Investigation of Raman spectroscopic signatures with multivariate statistics: an approach for cataloguing microbial biosignatures

Mitch W. Messmer, Markus Dieser, Heidi J. Smith, Albert E. Parker, Christine M. Foreman

Final publication is available from Mary Ann Liebert, Inc., publishers https://dx.doi.org/10.1089/ast.2021.0021

Investigation of Raman spectroscopic signatures with multivariate statistics: an approach for cataloguing microbial biosignatures

Mitch W. Messmer[1,2], Markus Dieser[1,2], Heidi J. Smith[1,3], Albert E. Parker[1,4], Christine M. Foreman[1,2,*]

[1] Center for Biofilm Engineering, Montana State University, Bozeman, MT 59717, USA
[2] Department of Chemical & Biological Engineering, Montana State University, Bozeman, MT 59717, USA
[3] Department of Microbiology and Cell Biology, Montana State University, Bozeman, MT 59717, USA
[4] Department of Mathematical Sciences, Montana State University, Bozeman, MT 59717, USA

[*] Corresponding Author:
Montana State University
Center for Biofilm Engineering
366 Barnard Hall
Bozeman, MT 59717, USA
Email: cforeman@montana.edu
Phone: 406-994-7361

Running title: Cataloguing Biosignatures

Manuscript Keywords:
Biosignatures, Raman spectroscopy, environmental isolates, Polar microbes, extraterrestrial life, unsupervised K-means clustering

Abstract:

Spectroscopic instruments are increasingly being implemented in the search for extraterrestrial life. However, microstructural spectral analyses of alien environments could prove difficult without knowledge on the molecular identification of individual spectral signatures. To bridge this gap, we introduce unsupervised K-means clustering as a statistical approach to discern spectral patterns of biosignatures without prior knowledge of spectral regions of biomolecules. Spectral profiles of bacterial isolates from analogous polar ice sheets were measured with Raman spectroscopy. Raman analysis identified carotenoid and violacein pigments, and key cellular features including saturated and unsaturated fats, triacylglycerols, and proteins. Principal component analysis and targeted spectra integration biplot analysis revealed that the clustering of bacterial isolates was attributed to spectral biosignatures influenced by carotenoid pigments and ratio of unsaturated/saturated fat peaks. Unsupervised K-means clustering highlighted the prevalence of the corresponding spectral peaks, while subsequent supervised permutational multivariate analysis of variance provided statistical validation for spectral differences associated with the identified cellular features. Establishing a validated catalog of spectral signatures of analogous biotic and abiotic materials, in combination with targeted supervised tools could prove effective at identifying extant biosignatures.

Introduction:

As mankind ventures further into the enigmatic depths of our solar system, the search for life beyond Earth is crucial for defining the context of human civilization and finding answers to the intrinsic question of Earth's biological solitude and the origins of life on our planet. Compelling evidence for the existence of ancient oceans and ice deposits on Mars and the presence of liquid oceans beneath the ice shell of moons has driven the search for biosignatures relevant to astrobiological investigations. Biosignatures range from microbial mineralization (e.g. Gleeson et al., 2012; Gaboyer et al., 2017; Levett et al., 2020), gases (e.g. Seager et al., 2012; Olson et al., 2018; Sousa-Silva et al., 2020), and stable isotopes (e.g. Anbar, 2004; Ono, 2008; Brady et al., 2013), to organics (e.g. Watanabe et al., 2000, 2004; Rye and Holland, 2000), among others. If the 'Rare Earth' hypothesis is correct (Ward and Brownlee, 2000), the search for extraterrestrial life is essentially limited to microorganisms and microbially-derived organic material such as lipids, hydrocarbons, or polysaccharides. The exploration for organic biosignatures on extraterrestrial bodies has been an ambitious endeavor, and the recent advancements in single-cell spectral analyses techniques may provide the resolution needed for unambiguous biosignature identification.

Spectral analysis tools are foundational to the exploration of extraterrestrial bodies. Landers, orbiting satellites, and observatories equipped with different spectrometers have provided invaluable information about planets and moons in our solar system and beyond (e.g. Giuranna et al., 2019; Lopes et al., 2019; Rapin et al., 2019; Hansen et al., 2020; Kunimoto et al., 2020; Paganini et al., 2020). While NASA's previous missions to Mars have exploited various forms of spectroscopy for geological applications, these approaches have uncovered details about the history of liquid water on the planet (Schmidt et al., 2008); an essential prerequisite for life

(Schwieterman et al., 2018). Moving forward, future missions will use spectroscopic technologies to search directly for signs of microbial life (Rull et al., 2017; Williford et al., 2018) and the positive, unambiguous identification of biomaterials indicative of extinct or extant life is intimately tied to this undertaking (Edwards et al., 2014).

Raman spectroscopy is a common spectral analysis technique that can be used to non-destructively identify biological compounds (Jehlicka et al., 2013; Rull et al., 2017). An important attribute of Raman spectroscopy when applied to the analysis of biomarkers is its ability to differentiate between molecular species based on intrinsic spectral signatures. Raman spectroscopy has been used to detect cellular components such as lipids, proteins, nucleotides, carbohydrates, and pigments in both algal and bacterial cells (Ermakov et al., 2005; Tuma et al., 2005; Huang et al., 2010; Wu et al., 2011; Jehlicka et al., 2013, 2014, 2015; Oren et al., 2015; Serrano et al., 2015; Yakubovskaya et al., 2019) to profile organic and inorganic substances (De Gelder et al., 2007), and to locate the presence of specific organic compounds in heterogeneous environmental samples (Wei et al., 2015; Tan et al., 2018).

The recently launched mission of NASA's Mars 2020 *Perseverance* rover includes both a Raman spectrometer with a remote pulsed 532 nm laser on the SuperCam instrument and a UV Raman spectrometer on the Scanning Habitable Environments with Raman and Luminescence for Organics and Chemicals (SHERLOC) instrument to specifically look for signs of ancient life (Wiens et al., 2021). Previous studies have determined the feasibility of detecting biosignatures within Martian geologic analogs using Raman spectroscopy techniques (Edwareds et al., 2013). However, even in defined samples containing a mixture of abiotic and biotic components, the unambiguous identification of biological material can be extremely difficult. This identification becomes even more challenging when analyzing extraterrestrial samples with unknown

composition. Thus, it is both timely and critical to develop an understanding of Raman spectral biosignatures representative of microbial life in order to accurately interpret spectral data.

The objective of this research was to provide a robust statistical approach to the analysis of microbial biosignatures generated by Raman spectroscopy. Bacteria isolated from the Greenland and Antarctic Ice Sheets were selected as model organisms, as these environments have long been recognized as analogs to habitable icy worlds (Price, 2000; Olcott Marshal and Cestari, 2015; Rodriguez e Silva et al., 2021). To identify the important spectral information from microbial cells, a principal component analysis (PCA) of cellular components was conducted. Spectral differences with respect to specific cellular features were visualized by a targeted spectra integration (TSI) biplot and statistically validated with supervised permutational (*i.e.*, non-parametric) multivariate analysis of variance (PERMANOVA). Ultimately, by not assuming any prior spectral knowledge of cellular components, an unsupervised K-means clustering was performed. Our statistical approach was subsequently validated using publicly available datasets. This multivariate statistical approach allowed for identification of trends in spectral data based on the presence of cellular biosignatures and determined statistically significant differences between spectra. The statistical affirmation of the spectral "fingerprints" of biogenic material may be used to guide the search for life beyond Earth in unknown materials.

Materials and Methods:

Microbial Isolates and Growth Conditions:

Twenty-five selected bacterial isolates (Supplemental Table 1) were studied from the Foreman Lab Isolate Collection, which were previously isolated from Pony Lake (PL) located at Cape Royds on Ross Island, Antarctica (77°33′S, 166°00′E), the supraglacial Cotton Glacier

stream (CG) in the Transantarctic Mountains, Antarctica (77°07′S, 161°40′E), and cryoconite holes on Russell Glacier (GRIS), Greenland (67°5'N, 50°0'W) (Figure 1). Bacterial strains were isolated on R2A agar media (Difco™) at 4°C. Single colonies from each site were selected according to morphological characteristics including color, size, and shape. Purified isolates were stored in 40% glycerol at -80°C and archived as part of the Foreman Lab Isolate Collection.

Organisms from freezer stocks were streaked on R2A agar plates and incubated at 4°C. Additionally, *Escherichia coli* AW3110 and *Bacillus subtilis* were grown at 37°C, respectively. A sporulation medium (0.8% w/v Nutrient Broth, 0.1% w/v KCl, 1 mM $MgSO_4 \cdot 7H_2O$, 0.01 mM $MnCl_2$, 0.5 mM $CaCl_2$, 1 mM $FeSO_4 \cdot 7H_2O$; Schaeffer et al., 1965) was used to induce spore formation of *B. subtilis*. Spores were purified by a lysozyme treatment (15 mg/mL) for 1 hour at 22°C, followed by centrifugations at 13,000 g for 1 minute and six washes with 1X phosphate buffered saline (PBS). The final spore pellet was resuspended in 1 mL PBS buffer and stored at -20°C prior to analysis.

