueller, A., Schäberle, T.F., Hughes, D.E., Epstein, S., et al. (2015). A new antibiotic kills pathogens without detectable resistance. Nature *517*, 455–459.

Lorenz, B., Wichmann, C., Stöckel, S., Rösch, P., and Popp, J. (2017). Cultivation-free Raman spectroscopic investigations of bacteria. Trends Microbiol. *25*, 413–424.

Lu, W., Chen, X., Wang, L., Li, H., and Fu, Y.V. (2020). Combination of an artificial intelligence approach and laser tweezers Raman spectroscopy for microbial identification. Anal. Chem. *92*, 6288–6296.

Ma, L., Kim, J., Hatzenpichler, R., Karymov, M.A., Hubert, N., Hanan, I.M., Chang, E.B., and Ismagilov, R.F. (2014). Gene-targeted microfluidic cultivation validated by isolation of a gut bacterium listed in Human Microbiome Project's Most Wanted taxa. Proc. Natl. Acad. Sci. U S A *111*, 9768–9773.

Marshall, C.P., Leuko, S., Coyle, C.M., Walter, M.R., Burns, B.P., and Neilan, B.A. (2007). Carotenoid analysis of halophilic archaea by resonance Raman spectroscopy. Astrobiology *7*, 631–643.

Matsuda, A., Sakaguchi, N., and Shigeto, S. (2019). Can cells maintain their bioactivity in ionic liquids? A novel single-cell assessment by Raman microspectroscopy. J. Raman Spectrosc. *50*, 768–777.

Miyaoka, R., Hosokawa, M., Ando, M., Mori, T., Hamaguchi, H.-o., and Takeyama, H. (2014). *In situ* detection of antibiotic amphotericin B produced in *Streptomyces nodosus* using Raman microspectroscopy. Mar. Drugs *12*, 2827–2839.

Nichols, D., Cahoon, N., Trakhtenberg, E.M., Pham, L., Mehta, A., Belanger, A., Kanigan, T., Lewis, K., and Epstein, S.S. (2010). Use of ichip for high-throughput *in situ* cultivation of "uncultivable" microbial species. Appl. Environ. Microbiol. *76*, 2445–2450.

Noothalapati Venkata, H.N., Nomura, N., and Shigeto, S. (2011). Leucine pools in *Escherichia coli* biofilm discovered by Raman imaging. J. Raman Spectrosc. *42*, 1913–1915.

Noothalapati Venkata, H.N., and Shigeto, S. (2012). Stable isotope-labeled Raman imaging reveals dynamic proteome localization to lipid droplets in single fission yeast cells. Chem. Biol. *19*, 1373–1380.

Novelli-Rousseau, A., Espagnon, I., Filiputti, D., Gal, O., Douet, A., Mallard, F., and Josso, Q. (2018). Culture-free antibiotic-susceptibility determination from single-bacterium Raman spectra. Sci. Rep. *8*, 3957.

Oshima, T., and Imahori, K. (1974). Description of *Thermus thermophilus* (Yoshida and Oshima) comb. nov., a nonsporulating thermophilic bacterium from a Japanese thermal spa. Int. J. Syst. Bacteriol. *24*, 102–112.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Puppels, G.J., de Mul, F.F.M., Otto, C., Greve, J., Robert-Nicoud, M., Arndt-Jovin, D.J., and Jovin, T.M. (1990). Studying single living cells and chromosomes by confocal Raman microspectroscopy. Nature *347*, 301–303.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., et al. (2013). Insights into the phylogeny and coding potential of microbial dark matter. Nature *499*, 431–437.

Schuster, K.C., Reese, I., Urlaub, E., Gapes, J.R., and Lendl, B. (2000). Multidimensional information on the chemical composition of single bacterial cells by confocal Raman microspectroscopy. Anal. Chem. *72*, 5529–5534.

Serrano, P., Hermelink, A., Lasch, P., de Vera, J.-P., König, N., Burckhardt, O., and Wagner, D. (2015). Confocal Raman microspectroscopy reveals a convergence of the chemical composition in methanogenic archaea from a Siberian permafrost-affected soil. FEMS Microbiol. Ecol. *91*, fiv126.

Solden, L., Lloyd, K., and Wrighton, K. (2016). The bright side of microbial dark matter: lessons learned from the uncultivated majority. Curr. Opin. Microbiol. *31*, 217–226.

