Published on Jun 15, 2023 by Arcadia Science **DOI**: 10.57844/arcadia-x8wk-sf94

Raman spectra reflect complex phylogenetic relationships

Even with many tools available, categorizing species is tough. We used data from Raman spectroscopy, a form of label-free imaging, to infer phylogenetic patterns among several dozen diverse microbial taxa, offering a non-destructive and rapid way to dissect species relationships.

Contributors (A-Z)

Prachee Avasthi, Feridun Mert Celebi, Megan L. Hochstrasser, Austin H. Patton, Ryan York

Version 2 · Mar 31, 2025

Purpose

Figuring out the relationships between organisms is an essential part of biological investigation. To do so, researchers often rely on methods that are destructive (e.g. DNA sequencing), require extensive tools (e.g. label-based imaging), or prior knowledge (e.g. expert classification).

In this pub, we show that we can use Raman spectroscopy — a form of non-destructive, label-free imaging — to infer complex phylogenetic relationships between microbial organisms. Specifically, we find that distinct portions of Raman spectra

reflect phylogenetic signal and that this relationship is reflective of genomic components.

These observations should be of interest to evolutionary biologists, ecologists, and, broadly, researchers interested in extending the capacities of label-free imaging methods.

- This pub is part of the **platform effort**, "Genetics: Decoding evolutionary drivers across biology." Visit the platform narrative for more background and context.
- All associated **code** is available in this GitHub repository.
- A full **walkthrough of the code base** for the framework appears in a companion notebook.

Background and goals

Many roadblocks in biological research boil down to a single problem: not knowing what you're looking at. Meaningful comparisons — be it microbes within a mixed community or the cells of a heterogeneous tissue — are hard when samples are morphologically indistinct, difficult to access, or exist in dense arrangements. To get around this, biologists often employ next-generation sequencing (NGS) or label-based imaging. However, these methods come with drawbacks. NGS-enabled phylogenetic analyses can require significant time investment, are prone to various types of systematic errors, and can be difficult to use when samples are mixed or composed of hard-to-sort and/or uncharacterized organisms **[1]**. On the other hand, label-based imaging can be destructive and is often limited to well-known molecules or species that require prior characterization or evidence for use (which is often lacking in evolutionary or ecological research) **[2][3]**.

Label-free imaging methods using vibrational spectroscopy, such as Raman imaging, offer promising alternatives for addressing a number of basic problems in biology **[3] [4]**. Raman methods detect the presence of various chemical bonds via light scattering, providing biochemical fingerprints that can be reflective of metabolism, physiological state, cell type, or species **[3]**. Accordingly, it has been proposed that

Raman methods could become important tools for identifying the provenance of living organisms **[5][6][7][8]** and have already been leveraged to detect taxonomic patterns in certain biological materials, such as bivalve shells **[9]** and animal fossilization products **[10]**. Similarly, increasing numbers of studies have shown that Raman spectra are amenable to species-specific classification using machine learning approaches **[3][6][7]**. However, to our knowledge, no studies have explicitly tested the utility of Raman spectra for identifying phylogenetic patterns or relationships between species.

Establishing this link, or lack thereof, will be a crucial step if these tools are to be broadly applied to evolutionary and ecological problems. With this in mind, we used a publicly available dataset of Raman spectra from 30 clinically isolated microbial strains **[7]**, exploring the extent to which we could uncover phylogenetic relationships solely from spectral data.

The approach

Detailed methods

Data

All data analyzed here were previously published and publicly available. Details and experimental conditions can be found in the original publication, which used deep learning to classify 30 clinically isolated strains of pathogenic bacteria and fungi **[7]**. Briefly, they obtained Raman spectra using a Horiba LabRAM HR Evolution Raman microscope targeting monolayers of dried samples. They obtained spectra between 381.98 and 1792.4 cm^{-1} and normalized by the maximum intensity to vary between 0 and 1.

Analysis

> All associated **code** is available in this GitHub repository (DOI: 10.5281/zenodo.7872093) and we provide a **code base walkthrough** for the framework in a companion notebook.