DNA Extraction and Amplification:

Genomic DNA was extracted using the PowerSoil DNA Isolation Kit (Qiagen, U.S.A.) following manufacturer's recommendations. PCR amplification of the 16S gene was performed with primers 9F 5'- GAGTTTGATCCTGGCTCAG-3' and 1492R 5'-GGTTACCTTGTTACGACTT -3' (Stackebrandt et al., 1993). PCR products were sent to Functional Biosciences (Madison, Wisconsin, U.S.A.) for Sanger sequencing. Sequences were assembled using the BioEdit 7.2 software package (Informer Technologies, U.S.A.). The nucleotide sequences were compared with the National Center for Biotechnology Information (NCBI) nucleotide database using the Nucleotide Basic Local Alignment Search Tool (BLAST)

(Altschul et al., 1990). Sequences were deposited in the NCBI database under the accession numbers GU592435, GU592436, MW293983-MW294005.

Raman Spectroscopy:

Colonies from environmental bacterial isolates (Supplemental Table 1), *E. coli* AW3110, and *B. subtilis* were scraped with a sterile loop into microcentrifuge tubes containing 1 mL 1X PBS buffer. Cells were fixed with paraformaldehyde (final conc. 2%) at 4°C for 2 hours, as paraformaldehyde has not been shown to impact Raman signal (Berry et al., 2014). After fixation, samples were centrifuged at 12,000 g for 5 minutes and washed three times with 1X PBS buffer. The final cell pellet was resuspended in 1 ml of a 1:1 mixture of 1X PBS buffer and 100% molecular grade ethanol. All samples including spores (see above) were stored at -20°C untill subjected to Raman spectroscopy.

To analyze the samples on the Raman spectrometer, 3 µL of each sample were placed onto a clean CaF_2 Raman slide. Samples were dried at 40°C for 5 minutes. The slides were dipped in Milli-Q water to remove PBS buffer residues and immediately dried with compressed air. The samples were analyzed on a Raman spectrometer (LabRAM HR Evolution Raman Spectrometer, HORIBA) equipped with the LabSpec6 version 6.5.1 software package (HORIBA, Japan). Raman spectra (300-3300 cm^{-1}) of individual cells were recorded using a 532 nm laser at 45 mW of power with a 600 nm grating and a 100X objective (NA 0.90). The parameters used for each spectral measurement were two accumulations of ten seconds. Raman spectra were recorded for twenty individual cells for each isolate. The resulting spectra were processed using a fitted 4^{th} degree polynomial line for baseline subtraction and the LabSpec6 4^{th} degree data smoothing function (Savitsky-Golay algorithm) to reduce the noise to signal ratio of the measurements (Savitsky and Golay, 1964; Cai at el., 2018; Horiba, 2018).

Statistical Analyses:

After the baseline subtraction and data smoothing steps, the Raman spectra were subjected to multiple multivariate statistical analyses. First, a PCA of the Raman spectra was conducted using the Eigenvector SOLO MIA version 8.7.1 software package (Eigenvector Research, Inc., U.S.A.). Exclusively for the PCA, the spectra were preprocessed using multiplicative signal correction (MSC) (Martens and Stark, 1991). A PCA scatterplot and corresponding principal component graphs were created to illustrate the regions of the spectra that accounted for the most variance between isolates (Supplemental Figures 1-2).

Raman spectral regions, which were associated with key cellular features, were further analyzed in detail with a secondary analysis. Raman spectra were normalized by dividing all intensity values by the intensity of the lipid peak between 2900-2930 cm^{-1} to reduce the impact of the varying amplitude of this spectral feature for this step and all subsequent analyses. The areas under the Raman spectra that corresponded to carotenoid pigments, saturated and unsaturated fats, triacylglycerols, and proteins were integrated to calculate a single cumulative value for each feature (Table 1). It should be noted that spectral overlap can exist between cellular features, which elicits the use of multiple spectral features to identify some compounds. A second PCA was performed to create the TSI biplot with loadings associated with each of the aforementioned parameters. To identify differences based on carotenoids, saturated and unsaturated fats, proteins, and triacylglycerols (Table 1), a PERMANOVA with respect to colony color was performed using the “adonis” and “pairwise adonis” functions in the “vegan” version 2.5.6 package (Dixon, 2003; Oksanen et al., 2019) in the statistical software, R (R Core Team 2020). The normalized spectra were analyzed using PERMANOVA from 550-1800 cm^{-1}

to remove the peak from the CaF_2 Raman slide (321 cm^{-1}) and the noise from the characteristic "silent region" (1800-2800 cm^{-1}) of the spectra.

For the unsupervised analysis, a non-targeted analysis of similarity (ANOSIM) was performed on the normalized, full Raman spectra between 550 and 1800 cm^{-1} using the "vegan" package in R to quantify intraspecies dissimilarity in relation to the variance between isolates. The "kmeans" and "fviz_cluster" functions (Kassambara, 2017) were used to create an unsupervised K-means cluster plot that grouped isolates based on their spectral features. Other clustering methods like fuzzy c-means or information theoretic approaches that assign clusters with a probability could also be applied, warranting future research (Dunn, 1973; Parker et al. 2010). This statistical approach was subsequently applied to additional Raman spectra from literature to demonstrate the replicability and applicability of the model. Twenty individual spectra were selected from twelve carotenoid pigmented bacterial isolates and from endospores from seven *Bacillus* spp.. Results from these additional organisms were used not in the interpretation of the spectra from this project, but rather as another way to test the applicability and practically of these analytical modeling techniques.

Results:

Characteristics of Bacterial Isolates:

Bacterial isolates were closely related to the genera *Glaciihabitans, Polaromonas, Herminiimonas, Undibacterium, Janthinobacterium, Variovorax, Cryobacterium, Actimicrobium, Hymenobacter, and Flavobacterium* (Supplemental Table 1), and represented lineages of the phyla Actinobacteria (N=3) and Bacteroidetes (N=6), and the order Burkholderiales (N=16). When grown on R2A agar plates, isolates *Glaciihabitans* sp. GRIS 2-

15, *Variovorax* sp. GRIS 2-14, *Cryobacterium* sp. GRIS 2-6, *Hymenobacter* sp. GRIS 2-18, *Flavobacterium* sp. CG 9-1, *Cryobacterium* sp. CG 9-6, *Flavobacterium* sp. CG 23-5, *Flavobacterium* sp. PL 10, *Flavobacterium* sp. PL 11, and *Flavobacterium* sp. PL 12 formed bright yellow, orange, or red pigments (Supplemental Table 1). Of the eight bacterial isolates that were closely related to *Janthinobacterium* spp., only *Janthinobacterium* sp. GRIS 1-11 produced purple pigments; a common characteristic of this genus. All other isolates were either cream-colored or had colonies that appeared in a salmon-pink tone.

Raman Spectroscopy:

The Raman spectra for the orange pigmented isolates *Glaciihabitans* sp. GRIS 2-15, *Flavobacterium* sp. PL 11, and *Flavobacterium* sp. CG 9-1, as well as the red isolates *Hymenobacter* sp. GRIS 2-18 and *Cryobacterium* sp. CG 9-6, showed pronounced peaks between 988-1000 cm^{-1}, 1138-1150 cm^{-1}, and 1498-1510 cm^{-1} (Figure 2A). Smaller peaks in the spectra of these isolates were identified between 2006-2024 cm^{-1}, 2141-2153 cm^{-1}, 2288-2303 cm^{-1}, 2504-2513 cm^{-1}, and 2645-2657 cm^{-1}. These characteristic Raman peaks were absent in the spectra for all yellow pigmented, pink-colored, and non-pigmented environmental isolates (Figure 2B-D; Supplemental Figure 3), as well as the lab strains (*E. coli* AW3110 and *B. subtilis* and spores; Figure 2F). High autofluorescence of the purple isolate *Janthinobacterium* sp. GRIS 1-11 caused difficulty discerning key features of the full spectra without a high degree of baseline subtraction. For this reason, the spectra for *Janthinobacterium* sp. GRIS 1-11 were excluded from further analyses. Nonetheless, with focus on 400-1800 cm^{-1}, the Raman spectra for *Janthinobacterium* sp. GRIS 1-11 showed peaks at 1130 cm^{-1}, 1275 cm^{-1}, and 1550 cm^{-1} (Figure 2E).

Common spectral peaks highlighted by Raman spectroscopy that were related to lipids included a predominant peak between 2900-2930 cm^{-1} (Table 2; Figure 2), and distinguishable peaks at 1260 cm^{-1}, 1300 cm^{-1}, 1440 cm^{-1}, 1650 cm^{-1}, 2800-3000 cm^{-1}, and 3023 cm^{-1}. Peaks around 1300 cm^{-1}, 1460 cm^{-1}, and 1630-1680 cm^{-1} may also be attributed to triacylglycerols. Peaks assigned to proteins were observed around 1004 cm^{-1}, 1300 cm^{-1}, 1550 cm^{-1}, and 1600-1700 cm^{-1}. Small spectral features corresponding to the presence of nucleotides were observed across almost all samples from 600-800 cm^{-1}. *B. subtilis* spores maintained the same spectral features as their vegetative counterparts, albeit producing stronger signals (Figure 2F).