Spang, A., Poehlein, A., Offre, P., Zumbragel, S., Haider, S., Rychlik, N., Nowka, B., Schmeisser, C., Lebedeva, E.V., Rattei, T., et al. (2012). The genome of the ammonia-oxidizing *candidatus* Nitrososphaera gargensis: insights into metabolic versatility and environmental adaptations. Environ. Microbiol. *14*, 3122–3145.

Stepanauskas, R. (2012). Single cell genomics: an individual look at microbes. Curr. Opin. Microbiol. *15*, 613–620.

Tourna, M., Stieglmeier, M., Spang, A., Könneke, M., Schintlmeister, A., Urich, T., Engel, M., Schloter, M., Wagner, M., Richter, A., et al. (2011). *Nitrososphaera viennensis*, an ammonia oxidizing archaeon from soil. Proc. Natl. Acad. Sci. U S A *108*, 8420–8425.

Uysal Ciloglu, F., Saridag, A.M., Kilic, I.H., Tokmakci, M., Kahraman, M., and Aydin, O. (2020). Identification of methicillin-resistant *Staphylococcus aureus* bacteria using surface-enhanced Raman spectroscopy and machine learning techniques. Analyst *145*, 7559–7570.

van Manen, H.-J., Kraan, Y.M., Roos, D., and Otto, C. (2004). Intracellular chemical imaging of heme-containing enzymes involved in innate immunity using resonance Raman microscopy. J. Phys. Chem. B *108*, 18762–18771.

Whitman, W.B., Coleman, D.C., and Wiebe, W.J. (1998). Prokaryotes: the unseen majority. Proc. Natl. Acad. Sci. U S A *95*, 6578–6583.

Withnall, R., Chowdhry, B.Z., Silver, J., Edwards, H.G.M., and de Oliveira, L.F.C. (2003). Raman spectra of carotenoids in natural products. Spectrochim. Acta A *59*, 2207–2212.

Xie, C., Dinno, M.A., and Li, Y.-q. (2002). Near-infrared Raman spectroscopy of single optically trapped biological cells. Opt. Lett. *27*, 249–251.

Xie, C., Mace, J., Dinno, M.A., Li, Y.Q., Tang, W., Newton, R.J., and Gemperline, P.J. (2005). Identification of single bacterial cells in aqueous solution using confocal laser tweezers Raman spectroscopy. Anal. Chem. *77*, 4390–4397.

Yamakoshi, H., Dodo, K., Okada, M., Ando, J., Palonpon, A., Fujita, K., Kawata, S., and Sodeoka, M. (2011). Imaging of EdU, an alkyne-tagged cell proliferation probe, by Raman microscopy. J. Am. Chem. Soc. *133*, 6102–6105.

Yasuda, M., Takeshita, N., and Shigeto, S. (2019). Inhomogeneous molecular distributions and cytochrome types and redox states in fungal cells revealed by Raman hyperspectral imaging using multivariate curve resolution–alternating least squares. Anal. Chem. *91*, 12501–12508.

Yawata, Y., Kiyokawa, T., Kawamura, Y., Hirayama, T., Takabe, K., and Nomura, N. (2019). Intra- and interspecies variability of single-cell innate fluorescence signature of microbial cell. Appl. Environ. Microbiol. *85*, e00608–00619.

Zhao, Z., Shen, Y., Hu, F., and Min, W. (2017). Applications of vibrational tags in biological imaging by Raman microscopy. Analyst *142*, 4018–4029.

Zheng, Y.-T., Toyofuku, M., Nomura, N., and Shigeto, S. (2013). Correlation of carotenoid accumulation with aggregation and biofilm development in *Rhodococcus* sp. SD-74. Anal. Chem. *85*, 7295–7301.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
Escherichia coli	JCM	JCM 20135
Bacillus subtilis	JCM	JCM 1465^T
Thermus thermophilus	JCM	JCM 10941^T
Thermococcus kodakarensis	JCM	JCM 12380^T
Sulfolobus acidocaldarius	JCM	JCM 8929^T
Nitrososphaera viennensis	JCM	JCM 19564^T
Chemicals, peptides, and recombinant proteins		
LB broth, Lennox	Nacalai tesque	Cat# 20066-95
Bacto tryptic soy broth	Becton, Dickinson and Company	Cat# 211825
Bacto peptone	Becton, Dickinson and Company	Cat# 211677
Bacto yeast extract	Becton, Dickinson and Company	Cat# 212750
Deposited data		
Raman spectra of bacterial and archaeal cells	This paper; Mendeley Data	https://doi.org/10.17632/8cd34fckgt.1
Software and algorithms		
Igor Pro 8.04	WaveMetrics	https://www.wavemetrics.com/ RRID: SCR_014216
Python 3.7.6	Python Software Foundation	https://www.python.org/RRID: SCR_008394
scikit-learn version 0.22.1	Pedregosa et al. (2011)	https://scikit-learn.org RRID: SCR_002577
Adobe Illustrator	Adobe	https://www.adobe.com/products/illustrator.html RRID: SCR_010279