The suite of analyses presented here is available in a fully interactive and editable notebook on GitHub. This notebook walks through the relevant code and methodological considerations. Below is a brief, complementary methods overview:

We obtained data from the original publication via the Dropbox link provided on their GitHub **[7]**. Depending on the analysis type, we used all replicates per strain (n = 100) or strain-level means (see subsequent paragraphs). We excluded the species *Streptococcus agalactiae* from all analyses based on what appeared to be an aberrant spectral profile (see Figure 1).

First, we collected taxonomic classifications for each strain from the NCBI taxonomy database **[11]**. Strain-specific classifications were compiled into a matrix in which each column corresponds to a specific level of the taxonomic hierarchy (e.g. strain, species, genus, etc.). We then used this matrix as input to generalized linear models (GLMs) predicting spectral relationships among strains. We used PC1 from a principal component analysis (PCA) of spectra across all replicates (n = 100/strain) as the outcome variable given that it explained over 20% of variance in the data (explored in more depth in Notebook 1). In total, we constructed eight GLMs, each for a specific level of taxonomic classification. We compared model fits using the Bayesian information criterion (BIC). We complemented these analyses by measuring the cosine similarity among replicates within different taxonomic groupings. We measured cosine similarity using the `cosine` function in the R package LSA and calculated its variance among taxonomic groupings.

To enable phylogenetic comparisons, we obtained a time-calibrated, species-level (n = 19 species) phylogenetic tree from timetree.org **[12]**. We then used this tree to calculate phylogenetic signal as a function of spectral position. To do so, we used a sliding window approach (width = 25 wavenumbers, stepsize = 1 wavenumber). Within each window, we inferred phylogenetic signal of species-level mean spectra by calculating Pagel's λ **[13]** using the `phylosig` function from the R package phylosig **[14]**. We calculated the spectral distance between species using these same sliding

windows, but, in place of Pagel's lambda, we calculated the euclidean distance between species within each window. We then time-calibrated spectral distances by calculating the cophenetic distances between all species (essentially the dates at which species are estimated to have diverged given the phylogenetic tree) using the function `cophenetic.phylo` in the R package ape **[15]**. We then matched species-wise cophenetic distances with spectral distances, allowing two-dimensional comparisons of these values. We used the window-based approach to calculate the difference between window-based trees and the observed phylogenetic tree via the Robinson-Foulds metric. We used the function `TreeDistance` in the R package TreeDist to infer the Robinson-Foulds metric **[16]**.

Finally, we compared genome features to the patterns observed above by collecting data from the NCBI Genome Database (Figure 1, C). We inferred the relationship between genomic features and spectra using the window-based approach from above. Within each window, we performed PCA on mean spectra and generated a GLM using PC1 as the outcome and each genomic feature (e.g. GC content) as the predictor. The outcome of this analysis was thus a continuous value representing the similarity between spectral and genomic relationships. We then computed Pearson correlations between these GLM fits and computed phylogenetic signal.

The results

Strain-level Raman spectra associate with taxonomy

As mentioned in "The approach," we obtained a publicly available dataset of Raman spectra collected from 30 clinically isolated strains of bacteria and fungi (Figure 1). To enable evolutionary comparisons, we identified the taxonomic classification for each sample, from strain to domain (Figure 2, A). We reasoned that if spectra contain meaningful phylogenetic information, then the similarity of strain-level spectra should scale with taxonomy (i.e. genus-level spectra should be more similar than kingdom-level spectra).