PCA clustered spores, red/orange pigmented isolates, and non-pigmented/pink-colored/yellow pigmented bacterial isolates into three distinct groups (Supplemental Figure 1). Additionally, to determine whether spectral signatures are characteristic of certain bacterial isolates, a TSI biplot was generated (Figure 3) to visualize the variability among the 5 cellular features listed in Table 1. The first principle component (PC1) explained 94.97% of the total variation. Spectral signatures of the orange pigmented *Flavobacterium* sp. PL 11 were best explained by the presence of triacylglycerols. *Flavobacterium* sp. CG 9-1 and *Glaciihabitans* sp.GRIS 2-15 were described by their carotenoid pigment spectra and protein peaks. Spectra of *Hymenobacter* sp. GRIS 2-18 were moderately influenced by proteins. All other bacterial isolates produced Raman spectra largely described by fatty acid peaks. Due to the overlap of spectral peaks for saturated and unsaturated fats, the angle between these two loading vectors was very narrow.

For the supervised approach, a pairwise PERMANOVA confirmed significant differences in Raman spectra between isolates (Table 3). Overall, all pairings showed statistically significant differences across different regions of the Raman spectra, with the exception of certain pairings

253 with pink-colored, yellow pigmented, and non-pigmented organisms. All pigment groups were
254 statistically significantly different when comparing the full Raman spectra (550-1800 cm^{-1}), or
255 just the carotenoid regions (Table 3). For the protein and triacylglycerol regions, only the yellow
256 and non-pigmented bacterial isolates were not statistically significantly different ($P \geq 0.056$).
257 Collectively, Raman spectra for non-pigmented isolates were not statistically significantly
258 different than other groups across the saturated fat region ($P \geq 0.132$). In addition, pairwise
259 PERMANOVA revealed that Raman spectra were more similar with respect to saturated fat
260 features as opposed to unsaturated fats.

261 The non-targeted ANOSIM indicated dissimilarity between the Raman spectra of the
262 different isolates (R=0.891, P=0.001). Subsequent unsupervised K-means analysis grouped
263 Raman spectra into four clusters and reflected trends based on pigmentation and prevalent
264 macromolecules (Figure 4A and Supplemental Table 3). Raman spectra for *B. subtilis* cells
265 grouped as an individual cluster (Cluster 1). Overall, Clusters 1, 3, and 4 exhibited similar
266 spectral characteristics, but were distinguished by the relative ratios of the intensities of each
267 peaks. Cluster mean 1 was defined by spectral features corresponding to nucleotides around 800
268 cm^{-1} (Barhoumi et al., 2008; Parab and Tomar, 2012), proteins near 1004 cm^{-1} (De Gelder et al.,
269 2007; Wu et al., 2011), potential triacylglycerols at 1150 cm^{-1} (Motoyama, 2012), amide (II and
270 III) bands at 1550 cm^{-1} and 1300 cm^{-1} (De Gelder et al., 2007; Wu et al., 2011), saturated fats at
271 1440 cm^{-1}, and unsaturated fats at 1660 cm^{-1} (Samek et al., 2010; Czmara et al., 2014; Sharma et
272 al., 2015). Red and orange pigmented bacterial isolates formed Cluster 2. The cluster mean for
273 the orange and red cells showed three peaks characteristic for carotenoids at 1006 cm^{-1}, 1154 cm^{-1}
274 , and 1520 cm^{-1} (Ermakov et al., 2005; Wood et al., 2005; De Gelder et al., 2007) (Figure 4B).
275 Conversely, non-pigmented, pink-colored, and yellow pigmented cells spanned across Clusters 3

and 4 with differences in cluster means largely driven by a nucleotide peak (Barhoumi et al., 2008; Parab and Tomar, 2012) around 800 cm^{-1} (Figure 4B). Noteworthy, while other spectral tool such as UV-VIS and Fourier-transform infrared spectroscopy can reveal spectra of microbial features (Supplemental Figure 5-11), they perform bulk measurements on samples. As such, these tools lack replicate measurements on single cells across a sample necessary for the supervised and unsupervised approaches.

While differences in the acquisition of spectra did not allow for a direct incorporation of the literature data into the existing unsupervised statistical model, analysis of these external datasets supported the model for distinguishing isolates based on prevalent Raman features. The literature dataset exhibited moderate dissimilarity between Raman spectra of different isolates (ANOSIM; R=0.6311, *P*=0.001) and showed trends in the K-means clusters based on carotenoid pigmentation and prevalence of saturated and unsaturated fat peaks (Supplemental Figure 4 and Supplemental Table 4). Akin to the K-means analysis of the polar isolates, spectral peaks associated with carotenoids were the most prominent and overshadowed lesser features in Cluster means 2 and 4. The sporulated isolates separated into their own grouping in Cluster 3. Despite sharing peaks associated with nucleotides, amides, saturated and unsaturated fats, moderate carotenoid peaks at 1150 cm^{-1}, and 1500 cm^{-1} were the differentiating factors between Cluster means 1 and 3, which consisted of less pigmented isolates and non-pigmented spores, respectively (De Gelder et al., 2007; Wu et al., 2011). Cluster means 2 and 4 reflected similar spectral features and were predominantly defined by the relatively high intensities of the aforementioned carotenoid peaks.

Discussion:

Spectra measured by Raman spectroscopy encompassed many cellular features fundamental to microbial life. Though spectral features for proteins, saturated and unsaturated fats, and pigments were identified, peaks affiliated with carotenoid pigments were surprisingly not observed for the yellow or pink-colored isolates. This reduced resolving power of Raman spectroscopy could be in part attributed to intact cells analyses, as purified pigment extracts produce clearer Raman characteristics (Ermakov et al., 2005; Wood et al., 2005; De Gelder et al., 2007). Appropriate acquisition parameters are also paramount to measuring well-defined spectra, as excessive laser power or exposure time could have led to bleaching or lysing of the biological samples. Raman peaks for isolates *Glaciihabitans* sp. GRIS 2-15, *Hymenobacter* sp. GRIS 2-18, *Cryobacterium* sp. CG 9-6, and *Flavobacterium* sp. PL 11, and *Flavobacterium* sp. CG 9-1 around 1006 cm^{-1}, 1154 cm^{-1}, and 1520 cm^{-1} were suggestive of υ_3 in-plane rocking of CH_3 groups, υ_2 in-phase C-C bond stretching, and υ_1 in-phase C=C bond stretching, respectively; spectral bands characteristic for carotenoid pigments (Ermakov et al., 2005; Wood et al., 2005; De Gelder et al., 2007; Jehlicka et al. 2014). Small peaks between 2000-2650 cm^{-1} further affirmed the presence of carotenoids (Ou-Yang et al., 2015). Despite the high autofluorescence of the purple bacterium *Janthinobacterium* sp. GRIS 1-11 above 2000 cm^{-1}, Raman spectra clearly exhibited peaks distinctive for violacein (Jehlicka et al. in 2015). In addition to pigments, Raman analysis successfully detected key cellular features across all bacterial isolates (Table 2), including membrane phospholipids as the most prominent biosignature (Kint et al., 1992; Parab and Tomar, 2012), unsaturated and saturated fatty acids (Samek et al., 2010; Wu et al., 2011), triacylglycerols (Samek et al., 2010; Motoyama, 2012), and amides groups I-III (Wu et al., 2011). Spectral overlap between protein and pigment peaks (Ermakov et al., 2005; Wood et al., 2005; De Gelder et al., 2007) or fatty acids and triacylglycerols (Motoyama, 2012; Czmara et al.,

2014) likely contributed to either cellular compound when present concomitantly. The spectral ratios of saturated and unsaturated fatty acids would allow for an approximation of the degree of membrane lipid saturation (Czmara et al., 2014). However, while it was expected that bacteria growing at lower temperatures incorporate more unsaturated fats into their membranes (Vokt and Brody, 1985; Neidleman, 1987), no clear trend in fatty acid saturation was determined based on growth temperature (Supplemental Table 2). Using membrane lipids as a biomarker, Raman analysis was also able to differentiate between vegetative *B. subtilis* cells and spores. *B. subtilis* spores had much higher spectral ratios of unsaturated to saturated fats compared to vegetative cells, and the thick spore envelope likely intensified Raman signals (Soni et al., 2016).