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Shinsuke Shigeto (shigeto@kwansei.ac.jp).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Original, unprocessed Raman spectral data from Figures 2 and S4 have been deposited at Mendeley Data and are publicly available as of the date of publication. The DOI is listed in the key resources table.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Escherichia coli JCM 20135, *Bacillus subtilis* JCM 1465^T, *Thermus thermophilus* JCM 10941^T, *Thermococcus kodakarensis* JCM 12380^T, *Sulfolobus acidocaldarius* JCM 8929^T, and *Nitrososphaera viennensis* JCM 19564^T were obtained from Japan Collection of Microorganisms (JCM). *E. coli* and *B. subtilis* were cultured, respectively, in LB broth and tryptic soy broth in a shaking incubator (180 rpm) at 30°C. *T. thermophilus* was

cultured in modified DSMZ medium 74 (pH 7.0–7.5) containing 4 g/L Bacto yeast extract, 8 g/L Bacto peptone, and 2 g/L NaCl, in a shaking incubator (180 rpm) at 75°C. The precultures of bacterial strains were grown for 1 day, and 50 μL of these precultures were then inoculated to 5 ml of fresh media. The overnight cultures were used for Raman measurements. Cell cultures of *T. kodakarensis*, *S. acidocaldarius*, and *N. viennensis* were used as received from JCM for Raman measurements within 5 days of shipping.

Cells of each strain were harvested and washed three times with PBS (pH 7.4) by centrifugation (8,000–10,000 × *g*, 30–60 s) at room temperature, followed by resuspension in PBS. The cell suspension was either diluted or concentrated for better microscopic observation of individual cells. When acquiring Raman spectral data for the construction of RF classifiers, 200 μL of the cell suspension of each strain was transferred to a glass bottom dish. For spectral acquisition from a mixed population, a 200 μL mixture of the cell suspensions of *B. subtilis*, *T. thermophilus*, and *N. viennensis* was transferred to a glass bottom dish.

METHOD DETAILS

Confocal Raman microspectroscopy

Single-cell Raman spectra were measured using a laboratory-built confocal Raman microspectrometer, which has been described previously (Huang et al., 2011, 2012; Matsuda et al., 2019). In brief, the 632.8 nm output of a He–Ne laser (Thorlabs, HNL210L) was used as the excitation light. The laser beam was magnified by a factor of ∼3 and subsequently introduced into an inverted microscope (Nikon, TE2000-U, customized) by a pair of an edge filter and a dichroic mirror. The beam was focused onto the sample with an oil-immersion phase-contrast objective (Nikon, 100×, NA 1.3, CFI Plan Fluor DLL), and backward Raman scattered light was collected with the same objective. After passing through a 100-μm pinhole for confocal detection, it was analyzed with an imaging spectrometer (SOL Instruments, MS3504i) and detected with an electron-multiplying charge-coupled device (CCD) detector with 200 × 1600 pixels (Andor Technology, DU970P-BVF). The use of a 600 grooves mm^{-1} grating enabled us to record the entire spectral window (660–3022 cm^{-1}) covering both fingerprint and CH-stretching regions with a spectral resolution of ∼5 cm^{-1}. The laser power at the sample point was adjusted to 3 mW. The Raman excitation beam that was tightly focused to a nearly diffraction-limited spot size (∼1 μm) grabbed and immobilized a randomly selected single prokaryotic cell via the optical tweezer technique (Xie et al., 2002) (Figure 1A). Only in the case of *T. thermophilus*, which tend to grow singly, in pairs, and in a chain in the polypeptone yeast extract medium (Oshima and Imahori, 1974), single to short-chain cells were trapped at their upper part, thereby enabling alignment of a rod-shaped cell(s) along the axial direction. The Raman spectrum of the trapped cell was measured with a 30 s exposure time without electron-multiplying gain. For each strain in a pure population, 40 spectra were acquired from 40 different cells. Those measurements were performed in triplicate. For the three strains in a mixed population, 20 or 21 spectra were obtained per strain. All spectroscopic measurements were done at room temperature.