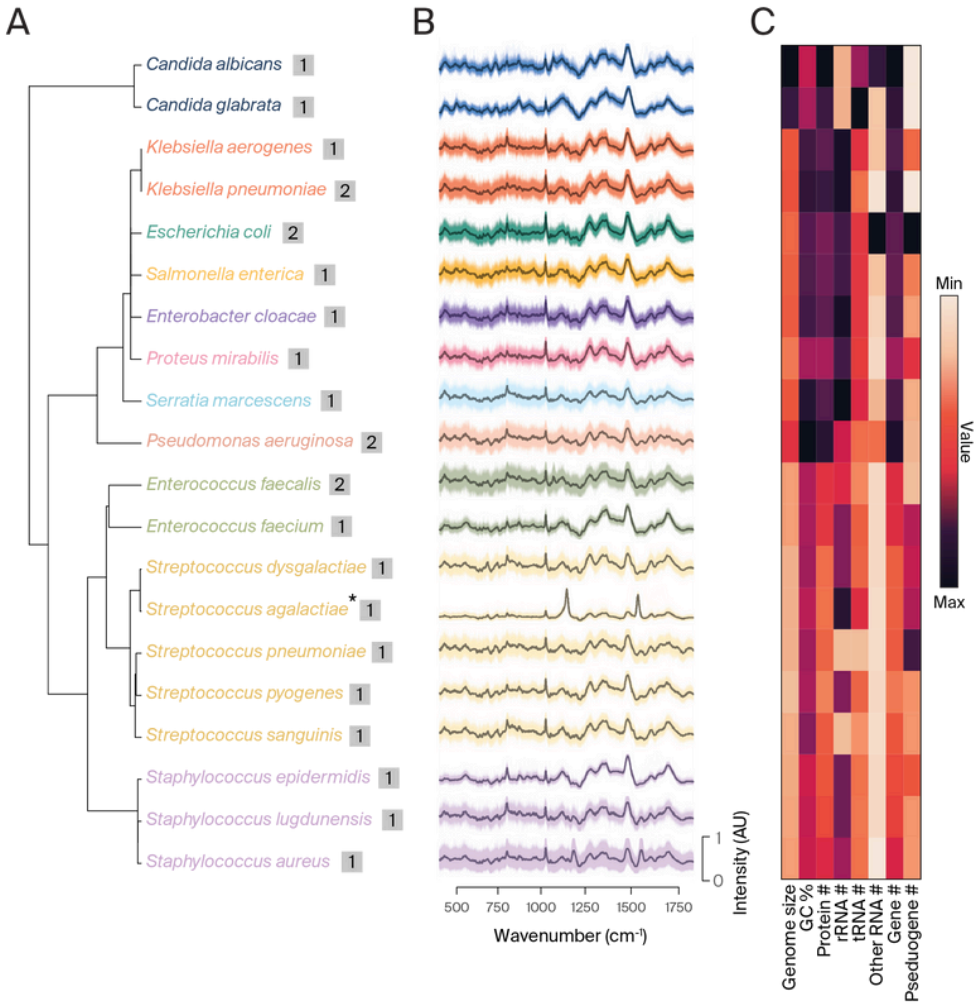

Figure 1

Phylogenetic context of the dataset.

(A) Time-calibrated phylogeny of species considered in this study. Species names are colored by genus. The number of strains per species in the data set is indicated by the number in the grey box.

* = species not included in statistical analyses.

(B) Spectra distributions for each species in the study. Mean spectra are indicated by the darker line, plotted over the spectra of all 100 replicated per species. AU = arbitrary units.

(C) Heatmap of genome statistics for each species.

To test this intuition, we analyzed how well taxonomic categories can predict spectral measurements (see The approach). Specifically, we used generalized linear models (GLMs) to assess the linear relationships between taxonomy and spectra and assessed model fit using the Bayesian information criterion (BIC), a common metric for comparing a set of models. Here, models with lower BIC are better able to predict spectra. Strikingly, we found that the range of BIC values exactly mirrored the taxonomic hierarchy (Figure 2, A–B). Strain identity best predicted the range of spectra (BIC = 9,034), followed by species (BIC = 10,974) (Figure 2, A). Interestingly, all other taxonomic predictors — genus to kingdom — displayed similar model fits. We also saw these patterns when analyzing spectral similarity (measured by cosine similarity) (Figure 2, C), observing increasing amounts of variance as taxonomic granularity decreased. These observations suggest that Raman spectra vary as a function of taxonomic relationship and that strain and species-level signals are most strongly encoded in spectral information.