While the TSI biplot and PERMANOVA were very effective at visualizing and statistically interpreting spectral differences, respectively, they are supervised approaches that require *a priori* knowledge of the dataset to reduce the analytical complexity and focus targeted regions of the Raman spectra (Maisa et al., 2013; Datta et al., 2017). Hence, the supervised statistical analysis is inherently limited to the selected spectral characteristics (Table 1). Conversely, K-means is an unsupervised algorithm that infers clustering of a set of samples without any initial knowledge on features of interest (Maisa et al., 2013; Datta et al., 2017; Szalontai et al., 2020). Importantly, this generalized analysis technique can be applied to any dataset without the need for assuming properties of the samples. Thus, the wide scope of the unsupervised analysis presents an advantage for the detection and identification of extraterrestrial biosignatures that are potentially hidden within the microscale of different molecules of unknown composition. In addition, an unsupervised approach could subsequently inform target regions for a supervised spectral analysis (Khodabakhshian and Abbaspour-Fard, 2020; Szalontai et al., 2020).

Conclusion:

Technological advances that drive astrobiology will undoubtedly shape the future as mankind continues to extend its reach beyond Earth into the solar system. As the use of spectral analyses for biological detection advances, it becomes increasingly imperative to have reference databases for spectroscopic signatures to interpret future data collected on Mars and beyond. Central to this goal is the need for analytical capabilities to (*i*) interpret large volumes of spectral profiles without prior knowledge of the sample, (*ii*) integrate multiple spectral datasets, and (*iii*) ultimately determine whether these signatures are representative of extinct or extant life. Measuring biosignatures of organisms from analogous environments through a variety of analytical techniques could inform site selection for future rover exploration and guide analyses for NASA's missions to Mars.

Overall, this study recapitulated the capabilities of Raman spectroscopy to detect spectral features that are indicative of the presence of life and expands upon the growing scientific field that may help guide more detailed analyses of extraterrestrial spectral samples. Among other cellular features, spectral analyses focused on carotenoid pigmented bacterial cells, spores, and model extraterrestrial organisms such as *Deinococcus radiodurans*, which have all been proposed as potential targets for extraterrestrial signs of life (Dartnell et al., 2010). Unsupervised K-means clustering successfully identified Raman spectral fingerprints of these astrobiologically-relevant biosignatures without requiring any prior knowledge of biological composition. This statistical model, however, also highlighted that not all carotenoid pigments and spores were distinct in their spectral information, requiring a more detailed chemical knowledge of cellular components to interpret the unsupervised clusters.

We propose unsupervised statistical analyses, such as K-means clustering, as an essential analytical method for spectral data, as this generic approach can be widely applied in extraplanetary applications. Further, establishing a catalog of microbial biosignatures from analogous environments, organic molecules, and inorganic substrates will aid the interpretation of spectral trends from these analyses and can allow for the use of supervised models (e.g. TSI biplot, PERMANOVA) in targeted, secondary analyses (Edwards et al., 2013; Lafuente et al., 2015; Mendili et al., 2019). Incorporating abiotic Raman signatures of relevant analog geological material should also be considered to further strengthen the model and our ability to unequivocally interpret the biogenicity of unknown materials. Building on this database will help with interpreting spectroscopic data, as well as shape the technology that pioneers the exploration of our neighboring worlds.

Acknowledgements:

We would like to express our gratitude to our colleagues at the Center for Biofilm Engineering, Montana State University, and all members of the Foreman Research Lab for their support. We would also like to thank Dr. Jürgen Popp for providing us with the additional Raman data.

Author Contributions:

CMF, MWM, and MD conceived the study. MWM, MD, HJS and CMF developed the analytical framework. MWM and AP analyzed the data. All authors wrote the manuscript.

Author Disclosure:

The authors declare no financial interests. The funders had no role in the design of the study, in the collection, analyses or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

Funding:

Funding was provided by NASA Exobiology grant 80NSSC18K0814 to C.M.F.. Initial support for collection of the polar microbial isolates came from National Science Foundation (NSF) grants 0338342 and 0838970 to C.M.F.. Montana State University's Confocal Raman microscope was acquired with support by the National Science Foundation (DBI-1726561) and the M.J. Murdock Charitable Trust (SR-2017331). The content is solely the responsibility of the authors and does not necessarily represent the official views of NASA and NSF.

References:

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of Molecular Biology* 1990;215:403-410.

Anbar AD. Iron stable isotopes: beyond biosignatures. *Earth and Planetary Science Letters* 2004;217:223-236.

Barhoumi A, Zhang D, Tam F, Halas NJ. Surface-enhanced Raman spectroscopy of DNA. *Journal of the American Chemical Society* 2008;130:5523–5529.

Berry D, Mader E, Lee KT, Woebken D, Wang Y, Zhu D, Palatinszky M, Schintlmeister A, Schmid MC, Hanson BT, Shterzer N, Mizrahi I, Rauch I, Decker T, Bocklitz T, Popp J, Gibson CM, Fowler PW, Huang WE, Wagner M. Tracking heavy water (D2O) incorporation for identifying and sorting active microbial cells. *Proceedings of the National Academy of Sciences of the United States of America* 2014;112:E194-203.

Brady AL, Druschel G, Leoni L, Lim DSS, Slater GF. Isotopic biosignatures in carbonate-rich, cyanobacteria-dominated microbial mats of the Cariboo Plateau, BC. *Geobiology* 2013;11:437-456.

Cai Y, Yang C, Xu D, Gui W. Baseline correction for Raman spectra using penalized spline smoothing based on vector transformation. *Analytical Methods* 2018;10:3525-3533.

Czmara K, Majzner K, Pacia MZ, Kochan K, Kaczor A, Baranska, M. Raman spectroscopy of lipids: a review. *Journal of Raman Spectroscopy* 2014;47:4-20.

Dartnell LR, Hunter SJ, Lovell KV, Coates AJ, Ward JM. Low-temperature ionizing radiation resistance of *Deinococcus radiodurans* and Antarctic Dry Valley bacteria. *Astrobiology* 2010;10:717-732.

Datta A, Ghosh S, Ghosh A. Unsupervised band extraction for hyperspectral images using clustering and kernel principal component analysis. *International Journal of Remote Sensing* 2017;38:850-873.

De Gelder J, De Gussem K, Vandenabeele P, Moens L. Reference database of Raman spectra of biological molecules. *Journal of Raman Spectroscopy* 2007;38:1133-1147.

Dixon P. VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science* 2003;14:927-930.

Dunn JC. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics* 1973;3:32-57.

Edwards HGM, Hutchinson IB, Ingley R, Jehlička J. Biomarkers and their Raman spectroscopic signatures: a spectral challenge for analytical astrobiology. *Philosophical Transactions of the Royal Society A* 2014;372:20140193

Edwards HGM, Hutchinson IB, Ingley R, Parnell J, Vitek P, Jehlička J. Raman spectroscopic analysis of geological and biogeological specimens of relevance to the ExoMars mission. *Astrobiology* 2013;13:543-549.

Ermakov IV, Sharifzadeh M, Ermokova M, Gellermann W. Resonance Raman detection of carotenoid antioxidants in living human tissue. *Journal of Biomedical Optics* 2005;10:064028.

Gaboyer F, Le Milbeau C, Bohmeier M, Schwendner P, Vannier P, Beblo-Vranesevic K, Rabbow E, Foucher F, Gautret P, Guégan R, Richard A, Sauldubois A, Richmann P, Perras AK, Moissl-Eichinger C, Cockell CS, Rettberg P, Marteinsson, Monaghan E, Ehrenfreund P, Garcia-Descalzo L, Gomez F, Malki M, Amils R, Cabezas P, Walter N, Westall F.