Data preprocessing

The recorded Raman spectra were subjected to a series of preprocessing procedures (Figure 1B) prior to training and testing in machine learning. In principle, all of them were used without selection. However, about 20% of *S. acidocaldarius* cells showed strong autofluorescence, and their spectra were rejected from the dataset because the objective of the present work was to classify cells based on Raman spectral patterns and not autofluorescence patterns (Yawata et al., 2019). (1) The PBS spectrum (average of 10 spectra) was subtracted from each single-cell spectrum in which cosmic rays, if any, were manually removed in advance. (2) Raman spectra measured on different days typically have slightly different wavenumber regions, resulting in a shift in the horizontal axis of the spectra. This shift was corrected for so that emission lines of a standard neon lamp recorded on different days appeared at the same CCD pixel with a tolerance of ±0.5 pixel. (3) The so-called silent region (1801.9–2773.7 cm^{-1}) was deleted where there are no Raman bands with few exceptions (e.g., C≡C and C≡N stretching (Yamakoshi et al., 2011; Zhao et al., 2017)); only the fingerprint (664.4–1800.4 cm^{-1}) and CH-stretching (2774.9–3018.5 cm^{-1}) regions were retained. This trimming process was required to deal with the fingerprint and CH-stretching regions separately in the subsequent steps. (4) To remove the slope of each spectrum due possibly to autofluorescence and subtle differences in measurement conditions, baseline subtraction was performed separately for the fingerprint and CH-stretching regions. For both regions, the higher and lower wavenumber edges of the region were fit to a linear function, and this linear baseline was subtracted from the spectrum. (5) After baseline subtraction, vector normalization was conducted on each spectral fragment. The norm of the spectrum, which is defined as the square root of the sum of the squared intensities of the spectrum, was calculated, by which each

intensity was divided. The above baseline subtraction and vector normalization were not directly applicable to the raw CH-stretching spectrum. Particularly in *N. viennensis*, the CH-stretching spectrum had markedly low intensities and suffered from large noise due to the smallest cell size among the prokaryotic species studied here. To cope with this problem, spectral smoothing using 15-point Savitzky–Golay polynomial filters of degree 2 was first performed. Subsequently, a linear baseline was determined in the same manner as described above for the smoothed spectrum, and the resulting baseline was subtracted from both unsmoothed and smoothed spectra. The norm of the baseline-subtracted, smoothed spectrum was calculated to vector-normalize the baseline-subtracted, unsmoothed spectrum. Finally, the normalized, unsmoothed spectra in the fingerprint and CH-stretching regions were merged, yielding a preprocessed single-cell Raman spectrum with 896 pixels. All of the above preprocessing was performed on Igor Pro 8.04 (WaveMetrics).

Random forest training and test

RF models were constructed for classification among the three bacterial and three archaeal species. A total of 240 Raman spectra (40 spectra per prokaryotic species) were used as the dataset for model construction. First, this dataset was split into 10 folds, and 10 patterns of training (9 folds) and test (1 fold) sets were generated (10-fold cross-validation). Then, a grid search for hyper-parameter optimization was performed with 5-fold cross-validation on each training set. The hyper-parameters (n_estimaters and max_features) that were frequently adopted in the 10 models were used to search best parameters that achieve high classification accuracy in the test datasets. Finally, a RF classifier was built using these optimized hyper-parameters and all of the 240 Raman spectra. Using this classifier, a total of 62 Raman spectra acquired from the mixed population of *B. subtilis* (21 cells), *T. thermophilus* (20 cells), and *N. viennensis* (21 cells) were classified. The importance score of a feature was calculated as the total reduction of the criterion brought by that feature (known as Gini importance) in each validation. An average of the scores over 10 validations was used to represent the importance of each feature.

To visualize the differences among the spectra of the six prokaryotic species, PCA was also carried out. Both RF modeling and PCA were performed using the scikit-learn package (Pedregosa et al., 2011) (version 0.22.1) in Python (version 3.7.6).

QUANTIFICATION AND STATISTICAL ANALYSIS

For comparison of the averaged class probabilities in the classification of the six prokaryotic species, p values were calculated using two-tailed Welch's *t* test and adjusted using Bonferroni correction. For comparison of the CH-stretching Raman intensities, a nonparametric Kruskal–Wallis test was performed, followed by a post hoc Dunn–Holland–Wolfe test. p values <0.05 were considered significant in both cases. The above statistical tests were performed on Igor Pro 8.04.