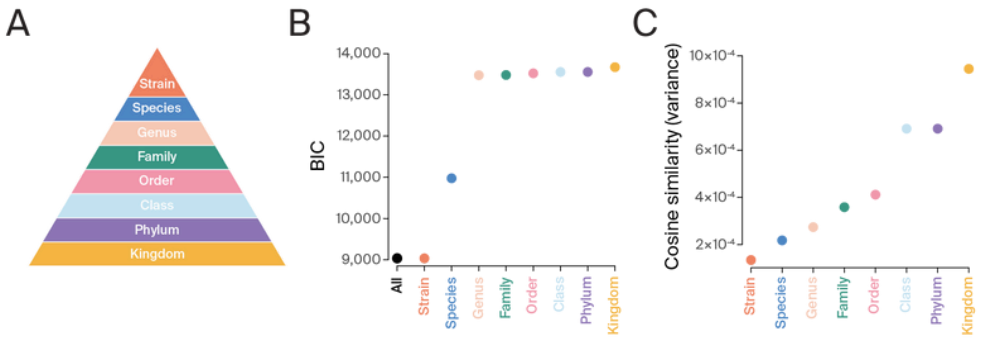

Figure 2

Raman spectra vary with taxonomy.

(A) Graphical depiction of the hierarchical relationships between taxonomic classes used here.

(B) Distribution of the Bayesian information criterion (BIC) values for generalized linear models (GLMs) comparing spectra and taxonomic categories.

(C) Distribution of cosine similarity variance as a function of taxonomic categories.

Evolutionary signals are position-specific within spectra

The above observations indicate that, when considered in their totality, Raman spectra vary as a function of taxonomy. Is this variation evenly distributed across spectra or restricted to specific portions? If the former is true, then it would appear that variations between species' spectra arise from biochemical signatures too complex or nonlinear to resolve solely from these data. In the latter scenario, specific molecular signatures may drive spectral differences, hinting at some possibility of identifying biological drivers of this measurement variation (via position-specific associations with taxonomy).

We explored these possibilities by calculating phylogenetic signal (Pagel's λ) **[13]** — a measure of how much species' phenotypic and phylogenetic relationships match each other — as a function of position in Raman spectra (for details see The approach). In this framework, higher values of phylogenetic signal indicate that closer-related species have more similar spectral measurements. Remarkably, we found increased phylogenetic signal in a series of clear bands (Figure 3, A–C). These bands were distributed across the spectral range (Figure 3, A), displayed an average width of 43 wavenumbers (standard deviation = 18 wavenumbers), and had maximum phylogenetic signal values between 0.25 and 0.79. These observations support the second scenario from above: Phylogenetic signal is unevenly distributed across Raman spectra.

Given that the amount of phylogenetic signal varied across the observed bands, we next wondered if this variation reflected the same, or different, evolutionary patterns. There were several possibilities. On one hand, relationships between species measurements could be identical across the spectrum. In this scenario, phylogenetic signal would vary simply as a function of measurement differences going up and down. On the other hand, it could be the case that species relationships change with position, either subtly or strongly. In that case, phylogenetic signal may be associated with a variety of species relationships, suggesting that Raman spectra reflect a more complex landscape of evolutionary relationships.