Mineralization and preservation of an extremotolerant bacterium isolated from an early Mars analog environment. *Scientific Reports* 2017;7:8775.
Giuranna M, Viscardy S, Daerden F, Neary L, Etiope G, Oehler D, Formisano V, Aronica A, Wolkenberg P, Aoki S, Cardesín-Moinelo A, Marín-Yaseli de la Parra J, Merritt D, Amoroso M. Independent confirmation of a methane spike on Mars and a source region east of Gale Crater. *Nature Geoscience* 2019;12:326-332.
Gleeson DF, Pappalardo RT, Anderson MS, Grasby SE, Mielke RE, Wright KE, Templeton AS. Biosignature detection at an Arctic analog to Europa. *Astrobiology* 2012;12:135-150.
Hansen CJ, Esposito LW, Colwell JE, Hendrix AR, Portyankina G, Stewart AIF, West RA. The composition and structure of Enceladus' Plume from the complete set of Cassini UVIS occultation observations. *Icarus* 2020;344:113461.
Horiba. Chapter 6 control panel. In *LapSpec 6 spectroscopy suit reference manual*, France, 2018, pp 68-71.
Huang YY, Beal CM, Cai WW, Ruoff RS, Terentjev EM. Micro-Raman spectroscopy of algae: composition analysis and fluorescence background behavior. *Biotechnology and Bioengineering* 2010;105:889-898.
Jehlicka J, Edwards HGM, Nemec I, Oren A. Raman spectroscopic study of the *Chromobacterium violaceum* pigment violacein using multiwavelength excitation and DFT calculations. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 2015;151:459-467.
Jehlicka J, Edwards HGM, Oren A. Raman spectroscopy of microbial pigments. *Applied and Environmental Microbiology* 2014;80:3286-3295.
Jehlicka J, Osterrothova K, Oren A, Edwards HGM. Raman spectroscopic discrimination of flexirubin pigments from two genera of *Bacteroidetes*. *FEMS Microbiology Letters* 2013;348:97-102.
Kassambara, A. Practical guide to cluster analysis in R: unsupervised machine learning, STHDA, 2017.
Khodabakhshian R, Abbaspour-Fard MH. Pattern recognition-based Raman spectroscopy for non-destructive detection of pomegranates during maturity. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 2020;231:118127.
Kint S, Wermer PH, Scherer JR. Raman spectra of hydrated phospholipid bilayers. 2. water and head-group interactions. *Journal of Physical Chemistry* 1992;96:446-452.
Kumar V, Kampe B, Rösch P, Popp J. Classification and identification of pigmented cocci bacteria relevant to the soil environment via Raman spectroscopy. *Environmental Science and Pollution Research* 2015; 22:19317-19325.
Kunimoto M, Matthews JM. Searching the entirety of Kepler Data. II. occurrence rate estimates for FGK stars. *The Astronomical Journal* 2020;159:248.
Lafuente B, Downs R T, Yang H, Stone N. The power of databases: the RRUFF project. In: *Highlights in Mineralogical Crystallography*, edited by T Armbruster and RM Danisi, Berlin, Germany, W. De Gruyter, 2015, pp 1-30.
Levett A, Gagen EJ, Rintoul L, Guagliardo P, Diao H, Vasconcelos PM, Southam G. Characterisation of iron oxide encrusted microbial fossils. *Scientific Reports* 2020;10:9889.
Lopes RMC, Wall SD, Elachi C, Birch SPD, Corlies P, Coustenis A, Hayes AG, Hofgartner JD, Janssen MA, Kirk RL, LeGall A, Lorenz RD, Lunine JI, Malaska MJ, Mastroguiseppe M, Mitri G, Neish CD, Notarnicola C, Paganelli F, Paillou P, Poggiali V, Radebaugh J, Rodriguez S, Schoenfeld A, Soderblom JM, Solomonidou A, Stofan ER, Stiles BW, Tosi F,

493 Turtle EP, West RD, Wood CA, Zebker HA, Barnes JW, Casarano D, Encrenaz P, Farr T,
494 Grima C, Hemingway D, Karatekin O, Lucas A, Mitchell KL, Ori G, Orosei R, Ries P, Riccio
495 D, Soderblom LA, Zhang Z. Titan as revealed by the Cassini Radar. *Space Science Reviews*
496 2019;215:33.
497 Martens H, Stark E. Extended multiplicative signal correction and spectral interference
498 subtraction: New preprocessing methods for near infrared spectroscopy. *Journal of*
499 *Pharmaceutical and Biomedical Analysis* 1991;9:625-635.
500 Masia F, Glen A, Stephens P, Borri P, Langbein W. Quantitative chemical imaging and
501 unsupervised analysis using hyperspectral coherent Anti-Stokes Raman scattering
502 microscopy. *Analytical Chemistry* 2013;85:10820-10828.
503 Mendili YE, Vaitkus A, Merkys A, Gražulis S, Chateigner D, Mathevet F, Gascoin S, Petit S,
504 Bardeau JF, Zanatta M, Secchi M, Mariotto G, Kumar A, Cassetta M, Lutterotti L, Borovin E,
505 Orberger B, Simon P, Hehlen B, Le Guen M. Raman Open Database: first interconnected
506 Raman-X-ray diffraction open-access resource for material identification. *Journal of Applied*
507 *Crystallography* 2019;52:618-625.
508 Motoyama M. Structure and phase characterization of triacylglycerols by Raman spectroscopy.
509 *National Agriculture and Food Research Organization (NARO) Institute of Livestock and*
510 *Grassland Sciences* 2012;12:19-68.
511 Neidleman SL. Effects of temperature on lipid unsaturation. *Biotechnology and Genetic*
512 *Engineering Review* 1987;5:245-268.
513 Oksanen J, Guillaume Blanchet F, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR,
514 O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagne H. Package 'vegan':
515 vegan: Community Ecology Package. R package version 2.5-6. 2019. https://CRAN.R-
516 project.org/package=vegan.
517 Olcott Marshall A, Cestari NA. Biomarker analysis of samples visually identified as microbial in
518 the Eocene Green River formation: an analogue for Mars. *Astrobiology* 2015;15:770-775.
519 Olson SL, Schwieterman EW, Reinhard CT, Ridgwell A, Kane SR, Meadows VS, Lyons TW.
520 Atmospheric seasonality as an exoplanet biosignature. *The Astrophysical Journal Letters*
521 2018;858:L14.
522 Ono S. Multiple-sulphur isotope biosignatures. *Space Science Reviews* 2008;135:203-220.
523 Oren A, Mana L, Jehlicka J. Probing singles cells of purple sulfur bacteria with Raman
524 spectroscopy: carotenoids and elemental sulfur. *FEMS Microbiology Letters*
525 2015;362:fnv021.
526 Ou-Yang S, Wu N, Tian, Y. Effect of high pressure on the molecular structure and π-electrons
527 delocalization of canthaxanthin as revealed by Raman spectra. *Journal of Spectroscopy*
528 2015;2015:467089.
529 Paganini L, Villanueva GL, Roth L, Mandell AM, Hurford TA, Retherford KD, Mumma MJ. A
530 measurement of water vapour amid a largely quiescent environment on Europa. *Nature*
531 *Astronomy* 2020;4:266-272.
532 Parab NDT, Tomar V. Raman spectroscopy of algae: a review. *Journal of Nanomedicine &*
533 *Nanotechnology* 2012;3:131.
534 Parker AE, Dimitrov AG, Gedeon T. Symmetry breaking in soft clustering decoding of neural
535 codes. IEEE Transactions on Information Theory 2010;56:901-927.
536 Price PB. A habitat for psychrophiles in deep Antarctic ice. *Proceedings of the National*
537 *Academy of Sciences* 2000;97:1247-1251.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Rapin W, Ehlmann BL, Dromart G, Schieber J, Thomas NH, Fischer WW, Fox VK, Stein NT, Nachon M, Clark BC, Kah LC, Thompson L, Meyer HA, Gabriel TSJ, Hardgrove C, Mangold N, Rivera-Hernandez F, Wiens RC, Vasavada AR. An interval of high salinity in ancient Gale Crater Lake on Mars. *Nature Geoscience* 2019;12:889-895.

Rodrigues e Silva T, Carlos Francelino Silva Júnior L, Cavalcanti de Queiroz A, Suzana Alexan dre Moreira M, Alberto de Carvalho Fraga C, Cunha Alves de Menezes G, Henrique Rosa L, Bicas J, Maia de Oliveira V, Wagner Fernandes Duarte A. Pigments from Antarctic bacteria a nd their biotechnological applications. *Critical Reviews in Biotechnology* 2021. DOI: 10.1080/ 07388551.2021.1888068

Rull F, Maurice S, Hutchinson I, Moral A, Perez C, Diaz C, Colombo M, Belenguer T, Lopez-Reyes G, Sansano A, Forni O, Parot Y, Striebig N, Woodward S, Howe C, Tarcea N, Rodriguez P, Seoane L, Santiago A, Rodriguez-Prieto JA, Medina J, Gallego P, Canchal R, Santamaria P, Ramos G, Vago JL. The Raman laser spectrometer for the ExoMars Rover ission to Mars. *Astrobiology* 2017;17:627-654.

Rye R, Holland HD. Life Associated with a 2.76 Ga ephemeral pond?: evidence from Mount Roe# 2 paleosol. *Geology* 2000;28:483-486.

Samek O, Jonas A, Pilat Z, Zemanek P, Nedbal L, Triska J, Kotas P, Trtilak M. Raman microspectroscopy of individual algal cells: sensing unsaturation of storage lipids *in vivo*. *Sensors* 2010;10:8635-8651.

Savitsky A, Golay MJE. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry* 1964;36:1627-1639.

Schaeffer P, Millet J, Aubert JP. Catabolic repression of bacterial sporulation. *Proceedings of the National Academy of Sciences* 1965;54:704-711.

Schmidt ME, Ruff SW, McCoy TJ, Farrand WH, Johnson JR, Gelert R, Ming DW, Morris RV, Cabrol N, Lewis KW, Schroeder C. Hydrothermal origin of halogens at Home Plate, Gusev Crater. *Journal of Geophysical Research* 2008;13:E06S12.

Schumacher W, Stöckel S, Rösch P, Popp J. Improving chemometric results by optimizing the dimension reduction for Raman spectral data sets. *Journal of Raman Spectroscopy* 2014;45:930-940.

Schwieterman EW, Kiang NY, Parenteau MN, Harman CE, DasSarma S, Fisher TM, Arney GN, Hartnett HE, Reinhard CT, Olson SL, Meadows VS, Cockell CS, Walker SI, Grenfell JL, Hegde S, Rugheimer S, Hu R, Lyons TW. Exoplanet biosignatures: a review of remotely detectable signs of life. *Astrobiology* 2018;18:663-708.