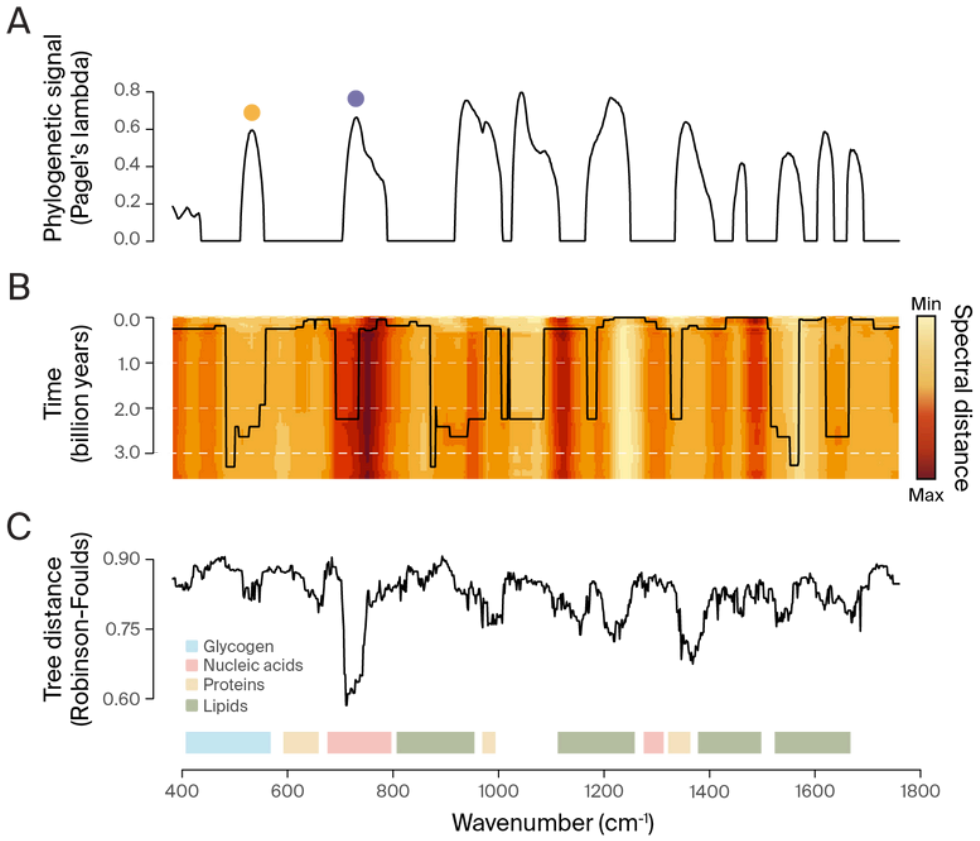

Figure 3

Evolutionary signals are position-specific within spectra.

(A) The phylogenetic signal distribution across the full Raman spectrum. Calculated in 25 wavenumber-wide windows. The yellow and purple dots mark example peaks discussed in the text.

(B) Heatmap of spectral distance. The y-axis corresponds to billions of years, darker color corresponds to greater average distance between species pairs as a function of divergence time. Black line represents the time point at which the maximum spectral distance for that position was measured.

(C) Distribution of distances between the phylogenetic tree and trees made from spectral relationships within windows along the spectrum. Tree distance corresponds to the Robinson-Foulds metric. Colored bands below reflect common biomolecular signatures in Raman spectra.

To explore these possibilities, we calculated the spectral distance between species as a function of evolutionary time (for details see The approach) and visualized the results as a heatmap (Figure 3, B). Color shows the distance among spectra as a function of evolutionary time (represented by the y-axis). We are essentially asking, for two species that diverged X million years ago, how different are their spectra? We then average these values over all of evolutionary time. We also plotted the time at which we saw the greatest spectral difference for each position along the spectrum, displayed as a black line. As may be expected, spectral distance within the bands was often elevated further back in time (reflecting phylogenetic structure; more distantly related species have more distant spectra) while regions with low phylogenetic signal displayed more recent spectral differences (Figure 3, B). However, despite these high-level patterns, we found a notable amount of diversity among the bands, both in the overall distance between spectra and specific relationships with time (Figure 3, B).

Certain bands reflected large overall distances between species (Figure 3, A and C; marked by purple dot) while others, though displaying increased phylogenetic signal, displayed spectral distance distributions more similar to that observed across the full spectrum (Figure 3, A and C; marked by yellow dot). Similarly, the conserved bands displayed variable relationships with the overall phylogenetic tree (Robinson-Foulds metric; Figure 3, C) wherein certain bands displayed strong similarities to the phylogeny (purple dot) while others did not (yellow dot). These findings suggest that the phylogenetic relationships of conserved bands are position-specific and reflect a complex evolutionary landscape.

This last observation is even more enticing when we consider the broader-scale molecular patterns present in Raman spectra (as represented by the colored boxes on the bottom of Figure 3, A–C). For example, the band between ~700–800 cm^{-1} overlapped strongly with a region known to reflect nucleic acid abundance while another at ~1,150–1,250 cm^{-1} appeared to correlate with lipids **[8]**. Interestingly, these two bands displayed quite different spectral and phylogenetic tree distance distributions (Figure 3, A–C). Might it be possible to detect evolutionary relationships unique to certain biomolecules from Raman spectral data?