Seager S, Schrenk M, Bains W. An astrophysical view of Earth-based metabolic biosignature gases. *Astrobiology* 2012;12:61-82.

Serrano P, Hermelink A, Lasch P, De Vera JP, König N, Burckhardt O, Wagner D. Confocal Raman microspectroscopy reveals a convergence of the chemical composition in methanogenic Archaea from a Siberian permafrost-affected soil. *FEMS Microbiology Ecology* 2015;91:fiv126.

Sharma SK, Nelson DR, Abdrabu R, Khraiwesh B, Jijakli K, Arnoux M, O'Connor MJ, Bahmani T, Cai H, Khapli S, Jagannathan R, Salehi-Ashtiani K. An integrative Raman microscopy-based workflow for rapid *in situ* analysis of microalgal lipid bodies. *Biotechnology for Biofuels* 2015;8:164.

584 Soni A, Oey I, Silcock P, Bremer P. Bacillus spores in the food industry: a review on resistance
585 and response to novel inactivation technologies. *Comprehensive Reviews in Food Science and*
586 *Food Safety* 2016;15:1139-1148.
587 Sousa-Silva C, Seager S, Ranjan S, Petkowski JJ, Zhan Z, Hu R, Bains W. Phosphine as a
588 biosignature gas in exoplanet atmospheres. *Astrobiology* 2020;20:235-268.
589 Stackebrandt E, Liesack W, Goebel BM. Bacterial diversity in a soil sample from a subtropical
590 Australian environment as determined by 16S rDNA analysis. *The Faseb Journal* 1993;7:232-
591 236.
592 Szalontai B, Debreczeny M, Fintor K, Bagyinka C. SVD-Clustering, A general image analyzing
593 method explained and demonstrated on model and Raman micro-spectroscopic maps.
594 *Scientific Reports* 2020;10:4238.
595 Tan J, Lewis JMT, Sephton MA. The fate of lipid biosignatures in a Mars-analogue sulfur
596 stream. *Scientific Reports* 2018;8:7586.
597 Tuma R. Raman spectroscopy of proteins: from peptides to large assemblies. *Journal of Raman*
598 *Spectroscopy* 2005;36:307-319.
599 Vokt JP, Brody S. The kinetics of changes in the fatty acid composition of *Neurosporu crussa*
600 lipids after a temperature increase. *Biochimica et Biophysica Acta (BBA) - Lipids and Lipid*
601 *Metabolism* 1985;835:176-182.
602 Ward PD, Brownlee D. *Rare Earth: Why complex life is uncommon in the universe.* Copernicus
603 Books, New York, NY, United States, 2000.
604 Watanabe Y, Martini JEJ, Ohmoto H. Geochemical evidence for terrestrial ecosystems 2.6
605 billion years ago. *Nature* 2000;408:574-578.
606 Watanabe Y, Stewart BW, Ohmoto H. Organic-and carbonate-rich soil formation ~ 2.6 billion
607 years ago at Schagen, East Transvaal District, South Africa. *Geochimica et Cosmochimica*
608 *Acta* 2004;68:2129-2151.
609 Wei J, Wang A, Lambert JL, Wettergreen D, Cabrol N, Warren-Rhodes K, Zacny K.
610 Autonomous soil analysis by the Mars microbeam Raman spectrometer (MMRS) on-board a
611 rover in the Atacama Desert: a terrestrial test for planetary exploration. *Journal of Raman*
612 *Spectroscopy* 2015;46:810-821.
613 Wiens RC, Maurice S, Robinson SH, Nelson AE, Cais P, Bernardi P, Newell RT, Clegg S,
614 Sharma SK, Storms S, Deming J, Beckman D, Ollila AM, Gasnault O, Anderson RB, André
615 Y, Michael Angel S, Arana G, Auden E, Beck P, Becker J, Benzerara K, Bernard S, Beyssac
616 O, Borges L, Bousquet B, Boyd K, Caffrey M, Carlson J, Castro K, Celis J, Chide B, Clark K,
617 Cloutis E, Cordoba EC, Cousin A, Dale M, Deflores L, Delapp D, Deleuze M, Dirmyer M,
618 Donny C, Dromart G, George Duran M, Egan M, Ervin J, Fabre C, Fau A, Fischer W, Forni
619 O, Fouchet T, Fresquez R, Frydenvang J, Gasway D, Gontijo I, Grotzinger J, Jacob X,
620 Jacquinod S, Johnson JR, Klisiewicz RA, Lake J, Lanza N, Laserna J, Lasue J, Le Mouélic S,
621 Legett C 4th, Leveille R, Lewin E, Lopez-Reyes G, Lorenz R, Lorigny E, Love SP, Lucero B,
622 Madariaga JM, Madsen M, Madsen S, Mangold N, Manrique JA, Martinez JP, Martinez-Frias
623 J, McCabe KP, McConnochie TH, McGlown JM, McLennan SM, Melikechi N, Meslin PY,
624 Michel JM, Mimoun D, Misra A, Montagnac G, Montmessin F, Mousset V, Murdoch N,
625 Newsom H, Ott LA, Ousnamer ZR, Pares L, Parot Y, Pawluczyk R, Glen Peterson C, Pilleri
626 P, Pinet P, Pont G, Poulet F, Provost C, Quertier B, Quinn H, Rapin W, Reess JM, Regan AH,
627 Reyes-Newell AL, Romano PJ, Royer C, Rull F, Sandoval B, Sarrao JH, Sautter V,
628 Schoppers MJ, Schröder S, Seitz D, Shepherd T, Sobron P, Dubois B, Sridhar V, Toplis MJ,
629 Torre-Fdez I, Trettel IA, Underwood M, Valdez A, Valdez J, Venhaus D, Willis P. The

630 SuperCam instrument suite on the NASA Mars 2020 rover: body unit and combined system
631 tests. *Space Science* Reviews 2021;217(4). DOI: 10.1007/s11214-020-00777-5.
632 Williford KH, Farley KA, Stack KM, Allwood AC, Beaty D, Beegle LW, Bhartia R, Brown AJ,
633 De la Torre Juarez M, Hamran SE, Hecht MH, Hurowitz JA, Rodriguez-Manfredi JA,
634 Maurice S, Milkovich S, Wiens RC. The NASA Mars 2020 rover mission and the search for
635 extraterrestrial life, In *From Habitability to Life on Mars,* edited by N. Cabrol and E.A. Grin,
636 Elsevier, Cambridge, MA, 2018, Chapter 11, Pages 275-308.
637 Wood BR, Heraud P, Stojkovic S, Morrison D, Beardall J, McNaughton D. A portable Raman
638 acoustic spectroscopic system for the identification and environmental monitoring of algal
639 cells. *Analytical Chemistry* 2005;77:4955-4961.
640 Wu H, Volponi JV, Oliver AE, Parikh AN, Simmons BA, Singh S, Vermaas WFT. *In vivo*
641 lipidomics using single-cell Raman spectroscopy. *Proceedings of the National Academy of*
642 *Sciences of the United States of America* 2011;108:3809-3814.
643 Yakubovskaya E, Zaliznyak T, Martinez Martinez J, Taylor GT. Tear down the fluorescent
644 curtain: a new fluorescence suppression method for Raman microspectroscopic analyses.
645 *Scientific Reports* 2019;9:15785.
646

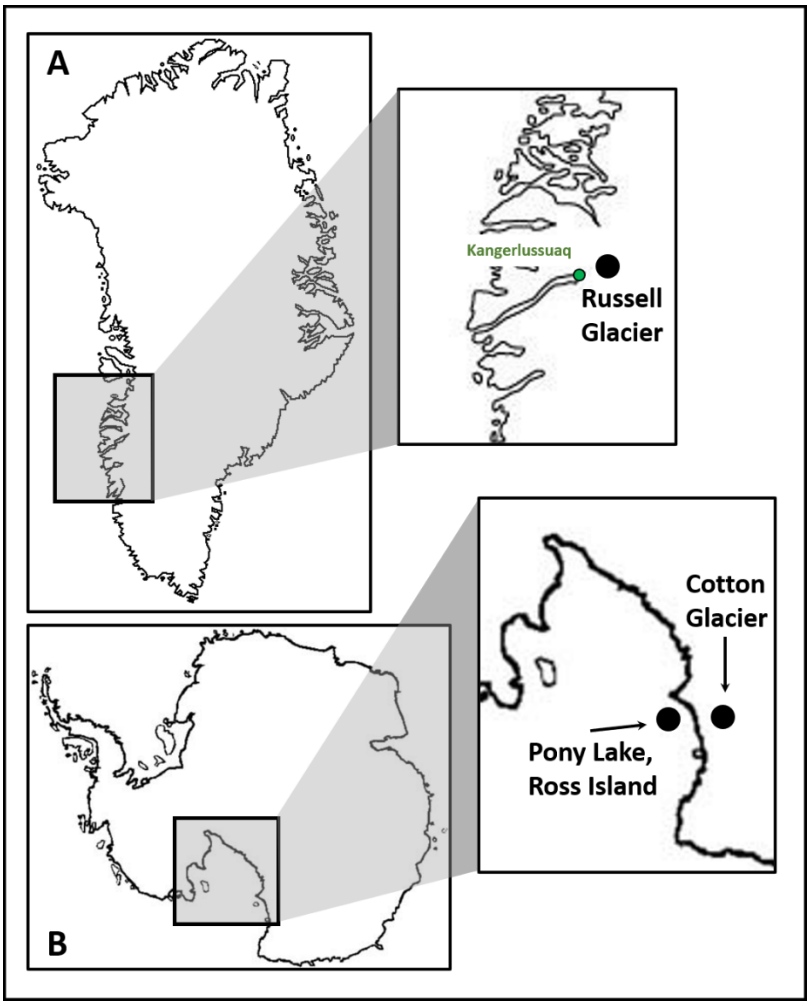
Figure Legend:

Figure 1: Sampling locations in (A) the western margin of the Greenland Ice Sheet and (B) the McMurdo Sound area, Antarctica.