Genomic features predict spectral variation across species

Finally, we compared high-level genomic features (e.g. genome size, number of genes, GC content; Figure 1, C) with spectral relationships. To do so, we calculated the association between a given genomic statistic and per-species spectral measurements within overlapping windows along the spectrum (width = 25 wavenumbers; see The approach and Notebook 1).

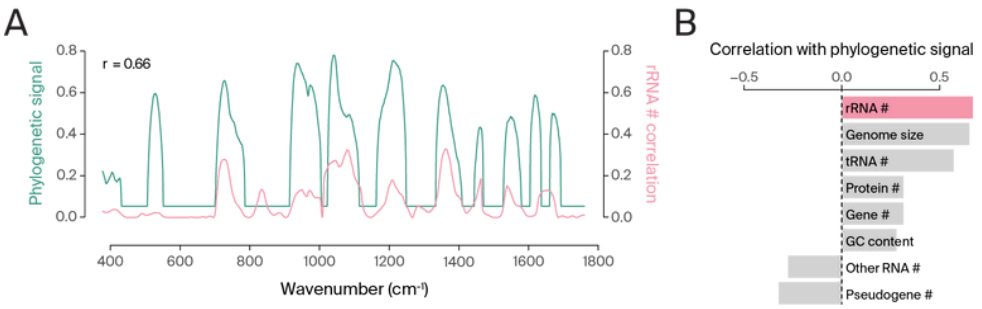

Figure 4

Association between genome features and phylogeny.

(A) Comparison of phylogenetic signal (green) and ribosomal RNA numbers along the spectrum. r = Pearson's correlation.

(B) Barplot of correlation coefficients (calculated with Pearson's correlation) between phylogenetic signal and ribosomal RNA number. GC content is highlighted in pink.

We found that several genomic features, such as the # of ribosomal RNAs (rRNAs), displayed clear peaks that mirrored those we observed for phylogenetic signal (Figure 4, A). All of these comparisons yielded moderate to strong correlations (Figure 4, B), the strongest being between rRNA # and phylogenetic signal (r = 0.66), followed closely by genome size (r = 0.65). Additionally, we found that a linear model using all genomic features could account for 76% of phylogenetic signal variation (R^2 = 0.76; for more details, see Notebook 1). These results suggest that basic genomic features can account for a substantial portion of phylogenetic information present in Raman spectra.

Key takeaways

- Raman spectra from clinically isolated bacteria and fungi vary as a function of taxonomic classification (Figure 2).
- Phylogenetic relationships are unevenly distributed across the Raman spectrum; specific spectral bands predict known phylogenetic relationships (Figure 3).
- Evolutionary diversification patterns vary as a function of Raman position (Figure 3).
- Phylogenetic signal in the Raman spectrum is strongly associated with high-level genomic features, suggesting that Raman methods directly detect biochemical information relevant to inferring phylogenetic relationships (Figure 4).

Implications

The set of analyses presented here support the idea that Raman spectral comparisons will be broadly useful for phylogenetic and evolutionary studies.

However, the conclusions from this study come with several caveats. First, these data are restricted to clinically isolated strains of bacteria and fungi. Future work is needed to assess how applicable these findings are to other taxa (including multicellular organisms). Furthermore, the Raman data we analyzed here came from researchers measuring pooled samples **[7]**. This strategy may limit the true dynamic range of species-level spectra, especially if the goal is to consider variation across individual organisms, since this strategy essentially averages out signals across individuals. Finally, the phylogenetic distances represented here are quite broad. It will be enlightening to test the outer limits of Raman capabilities in taxonomic classification, including but not limited to testing closely related species, measuring individual organisms, assessing the effect of optical variants (e.g. autofluorescence), or exploring variation in complex samples and tissues. These caveats also present many opportunities for substantial exploration and development. For example, it may be the case that we can uncover variable evolutionary patterns across spatially complex samples (e.g. between cells or in subcellular regions of interest).