Figure 2: Raman spectra of the environmental bacterial isolates. The spectra represent the average Raman spectra for 20 individual cells for each isolate. (A) orange and red, (B) yellow, (C) pink, (D) non-pigmented, and (E) purple bacterial isolates (non-baseline subtracted or data smoothed). (F) *E. coli* AW3110, vegetative cell and spores of *B. subtills*.

Figure 3: TSI biplot of the integrated regions of the Raman spectra that correspond to key cell features (*i.e.*, carotenoids, saturated fats, unsaturated fats, triacylglycerols, and proteins) for each environmental, bacterial isolate grown at 4°C and *B. subtilis, B. subtilis* spores, and *E.coli* AW3110 grown at 37°C. Data points reflect the color of the bacterial colonies. Non-pigmented isolated are shown in black. Each genus is displayed with a different marker shape.

Figure 4: (A) K-means Cluster plot of the normalized Raman spectra of the bacterial isolates from 550-1800 cm^{-1}. The list of the species comprising each cluster is listed in Supplemental Table 3. (B) Cluster means for the K-means Cluster plot of the normalized Raman spectra from 550-1800 cm^{-1}.



668

669 **Figure 1:**

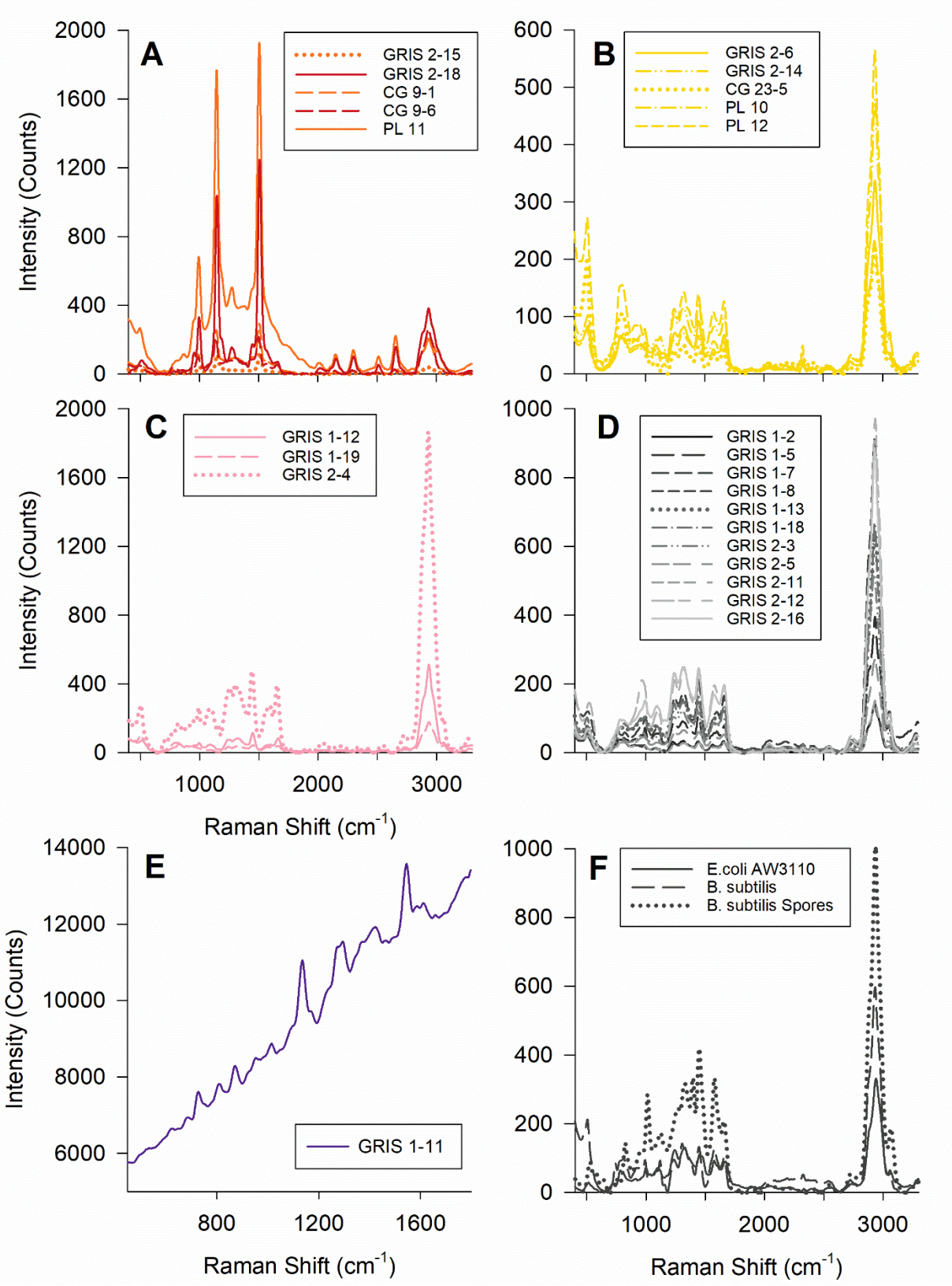
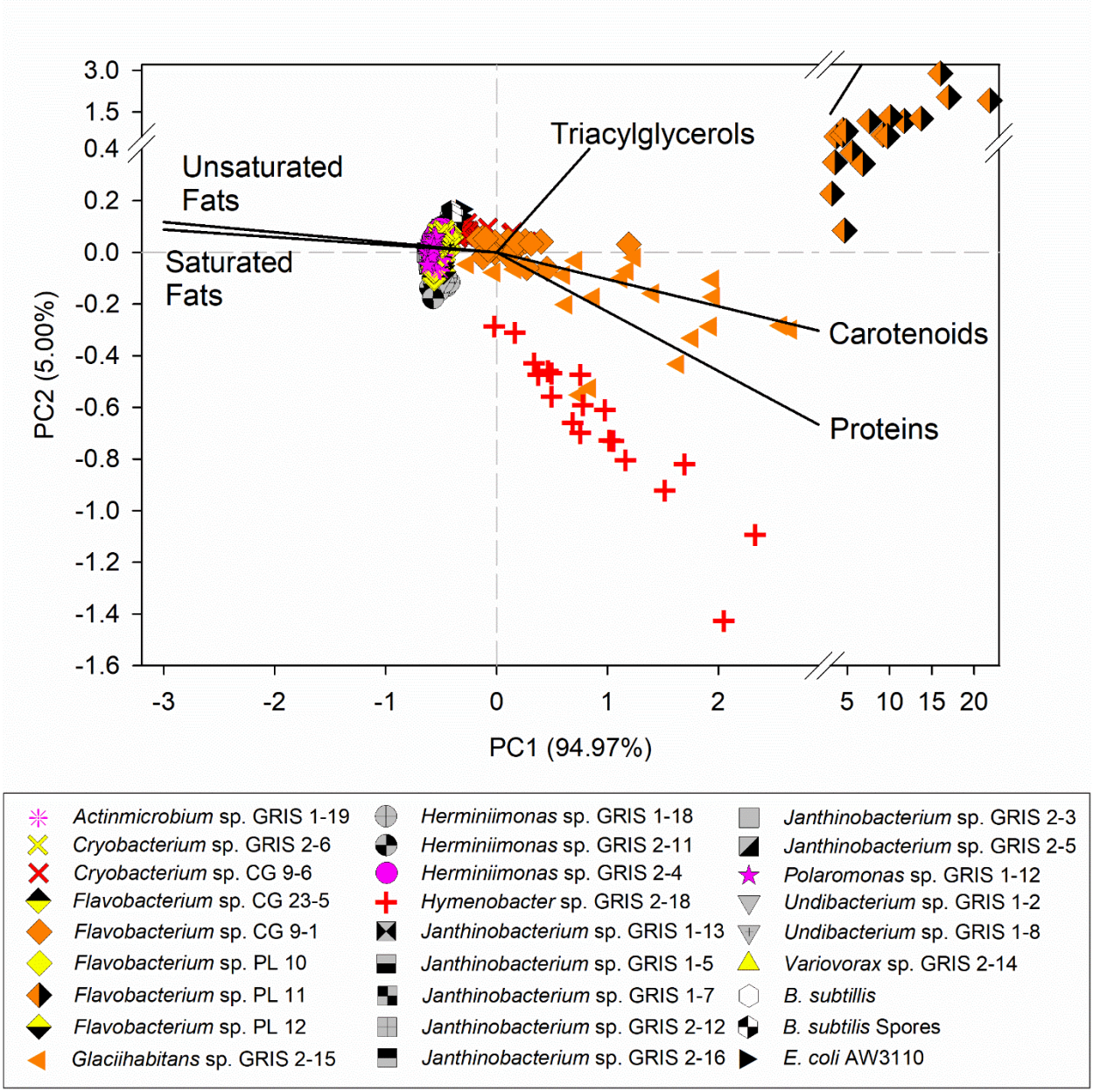
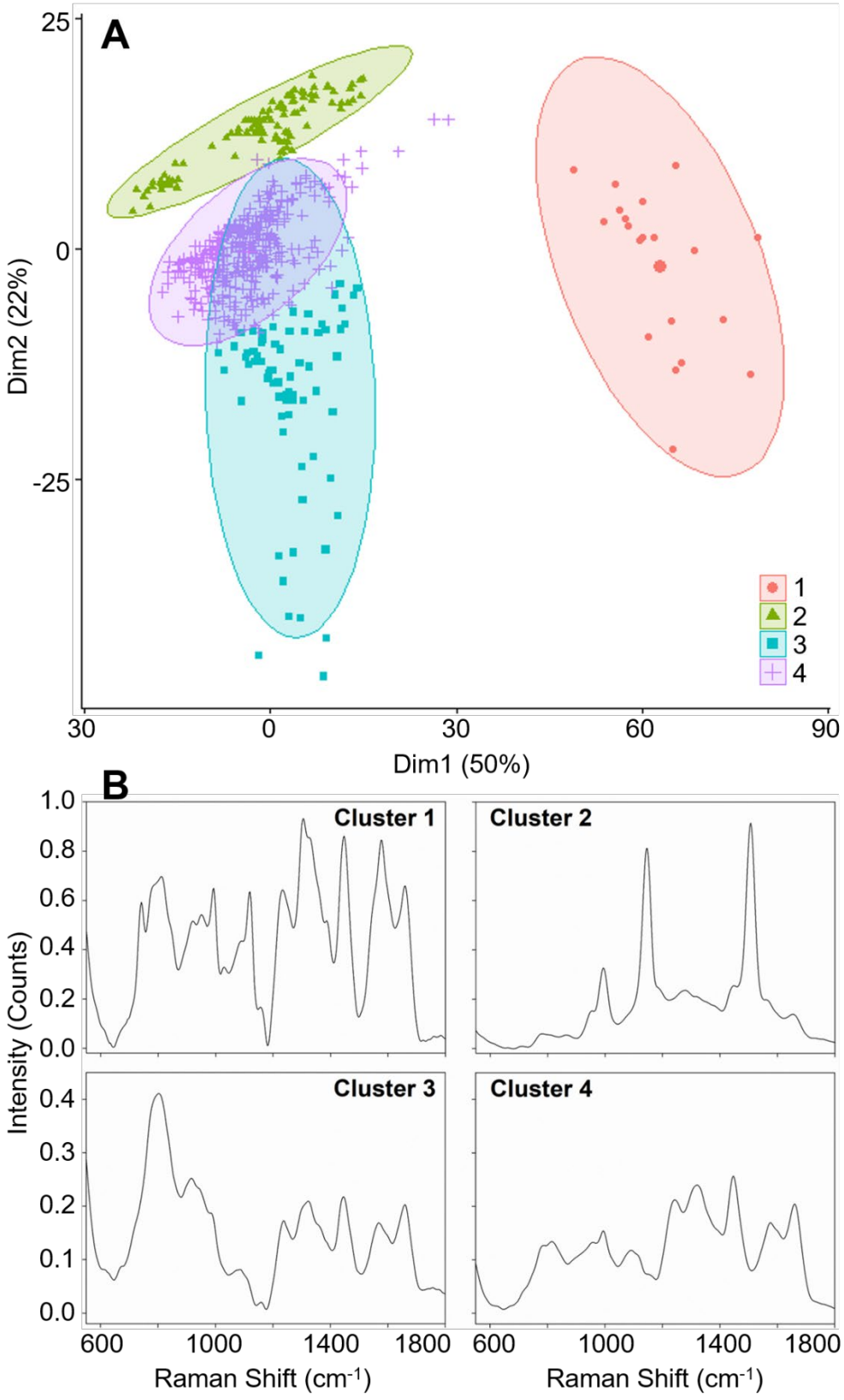