Finally, it is interesting to consider Raman as just one example of a certain type of high-content phenotype that is useful in dissecting complex biological processes. Raman spectra contain abundant information about the molecular structure and, as

we show here, phylogenetic context/evolutionary diversification patterns of biological samples. Even within a single Raman experiment, we should be able to extract insight into multiple dimensions of biology. Other types of biological measurements that quantify complex biophysical/chemical/molecular processes, such as chlorophyll fluorescence **[17]** or lifetime imaging, may also fit into this category. In general, we contend that these observations point toward the power of combining high-dimensional phenotypes with evolutionary inference to begin dissecting complex biology in a generalizable, scalable, and hypothesis-free framework.

References

1. Young AD, Gillung JP. (2019). Phylogenomics — principles, opportunities and pitfalls of big-data phylogenetics. https://doi.org/10.1111/syen.12406

2. Cheng J-X, Xie XS. (2015). Vibrational spectroscopic imaging of living systems: An emerging platform for biology and medicine. https://doi.org/10.1126/science.aaa8870

3. Kobayashi-Kirschvink KJ, Gaddam S, James-Sorenson T, Grody E, Ounadjela JR, Ge B, Zhang K, Kang JW, Xavier R, So PTC, Biancalani T, Shu J, Regev A. (2021). Raman2RNA: Live-cell label-free prediction of single-cell RNA expression profiles by Raman microscopy. https://doi.org/10.1101/2021.11.30.470655

4. Kobayashi-Kirschvink KJ, Nakaoka H, Oda A, Kamei KF, Nosho K, Fukushima H, Kanesaki Y, Yajima S, Masaki H, Ohta K, Wakamoto Y. (2018). Linear Regression Links Transcriptomic Data and Cellular Raman Spectra. https://doi.org/10.1016/j.cels.2018.05.015

5. Germond A, Kumar V, Ichimura T, Moreau J, Furusawa C, Fujita H, Watanabe TM. (2017). Raman spectroscopy as a tool for ecology and evolution. https://doi.org/10.1098/rsif.2017.0174

6. Lee KS, Palatinszky M, Pereira FC, Nguyen J, Fernandez VI, Mueller AJ, Menolascina F, Daims H, Berry D, Wagner M, Stocker R. (2019). An automated Raman-based platform for the sorting of live cells by functional properties. https://doi.org/10.1038/s41564-019-0394-9

7 Ho C-S, Jean N, Hogan CA, Blackmon L, Jeffrey SS, Holodniy M, Banaei N, Saleh AAE, Ermon S, Dionne J. (2019). Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. https://doi.org/10.1038/s41467-019-12898-9

8 Cui D, Kong L, Wang Y, Zhu Y, Zhang C. (2022). In situ identification of environmental microorganisms with Raman spectroscopy. https://doi.org/10.1016/j.ese.2022.100187

9 Wade J, Pugh H, Nightingale J, Kim J, Williams ST. (2019). Colour in bivalve shells: Using resonance Raman spectroscopy to compare pigments at different phylogenetic levels. https://doi.org/10.1002/jrs.5639

10 Wiemann J, Crawford JM, Briggs DEG. (2020). Phylogenetic and physiological signals in metazoan fossil biomolecules. https://doi.org/10.1126/sciadv.aba6883

11 Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, Sharma S, Soussov V, Sullivan JP, Sun L, Turner S, Karsch-Mizrachi I. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. https://doi.org/10.1093/database/baaa062

12 Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, Stecher G, Hedges SB. (2022). TimeTree 5: An Expanded Resource for Species Divergence Times. https://doi.org/10.1093/molbev/msac174

13 Pagel M. (1999). Inferring the historical patterns of biological evolution. https://doi.org/10.1038/44766

14 Revell LJ. (2011). phytools: an R package for phylogenetic comparative biology (and other things). https://doi.org/10.1111/j.2041-210x.2011.00169.x

15 Paradis E, Schliep K. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. https://doi.org/10.1093/bioinformatics/bty633

16 Smith MR. (2020). Information theoretic generalized Robinson–Foulds metrics for comparing phylogenetic trees. https://doi.org/10.1093/bioinformatics/btaa614

17 Murchie EH, Lawson T. (2013). Chlorophyll fluorescence analysis: a guide to good practice and understanding some new applications. https://doi.org/10.1093/jxb/ert208