A
GRIS 2-15
GRIS 2-18
CG 9-1
CG 9-6
PL 11
B
GRIS 2-6
GRIS 2-14
CG 23-5
PL 10
PL 12
C
GRIS 1-12
GRIS 1-19
GRIS 2-4
D
GRIS 1-2
GRIS 1-5
GRIS 1-7
GRIS 1-8
GRIS 1-13
GRIS 1-18
GRIS 2-3
GRIS 2-5
GRIS 2-11
GRIS 2-12
GRIS 2-16
E
GRIS 1-11
F
E.coli AW3110
B. subtilis
B. subtilis Spores
Intensity (Counts)
Raman Shift (cm-1)

Figure 2:



672

673 **Figure 3:**



674

675 **Figure 4:**

Table 1: Published regions of Raman spectra representative of cellular components used for statistical analyses.

Key Cellular Feature	Raman Spectral Region (cm^{-1})	Reference
TSI Biplot		
Carotenoids	980-1020, 1070-1170, 1220-1280, 1480-1550	Wood et al., 2005
Saturated fats	1290-1360, 1415-1465, 2880-2980	Sharma et al., 2015
Unsaturated fats	1220-1280, 1625-1675	Sharma et al., 2015
Triacylglycerols	1070-1170, 1290-1360, 1415-1465, 1625-1675	Motoyama, 2012
Proteins	980-1020, 1290-1360, 1480-1550, 1625-1675	Wu et al., 2011
PERMANOVA		
Carotenoids	1006, 1154, 1520	Wood et al., 2005
Saturated fats	1300, 1440	Sharma et al., 2015
Unsaturated fats	1260, 1650	Sharma et al., 2015
Triacylglycerols	1080, 1130, 1300, 1460, 1630-1680	Motoyama, 2012
Proteins	1004, 1300, 1550, 1600-1700	Wu et al., 2011

680 **Table 2:** Biosignatures of Polar, bacterial isolates determined by Raman spectroscopy.
681

Key Cellular Feature	Spectral Region (cm^{-1})	Reference
Carotenoids	1006, 1154, 1520, 2006-2024, 2141-2153, 2288-2303, 2504-2513, 2645-2657	Ermakov et al., 2005; Wood et al., 2005; De Gelder et al., 2007; Ou-Yang et al., 2015
Membrane lipids	2920	Kint et al., 1992; Parab and Tomar, 2012
Unsaturated fats	1260, 1650, 3023	Samek et al., 2010; Wu et al., 2011; Czmara et al., 2014; Sharma et al., 2015
Saturated fats	1300, 1440	Samek et al., 2010; Wu et al., 2011; Czmara et al., 2014; Sharma et al., 2015
Proteins	1004	De Gelder et al., 2007; Wu et al., 2011
Amide I-III	1300, 1550, 1600-1700	De Gelder et al., 2007; Wu et al., 2011
Triacylglycerols	1080, 1130, 1300, 1460, 1630-1680	Motoyama, 2012
Nucleotides	600-800	Barhoumi et al., 2008; Parab and Tomar, 2012

682

683 **Table 3:** Pairwise PERMANOVA for normalized Raman spectra from 550-1800 cm^{-1} of
684 bacterial isolates. Statistically significant results are shown in bold. NP (non-pigmented).
685 Significance level: * <0.050, ** <0.010, *** <0.001

Pairs	Full Spectra	Carotenoid Regions	Unsaturated Fat Regions	Saturated Fat Regions	Protein Regions	Triacylglycerol Regions
Orange vs. Pink	**	**	**	*	**	**
Orange vs. NP	**	**	*	0.501	**	**
Orange vs. Yellow	**	**	**	*	**	**
Orange vs. Red	**	**	**	**	**	**
Pink vs. NP	*	*	0.052	0.117	**	**
Pink vs. Yellow	**	**	1.000	0.533	**	*
Pink vs. Red	**	**	**	**	**	**
NP vs. Yellow	**	*	**	0.132	0.056	0.059
NP vs. Red	**	**	**	**	**	**
Yellow vs. Red	**	**	**	**	**	**

686